University of Oxford
Lincoln College

Thesis submitted for the degree of
Doctor of Philosophy

ADVANCED MACHINE LEARNING ALGORITHMS FOR HEAVY FLAVOUR JETS IDENTIFICATION AND HIGGS SEARCHES WITH THE ATLAS EXPERIMENT

Author
Maxence DRAGUET

Supervisor
Daniela BORTOLETTO

Trinity
MMXXIV

ABSTRACT

An essential element in many ATLAS analyses is the ability to identify the flavour of jets. This subject is extensively discussed in this thesis, with a review of the algorithmic developments carried out by the ATLAS Collaboration from 2020 to 2024. Increasingly sophisticated Machine Learning (ML) taggers have been developed for this specific purposed. The initial approach relies on a hierarchical construction combining low-level physically-motivated taggers physically with a Deep Set or a Recurrent Neural Network as inputs to a high-level tagger predicting the flavour. Recentely, a more nimble design leveraging a single network to deliver state-of-the-art performance has been introduced. The core of this network is either Graph Attention Network or a Transformer Encoder. Expert knowledge is passed to the architecture by optimising multiple tasks, with different physics input types used in a multimodal framework. The design and training of these taggers are reviewed, with a study of the Hyperparameter Optimisation (HPO) for large networks using techniques from the ML literature on Large Language Models (LLM).

Following the 2012 discovery of Higgs boson by the ATLAS and CMS experiments, increasingly refined measurements of the new particles have been performed. The leading production modes and the decay mode to third generation fermions and gauge vector bosons of the Higgs have been measured. Attention is now focused on the second generation fermions, such as the c-quark, and differential cross-section measurements. This thesis presents a combined search for the $H \to c\bar{c}$ coupled with a differential measurement of the $H \to b\bar{b}$ in the VH production mode. The analysis exploits the full 140 fb^{-1} proton-proton collision luminosity collected in Run 2 by the ATLAS experiment at a centre-of-mass energy of 13 TeV. The combination of these decay modes allow for a coherent joint analysis strategy with a better constraining of the shared backgrounds. The full p_T spectrum is covered, with the two candidate jets resolved at low momentum and a single merged boosted signature at high values. The previously reviewed flavour taggers are used to reconstruct the Higgs. Three leptonic channels are defined based on the number of electrons and muons. A fine categorisation is deployed with dedicated Boosted Decision Trees discriminants to increase the sentivity. The analysis is blinded with an expected 95% CL$_s$ upper limit for the $VH(H \to c\bar{c})$ process signal strength at 11.1 $\times$ the SM prediction. The $VH(H \to b\bar{b})$ signal strength is 7.9 σ over the background-only hypothesis, with the WH and ZH productions respectively measured with expected significances of 5.5 σ and 6.2 σ. A Standard Cross-Section Templates measurement is performed in stage 1.2 for $VH(H \to b\bar{b})$, in bins of p_T and number od additional jets.

Personal Contributions

The work presented in this thesis is inherentely collaborative, having been carried out as a member of the ATLAS Collaboration. This thesis present the two subjects to which I mostly contributed: the development of new heavy flavour taggers and the $VH(H \to b\bar{b}/c\bar{c})$ combined analysis using the full Run 2. While I produced some "ATLAS" labelled figures, others were taken from public results produced by other member of the Collaboration. Plots with no label have always been personally produced. This section aims to highlight my personal contributions to these different projects that are fully detailed in this thesis.

Flavour Tagging

I joined the flavour tagging group for my qualification task, and have contributed to the training of the new taggers. My main contributions are:

- Producing training samples with new ATLAS software release (R22)
- Modifications to the preprocessing to implement importance sampling when harmonising distributions of different flavour samples for full statistics use.
- Modifications to the training software to include taus, flexible input variables definition, and general debugging.
- Modifications to the postprocessing to implement new visualisation and grapics.
- Retraining DL1r on the new R22 release and comparing it to the previous DL1r.

- First training of DL1d with the DIPS subtagger on release R22. This was done for both PFlow and variable-radius (VR) jets.
- Studying the inputs and architecture of DL1d.
- Training the DIPS subtagger with VR jets (used in the above step).
- Studies of hyperparameter optimisation of GN1 and GN2, with modifications the software stack to leverage the μP parametrisation and implement the μTransfer algorithm.
- Adapting the codebase and developing a framework to train on CERN's KubeFlow server.

These different contributions lead me to significantly participate in the development of the Umami [1] and Salt [2] software used to train the network. My contributions have been part of different ATLAS publications, such as Ref. [3] using a DL1r model I trained, Ref. [4] using a DL1d model I trained, Ref. [5] for which I produced the DL1d input to the X_{bb}, as well as a coming ATLAS publication on GN2. I led the effort on hyperparameter optimisation of GN2, producing the public results in Ref. [6] presented at the 6th Inter-Experimental Machine Learning Workshop [7] and the Cloud-Native AI Day KubeCon conference [8].

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

I joined the $VH(H \to b\bar{b}/c\bar{c})$ analysis team in 2021, and my main contributions are:

- Study of the X_{bb} tagger versus a looser DL1r for the boosted $VH(H \to b\bar{b})$ regime, with the impact on the signal sensitivity after dedicated MVA trainings analysed.
- Studies of the Data-Monte Carlo agreement in $VH(H \to c\bar{c})$ with DL1r-based tagging.
- Contributing to different rounds of analysis samples productions.
- Design and study of a new top control region for the $VH(H \to c\bar{c})$ and $VH(H \to b\bar{b})$ resolved. Extensive list of studies to reach the final approach presented in this thesis. Additional study on the Higgs candidate reconstructions strategy.
- Derivation and harmonisation of the p_T^V-dependent ΔR_{cc} cut in $VH(H \to c\bar{c})$.
- Training and deployement of the CARL model for the single-top Wt- and t-channels of the top background for the resolved $VH(H \to b\bar{b})$.
- Development of the analysis modelling codebase to study the top backgrounds.

- Modelling studies of the top background in the resolved $VH(H \to b\bar{b})$. Derived shape and acceptance uncertainties for $t\bar{t}$, Wt, and t-channel, and studied the effect of the chosen modeling and the combination of $t\bar{t}$ with Wt.

- Numerous fit studies to validate new samples, the new top control region and top backgrounds normalisation scheme, the new Higgs candidate reconstruction strategy, as well as studying the impact of the introduction of CARL models and refinements to modelling.

- Modification to the fit framework to use the new ROOT version and help stabilise the fit, as well integrating an update to the results visualisation and plots made.

At the time of writing, the analysis is reaching its conclusion with final studies on the modelling and fit framework, aiming for publication by June 2024. The results presented here are therefore only temporary and still blinded.

Contents

1 Introduction **1**

2 Theoretical Particle Physics **4**
2.1 The Standard Model of Particle Physics . . . 4
2.1.1 Quantum Electrodynamics . . . 7
2.1.2 Electroweak Sector . . . 8
2.1.3 The Brout-Englert-Higgs Mechanism . . . 10
2.1.4 Quantum Chromodynamics . . . 12
2.1.5 Yukawa Interactions . . . 14
2.2 Experimental Higgs Phenomenology . . . 15

3 The LHC & ATLAS Experiment **19**
3.1 The Large Hadron Collider . . . 20
3.2 The ATLAS Detector . . . 22
3.2.1 The Inner Detector Tracker . . . 25
3.2.2 Electronic and Hadronic Calorimeters . . . 26
3.2.3 Muon Detection Systems . . . 28
3.3 Operation and Reconstruction with the ATLAS Detector . . . 29
3.3.1 Trigger System . . . 29
3.3.2 Low-Level Signatures . . . 30
3.3.3 Electrons . . . 31
3.3.4 Muons . . . 32
3.3.5 Jets . . . 32
3.3.6 Taus . . . 35
3.3.7 Missing Transverse Energy . . . 35

4 Machine Learning & Deep Learning **37**
4.1 Definitions . . . 37
4.1.1 Artificial Intelligence . . . 37
4.1.2 Machine Learning . . . 39
4.1.3 Deep Learning . . . 42
4.2 Machine Learning Methods for Physics . . . 43
4.2.1 Decision Trees . . . 45
4.2.2 Boosted Decision Trees . . . 48
4.2.3 Artificial Neurons . . . 53
4.2.4 Deep Neural Networks . . . 54

4.2.5 Recurrent Neural Networks . . . 62
4.2.6 Convolutional Neural Networks . . . 66
4.2.7 Graph Neural Networks . . . 68
4.2.8 The rise of the Transformers . . . 72
4.3 Training and Optimising Deep Learning Models . . . 76
4.3.1 Training Algorithms . . . 77
4.3.2 Regularisation . . . 79
4.3.3 Architecture & Hyperparameters Optimisation . . . 79
4.3.4 Acceleration Techniques . . . 80

5 Flavour Tagging **82**
5.1 Heavy-Flavour Jet Tagging . . . 82
5.1.1 Decay Topology . . . 83
5.1.2 Flavour Tagging at ATLAS . . . 85
5.1.3 Datasets . . . 86
5.2 DL1 Family of Taggers: DL1r & DL1d . . . 88
5.2.1 RNNIP . . . 89
5.2.2 DIPS . . . 91
5.2.3 Training of DIPS with Variable Radius Jets for Run 3 . . . 96
5.2.4 Training of DL1d & DL1r with PFlow for Run 3 . . . 100
5.2.5 Training of DL1d on Variable Radius Jets for Run 3 . . . 111
5.3 Graph Neural Network Family of Taggers . . . 114
5.3.1 GN1: Graph Attention Network for Flavour Tagging . . . 116
5.3.2 GN2: Transformer Encoder for Flavour Tagging . . . 126
5.3.3 GN2 Hyperparameters Optimisation . . . 135
5.3.4 GN2X: GN2 Variant for Boosted Higgs to Heavy Flavours Decays . . . 145
5.4 Calibration . . . 147
5.5 Conclusion . . . 149

6 $VH(H \to b\bar{b}/c\bar{c})$ Combined Analysis **150**
6.1 Introduction . . . 150
6.2 The $VH(H \to b\bar{b})$ and $VH(H \to c\bar{c})$ ATLAS Analyses . . . 151
6.3 Overview of the Combined $VH(H \to b\bar{b}/c\bar{c})$ Analysis . . . 154
6.4 Data and Simulated Samples . . . 155
6.4.1 Signal Processes . . . 157
6.4.2 Background Processes . . . 158
6.5 Selection and Categorisation . . . 162
6.5.1 Object Selection . . . 162
6.5.2 Event Selection . . . 166
6.5.3 Event Categorisation . . . 172
6.6 Tagged-jets corrections . . . 180
6.7 Discriminant Variables . . . 181
6.7.1 Multivariate Analysis . . . 182
6.7.2 Output Variable Transformation . . . 187
6.8 Experimental Uncertainties . . . 188
6.9 Signals & Backgrounds Modelling . . . 191
6.9.1 General Modelling Strategy . . . 193
6.9.2 Signal Modelling . . . 197
6.9.3 V+jets Modelling . . . 199
6.9.4 Top Modelling . . . 202
6.9.5 Diboson Modelling . . . 206
6.9.6 Multi-jet Modelling . . . 208
6.10 Statistical Analysis . . . 208
6.10.1 Likelihood Function Definition . . . 209

6.10.2 The $VH(H \to b\bar{b}/c\bar{c})$ Fit 212
6.11 Conclusion 221

7 Conclusion and Outlook **224**

Bibliography **226**

Appendices **241**

Common Abbreviations

μP Maximal Update Parametrisation
AI Artificial Intelligence
ANN Artificial Neural Network
AUC Area Under the Curve
BDT Boosted Decision Trees
BSM Beyond the Standard Model
CARL Calibrated Likelihood Ratio Estimator
CERN Centre Européen pour la Recherche Nucléaire
CKM Cabibb-Kobayashi-Maskawa
CL Confidence Level
CNN Convolutional Neural Network
CPU Core Processing Unit
CR Control Region
DIPS Deep Impact Parameter Sets
DL Deep Learning
DL1 Deep Learner 1 Model
DL1d DL1 with DIPS
DL1r DL1 with RNNIP
DNN Deep Neural Network
DT Decision Trees
EW Electroweak
FN Floating Normalisation
FPGA Field-Programmable Gate Array
FSR Final State Radiation
GAN Generative Adversarial Network
GAT Graph Attention Network
GN1 GN with GAT-core
GN2 GN with Transformer-core
GNN Graph Neural Network
GPU Graphics Processing Unit
GRU Gated Recurrent Unit
HEP High Energy Physics
HPC High Performance Cluster
HPO Hyperparameter Optimisation
IBL Insertable B-Layer
ID Inner Detector
IP Impact Parameter
ISR Initial State Radiation
JES Jet Energy Scale
JVT Jet Vertex Tagger
LHC Large Hadron Collider
LSTM Long-Short Term Memory
MC Monte Carlo
ME Matrix Element
ML Machine Learning
MLP Multilayer Perceptron
MS Muon Spectrometer
MSE Mean Squarred Error
MVA Multivariate Analysis
NLP Natural Language Processing
NN Neural Network
NP Nuisance Parameter
PCFT Pseudo-Continuous Flavour Tagging
PDF Parton Distribution Function
POI Parameter Of Interest
PS Parton Shower
PU Pile-up
PV Primary Vertex
QCD Quantum Chromodynamics
QED Quantum Electrodynamics
QFT Quantum Field Theory
ReLU Rectified Linear Units
RL Reinforcement Learning
RNN Recurrent Neural Network
RNNIP Recurrent Neural Network Impact Parameter
ROC Receiver Operating Characteristic
SCT Semiconductor Tracker
SF Scale Factor

SGD Stochastic Gradient Descent
SM Standard Model
SR Signal Region
STXS Simplified Template Cross-Section
SV Secondary Vertex
SV1 Secondary Vertex method 1
TRT Transition Radiation Tracker
UE Underlying Event
UFO Unified Flow Object
VAE Variational Auto-Encoder
VR Variable Radius
WP Working Point

CHAPTER 1

Introduction

Modern particle physics is built around a combined patchwork of theoretical models gathered into the so-called *Standard Model (SM)* [9, 10]. In the edifice of science, the SM is our current best approach at describing the foundation of Nature at its smallest scale. The elegance of its structure resides in its mingling of mathematical concept such as symmetries and gauge invariance to describe and predict the structure of matter and how it interacts under the strong, the weak, and the electromagnetic interactions. The Brout-Englert-Higgs mechanism plays a central role in the theory, allowing for the emergence of massive gauge vector bosons through spontaneous symmetry breaking [11, 12]. The success of the theory has been countlessly confirmed experimentally in hundreds of measurements, particularly by the ATLAS and CMS observation of the Higgs boson in 2012 [13, 14]. The SM is not however a complete theory of the elementary components of the Universe. Gravity is not included, neutrinos masses are observed but not accounted for, and the SM does not explain astronomical observations of the existence of dark matter. Mass is also introduced arbitrarily for fermions by introducing Yukawa couplings symbolising interactions between these particles and the Higgs boson. The observation of a strict mass hierarchy between the different generation is therefore not currently explained. Concerning the Higgs boson, while the leading production modes and the decay modes to third generation

fermions and vector bosons have been observed, the coupling to lighter-fermions have not yet been measured. In particular, the coupling to the second-generation c-quark can be probed for signs of physics Beyond the Standard Model (BSM) [15–21].

This thesis present, in its last chapter, an ATLAS search for the $H \to c\bar{c}$ decay mode coupled with a differential cross-section measurement of the $H \to b\bar{b}$. These analyses are performed jointly to leverage the same strategy and better constrain the shared background. The vector boson V (W or Z) associated production VH is used, with the V leptonically decaying to 0, 1, or 2 electrons or muons. This latter requirement reduces the important multi-jet background and provides an effective signature to select data to save through the triggers. The search is performed with the 149 fb^{-1} of proton-proton collision data collected by ATLAS during the Run 2 (2015-2018) of the Large Hadron Collider (LHC) at a centre-of-mass energy of 13 TeV.

The main ingredient for the presented analysis consists in identifying the b- and c-quarks flavour from the reconstructed structure of their decay called a *jet*. This is a challenging task due to the large event rate leaving to overlapping signatures and the complex structure of jets. Thankfully, the ATLAS experiment generates a tremendous amount of real and simulated data, leading to effective deployement of Machine Learning (ML) techniques. As such, the second central theme of this thesis is the elaboration of evermore complex neural networks called *tagger* to perform jet flavour classification. This thesis extensively describes the latest effort of the ATLAS Collaboration to improve these methods. A complete review of the development and training of DL1d, a hieararchical tagger using the DIPS subtagger built on a Deep Set network to replace the DL1r tagger using the RNNIP tagger. A recent breakthrough in performance has been obtained when adopting single complex neural network based on a Graph Attention Network (GAT) or a Transformer Encoder, in models respectively called GN1 and GN2. This document presents these new taggers in detail, as well as efforts to leverage techniques from the Artificial Intelligence (AI) community designed for Large Language Models (LLM) to optimise the hyperparameters of the large Transformer-based GN2 network.

The thesis aims to present a coherent and connected structure leading to the combined $VH(H \to b\bar{b}/c\bar{c})$ analysis presented last in Chapter 6. First, the SM, the Higgs mechanisms and Yukawa interactions are presented in Chapter 2. The experimental conditions of the Large Hadron Collider (LHC) and the ATLAS experiment are then outlined in Chapter 3. As machine

learning and artificial intelligence play an essential role in modern science, perhaps even more so in particle physics, Chapter 4 is entirely dedicated to a review of the field relevant to High Energy Physics (HEP). Building on this Machine Learning (ML) introduction, Chapter 5 presents the development of the modern flavour Taggers of the ATLAS Collaboration, reviewing the design and training of DL1r, DL1d, GN1, and GN2. Finally, the $VH(H \to b\bar{b}/c\bar{c})$ analysis is described in Chapter 6 before concluding with a look forward in Chapter 7.

CHAPTER 2

Theoretical Particle Physics

Particle Physics is the field of science dedicated to the study of the fundamental elements of matters and of their interactions. Nature at this scale is best represented by an intricate connexion between particles, the fundamental components of matter, and their interactions, themselves represented by particles. This framework is encapsulated into a mathematical foundation called quantum field theory. A major scientific achievement of the second half of the XXth century is the elaboration of the so-called Standard Model of Particle Physics, a unified patchwork of theories describing all known elementary particles and three of the four fundamental interactions affecting them. This chapter reviews relevant elements of the theory to contextualise the work presented in this thesis.

2.1 The Standard Model of Particle Physics

To date, the SM has been the most successful theory in describing the constituents and the dynamics of matter [10]. The SM stands at the centre of theoretical particle physics, ellaborated by combining different successful theories of quantum mechanis and special relativity in the second part of the XXth century. These different exploits have led to a total of 55 Nobel Prizes. The SM is often hailed as the most successful theory of science, with the unique ability to predict

properties of the Universe to a staggering degree of precision: most famously, the anomalous magnetic dipole moment predictions is in agreement with measurements to up to 10 significant decimals [22]. The SM is expressed in the language of the dynamics of quantised fields, or Quantum Field Theory (QFT). These fields play two roles: describing matter itself (*fermions*, such as the electron) and the different interactions (*bosons*, such as the W and the Higgs boson H) governing how this matter interacts under the electromagnetic, weak, and strong interactions. Particles are the results of local excitations of quantised fields that are defined as operator-values distributions over spacetime. Figure 2.1 displays the fundamentals particles of the SM.

Figure 2.1: Elementary particles of the Standard Model [23]. Elementary fermions (quarks on top, lepton on the bottom) are listed in the three left columns and elementary bosons in the two right ones, the gauge bosons in the first column and the Higgs scalar boson in the last one. The mass, electrical charge, and spins of the particles are indicated.

Particles are separated based on their intrinsic angular momentum or *spin*, with half-spin particles following the Fermi-Dirac statistics called *fermions*, and integer-spin particles called *bosons* following the Bose-Einstein statistics. The elementary fermions are evenly split into 6 quarks and 6 leptons, with only the first generation of each being stable. The distinction between these two types stems from the different quantum numbers categorising them. Quarks

carry a fractional electromagnetic charge as well as a colour charge making them sensitive to the strong interaction. On the contrary, leptons are colour-neutral and either have an electrical charge of -1e^+, in units of the anti-electron (positron) charge, or are neutral. The charged leptons include the electron e^- - the lightest and only stable one -, the muon μ^-, and the tau τ^-. The neutral leptons are called neutrinos, with one neutrino ν_ℓ associated per charged lepton *ell*, e.g., the electron-neutrino ν_e. For the quarks, the electromagnetic charge is fractional, dividing them evenly between *up*-type quarks with charge $+\frac{2}{3}$ consisting of the up u, charm c, and top t flavours, and the *down*-type quarks with charge -$\frac{1}{3}$ and the flavours down d, strange s, and bottom b. To every particle corresponds an *anti*particle, with the same quantum numbers except for the electrical charge that changes sign.

The kinematics and dynamics of the fields representing the particles in the theory are expressed through a Lagrangian density $\mathcal{L}$, a spacetime discretisated element of the general Lagrangian. Symmetries of the Lagrangian density play an essential role as they define conserved quantities through Noether's theorem. The construction of the SM Lagrangian id dictated by the expression of these symmetries to satisfy the experimentally observed conserved quantities, such as the electromagnetic charge. Two types of symmetries can be considered: global ones that are valid across spacetime and local ones, the so-called gauge symmetries valid for localised transformations. The SM Lagrangian must satisfy the global Poincaré symmetry, encapsulating the symmetry required for special relativity, and a local non-Abelian $SU(3)_C \otimes SU(2)_L \otimes U(1)_Y$ gauge symmetry. Non-Abelian groups are such that their generators do not commute. The Lagrangian density of a field ψ is a function of ψ and its spacetime partial derivative $\partial_\mu \psi$, where μ indexes the time and space dimenions in the 4-vector formalism. The full SM Lagrangian $\mathcal{L}_{\text{SM}}$ can be decomposed into 4 terms:

$$\mathcal{L}_{\text{SM}} = \mathcal{L}_{\text{EW}} + \mathcal{L}_{\text{QCD}} + \mathcal{L}_{\text{Higgs}} + \mathcal{L}_{\text{Yukawa}}. \tag{2.1}$$

These different terms represent different effects encoded within the unified framework of the SM. Three of the four known interactions of nature are encapuslated in the SM: the strong, the electromagnetic, and the weak forces, with the gravitational force considered aside due to the weakness of its influence at sub-atomic scales. The mediators of the three included interactions are the gauge bosons, spin 1 particles with different properties arising from the nature of the interaction they carry. The electromagnetic and weak force have been successfully unified as a single Electroweak (EW) interaction, while the theory of the strong force is Quantum Chromo-

dynamics (QCD). One essential element in the SM is the so-called Higgs interaction, a special force through which some gauge bosons acquire mass. This Higgs interaction is underpinned by a Higgs field, an excitation of which is called a Higgs boson H. Interactions between the Higgs field and quark can be introduced to assign masses to the latter through Yukawa couplings. The latter part of this thesis is dedicated to a measurement of such couplings to b- and c-quarks. The different interactions and their respective gauge bosons are further reviewed in this chapter.

2.1.1 Quantum Electrodynamics

Quantum Electrodynamics (QED) is the theory underpinning the behaviour of free fermions and their electromagnetic interactions, for which the gauge carried are the photons. Fermions are represented by a Dirac spinor field $\psi(x)$ defined over spacetime x. The Dirac equation of quantum mechanics is a first order partial derivative equation modelling the free dynamics of such a spin-$\frac{1}{2}$ fermion with:

$$(i\gamma^\mu \partial_\mu - m)\psi = 0, \tag{2.2}$$

where γ^μ are the Dirac γ-matrices generalising the Pauli spin matrices to spacetime dimension, and the Einstein notation is adopted, summing over indices repeated as covariant and contravariant. For clarity, any contraction $\gamma^\mu \partial_\mu$ is denoted as $\not{\partial}$. A Lagragian density can be constructed to result in the dynamics described by Equation 2.2 through application of Euler-Lagrange:

$$\mathcal{L}_{\text{Dirac}} = \bar{\psi}(i\not{\partial} - m)\psi, \tag{2.3}$$

Such a Lagrangian models the free dynamics of any spin-$\frac{1}{2}$ fermion, such as an electron e^-. Electrons have a $q = -1$ charge that is conserved by every known interactions. This conservations must be the result of a symmetry: the Dirac Lagrangian must be made invariant under a local gauge $U(1)$ transformation:

$$\psi \to \psi' = e^{-iq\alpha(x)}\psi, \tag{2.4}$$

which corresponds to a rotation in the complex spacetime by a phase $q\alpha(x)$. For the Lagrangian of Equation 2.3 to satisfy this symmetry, the partial derivative ∂_μ must be replaced by the *gauge covriant derivative* D_μ:

$$\partial_\mu \to \partial_\mu + iqA_\mu, \tag{2.5}$$

where a new vector field $A\mu$ is introduced and required to transform under the $U(1)$ symmetry as $A_\mu \to A'_\mu = A_\mu + \partial_\mu \alpha(x)$. The elegance of this approach is the possibility to give this

fauge field A_μ its own dynamic, modifying the Lagrangian of Equation 2.3 into the Quantum Electrodynamics (QED) Lagrangian:

$$\mathcal{L}_{\text{QED}} = \bar{\psi}(i\not{\partial} - m)\psi + q\bar{\psi}\not{A}\psi - \frac{1}{4}F_{\mu\nu}F^{\mu\nu}$$

$$\mathcal{L}_{\text{QED}} = \bar{\psi}(i\not{D} - m)\psi - \frac{1}{4}F_{\mu\nu}F^{\mu\nu} \tag{2.6}$$

where $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$ is the electromagnetic field tensor. The last term introduces a kinetic term for the gauge field and the interaction between the fermion ψ and the gauge field A is represented by the term combining them. In the present case, the charge q is the conserved quantity of the gauge symmetry, which required the introduction of a new gauge field A_μ that can be interpreted as the photon field and given the dynamic of the electromagnetic interaction through $F_{\mu\nu}$. Fermionic fields can be introduced in this approach for the different known fermions, ψ_e, ψ_μ, ψ_u, ψ_c, etc. Their interactions with A_μ defining each time a unique conserved electromagnetic charge q_e, q_μ, q_u, q_c, etc. This procedure is however general: the gauge invariance of a Lagrangian introduces a spin-1 gauge vector boson. The required $U(1)$ invariance forbids the presence of mass terms of the form $m^2 A^\mu A_\mu$, seemingly condemning the gauge bosons to be massless.

2.1.2 Electroweak Sector

The weak force is described by two massive gauge vector bosons: the $W^\pm$ of mass $m_W \approx 80.4$ GeV[1] and the Z^0 of mass $m_Z \approx 91.2$ GeV. This apparent contradiction with the massless requirements of a $U(1)$ symmetry is elegantly solved by the Brout-Englert-Higgs mechanism [11, 12]. This mechanism, described in the next section, is applied to a joint expression of the electromagnetic and weak forces known as Electroweak (EW) theory in the Glashow-Weinberg-Salam (GWS) model [24–26]. The fundamental symmetry group the theory is built upon is the non-Abelian $SU(2)_L \otimes U(1)_Y$, where $SU(2)_L$ is the weak isospin and $U(1)_Y$ the weak hypercharge. A local $SU(2)$ transformation acts as:

$$\psi \to \psi' = e^{ig\alpha^a(x)T^a}\psi, \tag{2.7}$$

where $T^a = \sigma^a/2$ are the generators of the $SU(2)_L$ group, built from the σ^a Pauli matrices ($a = 1, 2, 3$). Each generator corresponds to a gauge field. The gauge field linked to $SU(2)_L$ leads to covariant derivative, simularly to Equation 2.8, to ensure gauge invariance as expressed

[1] The unit system adopted throughtout this thesis is to set the speed of light in vacuum c at 1, leading to masses expressed in GeV. To convert to mass units, one simply needs to divide by the standard unit c^2.

by

$$D_\mu = \partial_\mu + igT_a W^a_\mu, \tag{2.8}$$

with three gauge fields W^1_μ, W^2_μ, W^3_μ with interaction strength g. The particularity of the weak interaction is that the charged current interactions described by the symmetry group $SU(2)_L$ only apply to left-handed L particles states and not the right-handed R. Consequently, fermionic fields are decomposed into

$$\psi = \psi_L + \psi_R$$

with left-and right-handed particles represented by isospin doublets. The weak isospin I_W charge of left-handed particles is $I_W = 1/2$, with a third component $I^3_W = \pm 1/2$. For the right-handed part, $I_W = 0$ with $I^3_W = 0$, decoupling it from the gauge bosons W^a_μ. Physically, the observed weak charge current interaction corresponding to the $W^\pm$ bosons are the linear combinations of gauge fields:

$$W^\pm_\mu = \frac{1}{\sqrt{2}}\left(W^1_\mu \mp W^2_\mu\right) \tag{2.9}$$

These W bosons only couple to left-handed particles, but an experimentally observed Z boson couples to both left- and right-handed particles. This represented by the additional $U(1)_Y$ symmetry of the weak interaction in the SM, with weak hypercharge Y, coupling g', and an additional gauge field B_μ. The weak hypercharge is

$$Y = 2(Q - I^3_W),$$

with Q the electromagnetic charge. The total covariant derivative of the electroweak sector of the SM is therefore expressed by the GSM model as

$$D_\mu = \partial_\mu + igT_a W^a_\mu + ig'\frac{Y}{2}B_\mu. \tag{2.10}$$

where W^a_μ and B_μ are respectively the $SU(2)_L$ and $U(1)_Y$ gauge bosons. The GSM model re-expresses the electromagnetic photonic field A_μ and the Z-boson field Z_μ as a linear combination of B_μ and W^3_μ depending on the *weak mixing angle* θ_W, a fundamental parameter of the SM such that:

$$\cos\theta_W = \frac{g}{\sqrt{g^2 + g'^2}}. \tag{2.11}$$

This establishes a connexion between the coupling strengths of the weak interaction and the electromagnetic interaction coupling e as

$$e = g \sin\theta_W = g' \sin\theta_W \cos\theta_W.$$

The intrinsic strength of the weak interactions is indeed of similar ordar to that of the electromagnetic interaction but is weak in apparance due to its gauge bosons being massive. The weak force is the only known fundamental interaction to violate parity conservation. Neutrinos can only interact throught the weak force, which can itself only apply to left-handed particles. As a result, there are no right-handed neutrinos in the SM.

A significant achievement of modern particle physics is the unification of interactions that are perceived as different at low-energies. The problem of the measured mass of the vector gauge bosons remains. Additionaly, the split of fermionic fields into a left- and right-handed componets lead the mass terms to violate the gauge invariance. Both issues are resolved by the mechanisn of the Brout-Englert-Higgs and Yukawa interactions, introduced in the next sections.

2.1.3 The Brout-Englert-Higgs Mechanism

The Brout-Englert-Higgs mechanism, abbreviated BEH henceforth, offers a solution to introduce mass terms to the gauge fields $W_\mu^\pm$ and Z_μ [11, 12]. It introduces a new scalar Higgs field, permeating the Universe. The field is mathematically defined as a weak isospin doublet, with a neutral component ϕ^0 and a charged one ϕ^+. They are jointly expressed as complex scalar field with 4 degrees of freedom:

$$\phi = \begin{pmatrix} \phi^+ \\ \phi^0 \end{pmatrix} = \frac{1}{2} \begin{pmatrix} \phi_1 + i\phi_2 \\ \phi_3 + i\phi_4 \end{pmatrix} \tag{2.12}$$

This complex scalar field is made to interact with the electroweak gauge fields as

$$\mathcal{L}_{\text{Higgs}} = (D_\mu \phi)^\dagger (D^\mu \phi) - V(\phi), \tag{2.13}$$

where the first term described the kinetic of the ϕ and the second term is the Higgs potential energy:

$$V(\phi) = -\mu^2 \phi^\dagger \phi + \lambda (\phi^\dagger \phi)^2. \tag{2.14}$$

where the expression of this potential is constrained by the need for the theory to be renormalisable. Two scalars constants govern the Higgs potential: μ and λ describing, respectively, the quadratic and quartic interaction of the complex Higgs field ϕ. The former manifests the interaction with the gauge bosons, while the latter introduces self-interactions. The minimum of this potential corresponds to the vacuum state. The requirement that the vacuum be stable demands $\lambda > 0$. For a positive $\mu^2 > 0$, a degenerate minimum is found at field values such that

$$\phi^\dagger\phi = \frac{1}{2}(\phi_1^2 + \phi_2^2 + \phi_3^2 + \phi_4^2) = \frac{-\mu^2}{2\lambda} = \frac{v^2}{2} \tag{2.15}$$

introducing in the last equality the so-called *vacuum expectation value* $v = \sqrt{\frac{|\mu^2|}{\lambda}}$ of the field ϕ. This infinite degeneracy of the Higgs potential minimum states underlines a special $SU(2)$ symmetry such that $\phi^\dagger\phi = v^2/2$. Through *spontaneous symmetry breaking*, the BEH mechanism crumbles this degeneracy into one single vacuum state, typically chosen by setting the components $\phi_1 = \phi_2 = \phi_4 = 0$ and $\phi_3 = v$ so that the vacuum expactation is simply

$$\langle 0|\phi|0\rangle = \frac{1}{\sqrt{2}}\begin{pmatrix} 0 \\ v \end{pmatrix}. \tag{2.16}$$

The breaking of the symmetry enforces

$$SU(2)_L \otimes U(1)_Y \to U(1)_Q,$$

with the final vacuum state correctly set as chargeless. Expanding the full field dynamic around the chosen vacuum state with a particular gauge choice to absord unphysical Goldstone fields into the vector fields called *unitarity gauge* [27], the expansion can be simplified to

$$\phi = \frac{1}{\sqrt{2}}\begin{pmatrix} 0 \\ v + h(x) \end{pmatrix}, \tag{2.17}$$

where h is the real neutral Higgs scalar field. Introducing this expression into the Higgs Lagrangian of Equation 2.15 gives

$$\begin{aligned}\mathcal{L}_{\text{Higgs}} = &\frac{1}{2}(\partial_\mu h)(\partial^\mu h) + \frac{\mu^2}{2}(v+h)^2 - \frac{\lambda}{16}(v+h)^4 \\ &+ \frac{1}{4}g^2(W_\mu^+ W^{\mu-})(v+h)^2 + \frac{g_W^2 + g'^2}{8}(Z_\mu Z^\mu)(v+h)^2\end{aligned} \tag{2.18}$$

where one can clearly identify mass terms for the physical gauge fields $W_\mu^\pm$ and Z_μ in the last line, but not for A_μ as required from observations that photons are massless. The vector gauge bosons masses are:

$$m_W = \frac{v}{2}g,\ m_Z = \frac{v}{2}\sqrt{g^2 + g'^2} \tag{2.19}$$

or equivalenty expressing the mass of the Z-boson in terms of the W-boson mass

$$m_Z = \frac{m_W}{\cos\theta_W}.$$

The Higgs field is massive[2] with mass

$$m_H = \sqrt{2|\mu^2|}.$$

The BEH mechanism elegantly assigns mass to the gauge vector boson, with the photon remaining massless and the addition of the Higgs boson H as a massive elementary particle. Furthermore, the introduction of the related Higgs field permits the expression of mass terms for the fermions in the standard model, as explained in Section 2.1.5 on Yukawa interactions.

2.1.4 Quantum Chromodynamics

The strong interaction is described by the theory of *Quantum Chromodynamics (QCD)*, underpinned by an $SU(3)_C$ symmetry with conserved quantum number called *colour*. The only particles having a colour charge in the SM are quarks and gluons g, the gauge mediators of the strong interactions. There are three colour charges typically labelled *red*, *blue*, and *green*, each coming into its direct or anti-colour (e.g., anti-red). Quarks carry one such charge and gluons two. Similarly to the electroweak sector, the symmetry leads to a covariant derivative under the $SU(3)_C$ group of

$$D_\mu = \partial_\mu + ig_s\lambda_a G_\mu^a. \tag{2.20}$$

where the coupling constant g_s of the strong interaction is often re-expressed as $\alpha_s = \frac{g_s}{4\pi}$, and the generator of the $SU(3)_C$ group are the set of eight λ_a Gell-Mann matrices. The gauge field introduced here are the G_μ^a corresponding to the associate mediator of the strong field: the gluons g. Gluons carry 2 colour charges, leading to the 8 Gell-Mann matrices λ^a and 8 gauge vector

[2] This requires to isolate the terms in h^2 in Equation 2.18.

fields G^a_μ. The generators of the $SU(3)_C$ group do not commutate:

$$[\lambda^a, \lambda^b] = if^{abc}\lambda^c,$$

where f^{abc} are the $SU(3)_C$ structure constants. Similarly to the electromagnetic strength tensor, strength tensors are built from the gluon fields as

$$G^a_{\mu\nu} = \partial_\mu G^a_\nu - \partial_\nu G^a_\mu - g_s f^{abc} G^b_\mu G^c_\nu.$$

This leads to a total QCD Lagrangian of:

$$\mathcal{L}_{\text{QCD}} = -\frac{1}{4} G^a_{\mu\nu} G^{\mu\nu}_a + \sum_k \bar{\psi}_k (i\not{D} - m_k)\psi_k, \tag{2.21}$$

where ψ_k are the six quarks fields, one per flavour, transforming as an $SU(3)_C$ triplet with one component per colour quantum number. Cubic and quartir terms in the gluon gauge fields G^a_μ are included, introducing self interaction of gluons. The non-commutation of the $SU(3)_C$ generators means the $SU(3)_C$ part of the SM is non-Abelian group, and therefore a case of a Yang-Mills theory requiring this self-interacting gauge fields [28].

Like every coupling constant, α_s varies with energy. At low energies, the interaction is so strong that perturbative calculation breaks and the behaviour of *colour confinement* is observed: any attempt to isolate a quark requires so much energy that a quark-antiquark pair is spontaneously produced. The strength of this interaction explains it short propagation distance despite the fact its gluonic mediatiors are massless. At higher energies ~ 100 GeV, asympotic freedom and perturbative calculations are possible thanks to the smallness of the coupling strength. This typically requires higher-order corrections for calculation to converge, with some terms, such as quark self-energy loops, diverging to infinity. This type of diverge is called *ultraviolet divergence*, and it is fixed by renormalising fields and parameters so that the divergence are absorbed away. This correction requires two parameters to arbitrarily define the scale of the process: the *renormalisation scale* μ_R and *factorisation scale* μ_F [29]. The former is introduced to deal with ultraviolet divergences in the running of α_s. The latter addresses the so-called *infrared divergences* due to massless particles radiating further massless particles at low-energies by entering the parton distribution and fragmentation functions, introduced latter in this chapter.

Quarks must combine to form colour-less aggregate of matters called hadrons, with either a 2-quark system combining a quark and an antiquark into a *meson*, or a 3-quark system forming a *baryon* of which the proton (uud, $q_p = +1$) and the neutrons (udd, $q_n = 0$) are prime examples. The content of hadrons are called partons. The process leading to the neutralisation of the colour-charge of an asymptotically freed quark is called *hadronisation*.

2.1.5 Yukawa Interactions

In the QCD Lagrangian of Equation 2.21, the introduction of the mass terms for the quarks breaks the gauge invariance of the theory to $SU(2)_L \otimes U(1)_Y$ and must be therefore further modified. The masses of all fermions can however be included in the SM by introducing so-called Yukawa interactions between the Higgs and fermionic fields. Such terms are expressed as the following Lagrangian:

$$\mathcal{L}_{\text{Yukawa}} = -\frac{1}{\sqrt{2}} \sum_f y_f \bar{\psi}_f (v + h) \psi_f, \tag{2.22}$$

where ψ_f are the fermionic fields and the fundamental *Yukawa coupling* parameters $y_f = \sqrt{2}\frac{m_f}{v}$ for each flavour of fermion f are introduced as coupling strengths. Picking the v component in the sum in parenthesis gives a clear mass term to the fermion, with the h terms leading to Higgs-fermion interactions. The vacuum expectation value plays the role of a mass scale setting, with Yukawa coupling refining the specific strength for each fermionic species. For the quark sector, a futher correction is required as the weak interaction eigenstate basis is different from the mass basis in which physical particles are detected. The transformation from the mass eigenstates basis is specified by the complex unitary *Cabibb-Kobayashi-Maskawa (CKM) matrix* [30]:

$$V_{CKM} = \begin{pmatrix} V_{ud} & V_{us} & V_{ub} \\ V_{cd} & V_{cs} & V_{cb} \\ V_{td} & V_{ts} & V_{tb} \end{pmatrix}, \tag{2.23}$$

where the probability of a transition $p \to q$ is given by the magnitude $|V_{pq}|^2$ of the associated element. Through this quark mixing matrix, weak charged currents interaction allows for flavour-changing processes through charged currents interactions. The matrix is almost diagonal, hence transitions between quarks of the same generation are preferred (e.g., $t \to b$ preferred over $t \to d$).

2.2 Experimental Higgs Phenomenology

The experimental process to observe the Higgs bosons at the LHC is to collide two proton beams head-on, as described in Chapter 3. The accelerator is primarily designed to achieve these measurements, targeting different production and decay channels. Protons are composite particules (hadrons) and, at high energies, the main *hard-scattering* interaction is beteen components of the protons called *partons*. These partons are primarily the *valence* quarks (*uud* for a proton) but also contribution from *sea* quarks, as well as gluons and photons present within the hadron due to quantum interactions. In a pp collision, two interacting partons ab from within the protons undergo the main event $ab \to X$, with the activity from the rest of the protons assigned to the so-called *Underlying Event (UE)*. The cross-section for the global process $pp \to X$ is expressed using the factorisation theorem [29] as

$$\sigma_{pp\to X} = \sum_{a,b} \int_0^1 dx_a \int_0^1 dx_b f_a(x_a, \mu_F) f_b(x_b, \mu_F) \int d\sigma_{ab\to X} \left(x_a P_a, x_b P_b, \mu_R, \mu_F\right), \tag{2.24}$$

where $f_i(x_i, \mu_F)$ is the *Parton Distribution Function (PDF)* giving the probability for the parton i to undergo a hard scattering with momentum $p_i = x_i P_i$, where x_i is the fraction of the proton momentum P_i, μ_F is the previously introduced factorisation scale, as the PDF depend on the energy scale of the underlying process. The interaction is effectively split into two steps: picking up the interacting partons and their fraction of momentum, then considering the main $ab \to X$ event.

As introduces in the previous section, the Higgs couples to particles proportionally to their measured mass, which influences its production and decay modes. The leading order production modes of the Higgs boson are schematised in Figure 2.2. At the LHC with a centre of mass energy $\sqrt{s} = 13$ TeV and a Higgs boson mass $m_H = 125$ GeV, the main processes presented are by decreasing cross-sections:

- *Gluon-gluon fusion* ggF: two gluons fuse into a quark loop with a radiated Higgs boson: $pp \to H$. The quarks in the loop must couple to the Higgs, hence top t-quarks are preferred, followed by bottom b-quarks. The cross-section for this process is $\sigma_{ggF} = 48.6 \pm 2.4$ pb [31]. This process is favoured thanks the large contributions of gluons to the protonic PDFs.

- *Vector boson fusion* VBF: two off-shell vector bosons V (W or Z) radiated from partonic quarks fuse to form a Higgs $pp \to qqH$, with cross-section $\sigma_{VBF} = 3.77 \pm 0.09$ pb [31]. The

quarks leave a characteristic twin forward jets in the event.

- *Associated production with a vector boson VH*: the Higgs boson is produced in association with a vector boson V (W or Z): $pp \to VH$. This process is studied in greater detail in Chapter **??**, dedicated to an analysis of Higgs decaying to b- or c-quarks from this production mode. It has a cross-section of $\sigma_{VH} = 2.24 \pm 0.14$ pb [31]. Leptonic decays of the associated vector boson give clean event signatures.

- *Associated production with a quark pair $q\bar{q}H$*: an "open quark loop" is produced from a pair of partonic gluons, with a Higgs radiated $pp \to q\bar{q}H$. The dominating contributions come from the $t\bar{t}H$ associated production with cross-section $\sigma_{VH} = 0.51 \pm 0.04$ pb, followed by the $b\bar{b}H$ [31].

Figure 2.2: The leading order Feynman diagrams for Higgs production at the LHC, from left to right: gluon-gluon fusion (ggF), vector boson V fusion (VBF), associated production with a vector boson (VH), and associated production with a $q\bar{q}$ pair ($q\bar{q}H$).

The dependency of the Higgs boson production from proton-proton collisions at $\sqrt{s} = 13$ TeV are represented in the left of Figure 2.3 as a function of the Higgs boson mass m_H. The total width of the SM Higgs boson is $\Gamma_H = 4$ MeV, implying a short lifetime of τ_H 10^{-22} s and restricting measurement to its decay products. The decay branching ratios are displayed in the right of the figure, with decays to heavier particles favoured due to the proportionality of the Higgs coupling strength to the mass. Decays to the massless gluons g and photons γ are possible thanks to intermediate quark loops, similarly to ggF. Relative decay rates are quantified by their branching ratio BR as

$$BR(H \to X) = \frac{\Gamma(H \to X)}{\Gamma_H} \tag{2.25}$$

, where the total Higgs width Γ_H is the sum of all partial decay width $\Gamma(H \to X)$, for all possible X. The right of Figure 2.3 displays the SM Higgs branching ratio at $\sqrt{s} = 13$ TeV. The most likely Higgs decay mode is to a pair $b\bar{b}$ ($BR \approx 58$ %), followed by the decay to a WW pair ($BR \approx 21$ %), and the $c\bar{c}$ decay branching ratio is 2.9%. All the decay branching ratios are displayed in the right of the figure, with decays to heavier particles favoured due to the proportionality of the Higgs coupling strength to the mass.

Figure 2.3: The Standard Model production cross-sections from proton-proton collisions (left) and decay branching ratio (right) of the Higgs boson as a function of m_H, at $\sqrt{s} = 13$ TeV, from [31].

Figure 2.4: Ratio of observed to the SM predicted event rate for different combinations of Higgs boson production and decay mode, from [32]. The horizontal bars denote the 68% confidence interval, with grey bands showing the theory uncertainties on the SM cross-section × branching ratio predictions.

The W^+W^- and ZZ decays can only be achieved by a virtual off-shell Higgs bosons. The fermionic decays are hard to observe due to the larger multi-jet background in a hadron collider. The vector bosons and di-photons leptonic decays benefit from advantageous experimental conditions, being easier to identify thanks to the presence of leptons and suffering from less background contamination. For these reasons, the ATLAS and CMS experiment both observed a particle of mass $m_H = 125$ GeV with the same properties as the Higgs boson in 2012 by combining the $H \to \gamma\gamma$, $H \to ZZ \to \ell^+\ell^-\ell'^+\ell'^-$, and $H \to W^+W^- \to \ell^+\ell^-\nu\bar{\nu}$ channels [13, 14].

This opened the way to many additional Higgs measurements, with some of the most sensitive ones summarised in Figure 2.4. Decay modes to third generations particles (t, b, τ) have been observed, with the sensitivity rising for the second generation (μ). In all analyses, the Higgs boson is measured to be in remarkably consistent with the predictions from the SM [32].

CHAPTER 3

The LHC & ATLAS Experiment

Modern particle physics is at the edge of the technological reach of science. To discover the Higgs boson, an extremely complex infrastuctre is required to probe physics at the required scale. The LHC at Centre Européen pour la Recherche Nucléaire (CERN) is consequently the largest and most powerful particle accelerator and has held this title since its construction in 2008 [33]. It is easily ranks as one of the mox complex machine ever created. Protons are accelerated to up to 99.9999991% of the speed of light, in a giant 27 km long ring-shaped accelerator, burried 100 m below the surface of the French-Swiss border near Geneva. Superconducting magnets cooled down with liquid helium to 1.9 K steer this energetic beam thanks to powerful magnetic fields of 8.33 Tesla. The beams are made of bunches of particles, collided precisely at four specific interactions points where large detectors are built and operated by dedicated experiments: ATLAS [34], CMS [35], ALICE [36], and LHCb [37]. The first two are multipurpose experiments with overlapping physics programs, while ALICE and LHCb are respectively dedicated to heavy ion and heavy flavour physics. This chapter described the experimental setup of the LHC and the ATLAS experiment, focusing on proton-proton collisions and introducing relevant elements to the work presented in this thesis.

3.1 The Large Hadron Collider

Figure 3.1: The accelerator complex of CERN during Run 3 [38]. The LHC is the top dark blue ring.

The last machine in a complex multi-stage accelerator complex of CERN displayed in Figure 3.1, the LHC is capable of frontally colliding proton or heavy ion beams. These beams are made of bunches of particles, collided precisely at four interactions points where large detectors are built and operated by dedicated experiments. The life of a proton beam starts innocuously ionised hydrogen H^- gas, passed to a linear accelerator called LINAC4[1] to reach energies of 160 MeV. After stripping the ionised hydrogen of its two electrons thus leaving a bare proton, the next acceleration stage is in the Proton Synchroton Booster (BOOSTER) bringing the beam energy to 2 GeV. The protons are then passed to increasingly larger synchrotons: the Proton Synchrotron (PS) to reach energies of 26 GeV, followed by the Super Proton Synchrotron (SPS) to reach energies of 450 GeV. The beam is finally injected into the LHC in two different beamlines

[1] Before 2020, it was LINAC2.

circulating the proton in opposite directions. Superconducting dipole magnets reaching 8.33 T steer the highly-energetic beams, while complex geometries of magnets (quadrupoles, sextupoles, ...) are deployed to refine the bunch shape through focusing effects. Powerful radiofrequency cavities accelerate them to their final energy of to 6.5 TeV, giving a total *pp* centre of mass energy of $\sqrt{s} = 13$ TeV in Run 2.

The operation of the LHC is split into dedicated *runs* of data taking, separated by *shutdowns* to maintain or upgrade the infrastucture. Some key metrics from these runs from the point of view of the ATLAS experiment are displayed in Table 3.1

	Year	$\sqrt{s}$ [TeV]	$\langle\mu\rangle$	Luminosity $\mathcal{L}$ [cm^{-2}s^{-1}]	$\int\mathcal{L}$ [fb^{-1}]
Run 1	2010 - 2012	7-8	18	0.8×10^{34}	26.4
Run 2	2015 - 2018	13	34	1-2 $\times 10^{34}$	140.1
Run 3	2022 - 2025	13.6	50	2×10^{33}	65

Table 3.1: Metrics on the accerelarator performance of the LHC in the different runs of data taking. The reported values correspond to those recorded by the ATLAS experiment [39–41]. Numbers for the ongoing Run 3 are preliminary, with the integrated luminosity listed considering events recorded until July 2023. The number of interactions per bunch crossing averaged over each run is displayed as $\langle\mu\rangle$.

Run 2 operated at a larger centre of mass energy ($\sqrt{s}$) and higher average instantaneous luminosity $\mathcal{L}$ than Run 1, with Run 3 again pushing up the limits. The average instantaneous luminosity measures the quantity of data collected from the relation

$$\mathcal{L} = \frac{1}{\sigma}\frac{dN}{dt} \tag{3.1}$$

relating the event rate of a particular process to its cross-section σ. The instantaneous luminosity is a machine parameter: it depends on the design and the operation of the accelerator. It is calculated from

$$\mathcal{L} = \frac{N_1 N_2 N_b f}{4\pi\sigma_x\sigma_y} \tag{3.2}$$

where N_1 and N_2 are the number of protons in each bunch, N_b the number of bunches, f is the collider revolution frequency, and σ_x and σ_y are the x and y geometrical extension of the beam density distributions. The integrated luminosity $\int\mathcal{L}dt$ is a direct measure of the quantity of data collected over a certain period, often expressed in units of inverse *barn* b^{-1}, where b $= 10^{-28}$ m^2. For Run 2, the total luminosity recorded corresponds to 140.1 ± 1.2 fb^{-1}, with a small uncertainty of 0.83 % [40] thanks to a complex measurement involving luminosity-dedicated detectors such

as LUCID-2 [42]. Figure 3.2 shows the cumulative total integrated luminosity during Run 2, jointly with another important machine parameter: the average number of interactions per bunch crossing $\langle\mu\rangle$.

Figure 3.2: The ATLAS cumulative integrated luminosity delivered, recorded, and approved for physics (left) and the average pile-up distribution weighted by luminosity (right) during Run 2 [43]. The luminosities indicated correspond to an early calculation that was corrected in Ref. [40].

The main event during the collision of the proton bunches is the hard scatter where most of the energy transfer occurs. Other protons in the bunches do have softer interactions leading to background activity called *Pile-up (PU)*. Two types of pile-up are distinguished: *in-time PU* when the soft interaction is from protons in the same bunch as the hard scatter recorded, and *out-of-time PU* when the protons are from another bunch crossing. Bunches in the LHC are separated by a 25 ns delay, corresponding to a machine frequency of 40 MHz. To control the luminosity, the angle of attack of the beams can be tweaked so that their overlap on the impact point is modified. Having more frontal collisions leads to a larger overlap and a higher luminosity at the price of more PU.

3.2 The ATLAS Detector

The ATLAS Collaboration built and operates the eponymous cylindrically shaped multi-layered detector lying 100 m underground with a length of 45 long and a 26 m diameter [34], as presented in Figure 3.3. The experiment is constructed to probe a braod range of physical phenomena, as required from the general purpose of the Collaboration. Aiming to be as hermetic as possible, the detector wraps around the interaction point with the barrel forming the central part of the cylinder and the end-caps closing the geometry at its extremity. Essential characteristics of the design had to be met to manage the extremely high event rate, requiring fast response,

radiation-hard sensors, and state-of-the-art readout electronics in combination with good spatial and temporal resolution to disentangle the effect of pile-up.

Figure 3.3: Cut-away view of the ATLAS detector [44].

The coordinate system adopted in ATLAS is described in Figure 3.4: the x-axis point to the centre of the LHC ring, the y-axis point upwards, and the z-axis in the direction of the beamline (anti-clockwise from the top). The azimuthal angle ϕ is defined in the transverse plane $x - y$ and the polar angle θ is measured upwards from the beam-axis (along y). Therefore, the transverse momentum p_T of a particle is obtained from its momentum vector $\boldsymbol{p} = (p_x, p_y, p_z)$ as: $p_T = \boldsymbol{p} \sin\theta = \sqrt{p_x^2 + p_y^2}$. This is of particular interest as the momentum's longitudinal component p_z is not measurable due to the openings for the beamline and the partonic interaction carrying only a fraction of the original proton momentum. Only the transverse momentum can be correctly assessed, as the initial parton is mostly boosted in the direction of the beamline. The rapidity y of a particle, playing a crucial role in the special relativity, is expressed as

$$y = \frac{1}{2} \ln \left(\frac{E + p_z}{E - p_z} \right) \tag{3.3}$$

with E and p_L the particle's energy and longitudinal momentum. In the ultrarelativistic limit, when $p >> m$, the rest mass is negligible and $E \approx p$. In such cases, the rapidity y is well

approximated by the experimentally reconstructable pseudo-rapidity η that is expressed as

$$\eta = -\ln\left(\tan\frac{\theta}{2}\right). \tag{3.4}$$

Like the rapidity, $\Delta\eta$ is an invariant under Lorents boosts along the longitudinal z-axis. It is often combined with the azimuthal angular apperture $\Delta\phi$ to define the angular separation ΔR between two objets as

$$\Delta R = \sqrt{\Delta\phi^2 + \Delta\eta^2} = \sqrt{\Delta(\phi_2 - \phi_1)^2 + \Delta(\eta_2 - \eta_1)^2}. \tag{3.5}$$

Figure 3.4: The ATLAS coordinate system, from [45].

As depicted in Figure 3.3, ATLAS combines different systems into one precision machine. These subdetectors are able to measure information in the range $|\eta| < 2.5$, with some specialised subdetectors extending further. An essential component in the detector system are its two types of superconducting magnets. Four central solenoid magnets wrap the point of impact and generate a powerful 2 T magnetic field within the inner detectors along the z-axis, while toroidal magnets are placed externally on the barrel and the endcaps muon systems to generate a 3.5 T magnetic field deflecting these leptons in the η-direction. A q-charged particles of momentum p is deflected by a magnetic field B due to the Lorentz force, leading to a connexion between the radius of curvature R of the trajectory and the momentum p such as:

$$p_\perp = 0.3\, qBR\, [\mathrm{GeV}/c] \tag{3.6}$$

where $p_\perp$ is the magnitude of the momentum perpendicular to the magnetic field $\boldsymbol{B}$, and q is expressed in unit of proton charge. Therefore, from the measurement of the curvature, experimentalists are able to infer the component of the momentum transversial to their generated B.

Higher magnetic fields lead to larger curvature simplifying the measurement of R and improving the resolution of $p_{\perp}$.

The rest of this chapter reviews the different subdetectors of ATLAS and introduces some common reconstruction methods that are relevant to the work presented in this thesis.

3.2.1 The Inner Detector Tracker

The detector placed closest to the point of interaction is the Inner Detector (ID) [46]. This is a tracker covering the range $|\eta| < 2.5$ in a radius of 3 cm to 1 m, designed to record hits in silicon semiconductors or straw-tubes when charged particles fly through so that their trajectory or *track* can be reconstructed from the collected hits. The powerful 2 T magnetic field of the central solenoids enable this detector to measure both charge and momentum. The ID combines three subsystems, as represented in Figure 3.5.

Figure 3.5: The Inner Detector of ATLAS [44].

First, the high-granularity *Pixel Detector* covers the innermost region with three barrel and three endcap layers, for a total of 80 million sensitive semiconductor-based pixels [46, 47]. During Run 2, an additional *Insertable B-Layer (IBL)* was added at 33 mm from the centre, with 12 million pixels [48]. This detector gives robust and precise tracking performance and plays a major role in flavour tagging, as described in Chapter 5. Pixels are 50 $\mu\times$ 400 μm in the $R\phi \times z$, with a smaller 50 $\mu\times$ 250 μm for the IBL. The current resolution delivered is of 10 μm (67 μm) in the transverse $R\phi$ plane (z-direction) [49, 50].

The *Semiconductor Tracker (SCT)* is the next detector, constructed by arranging pairs of

silicon microstrips layers into modules assembled into 4 concentric barrel layers and 9 disks in each endcap [51, 52]. The resolution is typically of 17 μm in $R\phi$ and 580 μm in z [53].

The final system is the *Transition Radiation Tracker (TRT)*, a gas-based straw-tube tracker aiding track reconstruction by delivering numerous hits [54]. Approximately 300 000 drift tubes of a 4 mm diameter filled with a mélange of argon and xenon are arranged along the beamline in the barrel and radially in the endcaps. Each tube has a conducting wire at its centre and its surface is electrically charged, so that a charged particle flying through ionises the gas, leading to a measurable discharge. Polyethylene is place between the tubes to encourage the emission of transition radiation from relativistic particles proportionally to their Lorentz boost γ E/m. Consequentely, the TRT is used both for tracking and electron and pion identification, by reconstructing the mass of the charged particles from the amount of γ-radiation. For tracking, the position resolution provided is of 130 μm in the $R\phi$ plane for the barrel and the $z - \phi$ plane in the endcaps [55]. Altogether, the inverse momentum resolution of the ATLAS ID is

$$\sigma(1/p_T) = 0.36 \oplus \frac{13}{p_T \sin\theta} \mathrm{TeV}^{-1} \tag{3.7}$$

where $\oplus$ is the sum in quadrature [34]. This corresponds to a relative error of 0.01% for a track with $p_T \sim 500$ MeV, and 4% at a $p_T \sim 100$ GeV.

3.2.2 Electronic and Hadronic Calorimeters

Covering the $|\eta| < 4.9$ region, calorimeters collect the energy of all interacting particles, neutral and charged, except for the muons. They are separated into an electromagnetic calorimeter (ECAL) and a hadronic calorimeter (HCAL) covering jointly the $|\eta| < 3.2$ region and a forward calorimeter for the $3.2 < |\eta| < 4.9$ region, each interlaying specific layers of active and passive materials, as displayed in Figure 3.6 [**cite {TheATLASCollaboration_2008}**]. The passive material typically has a large atomic number to induce cascade of particles called *shower*. The active material, tpyically liquid argon (LAr) collects the energy from these showers through ionisation or scintillation light.

The ECAL is designed to collect the energy of electrons and photons, as well as contributing to the measurement of the energy of jets. The active material is LAr and absorbing plates of lead are used as passive material, to encourage bremsstreahlung $e \to e\gamma$ and pair production from photons $\gamma \to e^+e^-$. The ECAL have a depth of 22 X_0, where the units of *radiation length*

X_0 track the distance for an electron to have only $1/e$ its original energy. The energy resolution is parametrised into three terms for the sampling, noise, and a constant contribution, summed in quadrature as [56]

$$\frac{\sigma_E^{\text{ECAL}}}{E} = \frac{2.8\%}{\sqrt{E}} \oplus \frac{0.12[\text{GeV}]}{E} \oplus 0.3\%, \tag{3.8}$$

giving an energy resolution between ∼0.5% for 10 GeV electrons, and ∼ 0.6% for 60 GeV photons.

The HCAL is designed to capture the energy of hadronic shower, with LAr as active material for the endcap and forward calorimeters and scintillating plastic tiles for the barrel. As passive material, the endcaps use copper plates, the forward calorimeters use copper and tungsten, and the tile calorimeter in the barrel uses steel. The depth of the hadronic calorimeter is $\sim 10X$, where X is the nuclear interaction length tracking the average distance before a hadron interacts with a nucleus. The calorimeters collect the majority of the energy, with an HCAL resolution expressed as [56]

$$\frac{\sigma_E^{\text{HCAL}}}{E} = \frac{52\%}{\sqrt{E}} \oplus \frac{1.6[\text{GeV}]}{E} \oplus 3\%. \tag{3.9}$$

Figure 3.6: The calorimeter systems of ATLAS [44].

3.2.3 Muon Detection Systems

Muons require dedicated detection systems to be efficiently and precisely reconstructed in ATLAS. Their high-mass coupled with long lifetime at the energy studied mean they barely interact with the previously described system, only leaving a few hits in the inner detector and flying largely undisturbed through the calorimeters. For this reason, the outmost subdetectors of ATLAS are specially designed to interact with muons. The Muon Spectrometer (MS), shown in Figure 3.7, is a dedicated muon tracking system which also provides an efficient triggering hardware, as described later in this chapter.

Figure 3.7: The muon detectors of ATLAS [44].

The muon tracker is composed of drift tubes split between a barrel region for $|\eta| < 1.2$ and the endcaps $1.2 < |\eta| < 2.7$, with cathore strip chambers in the inner-layers of the endcaps to resist the high activity. The trigger system relies on resitive-plate chambers in the barrel and thin gap chamber in the end-caps. For muon momentum and charge measurements from the tracks, powerful superconducting toroidal magnets are used to deflect muons in the MS. The resolution on p_T is measured to be of $\sim 1.7\%$ (2.3%) for muons from J/ψ decays in the central (forward) region [57].

3.3 Operation and Reconstruction with the ATLAS Detector

For physics-quality data taking, the different subdetectors of ATLAS must be performing according to specifications. In operation, the event rate produced by the LHC in the heart of the ATLAS detector is of 40 MHz, from the 25 ns bunch-crossing. This unfortunately leads to a data generation rate that is too high for the electronics and computing resource available, requiring the Collaboration to design specific approaches to reduce the rate to a manageable level [58]. This is the task of the trigger system, which is described first in the section. Events that pass the trigger thresholds are stored and must be further processed to reconstruct the physics processes from the low-level measurements performed by the different subdetectors: this is the task of reconstruction, the last subject described in this chapter only for objects types relevant to the presented work. This latter step is performed thanks to the extensive ATLAS software [59, 60], exploiting the specific signatures of the different detected particles as schematised in Figure 3.8.

Figure 3.8: Schematics of different particles signatures in the ATLAS detector [61].

3.3.1 Trigger System

The ATLAS trigger system relies on a hierarchical approach to progressively reduce the data rate and select events deemed interesting for physics. Firstly, the *low-level* Level-1 (L1) trigger is built on fast electronic hardware accessing only coarse information to reduce the rate to 100 kHz

in ~ 2.5 μs. This is followed by the High Level Trigger (HLT) that runs on a farm of 40,000 Core Processing Units (CPUs) to implement a finer software-based selection, bringing the rate down to a 1.2 kHz or 1.2 GB/s for data storage [62]. In this process, gradually more complex information is accessed by dedicated readout and measurement systems. Some commonly used triggers are based on signatures of electrons, muons, missing transverse energy, and b-jets. Different trigger menus are designed by the Collaboration, with dedicated data taking period for each setup. Analyses are then able to select data from the whole collected data based on the specific signatures they search.

3.3.2 Low-Level Signatures

Low-level signatures are used in higher-level reconstruction processes to identify physics objects, such as electrons and jets. Three types are described here: the trajectory of charged particles called *tracks*, the construction of vertices, and the formation of calorimeter clusters.

Tracks are the reconstructed trajectory of charged particles through the detector from collected *hits*: localised energy deposits. With denser PU activity, the number of hits in a single event becomes significant turning track reconstruction into a challenging computational problem [63, 64]. The trajectories are curved thanks to the previously described superconducting magnets. From a set of hits, tracks are fitted inside-out [64]: clusters of three hits in the Pixel or SCT detectors are first identified as *seeds*, with and additional hits associated by a combinatorial Kalman Filter [65] based on compatibility criteria on the initial track. Hits can initially be shared by several tracks, with the ambiguity resolved later when the reconstructed tracks are ranked by quality and χ^2-fits are performed to quantify the best possible association, favouring high p_T tracks. The process is then extended to the TRT from the oustide-in, and followed by additional quality criteria such as requiring tracks to have a $p_T > 500$ MeV in $|\eta| < 2.5$, a minimum of 7 hits in the Pixex and SCT, at most one hole, and at most two shared hits. Tracks are parametrised by the longitudinal (along z) and transversal (in the $x-y$ plane) Impact Parameter (IP), respectively z_0 and d_0: these measure the distance from the primary vertex to the perigee of the track[2].

Tracks emenate from the decay of particles or radiated through physics processes. If a reconstructed charged particle is produced in the main event, its trajectory leads back to this special location called the *pv* [66]. If the particle is produced in another decay, the point of emenation

[2]The point of closest approach of the track with respect to the Primary Vertex (PV).

can sometimes be distinguished and is labelled *Secondary Vertex (SV)*. Identifying vertices is crucial for the physics programme of the Collaboration. The primary vertex is identified from a seed vertex first identified from the set of all well reconstructed tracks. The vertex position is then iteratively refined by removing tracks incompatible with the reconstructed vertex and re-fitting, until some quality criteria are met. Discarded tracks are then used to identify secondary or tertiary vertices. The primary vertex is the one with the largest sum of squared track-p_T.

The calorimeters of atlas are composed of many granular layers. So-called clusters are identified by grouping clusters with certain energy deposit measured, either from the *sliding-window* or *topo-clusters* algorithm [67]. The former one generates fixed-size rectangles by translating the window to maxime the transverse energy E_T measured. The latter clusters neighbouring cells based on a signal to noie criterion. The advantage of the sliding-window is that it is easier to calibrate and is hence used in electron, photon, and hadronic τ's reconstruction. Topo-clusters are robust against noise and therefore used for jet and missing transverse energy reconstruction.

3.3.3 Electrons

Electrons leave a rich signature in the ATLAS detector, in the ID and ECAL. For the central region $|\eta| < 2.5$, eletrons are identified and reconstructed with both subdetectors, while the forward region $2.5 < |\eta| < 4.9$ is only covered by the calorimeters with the shape of the shower used to discriminate electrons. Here, only centrally produced *prompt*-electrons are described, where *non-prompt*-electrons are not produced from the main physics process but through later decays or interactions with the detector itself. Photons, pions, and sometimes jets can be mistaken as electrons, with identification and isolation criteria derived to provide high purity and efficiency samples to analyses. Reconstruction relies on calorimeter clusters and track information. When relevant, tracks are matched to the clusters based on the expected energy loss. The track is also extrapolated to insure compatibilty with the cluster barycentre. The process is re-ran with more stringent conditions after refitting the tracks matched to clusters, so that energy loss is taken into account. A prompt-electron is required to have a track matched to the PV. The absence of a precision hits or a matched track leads to consider the calorimeter clusters as a photon deposit.

To further discriminate prompt-electrons from non-prompt electrons and photons, which act as a background due to photon conversion to e^+e^-, a likelihood-based identification algorithm built on a Multivariate Analysis (MVA) is deployed [68]. Parameters used include the number of

hits in each tracker layer, the IPs, a calorimeter cluster parameters. Several working points that are progressively more selected are defined on the discriminant, from *Very Loose*, *Loose*, *Medium*, to *Tight*. Prompt-electron candidates are required to be isolated from other tracks and other energy deposits, with specific criteria that are either ID- or calorimeters-based. In the former case, the sum of tracks p_T in a ΔR cone around the electron is used, while the latter analyses the sum of energy deposit in a calorimeter cone around the electron cluster. As further described in Chapter 6.8, the efficiency of the electron reconstruction, including identification and isolation, are computed by comparing the measured and simulated measurements of the $Z \to e^+e^-$ and $J\psi \to e^+e^-$. Photons are reconstructed similarly to electrons, with no tracks being matched for unconverted photons. Converted photons are allowed hits in the ID.

3.3.4 Muons

The MS is the main detector to reconstruct muons, with other subdetectors playing a smaller role: muons can indeed leave a track in the ID and some energy deposits in the calorimeters. These fragmented signatures in the subdetectors are combined to identify muon candidates. In the MS, tracks are constructed from a fit of the sucessive hits in the different chambers. *Combined muons* are defined by matching a track in the MS to a track in the ID, with additional information from the calorimeters. Prompt-muons are rejected from background-produced muons (such as in the decay of a b-hadron) by specific criteria targeting discrepancies in the $p_T$between the MS and ID. Increasingly selective working points are defined to identify muons as *Loose*, *Medium*, *Tight*, *High-*p_T. Isolation requirements are applied similarly to the electron case, either track- or calorimeter-based in a ΔR cone around the candidate muons. The calibration of muons is performed similarly to the electrons, on a $Z \to \mu^+\mu^-$ and $J\psi \to \mu^+\mu^-$ samples.

3.3.5 Jets

Quarks and gluons are the most commonly produced particles in a hadron collider setting. As described in Chapter 2, these particles carry colour-charges and undergo hadronisation when produced to neutralise their free colour. This complex process leads to a unique signature in the detector: a spray of particles emitted within the original parton direction called a *jet*. Electrically charged and neutral particles are contained within jets, with most of their energy deposited in the hadron calorimeters. These aggregated objects are therefore constructed by applying a clustering algorithms on tracks and/or calorimeter clusters, depending on the jet defintiion. The most notarious clustering method is the anti-k_T algorithm, thanks to the robustness of the defined jet

to collinear splitting or additional soft emissions [69]. The algorithm starts by considering high momentum object, to which softer objects are added thus growing the clusters. Two measures of distances d are used as parameters when considering whether to cluster object i and j:

$$d_{ij} = \min\left(\frac{1}{k_{Ti}^2}, \frac{1}{k_{Tj}^2}\right)\frac{\Delta R_{ij}^2}{R^2}, \tag{3.10}$$

and

$$d_{iB} = \frac{1}{k_{Ti}^2}. \tag{3.11}$$

The first distance, d_{ij}, combines the angular apperture ΔR_{ij} between i and j[3] with the transverse momentum k_T of the two objects and a fixed *radius* parameter R. This distance defines a jet radius limiting the size of the associated cone. It is compared to the second distance, d_{iB}, assessing the size of the already formed cluster i. If $d_{ij} < d_{iB}$, j is clustered with i, otherwise i is identified as a jet and removed from consideration. The algorithm proceeds after updating the distances, until all constituents are assigned a cluster. Typical radius values for ATLAS are $R = 0.4$ and $R = 1.0$, defining respectively small-R and large-R jets. The former is commonly used for quark and gluon jets, while the latter is employed by analysis to identify boosted heavy objects decay, such as W or Higgs bosons.

Jets can be constructed from tracks, calorimeter clusters, or both. In ATLAS, several types of jets are deployed, depending on the desired properties and the objects they are meant to represent. The following types are all reconstructed with the anti-k_T algorithm, clustering different objects:

- PFlow jets are reconstructed from particle-flow objects [70] with a radius $R = 0.4$. These objects combine tracking information from the ID to the calorimeter clusters, leading to a better energy resolution at low p_T and lower PU contamination after calibration [71].
- EMTopo jets are constructed from denoised topological calorimeters clusters called *topoclusters*, based on the per cell energy sigificance $S_{\text{cell}} = E_{\text{cell}}/\sigma_{\text{cell}}$, where E_{cell} is the energy and σ_{cell} the expected noise level in the cell [72]. The topoclusters are then used with the anti-k_T method with a small 0.4 or large 1.0 radius.
- Large-R jets are built from topological calorimeter clusters with a radius $R = 1.0$. These jets are applied trimming to remove the contribution from soft contamination, which is mainly due

[3] $\Delta R_{ij}^2 = (\eta_i - \eta_j)^2 + (\phi_i - \phi_j)^2$.

to PU and UE activity, leading to an improved mass resolution [73]

- Track-jets are constructed are based on the anti-k_T clustering algorithm with a variable radius R depending on jet p_T, such that the wide cone used at low p_T ($R \approx 0.4$) becomes narrower at high p_T ($R \approx 0.02$). They are typically matched as sub-jets of a large-R jet, to deploy the flavour tagging techniques described in Chapter 5.

PFlow and variable-radius jets are used to train the algorithms of Chapter 5, while EMTopo, large-R, and track jets are used in the analysis of Chapter 6. Jets can be assigned a flavour based on the presece of an original parton within a $\Delta R = 0.3$ cone around the jet axis. Experimentally, the flavour is often determined based on the hadrons associated to the jet, as described in detail in Chapter 5.

Jets benefit from extensive calibration to correct the reconstructed properties such as the mass, the energy, and the jet axis direction. In particular, corrections to account for PU activity and out-of-cone emissions and deposits are considered. Detector effects are also taken into account, such as differences between the electromagnetic and hadronic calorimters and leakage out of the active regions. The so-called *Jet Energy Scale (JES)* calibrations corrects the jets in successive steps [74]:

- *Origin*: the jet axis, initially constructed from the centre of ATLAS, is corrected to point from the PV, and the p_T is updated.
- *Pile-up*: both in-time and out-of-time pile up leave additional energy deposit in the calorimeters. This is substracted from the jet, from first an overall estimation based on the average pile-up and then from the actual number of interactions and vertices in the event.
- *Absolute*: absolute energy corrections dependent of the energy E and η are derived toto match the data energy scale to the particle-level energy scale of dedicated simulated samples.
- *Eta inter-calibration*: the detector is not homogenous and the forward regions measurements are typically less accurate. Corrections to these forward jets are applied based on central jets ($|\eta| < 1.4$).
- *Global sequential calibration*: energy leakage in the calorimeters is accounted for with a set of four-momentum Corrections based on five different observables tracking the shape and the energy of the jet.

- *In-situ calibration*: corrects any possible differences due to an incorrect description of the detector in the simulations by performing a fit to data a dedicated measurement of a well reconstructed object. Events from the following processes are used in increasing p_T scales: Z+jet events with the Z leptonically decaying, γ+jet, and QCD multi-jet events at high momentum.

The JES is parametrised by p_T, and dedicated uncertainties are derived for analyses to use. The *Jet Energy Resolution (JER)* is then defined as σ_{p_T}/p_T, and also calibrated with dedicated uncertainties derived from a fit to di-jet events [74]. Despite the JES correction procedure, PU jets can be still significant and a dedicated Jet Vertex Tagger (JVT) is used to reject this background [75]. This implements a 2D likelihood method built from tracking variables from which different selection criteria are derived as working points with specific PU jet rejections.

3.3.6 Taus

Taus make the heaviest generation of the charged lepton, with a mass of 1.8 GeV slightly higher than that of c-quarks [30]. Their lifetime is so short they mostly decay within the beampipe. They leptonically decay to neutrinos and a e or a μ 35% of the time, while the rest occurs hadronically. The leptonic decays are hard to disentangle from prompt electrons and muons. Hadronic decays however leave a signature reconstructed as a small-R jet in ATLAS. These hadronic taus are identified by a Recurrent Neural Networks (RNNs), to disentangle them from PU and QCD jets [76]. This discriminant is used with specific working points derived to define different efficiencies.

3.3.7 Missing Transverse Energy

Some physics objects do not interact with the ATLAS detector, such as neutrinos and hypothetical Beyond the Standard Model (BSM) particles. Consequentely, their presence is not detectable but can be inferred thanks to the negligible initial p_T of the two interacting partons in the proton-proton collision. Requiring the transverse energy and momentum to be balance to 0, the so-called *Missing Transverse Energy* E_T^{miss} or *MET* is calculated as the negative vectorial sum of all physics object transverse momentum

$$\boldsymbol{E}_T^{\text{miss}} = -\sum_{\text{hard}} \boldsymbol{p}_T - \sum_{\text{soft}} \boldsymbol{p}_T, \tag{3.12}$$

where the sum is decomposed into a *hard* term englobing all physics objects (photons, electrons, muons, taus, and jets) and a soft term including good-quality ID tracks associated to the PV but not to a physics object [77]. The performance of the reconstruction is measured by comparing simulations to data, with scale and resolution derived with uncertainties to be used by physics analyses.

CHAPTER 4

Machine Learning & Deep Learning

This chapter is dedicated to a review of relevant machine learning and deep learning methods in the context of High Energy Physics (HEP). As for other fields of science and technology, the recent advancements in artificial intelligence have introduced many useful techniques that can be leveraged in particle physics. Before starting the review, some definitions of the different terms are presented. This is followed by introducing the most commonly deployed approaches in particle physics: decision trees and deep neural networks. A final word on optimisation techniques is given at the conclusion of this chapter.

4.1 Definitions

4.1.1 Artificial Intelligence

Artificial Intelligence encapsulates any piece of software, any *program*, that aims to mimic an aspect of human intelligence, a non-exhaustive list of which includes:

- *Reasoning*, the ability to conduct logical thoughts and establish their validity.
- *Inferring*, the ability to connect logical statements to induce or deduce new statements.
- *Creativity*, the ability to generate new information.

- *Acting*, the ability to perform a task or to change the environment.

Artificial Intelligence (AI) research is a large field of investigation that studies these various aspects in many subjects such as robotics, Natural Language Processing (NLP), computer vision, generative modelling, and reinforcement learning. Artificial intelligence can be broadly separated into three levels according to the performance of the created system:

1. *Narrow Intelligence:* representing artificial intelligence capabilities on a specified problem, for which the underlying software is uniquely trained or designed. This field includes *reactive AI*, where a model is trained to output an optimal decision or prediction based on current conditions only and *limited memory AI*, where a model can draw knowledge from past data to build an internal understanding of the problem and take better-informed decisions later. An example of the former is the IBM chess player Deep Blue, while the latter is most famously demonstrated by OpenAI's GPT model powering its ChatGPT chat-box.

2. *General Intelligence:* representing an artificial intelligence capable of matching human problem-solving skills in multiple environments. In particular, this hypothetical setting would let a machine learn new tasks on its own and extrapolate from its pre-acquired knowledge, a process referred to as *transfer learning*. Such a model would have the ability to adopt and combine several of the traits of intelligence and to generalise its automated learning process to any situation.

3. *Super Intelligence:* describes a hypothetical type of intelligence able to exceed human abilities and exhibit independent control of thoughts.

Of these, currently, only the first one is accessible and widely deployed, with the second one being an ambitious focus of research for state-of-the-art research laboratories. The inception of reactive AI, the first approach attempted, can be found in the research into games in the 50s and 60s. This paradigm saw the rise of algorithms capable of searching for the optimal move in a large space of possible actions using *heuristics*, human-passed knowledge on useful features of the specific environment of the game. For example, in chess one can use the point system assigning arbitrary values to each piece to help decide what action is worthwhile (e.g., a queen is worth more than a simple pion). In this reactive approach, neither the rules of the game nor the decision process are learnt. The former is forced into the search logic and the latter is the outcome of the search process. Unfortunately, state space exploration of many realistic

problems scales asymptotically scales with the dimension of the input, quickly rendering reactive-based approaches unable to perform in a reasonable amount of time and computing power. Combined with the need for human-encoded insight into the problem, the potential of reactive AI is restricted to specific well-controlled settings with a high degree of human understanding and low environment complexity. Limited memory AI revolutionised the field by removing the need for complete human control of the data interpretation and state formulation, letting instead a well-crafted mathematical model abstract and represent the information internally. It opens the door to applications that are not otherwise realistically feasible, such as autonomous driving, speech recognition, seamless robotics, etc. For such problems, there does not exist a problem formulation that would be compatible with a reactive-based approach. The new paradigm of limited memory AI has also been observed to outperform reactive AI in all settings (e.g., in chess) and can be exploited in abstract scenarios where heuristics finding is impractical or intractable. For this reason, the focus of this chapter is on limited memory AI as exemplified by machine learning.

4.1.2 Machine Learning

Machine Learning (ML) underpins the field of narrow AI with limited memory capabilities. It represents a shift of paradigm in AI, moving away from human-declared logic-based rules written in a specific syntax, the wholemark of reactive AI that executes statements such as

$$\textit{If } x \textit{ happens, do } y,$$

for an input x and an output y. In limited memory AI, the state representation and update steps are encoded by mathematical models of both the dataspace ($\mathcal{D}$) and the learning process:

$$\forall\, x \in \mathcal{D},\ \text{do } f(x) = \hat{y};\ \text{update } f(x) \text{ given } (x, y),$$

where $\hat{y}$ is the prediction of the model. In this case, both the internal representation of the rules and the decision-making are underpinned by the trained mathematical model f. Essentially, two distinct steps are applied to the model underpinned by adjustable parameters:

1. *Inferring:* the model has to give its prediction $\hat{y}$ on a new data point x: $f(x) = \hat{y}$.

2. *Learning:* the parameters of the model are updated based on a specified training or fitting procedure, depending on whether the training will be progressively exposed to the data

points of a training dataset or directly exposed to the entirety of the set. The objective is to align the output of the model $\hat{y}$ with the expected behaviour y: given the couple (x, y), let $f(x) = \hat{y} \rightarrow y$ under training convergence - this means the model f has to become an accurate estimator of the label y. Note that not every ML model requires a declared target output y, a paradigm referred to as *unsupervised* learning as described later in this chapter.

The training process closely depends on the type of model being deployed. These can be broadly separated into two groups:

- *Classical machine learning:* covers models deploying specific algorithms to exploit the data in a pre-defined and fixed approach. They include linear regression, decision trees (Boosted Decision Trees (BDT) or Multivariate Analysis (MVA), random forest, ...), Support Vector Machine, logistic regression, kernel methods, k-Nearest Neighbours, ...

- *Deep Learning* (Deep Learning (DL)): these methods are based on a core logical module called the Artificial Neuron. This module is stacked into layers of given widths, meaning a given number of neurons, and several layers of such modules are then connected along depth. Within this category, the information flow through the network defines different types of DL.

DL is thus very much a part of ML, only constituting a specialised approach to building models on a core unit. Non-DL are often referred to as *classical* machine learning and still prove valuable in many applications thanks to their ease of use and their ability to be deployed in contexts with small dataset sizes. There are many possible tasks for a ML model:

- *Classification:* the task of assigning discrete variables, also called labels, to a datapoint: e.g., labelling a jet as a b-jet. The general case is multiclass, with n labels possible, and a particular common case is binary classification with $n = 2$.

- *Regression:* the task of predicting continuous variables of a datapoint: e.g., the momentum of the particle is 15 GeV/c.

- *Features extraction:* given a dataset with specific internal features, reconstruct or extract new features, e.g., given a set of tracks, reconstruct the secondary vertex. A special subcase of this category is embedding data points into a different space. The dimension of this final space can be smaller - the case of dimensionality reduction, e.g. performing a projection

to a subspace spanned by the principal eigenvectors (those of largest eigenvalues) with the Primary Component Analysis - or larger when embedding the data into a richer space.

- *Generation:* output samples from a distribution matching the training dataset distribution. For example, given a sample of 1 million $t\bar{t}$ events, sample 10 new data points from the underlying statistical model.

- *Anomaly detection:* identify and flag rare events in an unlabelled dataset.

To perform these different tasks, models are constructed following different paradigms of ML, divided mostly along the lines of the amount of human intervention [78]:

- *Supervised learning:* the data used for training is endowed with the information the model must predict. In the learning step, the model is therefore optimised to make predictions that closely align with the target. Classification and regression are the most common tasks that fall under this realm.

- *Unsupervised learning:* the data is not endowed with extra information the model must learn to predict but rather has underlying features that must be extracted. The model is therefore trained with an objective to optimise without explicit targets and should discover patterns and insights without any guidance. Generative models and clustering are prime examples.

- *Semi-supervised learning:* also called *weak supervision*, is a paradigm combining the supervised and unsupervised approaches. The model is mostly unsupervised but can benefit from some labelled cases or human input (a technique also named *active learning*). A prime example is an unsupervised clustering task followed by a classification of the formed clusters. This is particularly fruitful when the cost of labelling the data is expensive, as is the case with real human-linked data but thankfully not so in the case of particle physics data.

- *Self-supervised learning:* a machine instructs itself on what tasks should be learnt. The overarching goal of the model is loosely defined and the learning process should include superficial objectives to learn.

- *Reinforcement Learning:* this paradigm of ML is dedicated to the setting of a game-theoretic environment. An agent must explore and interact with its environment by choosing a specific action policy - a method by which the agent selects an action given its current

situation and expected reward. In Reinforcement Learning (RL), the objective is for the agent to learn how to construct the best policy to satisfy a reward function and therefore obtain the best outcome for itself.

These different settings can be explored in ML and more specifically in deep learning. The latter is widely considered to be the most performant one thanks to its ease of scaling and is the main focus of research at the moment.

4.1.3 Deep Learning

Deep Learning (DL) combines a family of methods predominantly derived in the 1980s that have quickly grown in popularity around 2010, with widely advertised results on competitive benchmark tasks in pattern recognition, as exemplified by the super-human performance of the *DanNet* model [79] based on Convolutional Neural Network (CNN) [80]. The basis of any DL method is the Artificial Neuron, a logical unit built to mimic the functioning of a human neuron. Several such neurons are then combined into layers of any number of neurons (the width of the layer) and the layers themselves are stacked into depth, with deeper layers receiving as input the output of earlier layers. Different DL models are constructed by modifying the structure of the layers - in particular, the input, output, and activation function used - and the transfer of information between neurons, be that between layers, depth-wise, or between neurons, width-wise. DL is specifically well-suited to the setting of the ATLAS experiment, because:

- Large datasets of both real and simulated data are available.
- Thanks to advanced Monte Carlo (MC)-based simulation programs of both the physics process and the detector reconstruction efforts, the simulated data points are faithful representations of the real data.
- The data and data model from which the data originates is well understood in physics, the former coming from measurements from well-calibrated detectors and the second from crafted theories of the field.
- The data exhibits reach features due to the collection of different detectors and the different scales of the underlying physics processes. The typical available representations span images, sequences, sets, and graphs, all four being the main data representations studied throughout the deep learning community.

Given how important this form of AI has become in all technological fields, this chapter is primarily dedicated to introducing some of its approaches most relevant to HEP.

4.2 Machine Learning Methods for Physics

High-energy physicists enjoy a special relationship with ML methods. Experimental particle physics largely relies on statistical analyses of complex and large datasets, be that simulated using MC methods or collected from sophisticated detector apparata. A typical physics analysis can be described as the combination of five main steps:

1. Data collection: real data is collected from a detector exposed to the underlying physics desired, e.g., at CERN placing and calibrating the ATLAS detector at an interaction point of the LHC to collect proton-proton collision data.

2. Simulated data is generated to match the condition of collection of the real data in terms of detector effects and operational conditions such as energy, PU and luminosity. This simulated data englobes the best of our current theoretical knowledge of the law of physics.

3. The detector of a modern particle physics experiment is a complex set of sub-detectors sensitive to different physical phenomena, as described in chapter REF. This low-level information collected by different devices must be processed and recombined to generate *objects*, aggregated information that often holds physical meaning. For example, from hits in the tracking detector, a track can be fitted and some of its physical properties, such as p_T, reconstructed. This task corresponds to a mapping *low*-level $\rightarrow$ *high*-level information to reconstruct interesting and physically meaningful features of the measured data.

4. An analysis strategy is established, with the objective to similarly restrict the full datasets of both simulated and real data to a portion of the data space that is most sensitive to the studied *signal* or process. The sensitivity aspect underlies the need to take into account limited knowledge of theoretical physics, limited precision of the apparata, limited statistics of both simulations and data collected, etc. To optimise the analysis, selection rules are derived based on physically accessible information, e.g., the centre-of-mass energy, the presence of leptons, the transverse momentum p_T, and other high-level objects reconstructed in the previous step.

5. With the optimally selected set of real and simulated data points, a statistical model is built to quantify the agreement of the measured data with the expectations from the theory under the conditions of the experiment. This is most often achieved through a profile likelihood computation, where the parameters of interest targeted by the analysis are measured to be those maximising the likelihood under the given measured data.

Modern advanced machine learning has the potential to improve all steps of this process:

1. The operational side of running the detector and the accelerators can benefit from RL methods for improved control of the different electronic devices and online data quality monitoring. Triggers, an essential component of the ATLAS experiment described later in this thesis, can be upgraded to use a sophisticated DL model running online thanks to a hardware backbone built on Field-Programmable Gate Array (FPGA) or Graphics Processing Unit (GPU).

2. Simulating a dataset using a MC approach is a computationally intensive task. Each event must pass through a selection of probabilistic steps, with only a simulated data point satisfying all requirements ending in the usable sample. This process can be sped up and optimised significantly with more advanced and refined MC methods, but the cost remains significant to generate datasets of sufficient statistics. Generative AI has the potential to accelerate this step by giving a statistical model that can be efficiently sampled. Generative Adversarial Network (GAN) and Variational Auto-Encoder (VAE) have been shown to perform the sampling step in a competitive amount of time. However, a key limitation of these statistical approaches is their limited ability to incorporate the sophisticated theoretical model required to simulate the data, and any discrepancy or non-closure introduces levels of disagreements that are counter-productive to the final objective of the physics analysis.

3. ML is particularly well-suited for the object reconstruction task. Broadly, ML-based method can offer scalable, efficient, and precise solutions for object reconstruction. Important examples in ATLAS are identifying particles in the detector (e.g., τ identification), reconstructing missing transverse energy (E_T^{miss}), and classifying heavy-flavour jets - as exemplified in the next chapter of this thesis dedicated to flavour tagging.

4. Historically, physicists have relied on a cut-base approach to select their data: they analyse each of the relevant variables for the physics problem at hand to try and identify the best features to use to restrict the data space through manually defined restrictions. For example, to measure an interesting physical quantity of the process of Z-bosons decaying to two charged leptons l^+l^-, restricting the invariant mass of the lepton pair $m_{l^+l^-}$ to lie close the Z-boson rest mass $m_Z \approx 91.19$ GeV$/c^2$ is beneficial, as most of the signal will be found there. Machine learning can entirely bypass this manual operation, learning directly from an appropriate set of signal and background datasets with given features a

transformation of the input data features to a discriminant optimising the separation of signal from background.

5. The likelihood function of the constructed statistical test, verifying the level of agreement between the real data and the theory through the simulated sample, can be directly learnt by a model given access to both sets. Furthermore, anomaly detection settings, such as those in the search for unknown resonances, can be derived using an ML model in an unsupervised setting, thereby automating the discovery process and requiring only real data.

One of the main focuses of this thesis can be broadly summarised as contributing to step 3 in the aforementioned list: developing DL-tools for improved object reconstruction. The analysis presented in the latter part of the text also introduces some classical ML techniques of data selection - corresponding to step 4. The rest of this chapter now addresses a more detailed review of the relevant ML methods.

4.2.1 Decision Trees

Decision Trees (DT), also called *Classification and Regression Trees* (CART), are the bread-and-butter of any data analysis. They are simple to train, give a good ground performance for both classification and regression tasks, and are white box models - meaning their decision process is easy to interpret. Underlying the model is a recursive partitioning approach of the input space [78]. Labelling a partition step as *node*, the tree structure emerges from a *root* state that is subsequently partitioned along different branches with one *leaf* per final region giving the outputs. The splits are done along features of the input space, and the method accepts both discrete categorical values (e.g., the label of a lepton as e, μ, τ) and continuous values (e.g., m_l). The following is a simple example of a classification tree outputting the predicted class as 0 or 1:

At each node, there is a learnt condition with x_i, x_j, x_k being continuous features of the dataset that are cut at the thresholds c_i, c_j, c_k and c_l is a categorical feature (e.g., is the lepton an electron). The leaves values are the output of the tree in different regions defined by the combination of successive selections - here a binary variable indicating a class. An example of a tree performing classification is shown in Figure 4.1, where a tree with two nodes can isolate most of the blue class from the red class with the region limited by green lines, corresponding to both conditions $x_1 \geq c_2$ and $x_2 \geq c_2$ being satisfied.

Figure 4.1: A binary classification problem with two features. A decision tree applies two successive cuts c_1 and c_2 to isolate most of the blue class from the red.

Finding the optimal set of partitions of a dataset is an NP-complete problem and therefore intractable for large datasets. To build a tree, a greedy approach must be adopted - meaning using a heuristic approach to find a satisfying solution. The most common approach is to successively choose the most optimal step at each stage with no guarantee to find a global optimum instead of a local one. The chosen split is selected based on a defined *cost* function as suggested in Equation **??**.

$$(j^*, t^*) = \arg \min_{j \in \{1,...,D\}, t \in T_j} \min \left(\text{cost}(\{x_i, y_i : x_{ij} \leq t\}) + \text{cost}(\{x_i, y_i : x_j > t\}) \right) \tag{4.1}$$

where T_j is the set of possible thresholds, and x_j and y_j are the features and labels (or regressive objectives). For categorical variables, the inequality $x_j >< t$ is converted into a value equality $x_j == t$. The *cost* function will depend on the objective of the tree, with the regression case typically using the Mean Squarred Error (MSE) error function

$$cost(D) = \sum_{i \in D} (y_i - \bar{y})^2,$$

and for a classification the loss would be one of the following:

- *Missclassification rate:* $\frac{1}{|D|} \sum_{i \in D} \mathbb{I}(y_i \neq \hat{y})$, where D is the data in the leaf of the tree and $\mathbb{I}$ is the identity function: $\mathbb{I}(x) = 1$ if x is True, else 0.
- *Statistical entropy:* defining the class-condition probability as $\pi_c = \frac{1}{|D|} \sum_{i \in D} \mathbb{I}((y_i) \neq c)$, the entropy over the (C) classes is defined in Equation 4.2

$$H(\vec{\pi}) = -\sum_{c=1}^{C} \pi_c \log \pi_c, \tag{4.2}$$

  with $\vec{\pi}$ being a vector $(\pi_1, \pi_2, ..., \pi_C)$ with the class-condition probabilities as components.
- *Information Gain:* an equivalent formulation to the entropy, where the gain in information that should be maximised is the relative change in entropy by adding a selection on feature X_j after the current selections:

$$\text{Gain}(X_j < t, Y) = H(Y) - H(Y|X_j < t)$$

- **Gini:** computes and minimises the expected error rate:

$$\sum_{c=1}^{C} \pi_c (1 - \pi_c). \tag{4.3}$$

The pseudocode algorithm to train a DT with the update rule of Equation 4.1 is summarised in Algorithm 1.

DT can overfit a dataset: the model may tune itself to specific features of the training set that are not generalisable to any set samples from the true original distribution. Regularisation

Algorithm 1 Recursive Procedure to Train a Decision Tree [78].

```
function FITTREE(node, D, depth)
    node.prediction ← mean({y_i : i ∈D})
    (j*, t*, D_L, D_R) ← split(D)
    if not worthSplitting(depth, cost, D_L, D_R) then
        return node
    else
        node.left ← FITTREE(node, D_L, depth + 1)
        node.right ← FITTREE(node, D_R, depth + 1)
        return node
    end if
end function
```

serves as an important step to contain this often undesirable behaviour. For trees, a natural procedure to avoid overtraining is to interrupt the growth of the tree when it is no longer worth doing so - a criterion that is hard to decide *a priori* - or to *prune* the tree - removing at the end nodes or branches that contribute little to the overall performance. A simpler way to regularise the performance by reducing the variance of the estimate of the model is to train several trees with different subsets of the data chosen randomly with replacement and aggregate the results into a single prediction. For example, for regression, one can take as output the average over each base learner:

$$y(x) = \frac{1}{N_l} \sum_{i=1}^{N_l} y_i(x),$$

for N_l base learner making prediction $y_i(x)$ given the input features x. For classification, the predicted class can be decided with majority voting by the group of base learners. This statistical technique of using an ensemble of predictors is referred to as *bagging* and *ensembling* and is common throughout Machine Learning. To further decorrelate the performance of the different predictors, these can be built on a subset of the input features and training datapoints, thereby forming a *random forest*.

4.2.2 Boosted Decision Trees

A popular extension to the simple decision trees approach is to introduce the concept of *boosting*, leading to models referred to as Boosted Decision Trees (BDT) or MVA in particle physics. Boosting is a greedy algorithm leveraging a weak learner or predictor (e.g., a DT) and applying it sequentially to weighted versions of the data, with a larger weight given to misclassified or miss-regressed datapoints, as per the use case. This method is hugely popular in data science, having earned the title *"best off-the-shelf classifier in the world"* [81]. Two particularly useful approaches are adaptive boosting (AdaBoost) [82] and gradient boosting [83], both combining

an ensemble of M weak learners f_i $(i = 1, ..., M)$ into a strong learner F:

$$F(x) = \sum_{i=1}^{M} f_i(x).$$

For the following discussion, the model is built using a training dataset $\{(x_1, y_1), ..., (x_N, y_N)\}$ with input vectors $x_i \in \{\mathbb{R} \otimes \mathbb{D}\}^d$ of d features that are real or discrete $(\mathbb{D})$ and $y \in \mathbb{R}^d$ is a d-dimension real vector that serves as output to be predicted by the model.

AdaBoost

AdaBoost combines the M weak learners f_i with adaptive weights α_i to improve the ensemble performance as

$$F(x) = \sum_{i=1}^{M} \alpha_i f_i(x),$$

where F is the boosted model, and the successive boosting stages $F_T = \sum_{i=1}^{T \leq M} \alpha_i f_i(x)$ define stronger and stronger boosted variants of the model that combine weak learners f_i with a weight $\alpha_i \in \mathbb{R}$. At each iteration m of the training process $(m = 1, ..., M)$, a weak learner f_m is fitted to the training set to minimise a loss function $L(y_i, F_m(x_i))$. The loss in AdaBoost is the exponential loss on the datapoints of Equation 4.4:

$$L(y, F_m(x)) = \sum_{i=1}^{N} \exp\left(-y_i F_m(x_i)\right) = \sum_{i=1}^{N} \exp\left(-y_i (F_{m-1}(x_i) + \alpha_m f_m(x_i))\right), \tag{4.4}$$

that the new weak learner $\alpha_m f_m$ added at step m has to minimise: $(\alpha_m, f_m) = \arg\min_{(\alpha_m, f_m)} L(y, F_m(x))$. The typical case for AdaBoost is binary classification with $y_i \in \{-1, 1\}$, but the algorithm is generalisable to other cases [78]. Equation 4.4 can be re-expressed as:

$$\sum_{i=1}^{N} w_{i,m} \exp\left(-\alpha_m y_i f_m(x_i)\right),$$

where $w_{i,m} = \exp\left(-y_i F_{m-1}(x_i)\right)$ is interpreted as a weight applied to the datapoint (x_i, y_i) indexed by i at step m proportionaly to the error of the current strong learner. One can show that the weak learner f_m minimising the optimisation objective at step m is the one minimising the miss-classified weights sum error ϵ_m of the reweighted version of the dataset with weights $w_{i,m}$ [78], where:

$$\epsilon_m = \sum_{i} w_{i,m} \mathbb{I}(y_i \neq f_m(x_i)).$$

For the first time step $m = 1$, these weights are initialised to $1/N$. The weights are then updated to

$$w_{i,m+1} = w_{i,m} e^{-\alpha_m y_i f_m(x_i)},$$

and renormalised so that $\sum_i w_{i,m+1} = 1$ before being assigned to each training sample i for the next step. The weak learner is combined with the strong learner using an optimal weight α_m found by minimising the loss L of the combined learner:

$$\alpha_m = \frac{1}{2} \log \frac{1 - \epsilon_m}{\epsilon_m},$$

giving the overall update rule of Equation 4.5:

$$F_m(x) = F_{m-1}(x) + \alpha_m f_m(x), \tag{4.5}$$

combining the new weak learners f_m with optimal weight α_m to the current strong learner. The AdaBoost algorithm is summarised in Algorithm 2.

Algorithm 2 Adaboost for Binary Classification with Exponential Loss [78]

Initialise weights: $w_{i,1} = \frac{1}{N}$, where N is the number of samples.
for $m = 1$ to M **do**
  Minimise $\epsilon_m = \sum_i w_{i,m} \mathbb{I}(y_i \neq f_m(x_i))$ on training set with weights $w_{i,m}$ to find $f_m(x)$.
  Compute $\alpha_m = \frac{1}{2} \log \left(\frac{1-\epsilon_m}{\epsilon_m} \right)$.
  Update weights: $w_{i,m+1} \leftarrow w_{i,m} \exp(-\alpha_m y_i f_m(x_i))$ and renormalised $\sum_i w_{i,m+1} = 1$.
end for
return $F(x) = \sum_{m=1}^{M} \alpha_m f_m(x)$

Gradient boosting

Gradient boosting is a generic approach which, contrary to AdaBoost, does not require a specific derivation for each loss function. The objective is to minimise the empirical risk, the expected value of the loss function L on the training set as shown in Equation 4.6

$$\hat{f} = \arg\min_f \mathbb{E}_{x,y} L(y, f(x)) \tag{4.6}$$

where $f(x) = (f(x_1), ..., f(x_N))$ is the output of the learner on the whole training set. As the name suggests, the approach leverages gradient descent to find the optimal $\hat{f}$. At step m, the

gradient of the loss L is evaluated at $f = f_{m-1}$ as

$$g_{i,m} = \left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f=f_{m-1}},$$

which is then used to update the learner with a step

$$f_m = f_{m-1} - \alpha_m g_m,$$

where $g_m = (g_{1,m}, g_{2,m}, ..., g_{N,m})$ and the step-length α_m is chosen to minimise the residual loss $L(y, f_{m-1} - \alpha_m g_m)$. This implements functional gradient descent and leads the model to fit the N datapoints of the set. This procedure naturally leads to overfitting the training set, an undesirable feature that is remedied by using a weak learner to approximate the negative gradient. In the specific case of gradient-boosted decision trees, at step m a decision tree $h_m(x)$ is fitted to the pseudo-residuals $g_{i,m}$. This DT h_m at step m defines J_m disjoint regions through its leaves with predictions b_{jm} in each $j = 1, ...J_m$ region:

$$h_m(x) = \sum_{j=1}^{J_m} b_{jm} \mathbf{1}_{R_{jm}}(x),$$

where $\mathbf{1}_{R_{jm}}(x)$ is the indicator function - equals to 1 when $x \in R_{jm}$ and 0 otherwise. The update to the model is chosen so that:

$$f_m(x) = f_{m-1} + \alpha_m h_m(x),$$

with α_m selected by minimising the empirical risk of the updated model:

$$\alpha_m = \arg\min_{\alpha} \sum_{i=1}^{N} L(y_i, f_{m-1}(x_i) + \alpha h_m(x_i)).$$

Algorithm 3 Gradient Boosting [78]

Initialise $f_0(x) = \arg\min_\alpha \sum_{i=1}^{N} L(y_i, \alpha)$
for $m = 1$ to M **do**
  Compute the gradient residual for each $i = 1, ..., N$: $g_{i,m} = -\left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f(x_i)=f_{m-1}(x_i)}$
  Train weak learner h_m on the dataset $\{(x_i, g_{i,m})\}_{i=1}^{N}$
  Compute α_m by minimising $\sum_{i=1}^{N} L(y_i, f_{m-1}(x_i) + \alpha h_m(x_i))$
  Update $f_m(x) = f_{m-1}(x) + \nu \alpha_m h_m(x)$
end for
return $f(x) = f_M(x)$

The full algorithm for Gradient Boosting is presented in Algorithm 3, where the update rule

is added a *learning rate* hyperparameter ν to introduce regularisation and reduce the risk of overfitting. By keeping $0 < \nu \leq 1$, the ability of the model to fully adapt to the training error is limited, thereby improving generalisation to unseen data. The price is a slower updating of the model and therefore a more demanding computational complexity. Further regularisation techniques are bootstrap aggregation - training each weak learner on a random subset of the data -, limiting the number of leaves, or more generally penalising model of larger complexity - pruning branches that do not reduce the loss by a minimal amount.

BDT resist better to overtraining thanks to the regularisation effect of boosting and the different techniques described in this section. Unfortunately, an undesirable feature of boosting is the loss of direct interpretability of the decision-making. This is however more than met by an appreciable gain in performance of the underlying model. An interesting property exhibited by all tree-based algorithms and many ML approaches is the ease of quantifying the impact of a specific feature on the result. This technique of *feature importance* assigns a score to each input feature, typically the Gini importance of Equation 4.3, quantifying the reduction in entropy obtained by adding a feature. Another popular technique taken from the field of cooperative game theory is the Shapley value, measuring the average marginal contribution of each feature to the objective function [84, 85].

Pros:

- *High Accuracy:* BDT easily achieve high accuracy in both classification and regression tasks, making them suitable for a wide range of applications.
- *Adaptability to Different Distributions:* Boosting algorithms, such as AdaBoost and Gradient Boosting, can adapt well to different types of data distributions and can capture non-linear relationships in the data.
- *Ensemble Learning:* The boosting technique leverages ensembling by combining multiple weak learners to create a strong learner, improving overall model performance.
- *Robustness to Overfitting:* Boosting helps mitigate overfitting, enhancing the generalisation of the model to unseen data.
- *Feature Importance:* BDT provide a measure of feature importance, aiding in feature selection and interpretability of the model.

- *Ease of Deployment:* BDT typically perform reasonably well out-of-the-box and are easy to train with few hyperparameters to optimise.

Cons:

- *Sensitivity to Noisy Data:* BDT can be sensitive to noisy data and outliers, potentially leading to overfitting.
- *Computational Complexity:* Training multiple weak learners sequentially can be computationally expensive, especially for large datasets or deep tree structures.
- *Parameter Tuning:* BDT still require some fine-tuning of the hyperparameters, such as learning rate and tree depth, for optimal performance.
- *Black Box Nature:* The ensemble nature of boosted decision trees make them somewhat of a black box, sacrificing the perfect interpretability of DT for the sake of performance.

4.2.3 Artificial Neurons

The Artificial Neuron or *perceptron*, as initially named by its inventor Frank Rosenblatt in his seminal 1958 paper [86], is the logical gate that underpins most DL models. Notably, the Multilayer Perceptron (MLP) or Deep Neural Network (DNN) are obtained by stacking layers of artificial neurons. Inspired by biological principles, the perceptron, as shown in Figure 4.2, accepts multiple inputs and gives as output 1 if the combination of inputs exceeds a certain modifiable threshold, otherwise giving 0. This combination accepts weights to scale the input which, remarkably, can be adjusted during a training phase if the output of the perceptron is incorrect. Artificial neurons are a direct generalisation of this very principle, with the output no longer being thresholded but applied a chosen function f after being added a learnable bias term b.

In this thesis, a *perceptron* shall refer to a single artificial neuron, which is equivalent to a logistic regression model. The interest in artificial neurons stems from a significant theoretical result: stacks of artificial neurons are *universal function approximator* [87, 88], as is shown in the next section on deep neural network. This theoretical result is built on a mathematically advantageous function choice for f: the sigmoid σ, defined in equation 4.7 and shown in Figure 4.3:

$$\sigma(x) = \frac{1}{1 + e^{-x}}. \tag{4.7}$$

Figure 4.2: Schematics of a perceptron or an artificial neuron: the inputs x_i $(i = 1, ..., N)$ are multiplied by learnable weights w_i, summed and added a bias b before being passed to an output function f.

Thanks to its property to map the set of real numbers to the [0, 1] range, this activation function is often used for numerical stability and to map some input to a probability distribution. A practical mathematical property of the sigmoid that is particularly relevant for DL is the ease to compute its derivative:

$$\sigma'(x) = \sigma(x)(1 - \sigma(x)).$$

The power of the artificial neuron comes from its ability to be efficiently combined with other such units into a well-ordered structure with powerful representation power. For an input $x \in \mathbb{R}^d$, a neuron individually applies an affine transformation $W_i^T x + b_i$, where $W_i \in \mathbb{R}^d$, $b_i \in \mathbb{R}$ are the weights and bias of the neuron i, that is passed through an activation function f for a total output of a single neuron $f(W_i^T x + b_i)$. Combining these operations, as shown in the next section, leads to a well-structured mathematical model that has the power to represent all well-behaved continuous functions.

4.2.4 Deep Neural Networks

A Deep Neural Network (DNN) - also called Multilayer Perceptron (MLP), Artificial Neural Network (ANN), feed-forward neural network, or sometimes even only Neural Network (NN) - is made by stacking layers of artificial neurons as shown in Figure 4.4. Each neuron in a layer receives as input the output of each neuron of the previous layer and connects to each neuron of the next layer. Layers of artificial neurons that are placed between the input and output ones are said to be *hidden layers*. The particularity of the design underpinning this architecture is that layers of neurons connect to all neurons of the next layers only, defining a feed-forward

computation graph from input x to output y. Mathematically, a single layers at depth i with m units given as input the previous layer of dim n at depth $i-1$ computes the transformation of Equation 4.8:

$$a^i = f^i\left(W^{i^T} a^{i-1} + b_i\right), \tag{4.8}$$

where $W^i \in \mathbb{R}^{m \times n}$ is the matrix of weight of layer i - one row per unit of layer i, one column per unit of layer $i-1$, $b_i \in \mathbb{R}^m$ is the vector of bias, one per unit of layer i, f^i is the activation function of layer i, and $a^{i-1} \in \mathbb{R}^n$ are the n activated output of layer $i-1$. Note that strictly speaking the activation can differ for the units of the same layer, but is often unique per layer to accelerate computations with vector operations.

Figure 4.4: A deep neural network with d layers of width m, l ..., k. Each artificial neuron, represented by a ball of darkening blue along the depth, computes an affine transformation of the input of the layer followed by an activation function. The input of the DNN is x and the output is y.

Neural networks implement a system computing combination of Equation 4.8. A powerful theoretical motivation for neural networks stems from the fact they are *Universal Function Approximators*. This theorem, or rather family of theorem depending on the type of activation f and network considered, establishes that neural networks built with appropriate activation functions and sufficient capacity are able to approximate most well-behaving functions. The most famous such theorem states [87, 88]:

Theorem: *Let $C([0,1]^n)$ denote the set of all continuous function $[0,1]^n \to \mathbb{R}$ and σ be any sigmoidal activation function. Then the finite sum $f(x) = \sum_{i=1}^{N} \alpha_i \sigma(w_i^T x + b_i)$ is dence in $C([0,1]^n)$. In other words, given any $f \in C([0,1]^n)$ and $\epsilon > 0$, $\exists$ a sum $G(x)$ of the above form for which:*

$$G(x) - f(x) < \epsilon \quad \forall x \in [0,1]^n.$$

The theorem essentially establishes that any function defined over the n-dimensional unit hypercube $[0,1]^n$ can be approximated by an arbitrarily wide neural network. This result only

requires σ to be sigmoidal or discriminatory in the sense:

$$\sigma(x) \to \begin{cases} 1 \text{ if } x \to \infty, \\ 0 \text{ if } x \to -\infty, \end{cases} \tag{4.9}$$

which, naturally, is satisfied by the sigmoid function. This result can be applied outside the unit hypercube by a homothetic transformation of the data space. Many flavours of DNN exist with different functions used. An important theoretical result is the requirement for the output function applied to the artificial neuron to possess some degree of *non-linearity*. A neural network with strictly linear functions can indeed be seen to collapse into a linear regression model. Such a function, when applied to the output of an artificial neuron, is said to *activate* it and is hence called *activation functions*. Interestingly, similar theorems were derived for the most important activation functions commonly used in DL, in particular for the Rectified Linear Units (ReLU) [89]. The most popular such functions, shown in Figure 4.5, are:

- The sigmoid function of Equation 4.7.
- The hyperbolic tangent function $tanh = \frac{e^x - e^{-x}}{e^x + e^{-x}}$.
- The ReLU function of Equation 4.10 [90]:

$$\text{ReLu}(x) = \max(0, x), \tag{4.10}$$

  note that the non-linearity here is strictly between positive and negative inputs, making this activation function one of the simplest that can be leveraged in DNN for universal function approximation. The derivative of this function is indeed particularly easy to compute. A generalisation of ReLU called leakyReLU introduces a linear function in the negative range as leakyReLu $= \max(\alpha x, x)$, with $\alpha \in [0, 1[$.

- The Exponential Linear Unit (ELU) function, shown in Equation 4.11, modifies the Leakage ReLU in the negative domain while keeping the saturation property:

$$\text{Elu}(x) = \begin{cases} x \quad \text{if } x >= 0, \\ \alpha(e^x - 1) \text{ otherwise}, \end{cases} \tag{4.11}$$

  where the hyperparameter $\alpha > 0$.

- The softmax function. For an $x \in \mathbb{R}^n$, the softmax return a vector $\text{softmax}(x) = [..., \frac{e^{x_i}}{Z}, ...]$

(for $i = 1, ..., N$), where $Z = \sum_i e^{x_i}$. For a 2-dimensional x, the softmax is equivalent to the sigmoid function. In n-dimension, it maps each entry of x to the range $[0, 1]$ and guarantees $\sum_i \text{softmax}(x)_i = 1$. The softmax is therefore helpful to define probability distributions over multidimensional outputs.

Many more functions have been defined in the field of AI, however, none offers a distinctive advantage over the ones listed here. The sigmoid is no longer the choice of reference, due to its tendency to quickly saturate - meaning its gradient for large positive or negative input values vanishes in the sense that it tends to 0. The hyperbolic tangent *tanh* offers larger gradients thanks to its [-1, 1]range with steeper curvature and is often the activation of choice for autoregressive architecture, such as the RNN. The ReLU function is the most widely used activation function as its derivative is extremely efficient to compute - 1 for positive inputs and 0 for negative ones - and does not suffer from vanishing gradients for positive values. Its fixed 0-value for negative input is a double edge sword: on one side, it helps the network regularise itself by deactivating neurons, on the other it could let some neurons be stuck in *off*-mode if the weights are such that its pre-activation value is always negative for the input data. For this reason, it is important to correctly initialise the weights of a DNN. An easy workaround for this limitation is the leakyReLu and the Elu activations, which both introduce some leakage in the negative value as controlled by a parameter α.

Figure 4.5: The most common non-linear activations used in deep learning.

While the Universal Function Approximator theorem gives a powerful endorsement of DL networks, it does not state how to derive such a network to fit a specific function. The task of choosing the right architecture depthwise and widthwise and the correct weights and biases must be approximated by a learning strategy updating the weights and biases to reduce the error between the output and a given target. Many other data models are also Universal Function Approximators and what sets apart DL-like models that are built on artificial neurons is the simplicity of the procedure to update their weights: thanks to their nice computational structure, under a reasonable choice of activation functions for each layer, the network is *differentiable*.

This means a gradient can be computed on a function measuring the difference between the target y and the output $\hat{y}$ - often called a loss function when it should be minimised or a reward function when it is to be maximised - and *backpropagated* across each neuron of each layer to give a local update to be applied to each weight and bias. The recent rise of DL in AI can be traced back to improvements in making this backpropagation of the gradients with publically available software, such as PyTorch [91] and TensorFlow [92], implementing efficient algorithms to perform this essential step.

There are two main difficulties encountered when optimising a neural network: the non-convexity of the objective function means saddle points and local minima are typically abundant and the computational complexity due to the large number of parameters makes a single update using a large dataset expensive. The large number of parameters typically implemented by neural networks requires a large dataset to correctly assign values to the parameters without suffering from overtraining - the property to only correctly describe the data of the training set without being able to generalise the performance to data not seen during the training phase. To circumvent these problems, the *backpropagation* algorithm of Algorithm 4 is used [93].

Algorithm 4 Backpropagation Algorithm

```
function UPDATE(x, y, N, L)
    Forward step: propagate input x through network N to get prediction ŷ
    Loss: compute the loss or reward of N: L(y, ŷ)
    while ∃ a layer with no gradients do
        Take the right-most layer required a gradient
        Take input of layer to its right
        Using the chain rule of calculus, propagate the gradient of the next layer to that of the
current layer
        Store the gradient at the layer
    end while
end function
```

In summary, the backpropagation algorithm serves as an effective way to compute

$$\frac{\partial \mathcal{L}(\theta)}{x, \partial \theta} = \sum_{i=1}^{N} \frac{\partial l_i(x_i, \theta)}{\partial \theta},$$

where θ encapsulates all parameters of the model and there are N datapoints x, with a per datapoint x_i loss of l_i. The backpropagation algorithm starts with a forward pass through the network:

$$x \xrightarrow{\times W^1 + b^1} z^1 \xrightarrow{f^1()} a^1 \xrightarrow{\times W^2 + b^2} z^2 \xrightarrow{f^2()} a^2 \xrightarrow{\times W^3 + b^3} \ldots \xrightarrow{\times W^d + b^d} a^d = \hat{y},$$

the loss is then computed $\mathcal{L}(y, \hat{y}|x, W^1, b^1, W^2, b^2, ..., W^d, b^d)$ before computing the gradient of each layer by starting from the output:

$$\text{grad}_{W^d,b^d} = \nabla_{W^d,b^d}\mathcal{L} \rightarrow \text{grad}_{W^{d-1},b^{d-1}} = \nabla_{W^{d-1},b^{d-1}}\text{grad}_{W^d,b^d} \rightarrow ... \rightarrow \nabla_{W^1,b^1}\text{grad}_{W^2},$$

where the operation $\text{grad}_{W^{d-1},b^{d-1}} = \nabla_{W^{d-1},b^{d-1}}\text{grad}_{W^d,b^d}$ implies a use of the chain rule to obtain the local gradient, based on information already obtained. In the context of DL, the chain rule of multivariate calculus offers a transformation:

$$\frac{\partial h}{\partial x} = \frac{\partial h}{\partial z}\frac{\partial z}{\partial x}, \tag{4.12}$$

for a function $h : \mathbb{R}^n \rightarrow \mathbb{R}^m$, composed by two functions $h = g \circ f$ defined as $f : \mathbb{R}^n \rightarrow \mathbb{R}^k$, $g : \mathbb{R}^k \rightarrow \mathbb{R}^m$, with $x \in \mathbb{R}^n$ and $z = f(x) \in \mathbb{R}^k$. Two types of updates are necessary:

- For a layer l, the activation can be unpacked:

$$\frac{\partial l}{\partial z^l} = \frac{\partial l}{\partial a^l}\frac{\partial a^l}{\partial z^l}$$

- For a layer l having access to next layer $l+1$ ($\frac{\partial l}{\partial z^{l+1}}$):

$$\frac{\partial l}{\partial z^l} = \frac{\partial l}{\partial z^{l+1}}\frac{\partial z^{l+1}}{\partial z^l} = \frac{\partial l}{\partial z^{l+1}}\frac{\partial z^{l+1}}{\partial a^l}\frac{\partial a^l}{\partial z^l},$$

  but since $z^{l+1} = W^{l+1}a^l + b^{l+1}$, this reduces to:

$$\frac{\partial l}{\partial z^l} = \frac{\partial l}{\partial z^{l+1}}W^{l+1}\frac{\partial a^l}{\partial z^l},$$

- Combining the two above, we can compute $\frac{\partial l}{\partial z^l}$ for each layer in backward order. One thus now requires the derivative with respect to the per-layer weights:

$$\frac{\partial l}{\partial W^l} = (a^{l-1}\frac{\partial l}{\partial z^l})^T,$$

  for the weight matrix - an outer product between a row vector $\frac{\partial l}{\partial z^l}$ and a column vector a^{l-1}. For the bias:

$$\frac{\partial l}{\partial b^l} = \frac{\partial l}{\partial z^l},$$

  the result is directly obtained from the row vector.

Once all the gradients are locally available, the next step is to update all the parameters to reduce the loss by taking a step in the direction opposite the gradient, giving the largest reduction in loss. For example, for a specific parameter w_{ij} at training step $t+1$:

$$w_{ij}^{T=t+1} \leftarrow w_{ij}^{T=t} - lr \times \text{grad}\left[w_{ij}^{T=t}\right], \tag{4.13}$$

where the *learning rate* lr controls how large a step is taken in the opposite direction of the gradient. Since the gradient of the earlier layers will be derived from the gradient of later layers, the gradients need to respect some numerical stability to avoid the risk of vanishing gradient ($\rightarrow 0$) or exploding gradients ($\rightarrow \infty$). This requires some care in the architecture choice and might motivate the use of a specific activation function over another. Note that strictly speaking, the activation function needs not be continuously differentiable (C^1), as the existence of a right or left derivative is sufficient, hence the function should at least be continuous (C^0). Concerning the loss function $\mathcal{L}$, there is a lot of freedom in how to set it up - although it should be differentiable - and some typical choices are:

- The cross-entropy loss function - also called the logistic loss - is based on the notion of entropy, as defined in Equation 4.2, and is often used to assign probabilities in a classification with $c \in C$ classes:

  $$-y_i \log \hat{y}_i,$$

  where y_i is the real label of the datapoint and $\hat{y} \in [0,1]^C$ is the model prediction, respecting $\sum_i \hat{y}_i = 1$. Given the requirements of the output, it is typically combined with a softmax.

- the Mean Squared Error (MSE):

  $$\frac{1}{N}\sum_i^N (y_i - \hat{y}_i)^2.$$

- the Mean Absolute Error (MAE):

  $$\frac{1}{N}\sum_i^N |y_i - \hat{y}_i|.$$

To regularise the model, it is common to add terms to the loss function called *regularisers* to restrict the model weights to a certain size. This can be achieved by adding to the loss $\mathcal{L}$ an

L2-penalty

$$\lambda \sum_i w_i^2,$$

on the sum of the squared values of the weights, or an L1-penalty

$$\lambda \sum_i |w_i|,$$

where this last approach using the absolute value has the nice additional feature to make the network sparse - meaning to set un-necessary weights at 0. The amount of regularisation is controlled by the hyperparameter λ. Further regularisation can be obtained by randomly dropping out some connexions of the network during training, a technique called *dropout* and controlled by the dropout probability p to include an artificial neuron or not.

Pros:

- *Universal Function Approximators:* Feedforward networks, particularly deep ones, are known to be universal function approximators. They can approximate any continuous function to arbitrary precision given sufficient hidden units and appropriate activation functions.
- *Flexible Architecture:* The architecture of feedforward networks is flexible, allowing for customization in terms of the number of layers, the number of neurons in each layer, and the choice of activation functions. This flexibility makes them suitable for various tasks.
- *Feature Learning:* Feedforward networks can automatically learn hierarchical representations of features from the input data. This is beneficial for tasks where the underlying patterns are complex and not easily captured by handcrafted features.
- *Non-linearity Handling:* By introducing non-linear activation functions, feedforward networks can capture and model non-linear relationships in data, enabling them to solve more complex problems compared to linear models.
- *Availability of Optimisation Techniques:* Various optimisation techniques, such as gradient descent and its variants, are well-suited for training feedforward networks. This makes it possible to efficiently update the network weights during the training process.

Cons:

- *Limited Modeling of Sequential and Geometrical Data:* Feedforward networks are not naturally designed for handling sequential data and temporal dependencies nor images. RNN, CNN, and Transformer architectures are often preferred for tasks involving such information.

- *Fixed Input Size:* Traditional feedforward networks have a fixed input size. While techniques like padding or resizing can be employed, they might not be suitable for handling inputs of varying lengths in tasks like natural language processing.

- *Lack of Memory:* Feedforward networks do not have an inherent memory mechanism, which can be a limitation when dealing with tasks requiring the model to retain information over a sequence of inputs.

- *Vanishing and Exploding Gradients:* Training deep feedforward networks can be challenging due to the vanishing and exploding gradient problems. Gradients may become too small or too large during backpropagation, leading to difficulties in learning deep representations.

- *Need for Sufficient Labelled Data:* Feedforward networks often require a large amount of labelled data for effective training. In domains where labelled data is scarce, the performance may be limited.

This section introduces deep neural networks, the simplest feed-forward architecture constituted of artificial neurons stacked into layers to generate an output y based on an input x. There are many refinements to this base architecture, and some are explored in the next sections.

4.2.5 Recurrent Neural Networks

The first modification to the DNN considered in this thesis are Recurrent Neural Networks (RNN). These models were derived to work with sequences, such as occurring in natural language processing. The main change from the DNN architecture is in the way information is passed through the network: RNN are autoregressive models. The information flow is bidirectional: the computation sequentially processes through the same structure the input at a given step with the output of the prior step. The advantage of this representation is that this cyclical flow can be unfolded into a direct acyclical computational graph that, for a given sequence length, is equivalent to a DNN pass but dynamically adapts to variable length inputs. Figure 4.6 presents the structure of an RNN-based network as well as its unfolding. The input x is a sequence of N tokens, and the length of different inputs x_i in the dataset can vary. The mathematical structures

implemented by this architecture to generate an output y of length equal to the input is that of Equation 4.14 for the step t ($t = 1, ..., N$):

$$y_t = W(h_t) = W(V(x_t) + U(h_{t-1})), \tag{4.14}$$

where U, V, and W are three, potentially different, DNNs mapping taking at timestep t respectively the previous hidden state h_{t-1}, the current input token x_t and the new hidden state $h_t = V(x_t) + U(h_{t-1})$. The initial hidden state h_0 is usually initiated from a special mapping from the whole input x. An interesting feature of such a network is its ability to build an internal memory of previous inputs up to a timestep T thanks to the chain of hidden states $h_{t<T}$. To avoid having too large (exploding) or small (vanishing) gradients, which would not correctly update the weights of the model during gradient descent, the *tanh* function is often used as activation in RNN thanks to its smooth distribution and limitation to the range $[-1, 1]$. Unfortunately, due to the network relying on repetitive multiplications of numbers in the range $[-1, 1]$, the effect of much earlier timestep ($h_{t<<T}$) can be lost when processing later input at T. This process is referred to as *memory loss*. This undesired property is remedied with several architectural modifications to RNN that improve their operational memory, of which the two main ones are Long-Short Term Memory (LSTM) and Gated Recurrent Unit (GRU).

Figure 4.6: A recurrent neural network, using 3 feed-forward neural networks (DNN) U, V, W, to map the input sequence $x = [x_1, x_2, ..., x_N]$ to the output $y = [y_1, y_2, ..., y_N]$ using the internal hidden state h^t evolving for each timestep t. h_0 would typically be obtained by a mapping of the whole input sequence x.

Long-Short Term Memory

As shown in Figure 4.7, Long-Short Term Memory (LSTM) cells implement a specific architecture to propagate information along the sequence, with the introduction of a new control state c [94].

Three gates covering the forget, the input, and the output regulate the flow of information from the cell. In particular, the forget gate F decides what information to keep from prior states, by multiplying these values by a factor 1 and discards the rest through a multiplication by a factor close to 0. The input gate I is tasked with creating the new internal state of the cell and what information to store in it. Finally, the output gate O decides what information in the cell should be brought to the output. This selectivity of the LSTM cell to decide what to use from memory, what to keep in memory, and what to output gives this architecture a far-improved memory for long sequences which results in a much-improved efficiency.

Figure 4.7: An LSTM cell. Arrows and lines that merge imply concatenation of the inputs, the $\times$, $+$, and tanh are element-wise operations and the σ are different layered transformations (1-layer feed-forward network). F_t is the forget gate of the memory cell c, I_t the input gate, and O_t the output gate.

Gated Recurrent Unit

The Gated Recurrent Unit (GRU) is another advantageous design to improve the memory of a recurrent-like network without using as many parameters as an LSTM [95]. There is no output gate and only one internal hidden state h_t is required. This architecture comes in different versions, with the fully gated version presented in Figure 4.8. It only requires two gates: the *update gate* Z and the *reset gate* R. The former lets the model decide how much of the past information should be kept and the latter is used to forget past information.

RNNs and their modification have been designed for ordered sequence analysis and have had great results in such settings. Ordered sequences are natural in language analysis, where a sentence such as *"the boy hit the car"* is meaningfully different to its permutation *"the car hit the boy"*. The choice to sequentially analyse the tokens of a sequence with memory lets RNN-based models operate similarly to Turing Machines, endowing them with the powerful representational

Figure 4.8: A Gated Recurrent Unit cell (fully-gated version). Arrows and lines that merge imply concatenation of the inputs, the $\times$, $+$, tanh, and "1 - " are element-wise operations - the last one outputting the vector 1 minus its inputs - and the σ are two different layered transformations (1-layer feed-forward network) with sigmoid-like activation function. Z_t is the update gate and R_t the reset gate.

flexibility of Universal Turing Machine [96]. A significant drawback however is the impossibility of fully parallelising the analysis of the sequence due to its strict ordering, making RNN expensive models to train even on highly parallelised hardware. The main motivation behind the Transformer design, introduced in section 4.2.8, is to fix this crucial weakness.

Pros:

- *Sequential Processing:* RNNs are designed to handle sequential data, making them suitable for tasks with temporal dependencies.
- *Flexibility:* RNNs can operate on input sequences of variable length.
- *Memory:* RNNs have a memory mechanism that allows them to retain information about previous inputs.

Cons:

- *Vanishing and Exploding Gradients:* Training deep RNNs can suffer from vanishing and exploding gradient problems, affecting the learning of long-term dependencies.
- *Limited Short-Term Memory:* Traditional RNNs struggle to capture long-range dependencies due to their limited short-term memory.

- *Complexity:* While LSTM and GRU architectures can mitigate the two points above, the cost is a more complex architecture that is harder to train and requires more resources.
- *Interpretability:* The internal workings of a RNN is challenging to interpret.

4.2.6 Convolutional Neural Networks

Convolutional Neural Networks (CNN) [80, 97] have emerged as a powerful class of deep learning models that are particularly effective in computer vision tasks, including image and video analysis. The architecture of CNNs is roughly based on human visual processing. It consists of convolutional layers - implementing the fundamental convolution operation-, pooling layers, and feed-forward layers (DNN). This architecture, presented in Figure 4.9, enables CNN to automatically learn hierarchical representations of features while respecting properties of image-based data: spaciality (pixels have a position), locality (pixel share information within their neighbourhood), and symmetries (rotational, reflection, translation, etc.).

Figure 4.9: A layer of a convolutional neural network, implementing a convolution with 4 kernels followed by a pooling operation. This design can be stacked to create deep architecture, and combined with a feed-forward neural network, after flattening the output at some depth, before reaching the final loss function layer. In this example, the receptive field size is 4 and there are 4 different kernels used to analyse an input image of size 7×7.

The functioning of CNN involves the use of convolutional layers to extract local patterns and features from input data. Convolutional operations are applied to the input data by multiplying entry-by-entry a learnable *kernel* or *filter*, represented by a matrix of weights of smaller dimension than the total image size, with an equal size subpart of the input and applying an activation function. The size of the filter or kernel restricts the processing of the input to a given *receptive field* dimension, and this window is passed over the full input image by moving it by a defined *stride* length. Pooling layers are then used to reduce spatial dimensions and retain important features. This process can be parallelised for an image with multiple channels: e.g., the pixels of a coloured image are represented by a tuple of 3 values (for example, using the RGB coding). For classification and regression, a CNN based model would typically stack several convolutional

and pooling layers before leading to a fully connected neural network "flattening" the last representation to make predictions. *Flattening* refers to the process of transforming the input image, represented as a matrix $\mathbb{R}^x \times \mathbb{R}^y$, into a vector $\mathbb{R}^{x \times y}$. CNN-based models, such as AlexNet [97] and ResNet [98], have demonstrated state-of-the-art performance in various computer vision tasks.

A main advantage of the convolution operation on an image of size $x \times y$ is the reduction of the number of artificial neurons required to process the image, which helps to regularise the network.

- A DNN given the flattened image would require $x \times y$ neurons.
- A CNN with k kernels of size $\alpha \times \beta$ would require $k \times \alpha \times \beta$ artificial neurons, that would be applied to different subparts of the image.

For example, for an image of size 100×100, a DNN would require 10,000 weights while a CNN could process the image with only 25 if a single kernel of size 5×5 is used. Typical pooling functions are the *maxPooling* or the *sumPooling*, which, respectively, take the largest element or the sum in each window of their input, with specific hyperparameters governing the size and the movements of the window.

Pros:

- *Feature Learning:* CNNs automatically learn hierarchical representations of features on multidimensional data, reducing the need for manual feature engineering.
- *Spatial Hierarchies:* Convolutional and pooling layers enable the model to capture spatial hierarchies in the input data while respecting the properties of images, namely locality - how close pixels are - and symmetries - invariance to translation, reflection, etc.
- *State-of-the-Art Performance:* CNNs have achieved state-of-the-art performance in image classification, object detection, and segmentation tasks.

Cons:

- *Computational Complexity:* Training deep CNNs can be computationally intensive, requiring substantial resources.
- *Large Datasets:* CNN often require large labelled datasets for effective training, which may not be available for every application.

- *Interpretability:* The internal workings of CNNs are challenging to interpret, making them somewhat of a "black box".

4.2.7 Graph Neural Networks

In the last few years, Graph Neural Network (GNN) have gained significant attention for their ability to model and analyse complex relationships within graph-structured data [99]. Originally designed for tasks such as node classification and link prediction, GNNs have found applications in diverse domains such as social networks modelling, recommendation systems, and physics, for modelling the dynamic of a N-body system, performing tracks reconstruction, and identifying particles.

GNNs operate on graph-structured data, where nodes or vertices represent entities, and edges represent relationships between these entities. The functioning of GNNs involves iterative aggregation of information from neighbouring nodes and updating of the edges, allowing them to capture local and global structures as defined by the graph. This is achieved through the use of message-passing mechanisms. An interesting feature of graphs is that the input information does not need to be given a rigid structure. Consequently, graph-based methods have a much greater representation power than image- or sequence-based ones. Graphs are in fact able to represent arbitrary relational structures, as defined by the graph through its directional weighted edges [100]. A particular feature arising from this property is that graphs are permutation equivariant: the order of nodes can be rearranged without impact.

Figure 4.10: A graph neural network update on a directed graph $G(V, E, u)$ with a global representation u, four initial nodes $\in V$ and edges $e_{ij} \in E$ connecting nodes $i \to j$ (in the notation of the text, each node is given a different integer index k and written (e_k, r_k, s_k), with $r_k = j$ and $s_k = i$). By analysing the neighbours of each node, the graph is updated to a new graph $G^*(V^*, E^*, u^*)$.

Typically, a GNN architecture consists of multiple layers of message-passing operations. Each layer updates the node representations by aggregating information from neighbouring nodes, as schematised in Figure 4.10. Different architectures implement different update processes for the graph, with popular GNN architectures being Graph Convolutional Networks (GCNs) [101] and Graph Attention Network [102]. In this thesis, the notation adopted is to represent a graph G has a tuple of three integers:

1. $E = \{(e_k, r_k, s_k)\}_{k=1:N^e}$ the set of edges, with each edge having a real vector of features $e_k \in \mathbb{R}^e$ and storing the index of the receiver (sender) as r_k (s_k).

2. $V = \{v\}_{i=1:N^v}$ the set of nodes, each node having a real vector of features $v_i \in \mathbb{R}^v$.

3. u, a global attribute of the graph modelled by a real vector of features $u \in \mathbb{R}^u$.

Algorithm 5 Steps of Computation in a Full Graph Network Block [100]

```
function GraphNetworkUpdate(E, V, u)
    for k ∈ {1 ... N^e} do
        e*_k ← φ^e(e_k, v_{r_k}, v_{s_k}, u)
    end for
    for i ∈ {1 ... N^n} do
        Let E*_i = {(e*_k, r_k, s_k)} for k = 1 : N^e where r_k = i
        ē*_i ← ρ^{e→v}(E*_i)
        v*_i ← φ^v(ē*_i, v_i, u)
    end for
    Let V* = {v*}_{i=1}^{N^v}
    Let E* = {(e*_k, r_k, s_k)}_{k=1}^{N^e}
    ē* ← ρ_{e→u}(E*)
    v̄* ← ρ_{v→u}(V*)
    u* ← φ_u(ē*, v̄*, u)
    return (E*, V*, u*)
end function
```

The most general graph update algorithm to describe an update stage of a full GNN block is described in Algorithm 5. Essentially, for a given step the input is a graph $G(E, V, u)$ that is updated into a new graph $G^*(E^*, V^*, u^*)$ by first updating the edges $e \in E$, then modifying the nodes $v \in V$, and finally the global representation u. The update rule leverages different neural networks ϕ and aggregation function ρ to update the graph. The aggregation function should accept a variable number of inputs with permutation invariance to output a single element per group and is typically implemented with the sum or max pooling. This global update is decomposed into successive steps:

- Updates the edges, with a DNN ϕ^e mapping each of the input edges, their respective receiver and sender nodes, and the global state u to output a new edge feature vector e_k^*

for each edge k: $e_k^* = \phi^e(e_k, v_{r_k}, v_{s_k}, u)$. The new edges are stored in a set E^*.

- Before updating a vertex i represented by v_i, the E_i^* updated edges connecting to i ($i == r_k$) are pooled locally over the node as $\overline{e_i^*} = \rho^{e \to v}(E_i^*)$.
- The vertex is then updated with a DNN ϕ^v, mapping the pooled representation of the edges $\overline{e_i^*}$ connecting to the vertex being updated, the input vertex feature v_i, and the global representation u to update $v_i \to v_i^* = \phi^v(\overline{e_i^*}, v_i, u)$. The new vertices are stored in a set V^*.
- The set of edges is updated by a global pooling $\overline{e^*} = \rho^{e \to u}(E^*)$.
- The set of vertices is updated by a global pooling $\overline{v^*} = \rho^{v \to u}(V^*)$.
- The global representation is updated by DNN ϕ^u mapping $u^* = \phi^u(\overline{e^*}, \overline{v^*}, u)$, with the globally pooled updated edges ($\overline{e^*}$) and vertices ($\overline{v^*}$).

This formulation of a graph as a message-passing with edges update device is the most complete architecture of a GNN. The design is however flexible: for example, RNNs or CNNs can be used instead of DNNs. Furthermore, many specialisations of the structures exist to reduce the degree of complexity of the model and avoid overfitting or convergence issues, as listed in Figure 4.11. A notable example for this thesis is the Deep Set architecture [103], designed to run specifically on sets where the ordering does not matter. It essentially simplifies the graph network by dropping altogether the edges and considering instead a fully connected graph with static edges, with an update of the global representation only based on pooled node information:

$$v_i^* = \phi^v(\overline{e_i^*}, v_i, u) = \phi^v(v_i, u),$$

$$\overline{V^*} = \rho^{v \to u} = \sum_i v_i^*,$$

$$u^* = \phi^u(\overline{e^*}, \overline{v^*}, u) = \phi^u(\overline{v^*}, u).$$

This is somewhat similar to PointNet, a GNN designed to analyse sets of 3D points, that uses an analogous update with max-aggregation instead of sum pooling after updating the nodes in two steps [104].

A different approach introduced in [105] defines the non-local neural network, unifying different types of *attention*-based architecture. Attention is an essential feature of modern deep learning: it refers to how an element is given a weighted version of the inputs, with weights

(a) Full GNN.

(b) Message-passing GNN.

(c) Deep Set.

(d) Non-local NN.

Figure 4.11: Different types of GNN update rules, defining different GNN architecture [100].

standing for the degree of attention to be given to each different part of the input. This concept is not restricted to GNN but is easily encapsulated in this formalism. As will be shown in the next section, the Transformer is a special case of this non-local NN family. In this section, only the Graph Attention Network (GAT) is introduced for the sake of conciseness [102]. GAT introduces the attention mechanism by having a learnable weighting of the neighbours of the node being updated. When updating node v_i, a score is computed for each of the connected neigbours v_j of v_i by a NN mapping:

$$e(v_i, v_j) = \phi(v_i, v_j) = a^T \text{ leakyReLU}([Wv_i, Wv_j]),$$

where leakyReLU is the modification of the ReLU with negative leakage, the nodes $v_i, v_j \in \mathbb{R}^d$ with j connected to i, and the operation implements an embedding of the two nodes to a dimension d' with two learnable parameters: $a \in \mathbb{R}^{2d'}$ and $W \in \mathbb{R}^{d' \times d}$. The operation $[,]$ stands for concatenation of the elements. These scores are then combined for each node i over its neighbours $\{j\}$ to give attention scores α_{ij}:

$$\alpha_{ij} = \text{softmax}_j(e(v_i, v_j)) = \frac{\exp(e(v_i, v_j))}{\sum_{j' \in \text{neighours of i}} e(v_i, v_{j'})}.$$

The final step is to leverage these attention weights when updating each node v_i:

$$v_i^* = \sigma\left(\sum_j \alpha_{ij} W^v v_j\right),$$

where the sum over j is taken over neighbouring nodes of i, σ is any activation function and W^v is another matrix of learnable parameters.

Pros:

- *Modeling Graph Structure:* GNNs naturally handle graph-structured data, making them well-suited for tasks involving relationships between entities.
- *Transferability:* Pre-trained GNN models on one graph can be fine-tuned for related tasks on another graph.
- *State-of-the-Art Performance:* GNNs have achieved state-of-the-art results in various graph-related tasks, including node classification and link prediction.

Cons:

- *Computational Complexity:* Training GNNs can be computationally expensive, particularly for large graphs.
- *Limited Global Context:* Some GNN architectures may struggle to capture long-range dependencies in graphs, limiting their ability to consider global context.
- *Interpretability:* Similar to other deep learning models, GNN lack interpretability, making it challenging to understand the learned representations.

4.2.8 The rise of the Transformers

The Transformer architecture, introduced in 2017 [106], has become a foundational design for NLP tasks. It has significantly impacted the field, enabling the development of state-of-the-art models such as BERT [107] and GPT [108]. More recently, the transformer is also spearheading a revolution in computer vision tasks thanks to the generalisation of the architecture into the Vision Transformer (ViT) [109].

The Transformer architecture is based on the mechanism of self-attention introduced in the last section on graphs. As mentioned in section 4.2.5 on RNN, it moves away from sequential

processing and adopts a fully parallelised approach, allowing for efficient computation on dedicated hardware. The key components of the Transformer are the self-attention mechanism and position-wise feedforward networks. Self-attention allows the model to weigh the importance of different words or tokens in a sequence when assessing a specific token. This mechanism enables the model to capture long-range dependencies in the input data without the added complexity of LSTM or GRU. Strictly speaking, the input of a Transformer is a sequence that has no strict order. For NLP, the important ordering of the sequence is built into the model using position-wise embedding. This lets the network decypher the index of the token in the input sequence and leverage this information for its tasks. For the computer vision case, the Vision Transformer first splits the input image x into patches of fixed size, flattens them into a vector and maps them with a learnable positional embedding before processing them as a regular Transformer would.

Figure 4.12: The full transformer architecture, combining an encoder and a decoder each made of an arbitrary number of layers. The input x is first embedded with a dedicated mapping which can, when order matters, be supplemented with positional embedding. The encoder generates an internal representation h that is passed to the decoder. This component, depicted on the right, produces the next output using the internal representation y and the output shifted to the right - to force output tokens to only access prior information.

As presented in Figure 4.12, the general Transformer architecture consists of an encoder and a decoder, the main feature of auto-encoder models. The decoder works in an autoregressive way, combining the current outputs $y_{t<T}$ with an internal representation h built by the encoder to generate the next output tokens y_T. Both the encoder and the decoder are composed of multiple layers, each containing a multi-head self-attention mechanism and position-wise feed-

forward networks (DNN). The decoder is further endowed with a masked attention layer, for the output to compute self-attention with information accessible before the token's position. The attention mechanism allows the model to focus on different parts of the input sequence, while the feedforward networks provide additional non-linear transformations. Residual connexions are added to let the gradients propagate efficiently in-depth and layer normalisation is used after each block to avoid vanishing or exploding gradients and improve training speed [110]. This type of normalisation scales each activation (each neuron) by subtracting the empirical mean and dividing by the standard deviation per data point.

The attention mechanism maps the queries and a set of key-value pairs to an output as defined in Equation 4.15 and schematised in Figure 4.13a, with query $Q \in \mathbb{R}^{d_k \times d_q}$, key $K \in \mathbb{R}^{d_k \times d_v}$, and value $V \in \mathbb{R}^{d_v}$ and the output is a vector $\mathbb{R}^{d_q \times d_v}$ assigned a vector of attention values to each value per query. This combines d_q different queries of the d_k keys mapping to d_v values. Therefore, Equation 4.15 implements for each query a weighted sum of the values, based on a compatibility function established by comparing the queries and keys:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{Q^T.K}{\sqrt{d_k}}\right) V, \tag{4.15}$$

where the scaling by $\sqrt{d_k}$ is implemented to reduce the magnitude of the dot product $Q^T K$ and avoid landing in regions of saturation of the following softmax, which is applied per row of the formed attention matrix. This scaled dot-product attention mechanism leverages the extensive research into numerical optimisation of matrix multiplications, making this operation less time and memory-demanding than using a DNN mapping to compute the attention - a technique referred to as *additive attention* [111]. As shown in Figure 4.13b, multi-head attention runs this dot-product attention in parallel for h different heads, each head h_i $(i = 1, ..., h)$ implementing a separate learnable projection from the input Q, K, V with linear transformations of respective weights $W_i^Q \in \mathbb{R}^{d \times d_k}$, $W_i^K \in \mathbb{R}^{d \times d_k}$, and $W_i^Q \in \mathbb{R}^{d \times d_v}$, where N is the length of the sequence, d is the model dimension and h the number of heads:

$$Q_i = QW_i^Q,$$

$$K_i = KW_i^K,$$

$$V_i = VW_i^V,$$

$$H_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V).$$

The multi-head module then concatenates the h different heads H_i outputs and applies another linear tranformation of parameters $W^O \in \mathbb{R}^{hd_v \times d}$:

$$\text{Multi-Head Attention}(Q, K, V) = [H_1, ..., H_h]\, W^o.$$

In multi-head attention, there are therefore 3 different learnable projections W_i^Q, W_i^K, and W_i^V per head and a single global projection W^O for the output of the cell. Self-attention is a special case in which the input of the cell is not a tuple (Q, K, V) of distinctive vectors but a single input $x \in \mathbb{R}^{N \times d_v}$ that is mapped out to the tuple with the learnable projections:

$$Q = xW_i^Q,$$

$$K = xW_i^K,$$

$$V = xW_i^V.$$

The self-attention operation is equivalent to graph attention on a fully connected graph with only node features.

Pros:

- *Versatility:* The Transformer architecture has been successfully applied to various NLP tasks - including machine translation, text summarization, and language modelling - and computer vision tasks.
- *Parallelisation:* The architecture allows for efficient parallelisation of training, speeding up the learning process.
- *Capturing Dependencies:* The self-attention mechanism enables the model to capture long-range dependencies in sequences.
- *Stability:* per design Transformers benefit from regularisation effects due to residual connexions and normalisation layers. Models based on this structure can therefore scale in complexity while resisting overtraining with stable training conditions.

Cons:

- *Computational Complexity:* Training large Transformer models can be computationally expensive, requiring powerful hardware.

(a) Scaled dot-product attention with optional masking.

(b) Multi-head attention module, where L stands for a linear transformation of the input.

Figure 4.13: Multi-head attention mechanism in a transformer. The core operation is an optionally masked scaled dot-product of the queries Q, keys K, and values V. A head consists of three (optionally different) linear projections of the tuple (Q, K, V), each leading to a separate scaled dot-product. The multi-head modules then concatenate all the different head results and finish with another linear transformation.

- *Interpretability:* The attention mechanism, while effective, is challenging to interpret, making the model somewhat of a "black box."
- *Data Dependency:* Transformer models may require large amounts of data for effective training, which may limit their application in domains with limited labelled data.

4.3 Training and Optimising Deep Learning Models

Training and optimising neural network models involve a combination of selecting appropriate architectures, fine-tuning hyperparameters - which cannot be learnt by backpropagation-, and employing acceleration techniques to improve efficiency and convergence. In this section, key aspects of the training process are explored, and acceleration techniques commonly used in practice are discussed.

4.3.1 Training Algorithms

When optimising the learnable parameters of a model, different training algorithms can be deployed to update the weights. All strategies are refinements of the base gradient descent rule of Equation 4.13, and each method has different advantages. The two main approaches are:

- Stochastic Gradient Descent (SGD): the update rule is the same as that of Equation 4.13, with the only difference being in how the gradient of the weights is computed. Instead of deriving it for the whole dataset (full-batch), the expectation over a random sub-batch of b elements is taken, hence the stochastic behaviour:

$$\nabla w = \frac{1}{b}\sum_{s=1}^{b} \nabla w_s,$$

  with ∇w_s being the gradients of the parameters computed for a single data point. A common observation is that for sub-batches - from now on referred to as just *batch* - of sufficient size b, the statistical estimator of the gradient based on the batch is unbiased. This greatly accelerates the time it takes to compute the gradient and naturally splits the loop over the dataset into different iterations called *steps*, at which a batch is passed through the network, a beneficial feature in the case of large datasets that would not fit in memory. This has also an effect on the regularisation of the model, as each computed gradient on a batch has a larger variance than a full batch, making it harder for the model to overtrain at the cost of slowing convergence to an optimum.

- Adam is an algorithm published in 2014 leveraging an adaptive moment estimation approach as well as batch processing [112]. The moment in this sense is analogous to the physical moment and encapsulates the dynamic of the optimisation as driven by gradient descent. The fundamental idea is that larger gradients indicate a steep slope that can be quickly traversed so that any slowing down due to a changing curvature of the objective function landscape can be mitigated due to the inertia of the previous gradients. This behaviour is implemented as an exponentially decaying moving average: the moment m_w^t of weight w at step t is updated with a gradient forgetting factor $\beta_1 \in [0, 1[$ such that:

$$m_t \leftarrow \beta_1 m_w^{t-1} + (1-\beta_1)\nabla_w \mathcal{L}^t,$$

  where the previous contribution is successively multiplied by β_1, reducing the importance of earlier gradients progressively. Additionally, another element is taken into account in

the gradient descent rule: the second moment $(\nabla_w \mathcal{L}^t)^2$. This tracks the magnitude of the gradient and, by multiplying the gradient update by a term inversely proportional to the second moment, accelerates the gradient updates in "flatter" regions of the objective landscape giving small gradient magnitudes with the term:

$$v_t \leftarrow \beta_2 v_w^{t-1} + (1-\beta_2)(\nabla_w \mathcal{L}^t)^2,$$

where a second-moment forgetting factor $\beta_2 \in [0, 1[$ is introduced. To avoid biasing the gradient update, both the momentum (first moment) and the second moment are corrected with

$$\hat{m}_w^t \leftarrow \frac{m_w^t}{1-\beta_1},$$

and

$$\hat{v}_w^t \leftarrow \frac{v_w^t}{1-\beta_2}.$$

The two contributions are then combined into a single gradient descent step following equation 4.16:

$$w^t \leftarrow w^{t-1} - lr \times \frac{\hat{m}_w^t}{\sqrt{\hat{v}_w^t} + \epsilon}, \tag{4.16}$$

where ϵ is added for numerical stability and is typically chosen to be a very small number to not bias the results.

A key hyperparameter in any gradient descent-based algorithm is the learning rate lr. There is no evident choice for this parameter and suitable values have to be derived on a case-by-case approach. A useful technique to let the training process converge to a good minimum of the loss function and avoid unsuitable local minima is to adopt a *learning rate schedule*: the learning rate is modified throughout training to resolve different parts of the loss function. Initially, having a relatively larger lr allows the model to quickly update its weights in the direction of the minimum. If the rate is kept too high, the weights will not be able to approach the minimum and will overshoot or "bounce' around the optimal set. To avoid this, the scheduler should reduce the learning rate so that smaller steps can be taken later in the training to approach the chosen optimum. At the beginning, the rate is typically not set to its maximum to start the gradient process in a valley of interest. An equivalent choice is to modify the batch size while keeping the lr fixed [113]. This has also a regularising effect on the gradient: small batch sizes capture large variances and let the optimisation make drastic changes of orientation in the optimising function landscape, thereby avoiding unsatisfactory local minima. Larger batch sizes stabilise the

direction of descent, thereby offering a lower variance but potentially biased estimates towards a minimum. Combining these two characteristics at different epochs of the training is therefore an effective way to improve the training performance. Some methods, such as Adam, have other specific hyperparameters that should be optimised with the procedure described later in this chapter for hyperparameters.

4.3.2 Regularisation

Regularisation techniques are applied in the architecture and training procedure to prevent overfitting. Common methods include *dropout*, which randomly drops connexions or neurons during training with Bernoulli probability distribution of parameter p, and L2 (L1) regularisations, which penalises large weights proportionally to a penalisation parameter λ times the sum of the squared (absolute value) of the weights. Both p and λ require careful optimisation as regularising the model can introduce bias and hit the overall performance. Additionally, batch normalisation is a technique that normalises the inputs of a layer over the batch, reducing internal covariate shifts. It helps stabilise and accelerate the training process. This is distinct from the layer normalisation used in the Transformer architecture of section 4.2.8 as the normalisation is carried over the batch samples rather than the activations.

4.3.3 Architecture & Hyperparameters Optimisation

Several characteristics of the network need to be optimised. Mainly:

- **Architecture Selection:** choosing the right architecture is crucial for the success of a NN. Factors to consider include the complexity of the task, the nature of the data, and the desired trade-off between model complexity and interpretability. Limits in computing power should be factored in. Elements of the architecture include the type of ML chosen (BDT, DNN, CNN, RNN, Transformer, ...), the choice of activation functions (ReLU, *tanh*, ...), and the number of layers and nodes as well as connexions between the units.

- **Hyperparameter Tuning:** optimising hyperparameters - parameters that cannot be optimised with backpropagation of the gradients of the loss - is essential for achieving good model performance. Key hyperparameters include optimiser-related parameters such as the learning rate, the batchsize, regularisation parameters, and initialisation of the weights and biases.

The optimisation process for both hyperparameter tuning and architecture selection is expensive: it requires training and testing models with different combinations of hyperparameters or model

architecture to evaluate their respective performance and uncover the best-performing options. Techniques such as grid search, random search, and Bayesian optimisation can be employed to efficiently explore the hyperparameter space. Architecture search is usually performed by trials and attempts, and the literature offers insights and guidance into what models might best perform in specific situations.

4.3.4 Acceleration Techniques

Training an ML model is often a computationally demanding task that can be carried out more effectively on specifically designed hardware and with some tricks in the process.

- **Parallel Dataloading:** loading data in parallel can significantly speed up the training process. Libraries like TensorFlow and PyTorch provide functionalities for efficient parallel dataloading. Instead of preparing a single batch, multiple batches can be loaded by different processing units in parallel to avoid this step becoming a performance bottleneck.

- **Early Stopping:** to prevent overfitting and save computation time, early stopping involves monitoring the model's performance on a validation set and halting training when performance ceases to improve.

- **Hardware Accelerators:** specialised hardware accelerators can further accelerate training. The GPU architecture is designed to perform simple mathematical operations in parallel. This is precisely the need for NN optimisation, as each computational step is a matrix operation. Utilising GPU for training and inference can give substantial speedups compared to CPU-based computing. Other types of hardware optimised for DL include the Tensor Processing Units (TPU) and the Neural Processing Units (NPU).

- **Transfer Learning:** transfer learning leverages pre-trained models on large datasets and applies them to a new task. The idea is that there are some fundamental similarities between the new task and the task used to train the foundational model that will let this latter one start with useful information. Fine-tuning these models on specific tasks can significantly reduce the required training time and data compared to training a model from scratch. This approach is becoming increasingly fashionable, as larger *foundational* models trained on multiple tasks with huge datasets can then be applied to a specific task, with the pre-trained weights either kept fixed or free to be modified to optimise the new task. New modules can also be added on top of the foundational model to specialise it to a specific use case. Such large foundational models are already available in NLP (e.g., the Mistral 7B

[114] - a 7 billion parameters Transformer capable of analysing English sentences and code) and computer vision (e.g., the Florence-2 model by Microsoft [115] - built on a Transformer with 0.5 billion parameters).

Training and optimizing deep learning models involve a combination of careful architectural choices, hyperparameter tuning, and the use of acceleration techniques. Selecting the most appropriate techniques is a task-specific challenge that will further depend on available resources and the desired trade-offs between training time and model performance as well as interpretability.

CHAPTER 5

Flavour Tagging

This chapter is focused on an essential task for the ATLAS experiment: identifying particles flying through the detector. This objective of assigning labels to reconstructed particles from measurements is called tagging. An important family of particles to be tagged are quarks, and disentangling which specific flavour of the quarks should be associated with an observed signal is called flavour tagging. Free quarks and gluons hadronise as per the rules of QCD, forming a large number of particles that can themselves further decay. Such a dynamic results in many particles radiating within a cone centred around the initial flavoured particle, a structure referred to as a jet. This chapter introduces a computational method to tag jets, as labelled by the flavour of the initial parton. In particular, the different algorithms and methods relevant to this task that have been developed contemporaneously to this thesis project are reviewed, including the DIPS, DL1d, GN1, and GN2 models as well as early studies on the hyperparameters optimisation of GN2.

5.1 Heavy-Flavour Jet Tagging

A fundamental ingredient in any ATLAS analysis is the ability to correctly identify particles in the aftermath of a collision, from τ-leptons, to b- and c-quarks, and gluons g. Having well-calibrated and optimally performing b- and c-tagging tools is of primary importance in studies of

the Higgs boson couplings to b- and c-quarks. It is also critical for top t-quark measurements and in many searches of physics Beyond the Standard Model (BSM). As described by the theory of QCD, colour-charged objects, such as a b- or a c-quark, undergo hadronisation to form collections of colourless hadrons. These hadrons, mostly B for b-quark and D for c-quark, are quasi-stable and further decay in the volume of the detector. Such a succession of decays leaves a collection of particles within a cone oriented in the direction of the original parton, an easily recognisable pattern referred to as a *jet*. From an analysis of the complicated structure of the jet, the flavour of the initially decaying particle can be reconstructed. The labelling scheme applied is based on the species of hadrons associated with the jets: a b-jet must contain at least one b-hadron, a c-jet at least one c-hadron and no b-hadron, and if none of these hadrons are found the jet is said to be a light-jet, thereby grouping u-, d-, and s-quarks with gluons g. This is the task of *flavour tagging*, and the tool to achieve this identification is called a *flavour tagger*. The focus of this chapter is on the development of novel taggers to identify b- and c-jet for the ATLAS experiment during the 2020-2024 period, overlapping with the end of Run 2 (2015-2018) and the beginning of Run 3 (2022-2026).

5.1.1 Decay Topology

When a b-quark is produced, such as in the aftermath of a hard scatter due to a proton-proton collision, it quickly undergoes the process of hadronisation to neutralise its free colour charge. This process leading free quarks and gluons to a final state of hadrons and leptons is intrinsically non-perturbative and can only be described with phenomenological models of fragmentation [116]. The family of b-hadrons is composed of different ensembles of a bottom quark b with one or more lighter quarks. These include the B-mesons, mainly $B^0 = d\bar{b}$, $B^- = \bar{u}b$, $B^+ = u\bar{b}$, and the strange and charmed B-mesons, and baryons, such as the $\Lambda_b^0 = udb$ [117]. For b-quarks, the hadronisation process is hard and most ($\sim 75\%$) of the quark momentum is passed to the b-hadron [116]. Tagging b-jets benefits from a particularly advantageous configuration: the b is the lightest element of the third generation of quarks and must decay through a weak interaction flavour-changing process. Because of the relatively small SM value of the $|V_{bc}|$ CKM matrix element, decay processes involving a transition $b \to c$ are suppressed, giving b-hadrons a characteristically long proper lifetime $\tau_B \approx 1.5$ ps, corresponding to a proper decay length $c\tau_B \sim 450$ μm [30]. In the laboratory frame and considering a boost of the b-hadron given by a

Lorentz γ factor ($\gamma > 1$) in the high energy limit $\beta = v/c \approx 1$, the distance travelled is

$$d = \gamma\beta c\tau_B \approx \gamma c\tau_B.$$

In this high energy limit, $\gamma \approx E_B/m_B$, where the B-hadron rest mass is in the range of 5 to 6 GeV. Consequently, a 50 GeV b-hadron decays at a distance of $d \approx 4.5$ mm from the primary vertex, which can be resolved using existing detector technology. This distance travelled, also called the b-hadron radius L_{xy}, further increases with rising jet p_T and at $p_T \sim 500$ GeV even overpasses the first detector layer of the IBL, located at a distance of ~33 mm from the centre of ATLAS, as shown in Figure 5.1b. The location of the b-hadron decay, called the SV, can therefore often be reconstructed with the ATLAS detector [118]. Some important variables describing the decay of b-hadrons are the IPs d_0 and z_0 of the tracks left by charged particles emanating from the SV. As shown in Figure 5.1a, d_0 and z_0 are the transversal and longitudinal distances from the primary vertex to the perigee[1] of the track. For a b-jet, the IPs can be large thanks to the long lifetime of the associated hadron. On average, a b-hadron decays weakly to four or five charged stable particles [117]. Another characteristic of b-jets is the likely presence of leptons in the jet cone, as ~40% of the b and c-hadrons decays are semi-leptonic [30].

(a) Representation of a b-jet [119].

(b) b-hadron decay radius as a function of jet p_T reconstructed for b-jets in a Z' events with the IBL and pixel layers indicated, from [120]. Error bars show the standard deviation of L_{xy} in each p_T bin.

While b-jets benefit from an advantageous topology, tagging c-jets at ATLAS proves more challenging as they are at an intermediate-mass scale between light and heavy flavour jets. A

[1] The point of closest approach.

c-jet must contain at least one c-hadron, from either a D-meson (e.g., $D^+ = c\bar{d}$, $D^- = d\bar{c}$, $D^0 = c\bar{u}$) or a c-baryon (e.g., $\Lambda_c^+ = udc$). The average decay length for charged (neutral) D-mesons, $c\tau_D \sim 300$ (100) μm [30], is smaller than for b-hadrons and is harder to resolve with the currently deployed tracker. The decay chain of b-hadrons often includes a c-hadron, making a clean separation of c-jets from b-jets harder. Compared to b-jets, c-jets have a lower final state average charged particle multiplicity of 4. This lets τ-jets easily be mistaken for c-jets, as these leptons can hadronically decay into a similar number of particles and appear as jets in the detector. Tagging c-jets is however becoming particularly important due to the focus on the challenging $H \to c\bar{c}$ search [121–123], as presented in the analysis of Chapter 6.

5.1.2 Flavour Tagging at ATLAS

In ATLAS, a choice was made to centrally develop a tagger to be used throughout the collaboration. The tagger simultaneously performs b- and c-tagging, and the software stack and methods are continuously improved to meet the requirements of the physics program. Currently, all studied approaches rely on DL methods, given their vastly superior effectiveness. This area of research has been evolving rapidly in recent times thanks to the community adopting advanced methods from the field of AI. As such, various models have been introduced, that can be split into two generations:

1. The DL1 family are DL models built in a hierarchical way. These DL methods rely on high-level features reconstructed by sub-algorithms based on physics variables, such as the tracks IPs, and the reconstruction of secondary vertices [124]. The most important models in this family are those including a DL sub-model to analyse tracks with either a RNN approach for DL1 with RNNIP (DL1r) [125], leveraging the Recurrent Neural Network Impact Parameter (RNNIP) sub-tagger [126], and a Deep Set approaches for DL1 with DIPS (DL1d), leveraging the Deep Impact Parameter Sets (DIPS) sub-tagger [127]. This last tagger is, at the moment of writing this thesis, the state-of-the-art calibrated tagger for ATLAS in the ATLAS software [59]. Algorithms from this family were mainly developed for the end of Run 2 of the ATLAS experiment [128], with DL1d behind developed before the start of Run 3.

2. The GN family of taggers are built on more advanced DL methods and moves away from the hierarchical approach of the DL1 family. These models directly analyse tracks and jet information with a unique powerful architecture. The GN family is based either on a full Graph Attention Network (GAT) for GN with GAT-core (GN1), or a Transformer

encoder for GN with Transformer-core (GN2) [3, 4, 129]. This streamlined algorithm pipeline greatly simplifies the maintenance and turnaround time for modifications, making the process of updating the taggers nimbler and easier to tailor to specific applications. The GN taggers greatly outperform the DL1 family and represent an exciting area of progress for future analysis requiring precise flavour jets tagging. GN2 is, at the time of writing, being calibrated and integrated into the ATLAS software stack, so that it can be used for Run 3 analyses [59].

Figure 5.2: Comparison of the performance of flavour tagging models introduced through the years, from [4]. Light and c-jet rejections (inverse of the mis-tag efficiency) are plotted for different taggers at a fixed b-jet tagging efficiency of 70% on a $t\bar{t}$ evaluation set. The multiplicative factors in the bars are with respect to the bare DL1 model performance.

A historical perspective on the evolution of performance attained with the different taggers mentioned is presented in Figure 5.2, showing a remarkable and consistent increase in light- and c-jet rejections at a fixed b-tagging efficiency of 70% evaluated a $t\bar{t}$-simulated dataset. The analysis presented in the latter part of this thesis was carried out from 2021 to 2024 and was therefore restricted to tools and methods available to the experimental team during this period. As such, due to the need to calibrate the GN taggers as explained later in Section 5.4 of this chapter, the analysis was constrained to use the DL1 family. The taggers described in this chapter have been integrated into the ATLAS software [59].

5.1.3 Datasets

ATLAS analyses scan a p_T spectrum that covers a wide range of energies as it extends up to 7 TeV. To train models able to perform on this large phase space, two training datasets are typically combined and described in this section. The datasets simulate proton-proton collisions

at a centre of mass energy $\sqrt{s} = 13$ TeV. The lower p_T phase space is filled with simulated SM top-antitop quark pair production $t\bar{t}$ events, where at least one W-boson produced decays leptonically. A BSM Z' process is used for the higher momentum region. The latter simulates a modified Z-boson with an increased mass to generate a flat jet p_T spectrum up to 6 TeV. These Z'-bosons decay in similar proportions to a pair of b, c, and light-jets. All simulations include realistic effects present in the real data such as PU, with an average value of $\langle\mu\rangle = 40$. Other effects included in the simulations are the detector response from prior and posterior bunch crossing (in-time PU), as well as the activity from the rest of the event (UE).

Events in the $t\bar{t}$ sample are simulated using POWHEGBOX V2 generators to Next-to-Leading Order (NLO) in the strong coupling constant α_s [130–133]. The hard scatter matrix element is computed for proton-proton collision with the NNPDF3.0NNLO set of parton distribution functions (PDF) [134], and the simulated hard scatter events are interfaced with PYTHIA 8.230 [135] using the A14 parameter tune [136] and the NNPDF2.3LO PDFs for the parton shower, hadronisation, and underlying event simulations [137]. Studies in Refs [138, 139] showed these choices correctly model the top quark p_T and the number of additional jets in the event, with the h_{damp} parameter set at 1.5 the mass of the top quark $m_{\text{top}} = 172.5$ GeV. The Z' events are fully simulated with PYTHIA 8.212, A14 tune and the NNPDF2.3LO PDFs. The decays of b- and c-hadrons are performed by EvtGEN v1.6.0 [140].

The detector reconstruction effect of ATLAS and the modelling of the interaction between long-lived hadrons and the detector material are simulated with a dedicated software [141] built on GEANT4 [142]. Jets are selected in the phase space region defined by $|\eta| < 2.5$ and $p_T > 20$ GeV, with no overlapping allowed with prompt generator-level e or μ from the W-decay. Pile-up contamination is further reduced by an additional selection using the Jet Vertex Tagger (JVT) algorithm at a tight working point for jets with $p_T < 60$ GeV and $|\eta| < 2.4$ [75]. Tracks are associated with jets using a ΔR association cone of width decreasing with p_T, such that $\Delta R \approx 0.45$ at $p_T = 20$ GeV and $\Delta R \approx 0.25$ at $p_T > 200$ GeV. Tracks within the cone of several jets are associated with the jet minimising the angular distance ΔR(track, jet). The label of the jet is inferred from the presence of a truth-level hadron within the cone ΔR(hadron, jet) < 0.3 centred around the jet axis.

5.2 DL1 Family of Taggers: DL1r & DL1d

This family of taggers is built with a hierarchical approach, combining low-level algorithms that are independently optimised into a final DNN network of a few layers to output the predictions. Not all low-level modules are based on DL, with some instead directly implementing physics-motivated algorithms. They consist of [128, 143]:

- IP likelihood discriminants: IP2D and IP3D (summarised IPxD) are likelihood-ratio templates in 2D and 3D to assign flavour-discriminating weights based, respectively, on the transversal and global impact parameters significances[2] S_{d_0} (35 bins) and S_{z_0} (20 bins) of the tracks, and 14 bins of tracks categorisation for IP3D [125]. For the three flavours, this results in $35 \times 20 \times 14 \times 3 = 29,400$ final bins, with each probability computed per track. The likelihood assigned to the jet assumes the tracks are independent, and is therefore calculated as the product of the per-track likelihoods. A discriminant is derived from the conditional log-likelihood, e.g., $D^b_{IP3D,f} = \sum_{i \in \text{tracks}} \log(p^i_b / p^i_f)$, to discriminate b-jets from f-jets ($f = c$ or light) [124].

- Track collection analyser: either with RNNIP [126] or DIPS [127]. These are DL approaches to extract discrimination information on the set of tracks associated with a jet. They importantly do not assume that tracks are independent. These taggers are further described later in this chapter.

- Secondary Vertex method 1 (SV1): combining a secondary vertex finder and a tagger to offer flavour discrimination information [128]. The former, based on the VKalVrt vertex reconstruction package [144], returns a list of candidate secondary vertices with measured quantities assigned to each vertex. The latter derives jet weights based on discriminative variables and computes properties of the SV, such as its mass.

- Jet Fitter: a vertexing algorithm based on a Kalman filter to reconstruct the topology and fit the decay chain PV $\rightarrow B \rightarrow D$, with the assumption that the vertices of the weakly decaying B- and D-hadrons tend to align with the PV [128, 145].

Figure 5.3: The algorithms for flavour tagging in the DL1 family. High-level taggers are in dark blue, track-based taggers in light blue, and vertex taggers in white.

[2] Corresponding to the reweighted IP variables by their respective uncertainties.

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

$$D_b = \log \frac{p_b}{f_c^b \times p_c + (1 - f_c^b) \times p_u}, \tag{5.1}$$

where f_c^b is the charm fraction, a parameter that can be modified to tweak the relative importance of the rejected flavours. The analogous c-tagging score D_c relying on the f_b^c parameter is

$$D_c = \log \frac{p_c}{f_b^c \times p_b + (1 - f_b^c) \times p_u}. \tag{5.2}$$

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

$$\epsilon_Y^X = \frac{N_{Y-jets}^{X-tagged}}{N_{Y-jets}} \quad \text{and} \quad \mathcal{R}_Y^X = \frac{1}{\epsilon_Y^X}, \tag{5.3}$$

where $N_{Y-jets}^{X-tagged}$ and N_{Y-jets} are respectively the number of X-tagged Y-flavoured jets and the total number of Y-flavoured jets. The f-rejection is the inverse mis-tag efficiency of flavour f.

These high-level models are trained on MC simulated data samples, as mentioned in Section 5.1.3, and need to be calibrated on real data to deliver an unbiased estimate, by deriving Scale Factors (SFs) weights correcting the predictions for each jet as described in Section 5.4. Uncertainties are derived on the predicted score and passed along to analyses using the tool. The novel algorithm of this family introduced in this work is the DL1d tagger, which relies on the DIPS sub-tagger to extract correlations between the tracks.

5.2.1 RNNIP

The RNNIP tagger runs on arbitrary-length input sequences made of track features, as ordered by the absolute transverse IP significance $|S_{d_0}|$, to extract tagging information from correlations

between tracks [126]. The vector of track features, described in greater detail in Table 5.1, includes the transverse and longitudinal impact parameter significances, the jet p_T fraction, the distance between the tracks and the jet axis, and a learned 2D embedding of the quality of the tracks [143]. RNNIP outputs a probability p_X for the jet to belong to flavour $X \in [b, c, \text{light}, \tau]$.

Figure 5.4: Diagram of the RNNIP tagger, from [143]. The input consists of track features augmented with an embedding of track categories. Tracks are then ordered by absolute transverse IP significance and fed through an LSTM core. The unrolled sequence from this LSTM is padded to a fixed size and processed by a DNN to output the per-flavour probabilities.

The architecture of RNNIP is a RNN-based model leveraging an LSTM core, as depicted in Figure 5.4. The arbitrary-length sequence fed as input is mapped by the LSTM cell with a 100-unit hidden layer into a 50-dimensional vector. This vector is then processed by a 20-unit fully-connected feed-forward neural network outputing the per-flavour probabilities by computing the softmax of the last layer's output. To avoid overfitting, a dropout value of 0.2 is applied to the LSTM cell.

RNNIP is designed to capture correlations between the tracks of a jet, an important insight explicitly missing from the IP-based discriminant of IP2D and IP3D due to the factorisation of the likelihood. Some degree of correlation is expected between tracks, as these can emerge from the same secondary or tertiary vertex of the displaced decays in b- and c-jets. RNNIP removes the cumbersome procedure to build likelihood templates, which demands a large amount of data to scale to finer bin resolution and is computationally expensive due to the number of bins scaling exponentially with the number of variables. RNNIP is effective at building a discriminant,

Track Variables	Description
S_{d_0}	Lifetime signed transverse IP significance d_0/σ_{d_0}, with d_0 the transverse IP and σ_{d_0} the error on d_0. If the perigee is in front (behind) the PV with respect to the jet direction, the sign is positive (negative).
S_{z_0}	Lifetime signed longitudinal IP significance z_0/σ_{z_0}, with z_0 the longitudinal IP and σ_{z_0} the error on z_0. A sign is assigned as per the prescription of S_{d_0}.
p_T^{frac}	Fraction of the reconstructed jet p_T^{jet} carried by the track $p_T^{\text{frac}} = p_T^{\text{track}}/p_T^{\text{jet}}$.
ΔR(track, jet)	Geometrical distance in 2D angle between the track direction and jet axis $\Delta R = \sqrt{(\phi_{\text{track}} - \phi_{\text{jet}})^2 + (\eta_{\text{track}-\eta_{\text{jet}}})^2}$.
Category	2D representation of the track quality learnt by an embedding layer. The categorisation is based on the number of observed, expected and missing hits in the different layers of the tracker (silicon pixel and strip detectors) [124].

Table 5.1: Track variables passed to the initial version of the RNNIP model [126]. Later versions removed the category embedding and added the per-track hit information shown for DIPS in Table 5.2.

delivering superior performances to the IP-based approaches with only ∼40 % of the parameters - 11,636 trainable parameters for RNNIP [143].

5.2.2 DIPS

The DIPS tagger, based on the Deep Set architecture [103] and depicted in Figure 5.5, is a GNN-based alternative approach to RNNIP to model the correlations between an arbitrary number of tracks [127]. As introduced in Chapter 4.2.7, such a model is composed of two fully-connected feed-forward neural networks. A first DNN called the *track network* Φ maps each track feature vector - similar to the input of RNNIP - to a latent space representing the nodes of a graph. The representations of each track in this latent space are then pooled by a simple summation operation - representing the unweighted edges of a fully connected graph - and given as input to a secondary DNN, called the *jet network* F. This latter network outputs the predicted probability p_X for the jet to belong to flavour $X \in [b, c, \text{light}, \tau]$. This last network represents the global attribute of the graph u, in the notation of Chapter 4.2.7. In summary, DIPS computes the following equation on the set of track features $\{p_i\}$, with $i = 1, ..., N$ for arbitrarily-sized jets of N tracks

$$DIPS(\{p_1, ..., p_N\}) = F\left(\sum_{i=1}^{N} \Phi(p_i)\right) \tag{5.4}$$

to output the per-flavour probabilities. The separation of computation into a per-track embedding and a per-jet processing after a size-independent pooling performed by the summation operator allows the model to process unordered sets of variable size. The track features used as inputs are described in Table 5.2, with only the top 15 tracks as ranked by decreasing S_{d_0}

considered.

Figure 5.5: Diagram of the DIPS tagger for flavour tagging, from [127]. The input consists of a set of N tracks, each represented by a feature vector. Each track is embedded by a DNN track network Φ into a fixed-dimension vector. All embedded track vectors are then pooled by summation to a fixed-size vector. The last step is to process this vector with another DNN jet network F outputing the per-flavour probabilities. The number and width of layers presented here correspond to the nominal architecture.

This approach has several advantages over RNNIP, mainly the physically motivated permutation-invariance of the input and the improved training and evaluation time thanks to a more parallelisable architecture, as the track embedding performed by Φ can be massively parallelised on GPU. These motivations are translated in an appreciable performance delivered by DIPS, which globally outperforms RNNIP while operating at a reduced computational cost [127]. The performance can be assessed from Figure 5.6, presenting the Receiver Operating Characteristic (ROC) curves for baselines trainings of DIPS and RNNIP in terms of light- and c-rejection for b-jet tagging on the same held-out $t\bar{t}$ evaluation sample.

The training times on the same GPU hardware for a 48k parameters DIPS model is 78 $\pm$ 4 seconds per epoch, while a 47k parameters RNNIP requires roughly thrice as much, 241 $\pm$ 14 seconds per epoch [127]. The faster training time allowed the Collaboration to focus on optimisation studies of the hyperparameters. The optimisation campaign focused on three aspects of the DIPS network: the architecture of the two NN, the track selection, and the set of track features used as input in addition to those of Table 5.2. Regarding architecture, a grid

Variables	Description
S_{d_0}	Lifetime signed transverse IP significance d_0/σ_{d_0}, with d_0 the transverse IP and σ_{d_0} the error on d_0. If the perigee is in front (behind) the PV with respect to the jet direction, the sign is positive (negative).
S_{z_0}	Lifetime signed longitudinal IP significance z_0/σ_{z_0}, with z_0 the longitudinal IP and σ_{z_0} the error on z_0. A sign is assigned as per the prescription of S_{d_0}.
$\log p_T^{\text{frac}}$	Logarithm of the fraction of the reconstructed jet p_T^{jet} carried by the track $\log p_T^{\text{frac}} = \log p_T^{\text{track}}/p_T^{\text{jet}}$.
$\log \Delta R(\text{track, jet})$	Logarithm of the geometrical distance in 2D angle between the track direction and jet axis $\log \Delta R = \log \sqrt{(\phi_{\text{track}} - \phi_{\text{jet}})^2 + (\eta_{\text{track}-\eta_{\text{jet}}})^2}$.
IBL hits	Number of hits recorded in the IBL - 0, 1, or 2.
PIX1 hits	Number of hits in the innermost pixel layer, after the IBL - 0, 1, or 2.
Shared IBL hits	Number of hits in the IBL that are shared by more than one track.
Split IBL hits	Number of split hits in the IBL, that are created by multiple charged particles.
nPixHits	Total number of hits in all the pixel layers.
Shared pixel hits	Number of shared hits in the pixel layers.
Split pixel hits	Number of split hits in the pixel layers.
nSCTHits	Total number of hits in the SCT layers.
Shared SCT hits	Number of shared hits in the SCT layers.

Table 5.2: Track variables passed to the DIPS model and later versions of the RNNIP model [127]. Compared to the initial RNNIP variables of Table 5.1, the p_T^{frac} and ΔR are passed as log values to reduce the magnitude of the long tail observed at large values and improve the training time. Shared hits are hits used by multiple tracks without being classified as split by a dedicated cluster-splitting NN [64].

(a) Light-rejection. (b) c-rejection.

Figure 5.6: Light- (left) and c-rejection (right) as a function of b-jet tagging efficiency for RNNIP (green) and DIPS (purple), taken from [127]. Each curve and error bands show, respectively, the mean and standard deviation of the rejections for 5 trainings per algorithm. The bottom panel shows the ratio to RNNIP, showing a clear performance gain for DIPS at all b-jet efficiency considered.

search over various possible values for the number of layers in Φ and F, the number of nodes, and the dimension of the track embedding space showed no significant performance change. The

selected architecture is:

- Track network Φ: three layers of 100, 100, and 128 units applied to each track.
- Jet network F: four layers of size 100, 100, 100, 30 before the final output of size dictated by the number of flavours to identify (3 or 4 typically).

To regularise and avoid overfitting, both batch normalisation and dropout were tested with the former observed to give better results.

(a) Light-rejection. (b) c-rejection.

Figure 5.7: Light- (left) and c-rejection (right) as a function of b-jet tagging efficiency for different DIPS model, with the baseline (nominal) DIPS in purple, the loosened track selection in blue, and the fully optimised DIPS in orange, from [127]. The curve and error bands show, respectively, the mean and standard deviation of the rejections for 5 trainings per algorithm with different initial random seeds. The bottom panel shows the ratio to the baseline DIPS, showing a clear performance gain from the two-step optimisation procedure at all b-jet efficiency considered.

The second optimisation step however uncovered that loosening the criteria of the track selection does improve performance. For RNNIP, IP2D, and IP3D, the selected tracks must pass the following quality selection: ≥ 7 hits in the silicon layers, ≤ 2 missing hits in the silicon layers, ≥ 1 hit in the pixel detector, ≤ 1 hit shared by multiple tracks, $p_T > 1$ GeV, $|d_0| < 1$ mm, and $|z_0 \sin(\theta)| < 1.5$ mm. For DIPS, a looser track selection increasing the acceptance of the last three cuts is preferable, modifying the nominal selection in the following way: $p_T > 0.5$ GeV, $|d_0| < 3.5$ mm, and $|z_0 \sin(\theta)| < 5$ mm [127]. Loosening the selection and keeping the top 25 tracks as ranked by decreasing S_{d_0} to capture more tracks from heavy flavour decays were observed to lead to a significant improvement in performance for jets with $p_T < 250$ GeV for DIPS. From an ML viewpoint, a larger set of input information with more noise can still prove beneficial if the underlying model is complex enough to capture useful features in the noisy data,

that would otherwise be erased by a more stringent selection. The performance gain from this loosened selection-trained DIPS is displayed in the ROC curves of Figure 5.7. Clear benefits are also obtained when adding additional track features as input on top of the looser selection, as shown by the orange curve of Figure 5.7, which plots the performance of a loose track selection DIPS trained with the per-track IP parameters d_0 and z_0 in addition to the features of Table 5.2.

Figure 5.8: Saliency map for b-tagging with 8 tracks sorted by $|S_{d_0}|$ and indexed by i, showing the gradient of the discriminant D_b with respect to the k track features x_{ik} [127].

How does DIPS work under the hood? The interpretability of machine learning models is an active area of research. Several effective approaches exist to gauge the importance of the input on the prediction. Figure 5.8 presents the result of applying the *saliency maps* technique [146]. Using the b-tagging discriminant D_b of Equation 5.1 at a fixed efficiency of 77%, the average importance of each feature in the track inputs is assessed by averaging the gradient of the discriminant with respect to the track features over a set of N jets with strictly 8 associated tracks failing the threshold:

$$\frac{\partial D_b}{\partial x_{ik}} = \frac{1}{N} \sum_{j=1}^{N} \frac{\partial D_b^j}{\partial x_{ik}^j}, \tag{5.5}$$

where i indexes the 8 tracks, j indexes the jet in the sample of size N, x_{ik} is the kth feature of the ith track [127]. This process effectively probes the linear sensitivity of the discriminant

on the track features. Using the saliency map, one can infer what features to modify to correct the failed tag assigned to the b-jets sample. The most sensitive parameters are measured to be the IP significances of the first five tracks, and the logarithm of the p_T^{frac} and ΔR of the track with largest $|s_{d_0}|$. This observation is physically motivated by the dynamic of the harder fragmentation of b-quarks, compared to light- and c-quarks. Negative gradients are measured for shared and split hits observables, translating into a further incorrect discriminant under a linear increase of these features. This is also physically motivated, as higher counts can be traced back to denser event environments where random combinations of hits to form tracks are more likely. However, total hit counts in the different tracker layers have a small positive impact, as these correlate with the reconstruction of the IP parameters.

5.2.3 Training of DIPS with Variable Radius Jets for Run 3

The physics program of the ATLAS Collaboration covers a wide range of analyses, targeting different topologies and processes at different energies. Concerning flavour tagging, a particularly relevant aspect is the energy or transverse momenta of the jets to label. Indeed, flavour taggers are extremely sensitive to the dynamic of the underlying events. At higher energies, corresponding to higher momenta of the hadronised quark or gluon, the jet constituents emanating from the decaying parton tend to be more collimated in the same direction, as they have to share a large initial amount of energy between themselves. This topology confounds tracks and blends the rich internal jet dynamics in the measured signature, making track separation and secondary or tertiary vertex identification more difficult. Analyses targeting jets from hadronic or semileptonic decays of heavy particles, such as the top t-quark, Higgs H, or the gauge vector W-/Z-bosons, can easily produce such highly energetic or *boosted* jets.

So far in this chapter, jets have always referred to the object as reconstructed by the anti-k_T algorithm with a fixed radius $R = 0.4$ applied to PFLow objects, as introduced in Chapter 3. This reconstruction method proves robust in the hadron collider setting as it both leads to suitably-shaped jet structure and PU-resistant properties. The fixed radius however becomes a hurdle to reconstruct boosted jets, as the average radius of a jet decreases with energy due to the collimation of the jet content. Indeed, the angular separation $\Delta R = \sqrt{\Delta\eta^2 + \Delta\phi^2}$ between the products of a decaying particle X of large mass m_X scales inversely to the transverse momentum [73]:

$$\Delta R \approx \frac{2m_X}{p_T^X}. \tag{5.6}$$

At low p_T^X, the individually produced particles from the decay are sufficiently separated to be reconstructed as individual objects, hence the *resolved* regime label [147]. For example, a non-boosted Higgs decaying to a $b\bar{b}$ pair can be reconstructed as two b-jets with small $R = 0.4$. At higher momentum, however, the content of the decay is collimated and overlaps: this is the *boosted* regime. The decaying particle X in such a regime is typically reconstructed as a single large-radius jets, to catch the different underlying jets, for example with the anti-k_T method with radius $R = 1.0$. Using such a fixed large radius overestimates the size of boosted jets which are easily contaminated by the PU, as well as the underlying event and initial-state radiations.

Another approach to reconstruct jets from boosted objects decay is the Variable Radius (VR) jet algorithm [148], as introduced in Chapter 3. VR jets have a size that scales with the inverse of the reconstructed jet momentum, thus correctly following the expected dynamic of Equation 5.6. Such a significant change to the jet reconstruction is bound to have an impact on algorithms learning structure from the jet contents, as is the case of all deep learning-based taggers presented in this chapter. These models must therefore be fine-tuned to this new jet type for optimal performance, which is the focus of this section.

For the VR-training, the dataset is composed of three samples simulating proton-proton collisions at $\sqrt{s} = 13$ with the following fractions:

- 85 % of jets are sampled from the $t\bar{t}$ with a maximal p_T of 400 GeV. At least one of the W-boson from the t-quark is required to decay leptonically.
- 7.5% are sampled from Z' events, where an exotic boson Z' decays as $Z' \to q\bar{q}$ or $\tau\bar{\tau}$, with a variable Z' mass to generate a flat p_T spectrum extending the p_T-range of the jets studied up to 4 TeV. These jets are required to have a $p_T > 150$ GeV.
- 7.5% are sampled from a simulated graviton process to also increase the range towards higher momenta. These jets are required to have a $p_T > 150$ GeV.

The simulation process is similar to that introduced in Section 5.1.3. Appendix Figure **??** displays the jet p_T and $|\eta|$ distributions for the hybrid sample as well as the individual samples it is based upon, for a total of 40×10^6 jets per flavour in $\{b, c, \text{light}\}$. To reach such high statistics, importance sampling with replacement is used to upsample the limited amount of c-jets while using all available b- and downsampling light-jets. A particularity of the processing is the requirement for the p_T and $|\eta|$ spectra to be equally distributed for all jet flavours so that

these features arising from inherent physics effects in the specific processes simulated cannot be used by the model to discriminate between flavours. Jets of different flavours are selected to match a specific target distribution. The importance sampling weights are derived by first computing the ratio of the targeted 2D distribution to the per flavour one. Weights above 1 indicate jets in the i, j bin have to be oversampled, while values lower than 1 indicate they should be downsampled. Jets are then iteratively sampled until the sampled distribution of each flavour individually matches the target distribution. As displayed in Appendix Figure **??** for which the target is b-jets, the thus constructed distributions have the same p_T and $|\eta|$ distributions for all flavours. This work introduced the first implementation of the importance sampling method, now widely used to develop flavour tagging tools leveraging the full statistical power of the simulations.

The optimised DIPS model with 62,167 learnable parameters from the previous section was trained for 200 epochs on 4 Quadro RTX 8000 GPUs. The learning rate started at 0.001 and was reduced by a factor of 0.8 on plateaus of 3 epochs, with a batch size of 15k jets, batch normalisation, and a dropout rate of 0.1 for the F network. The training proved stable with no signs of overtraining. The model at the epoch giving the smallest loss on a heldout validation set of 300k jets as well as the best light- and c-rejections at a fixed 77% b-tagging efficiency is selected. Figure 5.9 shows the ROC curves for b- and c-tagging of the best DIPS model on VR-jets (blue), as well as some comparison to the DIPS model trained on PFlow jets (orange) and RNNIP trained on VR-jets from the previous software release R21 (green). These ROC plots show, on the x-axis, the b-tagging efficiency (ϵ_b^b) versus, on the y-axis, the rejection $\mathcal{R}_Y^b$ for $Y \in [c, \text{light}]$, or equivalently for c-tagging swapping $b \leftrightarrow c$.

Training DIPS on a dedicated set of VR-jets improves performance compared to relying on the PFlow-trained version, as observed by comparing the blue (VR-trained DIPS) to orange curves (PFlow-trained DIPS). At a b-tagging efficiency of 77%, the light-rejection of the PFlow-trained DIPS is ∼40% lower. However, the c-rejection does not benefit as much, being either on par or even lower for the VR-trained DIPS on the $t\bar{t}$ samples. This difference in performance indicates an inappropriate choice of f_c value for the b-tagging discriminant of the VR-trained DIPS. A so-called *flavour fraction scans*, displaying the rejections at a fixed tagging efficiency for different values of the flavour fraction, can lead to a better choice for a balanced improvement in both background jet rejections. However, DIPS probabilities are not meant to be used directly in a discriminant but rather passed on to the high-level algorithm DL1d, hence this optimisation

(a) b-tagging on $t\bar{t}$.

(b) b-tagging on Z'.

(c) b-tagging on graviton.

(d) c-tagging on $t\bar{t}$.

(e) c-tagging on Z'.

(f) c-tagging on graviton.

Figure 5.9: ROC curves for b-tagging and c-tagging on 300k jets test samples of $t\bar{t}$ (left), Z' (centre), and graviton (right). Models displayed are the VR-jets DIPS in blue, the PFlow-trained DIPS in orange, and RNNIP trained on VR-jets from the previous software release in green.

is reserved for the final model as presented in Section 5.2.5. Figures 5.9d to 5.9f lead to similar conclusions for c-tagging.

5.2.4 Training of DL1d & DL1r with PFlow for Run 3

The ATLAS Collaboration continuously updates its software, updating specific methods to adopt new techniques, maintaining its many tools, and adding capabilities. In preparation for the current Run 3 of the LHC that started in 2022, ATLAS improved its reconstruction software from release 21 (R21) to release 22 (R22). As such, important elements used by flavour tagging methods have changed, requiring to retrain all taggers to ensure optimal performance under the new conditions. This work presents the first ATLAS study of the retraining of DL1r on the new release R22 and the first training of DL1d, including the DIPS sub-tagger in the high-level flavour tagging tool. Other important novelties of this work are the possible inclusion of τ-jets in the Deep Learner 1 Model (DL1) model's predictions and the importance sampling technique to process high-statistics training datasets introduced in the previous section. The interest in including τ stems from their tendency to be misclassified as c-jets when hadronically decaying, as both particles commonly leave three to four particles in the detector. The resulting taggers are observed to efficiently identify τ-jets thereby providing a new way to perform τ-identification and improving c-jet tagging.

Two samples, the $t\bar{t}$ and Z' from proton-proton collisions at $\sqrt{s} = 13$, are simulated and combined in the datasets, as described in Section 5.1.3. For both samples, PFlow jets are reconstructed using the anti-k_T algorithm with radius $R = 0.4$. These two samples are combined into a single *hybrid* sample to train the taggers, with 70% of the total number of jets coming from $t\bar{t}$ and the remaining from the Z'. The $t\bar{t}$ and Z' samples cover, respectively, a low- and high-p_T region based on a reconstructed b-hadron p_T separation threshold of 250 GeV for b-jets and a jet p_T of 250 GeV for non-b-jets. They are re-sampled to have the same $p_T - |\eta|$ distributions. The relative proportion of each sample was chosen to avoid any discontinuity in the p_T spectrum at their junction, as evidenced in Figure 5.10. The total statistics available for the R22 training is 25×10^6 jets per flavour. The final evaluation of the performance of a trained tagger is performed on separate test sets of both processes and unfolded over the flavours. The $t\bar{t}$ and Z' samples for validation and testing are each made of 1 million jets and are not downsampled to have the same $[p_T - \eta]$ distribution nor the same yield of different flavours: they represent a realistic distribution of the underlying processes.

Figure 5.10: The p_T (left - in MeV) and $|\eta|$ distributions of the resampled b-, c-, and light-jets in, respectively, blue, orange, and green. The three sets are resampled to have the same $p_T - |\eta|$ 2D distributions. The flat p_T spectrum extending up to several TeV is due to the exotic Z' process generated with varying mass, starting at 150 GeV. The large peak at lower p_T is the $t\bar{t}$-process. These sets have 8.3 million jets per flavour.

Training is performed with the UMAMI framework [1] based on TensorFlow [92] for 300 epochs with a variable learning rate schedule and the default network structure adopted in the previously released DL1r (R21): 8 fully connected NN of smoothly-decreasing sizes in [256, 128, 60, 48, 36, 24, 12, 6] with ReLU activation leading to a final softmax layer producing the predicted probabilities for each flavour. The models at an epoch offering the best combined results in terms of b-tagging efficiency and rejection from b-jets on the validation set are selected for further analysis. Every training converged to a fixed set of performance values, with no overtraining occurring.

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

- The preprocessing steps were revised to reduce the size of the evaluation sets for the benefit of the training statistics. A dual approach, downsampling light-jets and upsampling c-jets to the b-jets $[p_T - \eta]$, has also been implemented. As previously described, this approach based on importance sampling with replacement enforces the same p_T and $|\eta|$ distributions for the different flavours. While the performance of the majority classes was observed to improve, the efficiency at tagging the upsampled minority class (c-jets) was slightly lower. This trade-off can be compensated by modifying the flavour fractions and thus does not result in any significant performance change. This is likely due to the model saturating its performance given the large dataset already available. Other models, such as those from the GN family that have more parameters, have however been observed to make gains from

the importance sampling approach.

- Several modifications to the list of input features have been attempted, with no clear advantage uncovered. Adding pile-up information (the actual number of interactions per crossing and the number of primary vertices were tested) was not observed to have an impact on the tagging efficiency. Adding other variables from SV1 or JetFitter was also not observed to improve performance. However, a positive observation is that the IP2D and IP3D taggers can both be safely removed without impacting the performance, as the information they add is now better covered by the DIPS sub-tagger, thereby reducing the list of sub-taggers to maintain and simplifying the architecture.

- The structure of the network and its training procedure, leveraging transfer learning. Using samples produced with the previous ATLAS software release (R21) to pre-train the model was not observed to deliver a boost in performance when training on the new release (R22). Changing the size of the network and the batch size was also not observed to have a visible effect.

The conclusion driven by the lack of improvements from these three attempts is that models built on this simple DNN structure with such a large dataset are already likely saturating their performance from the set of inputs. The performance of the retrained DL1r tagger on the new release was found to be in good agreement with the then-released DL1r, despite using the same training of RNNIP from the previous release. To establish a meaningful benchmark for the newly trained taggers, the performance of the then recommended DL1r tagger, trained and evaluated on an analogous set of samples from the previous release (R21), is included in the following results under the label *Recom. DL1r*. A first look at the new family of taggers is also advertised by plotting the performance of a pre-release GN1 tagger, although this is discussed in further detail in the next Section 5.3.

Figure 5.11 presents the ROC curves on the $t\bar{t}$ (left) and Z' (right) test samples for b-tagging. These ROC plots are similar to those of Figure 5.9. The two bottom sub-plots present the ratio of the c-jet and light-jet rejection curves to the blue ones. This blue curve is the recommended DL1r performance and serves as the baseline of the comparison, while the new tagger DL1d is plotted in orange. Figure 5.12 shows the same plots for c-tagging, with respect to b- and light-jet rejections. The important observation is the clear gain obtained when replacing RNNIP with DIPS. Both the b- and c-tagging performance of DL1d dominate the DL1r versions, with a significant improvement in background flavour rejection for all tagging efficiency considered,

Figure 5.11: Performance for b-tagging with a flavour fraction of $f_c^b = 0.018$. Left: $t\bar{t}$; right: Z'. Top: ROC curves; centre: ratio of c-jets rejection from b-jets relative to the R22-retrained DL1r; bottom: same ratio for light-jets rejection. The recommended DL1r from the previous release is in blue. The new release DL1d is in orange and GN1 is in green.

Figure 5.12: Performance for c-tagging with a flavour fraction of $f_b^c = 0.2$. Left: $t\bar{t}$; right: Z'. Top: ROC curves; centre: ratio of b-jets rejection from c-jets relative to the R22-retrained DL1r; bottom: same ratio for light-jets rejection. The recommended DL1r from the previous release is in blue. The new release DL1d is in orange and GN1 is in green.

as summarised in Table 5.3. The largest performance improvement is obtained for b-tagging on the $t\bar{t}$ process, at lower jet momenta. This latter points to a dynamical behaviour of the DIPS subtagger that can be traced back to the looser jet selection. Higher momentum jets are more likely to have a larger set of tracks and these tracks tend to be closer to each other due to

relativistic boosting. The looser selection forces the DIPS model to sift through a larger set of noisy tracks. This brings lesser gains in performance at higher momentum, while an improvement is obtained at lower momentum from the good geometrical separation and smaller initial set.

b-tagging on $t\bar{t}$		
WP	c-rejection	light-rejection
60%	+26%	+73%
70%	+19%	+56%
77%	+12%	+41%
85%	+7%	+32%

b-tagging on Z'		
WP	c-rejection	light-rejection
60%	+19%	+43%
70%	+10%	+32%
77%	+9%	+26%
85%	+6%	+19%

c-tagging on $t\bar{t}$		
WP	b-rejection	light-rejection
25%	+26%	+5%
30%	+25%	+9%
40%	+22%	+12%
50%	+18%	+15%

c-tagging on Z'		
WP	b-rejection	light-rejection
25%	+12%	+22%
30%	+11%	+19%
40%	+8%	+14%
50%	+7%	+10%

Table 5.3: The change in background flavour rejections of DL1d relative to DL1r at various tagging efficiencies, both trained on the new release. Top: b-tagging ($f_c^b = 0.018$); bottom: c-tagging ($f_b^c = 0.2$); left: $t\bar{t}$; right: Z'.

The light-rejection from b-jets ROC curve in Figure 5.11 traces an elbow at high b-jet efficiencies. This effect is also present in the b-rejection from c-tagging in Figure 5.12. Both correspond to a set of, respectively, light-jets and b-jets that do not overlap with the b-jets b-tagging and c-jets c-tagging discriminants distributions, as shown in Figures 5.13 and 5.14. These "background" jets are easily removed from the core set of "signal" jets due to inherent differences between the flavours and the discrete nature of some sub-taggers used.

In Figures 5.11 and 5.12, a GN-like tagger trained on 20 million jets from the new family base on GNN that was in development at the time is introduced: GN1 [3]. This model is based on a graph attention network (GAT) directly processing low-level inputs, thereby diverging from the traditional ATLAS flavour tagging philosophy of combining several low-level sub-taggers into a high-level one, such as in DL1d. As exemplified in this plot, the method significantly improves the performance and is explored in further detail in Section 5.3.

The background rejections of the various taggers for b-tagging (c-tagging) as a function of the jet transverse momentum p_T at an inclusive b-efficiency of 70% (c-efficiency of 30%) per region displayed are shown in Figure 5.15 (Figure 5.16). Throughout the p_T range considered, DL1d outperforms the DL1r tagger. The low p_T b-rejection from c-jets is noticeably better for the newly trained tagger compared to DL1r. The discontinuity of the rejections between the two processes arises from the inclusive b-tagging efficiency being computed inclusively per region and not exclusively for the whole range.

Figure 5.13: Distribution of DL1d b-tagging discriminant with $f_c = 0.018$ for the different jet flavours, evaluated on $t\bar{t}$ (left) and Z' (right).

Figure 5.14: Distribution of DL1d c-tagging discriminant with $f_b = 0.2$ for the different jet flavours, evaluated on $t\bar{t}$ (left) and Z' (right).

Figure 5.15: Background flavour rejections at a fixed b-tagging efficiency of 70% (per region shown) for the various taggers. Top: $t\bar{t}$; bottom: Z'; left: c-rejection; right: light-rejection. For each plot, the bottom panel presents the ratio to the recommended DL1r.

Figure 5.16: Background flavour rejections at a fixed c-tagging efficiency of 30% (per region shown) for the various taggers. Top: $t\bar{t}$; bottom: Z'; left: b-rejection; right: light-rejection. For each plot, the bottom panel presents the ratio to the recommended DL1r.

(a) Flavour fraction f_c^b scan for b-tagging: left is $t\bar{t}$ and right Z' test samples.

(b) Flavour fraction f_b^c scan for c-tagging: left is $t\bar{t}$ and right Z' test samples.

Figure 5.17: The flavour fraction scans of the DL1d model. The chosen values are marked on the curves, displaying on the y-axis the c-rejection (b-rejection) for b-tagging (c-tagging) vs the light-rejection on the x axis at a fixed working point of 77% (33%). Increasing f_c or f_b shifts the marker upwards along the curves.

The DL1d model, integrating the Deep Set-based DIPS network in the classical DL1 hierarchical approach, was a valuable step in the development of a modern performant flavour tagger for ATLAS. Thanks to its similarities with the previous DL1r generation of taggers, built with the RNN-based RNNIP, it was smoothly integrated into the processing pipeline of the flavour tagger group. Its fast calibration led to its rapid introduction to the Collaboration and deployment to early Run 3 analyses. To exploit the full potential of the trained model and to cater to the specific needs of individual analyses, several working points are centrally defined and calibrated. An important parameter to control the relative importance of the jet classes to be rejected with the discriminants of Equations 5.1 and 5.2, light and c for b-tagging and light and b for c-tagging, are the flavour fractions f_c and f_b. Naturally, there is a trade-off: for b-tagging, a larger f_c-value favours a better c-rejection at the cost of a degraded light-rejection. To measure this dependency, flavour fractions scans are performed at a fixed b-tagging (c-tagging) efficiency of 77% (30%) in Figure 5.17a (Figure 5.17b).

About interpretability, it is of course challenging to explicitly detail the prediction process underneath the calculations of DL1d. An effective technique to measure the relative importance of the different variables is to quantify their contribution to the output using Shapley values. This technique for model explanation calculates the average contribution of each input to the output [85]. Figures 5.18 and 5.19 present the outcome of applying this framework, as proposed in Ref. [149] to approximate the Shapley values of the inputs to the b-tagging D_b and c-tagging D_c discriminants of DL1d respectively. These so-called *beeswarm* plots measure the impact of the evidence on the output of the model for each input feature. The plots display how each feature' Shapley value modifies the discriminant by moving from a prior background-data distribution expectation to the final model prediction using the real feature. A set of test datapoints of the targeted jet distributions[3] are sampled and, for each, a prior expectation was randomly sampled for the initial test. The impact of using the real value in the prediction was then measured. Positive Shapley values indicate variables having an increasing effect on the discriminant, thereby helping either b- or c tagging as per the plot considered. Each data point is coloured on a gradient scale from low feature value in blue to high feature value in red, and the dots pile up to indicate the density of the distribution. A feature that has more weight of its Shapley values distribution at larger values of the feature can be expected to help the model in identifying the main flavour of jets. Conversely, if the Shapley values are negative for large values of the feature, the feature value should be lowered for the model discriminant to improve.

Inspecting Figure 5.18 reveals some interesting patterns in the DL1d network for the task of b-tagging. The most important family of features for this task are the DIPS probabilities, with higher values of p_b correctly identifying the jet as b while higher values of p_c and p_{light} (noted p_u) have the opposite effect. The number of 2-track pairs from SV1 and some JetFitter variables - namely the mass of the vertex, the energy fraction and the number of tracks at the vertex - are also highlighted as important features. These observations are in line with a physics-based reasoning about the dynamic behind the jet: b-jets are expected to have a large charged particle multiplicity and the exchange of momentum is hard, with the b-hadron taking most of the b-quark momentum. Some other interesting features to consider are the ones formatted as "algoName_isDefaults": they encode whether the base-method "algoName" is activated (0 - blue) or not and thus defaulting (1 - red) for each jet. Interestingly, most of the occurrences of a defaulting behaviour of SV1 and JetFitter are associated with a negative Shapley value, demonstrating the validity of the physics reasoning behind these methods and their active contributions to b-

[3] e.g., b-jets for b-tagging.

Figure 5.18: Shapley values of the different inputs variables of DL1d for b-tagging, $t\bar{t}$ on the left and Z' on the right. High feature values are marked as red dots, while low feature values are blue.

Figure 5.19: Shapley values of the different inputs variables of DL1d for c-tagging, $t\bar{t}$ on the left and Z' on the right. High feature values are marked as red dots, while low feature values are blue.

tagging. IPxD variables generally score low in the ranking, indicating these methods contribute little to the model predictions and can be safely removed, an observation confirmed by direct optimisation of the input features set. Contrasting the Shapley values for $t\bar{t}$ (left) and Z' (right), the same variables roughly rank in the same order with the minimal differences found explained by the change in kinematic phase space between the two samples.

The same analysis can be carried out for c-tagging, with the results displayed in Figure 5.19. As discussed for b-tagging, the most important features are again the DIPS probabilities with p_c ranking first and contributing the most to D_c. Interestingly, the ranking of features is roughly the same as for D_b, with most features that had a positive impact on D_b when taking larger values now hurting D_c. This is the case for most of the JetFitter and SV1 variables. Defaulting behaviour of these algorithms, occurring when the conditions of a jet do not pass certain requirements, often has a positive effect on D_c as expected. Again, the IPxD family of features score low, indicating the limited importance of their contributions to the output. This anti-correlation behaviour of sub-algorithms to the D_c discriminant is expected, these methods having been primarily designed to help b-tagging.

5.2.5 Training of DL1d on Variable Radius Jets for Run 3

As for DIPS, changing the jet definition from PFlow to VR-jets is expected to have a large impact on the performance of the methods described here. Building on from the VR-trained DIPS model introduced in Section 5.2.3, this section presents the training of DL1d for VR-jets. The datasets are similar to those of Section 5.2.3. The VR-trained DL1d was trained for 300 epochs with no signs of overtraining. Its performance here is compared to the PFlow version introduced in the previous section, as well as the R21 DL1r version trained on VR-jets too and a pre-release GN1 trained on 20 million VR-jets.

A clear benefit from retraining on the dedicated VR-jet sets is observed in the ROC curves, with the VR-DL1d outperforming the PFlow version for all b- and c-tagging efficiencies considered. Introducing DIPS in the DL1 architecture has a significant impact on the performance of the tagger and greatly overmatches the RNNIP contribution. This is further highlighted by Table 5.4 reporting the rejections obtained at different WP of typical interest in analyses.

(a) $t\bar{t}$ test sample b-tagging, $f_c = 0.018$ for DL1d.

(b) Z' test sample b-tagging, $f_c = 0.018$ for DL1d.

(c) Graviton sample b-tagging, $f_c = 0.018$ for DL1d.

(d) $t\bar{t}$ test sample b-tagging, $f_c = 0.1$ for DL1d.

(e) Z' test sample b-tagging, $f_c = 0.1$ for DL1d.

(f) Graviton sample b-tagging, $f_c = 0.1$ for DL1d.

Figure 5.20: ROC curves for b-tagging. Top row uses $f_c = 0.018$ for DL1d, and bottom row $f_c = 0.1$ (GN1 $f_c = 0.05$ everywhere). The VR-jets DL1d model is in blue, a pre-release VR-trained GN1 in orange, R21 DL1r trained on VR-jets in green, and the PFlow DL1d in red.

As shown in Table 5.4, the specifically VR-trained DL1d outperforms the PFlow version with the flavour fraction parameter for b-tagging f_c^b changed from 0.018 (which is used for the PFlow model) to 0.1. For c-tagging, a clear gain in light-rejection comes at a cost of a lower b-rejection which can also be corrected by an appropriate change of the flavour fraction parameter for c-tagging f_b^c, currently set at 0.2 for both DL1d models. As highlighted in Figure **??** of Appendix **??**, displaying flavour fractions scans for b- and c-tagging, this choice of f_b^c is indeed suboptimal for the 30% WP.

b-tagging						
	$t\bar{t}$		Z'		Graviton	
WP	c-rej	light-rej	c-rej	light-rej	c-rej	light-rej
60%	+20%	+6%	+14%	+83%	+19%	+72%
70%	+18%	+9%	+14%	+65%	+16%	+57%
77%	+13%	+15%	+13%	+56%	+14%	+51%
85%	+1%	+25%	+11%	+45%	+12%	+40%

c-tagging						
	$t\bar{t}$		Z'		Graviton	
WP	b-rej	light-rej	b-rej	light-rej	b-rej	light-rej
25%	-20%	+137%	-17%	+90%	-17%	+80%
30%	-25%	+114%	-21%	+73%	-19%	+66%
40%	-29%	+99%	-23%	+53%	-22%	+48%
50%	-29%	+80%	-24%	+39%	-22%	+35%

Table 5.4: The change in background flavour rejection of VR-trained DL1d relative to the PFlow trained DL1d at various tagging efficiencies, both trained on the new release. Top: b-tagging ($f_c^b = 0.1$ and 0.018 for the VR and PFlow training); bottom: c-tagging ($f_b^c = 0.2$).

While this physics-motivated architecture optimisation moving from an RNN-based to a Deep Set-based track analyser improves the efficiency of the hierarchical model, a clear gain in performance is accessible through the more radical modification of the architecture that is adopted for the GN1 model. This is a classical observation in the world of machine learning: the vast amount of low-level noisy data can be better exploited by sophisticated architecture than by using a simple model fed a few highly engineered and reconstructed features, even when these are physically motivated. GN1 is not based on any physics principles. As shown in the next section, the tracks themselves contain enough of the rich physics signature required to unlock the label of the jet they compose.

5.3 Graph Neural Network Family of Taggers

As previously advertised in the PFlow- and VR-trained DL1d, the new generation of classifiers developed for flavour tagging at ATLAS introduces a fundamental shift in design. It moves away from the hierarchical approaches, using low-level specialised methods based on physics-inspired algorithms or neural networks as input to a high-level neural network. Instead, a single large neural network operates on a rich set of track information as well as some jet features to directly output the per-flavour probabilities. As suggested in Figure 5.21, this change to the flavour tagging software stacks greatly simplifies the maintenance and development effort, with all the attention being focused on a single network. A new framework called Salt [2] built on PyTorch [91] is introduced to simplify the definition and training of multitask multimodal models with multiple GPUs. This large network is built on a far more powerful and rich architecture with advanced expressive powers, thanks to a modified graph attention network (GAT) [102, 150] for GN1 and a transformer encoder for GN2 [106].

Figure 5.21: Comparison of the tagging scheme between the DL1 family (left) and the GN family (right), from [3]. Solid lines represent reconstructed information while dashed lines represent truth information only accessible from the simulations.

GN1 uses the information associated with charged tracks in a jet to directly output the flavour-tag probabilities, which are then combined into analogous discriminants to Equations 5.1 and 5.2. This constitutes the primary goal of the network. Alongside predicting the flavour of the jet, auxiliary objectives are also optimised to aid and guide the training. This so-called *multitask* framework is a common way to distil expert knowledge into the design of a ML method, focusing

the attention of the network on spelled-out metrics. In this case, two side tasks are passed along due to the physical insights they highlight:

1. *Track origin prediction*: a classification task aiming to assign a physical process from which the track arises, as per the prescriptions detailed in Table 5.5. The flavour of a jet is strongly correlated to the origin of the tracks. This task brings the attention of the network to this important information as a form of supervised attention [151].

2. *Vertex prediction*: a classification task predicting whether two tracks come from the same vertex. The decays of b- and c-hadrons include secondary and tertiary vertices inside a jet. Highlighting the compatibility of two tracks to share a vertex allows the model to infer the presence of such vertices. On the truth side, vertices separated by a distance < 0.1 mm are merged, and tracks labelled as Pileup or Fake are forced to not have any shared vertex.

These complementary objectives use truth information from the simulation and cannot therefore be predicted at inference time on real data. They improve performance during the training by providing useful information on the content of the jets. A modified approach in which a model is pre-trained on the auxiliary objectives and then fine-tuned on the primary objective is not observed to lead to a gain in performance, hence the objectives are optimised simultaneously.

Truth Origin	Description
Pileup	From a pp collision other than the primary interaction
Fake	Created from the hits of multiple particles
Primary	Does not originate from any secondary decay
fromB	From the decay of a b-hadron
fromBC	From a c-hadron decay, which itself is from the decay of a b-hadron
fromC	From the decay of a c-hadron
OtherSecondary	From other secondary interactions and decays

Table 5.5: Truth origins used to label the physics process leading to the produced tracks, from [3]. Charged particles and tracks are matched using the truth matching probability [64], and a value below 0.5 is taken to imply the reconstructed track parameters are mismeasured.

Being built around a graph computation, the GN1 and GN2 networks are directly adapted to work with a variable number of unordered inputs. The input is composed of 21 tracks with track features listed in Table 5.6. Each track is further decorated with 2 jet-level features: the jet transverse momentum p_T and signed pseudorapidity η. Tracks are selected from a set of requirements slightly modified from those used for DIPS: ≥ 8 hits in the silicon layers with < 2 shared hits, < 3 holes in the silicon layers, < 2 holes in the pixel detector, and tracks must have

Jet Inputs	
p_t	Jet transverse momentum
η	Signed jet pseudorapidity

Track Inputs	
q/p	Track charge divided by momentum (curvature)
$d\eta$	Pseudorapidity of the track, relative to the jet η
$d\phi$	Azimuthal angle of the track, relative to the jet ϕ
d_0	Closest distance from the track to the PV in the longitudinal plane
$z_0 \sin\theta$	Closest distance from the track to the PV in the transverse plane
$\sigma(q/p)$	Uncertainty on q/p
$\sigma(\theta)$	Uncertainty on track polar angle θ
$\sigma(\phi)$	Uncertainty on track azimuthal angle ϕ
$\sigma(d_0)$	Lifetime signed transverse IP significance
$\sigma(z_0)$	Lifetime signed longitudinal IP significance
nPixHits	Number of Pixel hits
nSCTHits	Number of SCT hits
nIBLHits	Number of IBL hits
nBLHits	Number of B-layer hits
nIBLShared	Number of shared IBL hits
nIBLSplit	Number of split IBL hits
nPixShared	Number of shared Pixel hits
nPixSplit	Number of split Pixel hits
nSCTShared	Number of shared SCT hits
nPixHoles	Number of Pixel holes
nSCTHoles	Number of SCT holes

Table 5.6: Input features of the GN family of models, from [3].

$p_T > 0.5$ GeV, $|d_0| < 3.5$ mm, and $|z_0 \sin\theta| < 5$ mm. A hole is a missing hit that was expected on a layer between two recorded hits of the same track. At most the first 40 tracks associated with a jet as ranked by transverse IP significance s_{d_0} are selected for processing. The input feature list includes missing information from the track and shared hits to specifically target high p_T jets, where tracks are more collimated and their separation can be unresolvable with the deployed detector technology. The GN1 and GN2 models shared the presented properties so far. They however differ in the architecture, which is explored in further detail in the next two sections.

5.3.1 GN1: Graph Attention Network for Flavour Tagging

The architecture of GN1, described in Figure 5.22, relies on a modified graph attention network [150] specifically designed for graph learning on sets, the so-called *Set2Graph* [152]. The design of the network architecture was subject to coarse hyperparameter optimisation. The first step takes all tracks, each represented by a vector of features composed of the 21 track features plus

the two jet features, and embeds each of these track vectors into a latent space of dimension 64 with a fully-connected feed-forward network of three hidden layers with 64 neurons. This is similar to the track neural network Φ of the DIPS model.

Figure 5.22: The architecture of the GN1 network, from [3]. The combined input is made of the set of tracks, each of which is given a copy of the two jet variables in addition to the track features, as described in Table 5.6. After a first embedding taking the input to an enriched latent representation, a fully connected graph is defined with the embedded tracks as nodes. The output of the graph is a conditional track representation used by the three training objectives.

A fully-connected graph is built with the embedded track representations as nodes. For this section, there is one node per track labelled h_i and represented by a feature vector of dimension 64. The graph network updates the defined graph $G(\mathcal{N})$ into a graph $G'(\mathcal{N}')$, with $\mathcal{N}$ and $\mathcal{N}'$ the set of edges, by aggregating the features of each node h_i and neighbouring nodes $\mathcal{N}_i$ to h_i using the operation of Ref. [150]. In the present case, the graph is fully connected, hence $\mathcal{N}_i = \mathcal{N}$. The following 4 steps are applied during a single graph update [3]:

1. Each node feature vector is passed through a fully connected layer W producing an updated representation $W\,h_i$ of size 64.

2. Pairwise scalar edge scores are computed for each pair of nodes $i, j \in \mathcal{N}$ by

$$e\left(h_i, h_j\right) = V^T \theta\left(\left[W h_i, W h_j\right]\right), \tag{5.7}$$

   where V is a second fully-connected feed-forward layer of size 128, θ is the ReLU activation function, and $[,]$ stands for the concatenation operation of two tensors.

3. Attention weights are derived from the pairwise edge scores, using a softmax over all j per node h_i:

$$a_{i,j} = \mathrm{softmax}_j\left(e(h_i, h_j)\right). \tag{5.8}$$

4. The final step is to aggregate the information to update each node $h_i \rightarrow h'_i$ by computing the attention-weighted sum over all node representations $\forall j \in \mathcal{N}$:

$$h'_i = \sum_j a_{i,j} \, . \, W h_j, \tag{5.9}$$

For GN1, applying 2 attention heads with 3 successive graph network layers is found to deliver optimal performance without any overtraining observed. The outputs of the graph network are *conditional track representations*, updating every track representation with information from other tracks. The ordering of the conditional tracks is kept similar to that of the original set to match processed tracks to their truth information. Furthermore, a global representation is derived by combining the conditional track representation with a pooling operation using learnable attention weights. These rich conditional and global representations can now be passed as inputs to three distinct feed-forward neural networks leading to the different objectives [3]:

1. *Jet flavour prediction*: performed by a graph classification network that is only fed the global representation. The primary objective of predicting the jet flavour is done by this network, composed of 4 hidden layers with 128, 64, 32, and 16 neurons respectively, finishing on an output of size 3 with softmax for b-, c-, and light-jet probabilities (4 if τ-jets are included).

2. *Track origin prediction*: performed by a nodes classifier processing each conditional track representation with the global representation. This network is built with three layers of reducing size 128, 64, and 32 to finish on the output layers of size 7 with softmax, matching to the 7 classes corresponding to the different truth origins considered in Table 5.5.

3. *Vertex prediction*: performed by a nodes pairs binary classifier that receives every possible combination of conditional track representations as well as the global representation. This network is made of 3 layers of size 128, 64, and 32 for a final output of size 1 with sigmoid, stating whether the pair of tracks have a common vertex or not.

The architecture of GN1 is an enhanced version of DIPS, with the track initialiser and graph classifiers corresponding to Φ and F. Added elements are the powerful GNN layers and conditional representation pooling layer with attention, as well as the auxiliary objectives.

Training GN1 involves minimising the combining objective $\mathcal{L}_{\text{total}}$ of Equation 5.10 [3]. $\mathcal{L}_{\text{flavour}}$ is the categorial cross-entropy loss, as defined in Equation 4.2, over the different jet flavours to

output the per flavour probabilities. $\mathcal{L}_{\text{track}}$ is the categorial cross-entropy loss for the track origin prediction averaged over all tracks in a batch. Due to intrinsic differences in the relative frequency of track origins, the contribution of each origin is weighted by its inverse frequency of occurrence. Finally, $\mathcal{L}_{\text{vertex}}$ is the binary cross-entropy of the track-pair compatibility averaged over all track-pairs in a batch. The importance of matching tracks from b- and c-hadrons is artificially increased by giving them twice the weight of track pairs.

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{flavour}} + \alpha\,\mathcal{L}_{\text{track}} + \beta\,\mathcal{L}_{\text{vertex}}. \tag{5.10}$$

In Equation 5.10, special weights are applied to combine the different tasks that are represented by distinct values, reflecting their specific loss functions and difficulties. Weights of $\alpha = 0.5$ and $\beta = 1.5$ [3] are found to lead the auxiliary objectives to converge to similar values, giving the different additional terms equal weighting in $\mathcal{L}_{\text{total}}$. The proposed choice for these parameters also lets the primary objective $\mathcal{L}_{\text{flavour}}$ dominate the global loss, and small variations of α and β do not significantly impact the performance. The results presented here come from Ref [3], where a GN1 model is trained for 100 epochs with a sample of 30 million jets made of 60% $t\bar{t}$ and 40% Z', as previously described in this chapter. The validation loss on a statistically independent sample of 500k jets is monitored, with the epoch minimising it selected for further analysis. The optimiser is based on Adam [112] with a learning rate of $1e-3$ and a batch size of 4000 jets spread across 4 GPUs.

The results of the training are presented in Figures 5.23 and 5.25 for b- and c-tagging respectively, where a DL1r model retrained on similar inputs to the GN1 with 75 million jets is presented as reference. The RNNIP sub-tagger used as input to DL1r is not re-trained on the new samples, but the performance of the retrained DL1r nonetheless matches the previously released DL1r (R21). The ROC curves of a GN1 model with an additional track input to those of Table 5.6 indicating whether a track was used in the reconstruction of an electron or a muon is also included as GN1 Lep. At the time of deriving these results, the DL1d tagger was yet not officially released and is thus not included here but discussed in the next section. The performance of DL1d is approximately 20% to 50% above DL1r at the 70% working point, far from the observed gains made by the GN1 models - as was highlighted in Figures 5.19 and 5.18. Most of the improvement in rejection from GN1 models is found a lower tagging efficiencies. At the typical working point of 70% on the low p_T region defined by $t\bar{t}$, the c-jet (light-jet) rejection is 110% (80%) above that of DL1r. Gains are made across the p_T spectrum, with a 180% (500%)

(a) GN1 and DL1r performance on the $t\bar{t}$ test sample, with $20 < p_T < 250$ GeV.

(b) GN1 and DL1r performance on the Z' test sample, with $250 < p_T < 5000$ GeV.

Figure 5.23: ROC curves tracing the b-tagging efficiency versus the c-jet (left) and light-jet (right) rejections for the $t\bar{t}$ (top) and Z' (bottom) test samples, from [3]. Models compared are DL1r in red, GN1 in blue, and GN1 Lep in purple. The bottom panels show the ratio to DL1r. The flavour fraction is set at $f_c^b = 0.018$ for DL1r and 0.05 for GN1 and GN1 Lep. The binomial error bands are shown as shaded regions.

increase in rejection at a 30% working point on Z', which roughly corresponds to applying the 70% working point from $t\bar{t}$. The GN1 version with lepton information further improves the performance, to a c-rejection (light-rejection) of 180% (150%) at the 70% WP on $t\bar{t}$ and 180% (600%) on the Z' at the 30% WP. A factor behind the observed performance improvement is the looser track selection leveraged by GN1 and the more sophisticated exploitation of the noisy set of low-level track information. The GN1 and DL1r discriminants for b-tagging are presented in Figure 5.24. The distributions for GN1 move the b-jet distribution to higher values of the discriminants, indicating higher confidence in the associated predicted p_b.

Figure 5.24: Comparing the GN1 and DL1r b-tagging discriminants D_b normalised distributions on the $t\bar{t}$ (top) and Z' (bottom) test samples, from [3]. Models compared are DL1r in dashed lines and GN1 in continuous lines. Each flavour is indicated by a different colour: green for b-jets, yellow for c-jets, and blue for light-jets. The flavour fraction is set at $f_c^b = 0.018$ for DL1r and 0.05 for GN1.

The c-tagging performance is presented in Figures 5.25 and 5.26, displaying the ROC curves and c-tagging discriminant distributions D_c. GN1 significantly outperforms DL1r for c-tagging: both background rejections are doubled on the $t\bar{t}$ sampled at a c-tagging WP of 25 %, with a more modest increase on the Z' sample of 60% for b-rejection and 100% for light-rejection at the same c-tagging WP.

As previously highlighted, the tagging performance is strongly anti-correlated with the jet energy considered, explaining the observed rejection differences between the $t\bar{t}$ and Z' samples. Higher energies correlate with higher transverse momentum p_T. More energy in the system introduces a higher multiplicity of fragmentation particles challenging the reconstruction process. The direction of emission of the particles is more collimated and approaches the resolution power of the tracking detector granularity. Different tracks are no longer individually resolvable and their hits are merged. Due to relativistic effects, at higher p_T the time of flight of heavy-hadrons increases,

(a) GN1 performance on the $t\bar{t}$ test sample, with $20 < p_T < 250$ GeV.

(b) GN1 performance on the Z' test sample, with $250 < p_T < 5000$ GeV.

Figure 5.25: ROC curves tracing the c-tagging efficiency versus the b-jet (left) and light-jet (right) rejections for the $t\bar{t}$ (top) and Z' (bottom) test samples, from [3]. Models compared are DL1r in red, GN1 in blue, and GN1 Lep in purple. The bottom panels show the ratio to DL1r. The flavour fraction is set at $f_b^c = 0.2$. The binomial error bands are shown as shaded regions.

delaying their decay further into the depth of the detector. Traces left by the heavy-hadrons paths and fragmentation particles introduce inaccuracies in the reconstructed track parameters [64]. This degradation of the track quality impacts the jet tagging performance significantly, as displayed in Figure 5.27 showing the b-tagging efficiency as a function of jet p_T for a fixed light-jet rejection of 100 in each bin. GN1 outperforms DL1r across the studied p_T spectrum, with a very significant b-efficiency improvement of a factor ~ 2 at high values of p_T, above 2 TeV.

To conclude this section on GN1, the importance of the auxiliary tasks is discussed by presenting ablations studies removing them iteratively from the full GN1 model. For this purpose,

Figure 5.26: Comparing the GN1 and DL1r c-tagging discriminants D_c normalised distributions on the $t\bar{t}$ (top) and Z' (bottom) test samples, from [3]. Models compared are DL1r in dashed lines and GN1 in continuous lines. Each flavour is indicated by a different colour: green for b-jets, yellow for c-jets, and blue for light-jets. The flavour fraction is set at $f_b^c = 0.2$.

Figure 5.27: Comparing the GN1 and DL1r b-tagging efficiency as a function of jet p_T at a fixed 100 light-jet rejection in each bin on the $t\bar{t}$ (left) and Z' (right) test samples, from [3]. Models compared are DL1r in dashed lines and GN1 in continuous lines. The flavour fraction is set at $f_c^b = 0.018$ for DL1r and 0.05 for GN1 and GN1 Lep.

three variants of GN1 are trained equivalently to the full GN1 but without:

- Any auxiliary objectives, leading to a model label "GN1 No Aux" only optimising the jet classification objective.
- The vertexing objective but not the track classification one, for the model labelled "GN1 Vert".
- The track classification objective without vertex, referred to as "GN1 TC".

Figure 5.28 displays the ROC curves of these modified models compared to the previously introduced DL1r and the full GN1. Removing the auxiliary objectives has a large impact on performance. The GN1 No Aux model is effectively similar to a DL1d model, having similar performance gains with respect to DL1r. Remarkably, this performance is obtained from a single network processing track without any of the sub-tagger nor methods used by the DL1 family, effectively underlying the powerful representation power of GAT. Adding either of the auxiliary tasks has the same beneficial impact on performance, as GN1 TC and GN1 Vert performs similarly and each is enough to significantly outmatch DL1r. The real gain is obtained by adding both auxiliary tasks, which further boosts the effectiveness of the model.

So far, the performance of GN1 on the primary objective of jet flavour classification has been discussed. The performance on the auxiliary objectives was not initially intended to be leveraged on real data but only to distil information to help the primary goal. The track-pairs vertexing performance is assessed by leveraging the information to perform vertex finding: grouping sets of tracks that are found to share a vertex into a single reconstructed vertex. The result is compared to the truth vertex label available in the simulations. Vertices identified by GN1 as containing tracks coming from a b-hadron decay are grouped, and the same procedure is applied to the truth information. To measure performance, the reconstructed and true vertices are compared as well as the number of tracks correctly assigned. A vertex is correctly identified when it contains at least 65% of the correct tracks with a purity of at least 50%. The comparison is only carried out for reconstructed tracks, meaning a 100% GN1 efficiency corresponds to correctly identifying all possible secondary vertices within the limit of the track reconstruction efficiency. An inclusive reconstruction efficiency in b-jets of $\sim$80% is measured for GN1, effectively proving that the model can identify b-hadron decay vertices. An important caveat is the current restriction is only on finding such vertices, not on reconstructing them. To implement a fully-fledged secondary vertex fitter as an auxiliary objective, the fitting of the vertex must be produced by a

Figure 5.28: ROC curves tracing the b-tagging efficiency versus the c-rejection (left) and light-rejection (right) for the $t\bar{t}$ (top) and Z' (bottom) test samples, from [3]. Models compared are DL1r in red, GN1 in blue, and versions of GN1 with missing auxiliary tasks. GN1 No Aux in green has none of the auxiliary, GN1 Vert in purple only the vertexing task, and GN1 TC in orange only the track classification. The flavour fraction is set at $f_c^b = 0.018$ for DL1r and 0.05 for GN1. The binomial error bands are shown as shaded regions.

differentiable algorithm to allow for backpropagation. This is a promising area of research, given the physics-based interest in accessing this important SV information. Promising work from Ref. [153] is under study to introduce an auxiliary differentiable single vertex fitting task.

Concerning the track origin classification performance, Figure 5.29 presents the traditional ROC curves, comparing the false positive rate (tracks wrongly assigned a label) versus the true positive rate (correctly assigned the label), for the different track origin classes of Table 5.5. Some classes are combined with weights dictated by the subclass relative abundance: this is the case of the FromB, FromBC, and FromC classes that are combined as Heavy Flavour, and the

Primary and OtherSecondary labels. The Area Under the Curve (AUC) of the ROC of all groups is above 90%, indicating good classification performance. The most challenging categories are the Heavy Flavour, Primary, and OtherSecondary tracks, while the Fake and Pileup tracks are effectively identified. The global mean (weighted) AUCs are of 92% (95%) on $t\bar{t}$ and 94% (96%) on Z' [3]. This performance ranking is in line with a physics-based intuition, and the p_T effect can be noted by the reduction in AUC for the Heavy Flavour tracks on the Z' sample.

Figure 5.29: ROC curves tracing the false positive rate versus the true positive rate of the truth origin classification on the $t\bar{t}$ (left) and Z' (right) test samples, from [3]. Heavy Flavour is a weighted combination of the FromB, FromBC, and FromC by their relative abundance.

GN1 was an exciting direction of development for flavour tagging at ATLAS. It showed clear benefits in moving away from the previous recipe to build taggers by combining several sub-algorithms and methods with physics meaning. Embracing modern advanced machine learning, it showed the superiority of deploying a single network built around an advanced unit, in this case, a GAT. While the functioning of the model is somewhat less interpretable than the previous DL1 family of taggers, expert knowledge is still passed to the network thanks to the multitask paradigm. Building on this success, an upgraded architecture was quickly developed to accelerate the speed of training and continue pushing the performance of the method ever higher: GN2.

5.3.2 GN2: Transformer Encoder for Flavour Tagging

GN2 is not a radical change on the architecture of GN1. Rather, it is a fine-tuned modified model aiming to reproduce the same conceptual processing chain as GN1, only with an easier-to-train and simpler-to-scale design. The main modification with respect to GN1 is the replacement of the computationally complex and expensive graph attention operators by a now ubiquitous

architecture in machine learning: the transformer [106]. As described in Chapter 4.2.8, the transformer is a remarkably effective and expressive design, both able to extract fine-grained correlations between ordered and unordered tokens in a sequence through the mechanism of attention and to scale to very large network size without suffering from overtraining. By design, transformers combine rich attention computing and regularisation-inducing steps which let such networks scale significantly their number of parameters while guaranteeing effective parallelisable training on GPU hardware.

Modification	Parameter	GN1	GN2
Hyperparameter	Trainable parameters	0.8M	1.5M
Hyperparameter	Learning rate	Fixed 1e-3	One-cycle scheduler
Hyperparameter	Core unit layers	3	6
Hyperparameter	Attention heads	2	8
Hyperparameter	Embedding dimension	128	192
Architecture	Attention Type	GATv2	Scaled dot product
Architecture	Dense update	No	Yes (dim 256)
Architecture	Separate value projection	No	Yes
Architecture	LayerNorm + Dropout	No	Yes
Inputs	Number of training jets	30M	192M

Table 5.7: Main modifications between the last generation of GN1 and the first generation of GN2, taken from [4].

In the case of GN2, the design only requires building a global representation of the sets of tracks composing a jet, hence only the encoder part introduced in Ref [106] and modified in Ref [154] is deployed to replace the GAT component of Figure 5.21. A summary of the modifications adopted when switching from GN1 to GN2 is presented in Table 5.7. The reference to GN1 corresponds to the last version of the model that was developed, which already adopted some minor modifications to the GN1 model previously described. Similarly, the GN2 model described here corresponds to the first publically released model, and this generation is also being refined and improved at the time of writing this thesis. Some significant changes adopted for GN2 are a learning rate scheduler, a larger embedding space dimension giving a wider and deeper - thanks to the doubling of the number of layers - core transformer unit, and the introduction of regularising effects from layer normalisation and dropout [110]. The learning rate scheduler is based on the one-cycle scheduler of Ref. [155], with some important parameters described in Table 5.8. This scheduler speeds up the training by initially growing the learning rate to large values, corresponding to large steps in the parameters' optimisation landscape, before annealing progressively the learning rate to small values, helping the optimiser to converge to a specific minimum [156]. The attention computation implemented by the transformer allows

similar physics performance to the GAT at a reduced memory footprint and training time [129]. The improved computational performance of GN2 makes it possible to scale up the number of parameters of the network and the training dataset size. Consequently, GN2 has roughly twice as many parameters as GN1 and was trained on a much larger training dataset. GN2 can indeed be trained on roughly $\times 6$ more jets than GN1 with the same computing resources. The datasets for the GN2 training presented here are derived similarly to those previously introduced for DL1d and GN1, using importance sampling to fully utilise the b- and light-jets statistics.

Parameter	Description
LR initial	Initial value of the learning rate
LR maximal	Maximal value of the learning rate reached at the end of warm-up
LR final	Value of the learning rate reached at peak epoch
Warm-up	Period covering the increase from initial to maximal
Peak epoch	Epoch at which LR maximal should be reached

Table 5.8: The five parameters of the one-cycle scheduler.

The attention mechanism in the transformer is subtly different from the GAT and corresponds to the multihead self-attention process described in Chapter 4.2.8. The nodes are updated in two steps: first attention is computed and applied, then a dense layer updates the set of nodes. In more detail, the transformer implements the following update on the set of nodes $h_i \in \mathcal{N}$ defining the fully connected graph $G(\mathcal{N})$:

1. Layer normalisation is applied to the input set of nodes $\mathcal{N}$.

2. For each attention head, 3 individual mappings represented by layers W_q, W_k, and W_v map each node $h_i \in \mathcal{N}$ to three independent representations $W_q h_i$, $W_k h_i$, and $W_v h_i$.

3. For each node $h_i \in \mathcal{N}$, edge scores are computed with all nodes h_j using the scaled dot product attention
$$e\left(h_i, h_j\right) = \frac{W_q h_i \,.\, W_k h_j}{\sqrt{s}},$$
where the s parameter representing the scaling weight is typically taken to be the dimension of matrix W_k.

4. The edge scores are turned into attention scores for node i, by taking the softmax over all nodes:
$$a_{i,j} = \mathrm{softmax}_j\left(e(h_i, h_j)\right).$$

5. Each node $h_i \in \mathcal{N}$ is updated into a node $h'_i \in \mathcal{N}'$ as:

$$h'_i = \sum_j a_{i,j} \, . \, W_v h_j$$

6. Using a skip connexion, the updated nodes $\mathcal{N}'$ are added their original $\mathcal{N}$ values.

7. Layer normalisation is applied to the updated nodes $\mathcal{N}'$.

8. The updated nodes are passed through a DNN.

9. The output of the DNN is summed to the updated nodes by a skip connexion, given the final updated set of nodes $\mathcal{N}'$.

The GN2 model presented here combines 6 such transformer layers with 8 attention heads in total. A comparison of the global performance of this PFlow-trained GN2 model to the already introduced PFlow-trained DL1r, DL1d, and GN1 models is displayed in the b-tagging ROC curves of Figures 5.30. For this comparison, the GN2 and DL1d models have been retrained on the same datasets, with the DL1r and GN1 models equivalent to those presented in the previous Chapter 5.3.1. GN2 delivers yet another significant boost in performance, drastically surpassing the GN1 rejections at all efficiencies considered. The largest improvement is obtained at lower b-jet efficiencies. Compared to GN1, GN2 delivers $\times 1.5$ ($\times 1.7$) the c-rejection (light-rejection) on $t\bar{t}$ at the 70% b-tagging WP and $\times 1$ ($\times 1.7$) on Z' at 30% WP. With respect to DL1d, the gains in c-rejection (light-rejection) are respectively close to $\times 3$ ($\times 2$) for $t\bar{t}$ and $\times 3$ ($\times 4$) on Z' at the same working point. The c-rejection on Z' of the GNN models is essentially equivalent, although the significantly improved light-rejection of GN2 indicates its c-rejection can be boosted by further increasing its flavour fraction f_c^b above 0.1.

Turning to c-tagging, as displayed in Figure 5.31, a similar large performance gained is obtained by the new GNN family over the DL1 one, both in terms of b- and light-rejection. GN2 again introduces a large improvement on top of GN1, although their b-rejection performance is equivalent on Z'. The gains from GN2 with respect to GN1 are of a factor $\times 1.3$ ($\times 1.3$) for b-rejection (light-rejection) on $t\bar{t}$ at the 30% WP, while they are $\times 1$ ($\times 1.2$) on Z'. The comparison to DL1d is of $\times 1.9$ ($\times 2.1$) on $t\bar{t}$ and $\times 1.3$ ($\times 1.8$) on Z' at the same WP.

Fixing the b-tagging performance at the 77% working point for both the $t\bar{t}$ and Z', Figure 5.32 scans the f_c^b flavour fractions for the different models. A clear hierarchy of performance

Figure 5.30: The c- and light-rejections as a function of the b-jet tagging efficiency in the $t\bar{t}$ with $20 < p_T < 250$ GeV (left) and Z' with $250 < p_T < 6000$ GeV (right) test samples. Models compared are DL1r in purple, DL1d in green, GN1 in blue, and GN2 in yellow. The bottom plots show the ratio to the DL1d performance. Flavour fractions are set at $f_c^b = 0.018$ for DL1r and DL1d, 0.05 for GN1, and 0.1 for GN2. Shaded regions represent the binominal error band.

Figure 5.31: The b- and light-rejections as a function of the c-jet tagging efficiency in the $t\bar{t}$ with $20 < p_T < 250$ GeV (left) and Z' with $250 < p_T < 6000$ GeV (right) test samples. Models compared are DL1r in purple, DL1d in green, GN1 in blue, and GN2 in yellow. The bottom plots show the ratio to the DL1d performance. Flavour fractions are set at $f_b^c = 0.2$ for all models. Shaded regions represent the binominal error band.

is observed: GN2 is orders of magnitude above the DL1 family and occupies undisputedly the highest rejections regions, followed by GN1, DL1d, and finally DL1r. For b-tagging on Z', the c-rejection can be further improved with limited impact on light-rejection by increasing f_c^b. However, the flavour fractions are optimised for an improved c-rejection on $t\bar{t}$, with limited change to the light-rejection across tagger generations. If desired, the light-rejection on $t\bar{t}$ of a GN2 taggers could be pushed upwards by lowering the f_c^b, reaching values as high as 1800 at a c-rej

of 4.8. The maximal DL1d light-rejection is 450 for a c-rejection of 4.5, thus a mere 25% of the GN2 light-rejection. Similarly, GN2 can reach a c-rejection of 19.5 at a light-rejection of 110, compared to a maximal c-rejection of 9.7 for a light-rejection of 40.

Figure 5.32: The flavour fraction f_c^b scans for b-tagging at a fixed working point of 77% of the different models considered evaluated on the $t\bar{t}$ (left) and Z' (right). The chosen values are marked on the curves, displaying on the x-axis the c-rejection vs the light-rejection on the y axis. Increasing f_c^b shifts the marker rightwards along the curves.

Figure 5.33 displays the flavour fraction f_b^c scans for c-tagging at the 30% working point. The same conclusions as for b-tagging hold, underlying the overall superiority of GN2. The f_b^c scans for c-tagging show a different shape than the b-tagging ones: at large f_c^b, the b-rejection rapidly increases while for b-tagging the c-rejection was saturating. This behaviour is due to the clear identification of b-jets giving them an outlying distribution compared to the overlap of c- and light-jets.

Figure 5.33: The flavour fraction f_b^c scans for c-tagging at a fixed working point of 30% of the different models considered evaluated on the $t\bar{t}$ (left) and Z' (right). The chosen values are marked on the curves, displaying on the x-axis the b-rejection vs the light-rejection on the y axis. Increasing f_b^c shifts the marker rightwards along the curves.

Figure 5.34 displays the effective per bin b-tagging efficiency for inclusive b-tagging efficiency

of 70% for $t\bar{t}$ and 30% for Z' in each p_T region considered. The performance is visibly not uniform across p_T, with the model accommodating specific parts of the p_T spectrum more easily. The region [100, 800] GeV overlapping the two samples is a sweet spot for performance, with more challenging results at lower and higher p_T. The performance for Z' in particular reduces dramatically with larger momentum, due to the physics reasons previously explained. Figure **??** in Appendix **??** displays the same information for c-tagging, leading to the same conclusions.

Figure 5.34: Comparing the different models b-tagging efficiency as a function of jet p_T for the inclusive b-tagging 70% working point on the $t\bar{t}$ (left) and 30% working point on Z' (right). The flavour fraction is set at $f_c^b = 0.018$ for DL1r and DL1d, 0.05 for GN1, and 0.1 for GN2.

To avoid biasing the analysis of the results with this per bin performance dependency, Figure 5.35 displays the b-tagging efficiency distribution across p_T at a fixed per bin light-rejection of 100. The superior capabilities of GN2 are exhibited across the p_T spectrum. The same conclusion holds for c-tagging, as displayed in Figure **??** of the appendix. Inspecting the rejections at a fixed b-tagging efficiency of 70% per bin also leads to concluding the clear superiority of GN2. Figures 5.36 and 5.37 respectively display the c- and light-rejection for a 70% b-efficiency per bin, showing that most of the improvement from GN2 and GN1 is in the [100, 800] GeV p_T sweetspot. The same distribution with an inclusive 70% b-tagging efficiency, over the entire p_T regions, is displayed in Figures **??** and **??** of the appendix.

To conclude this section, the b- and light-rejection at the 30% c-tagging per bin working point are displayed in Figures 5.38 and 5.39 respectively. Most of the improvements unlocked by GN2 and GN1 are to be found in the [100, 800] GeV sweet spot of the p_T spectrum.

These results, albeit intermediary as the development of the new tagger was still underway at the time of writing, are highly suggestive of the promised performance unleashed by the state-of-the-art GN2 model. Leveraging a simpler design and a more parallelisable architecture, GN2

Figure 5.35: Comparing the different models b-tagging efficiency as a function of jet p_T at a fixed 100 light-jet rejection per bin on the $t\bar{t}$ (left) and Z' (right) test samples. The flavour fraction is set at $f_c^b = 0.018$ for DL1r and DL1d, 0.05 for GN1, and 0.1 for GN2.

Figure 5.36: Comparing the different models c-rejection as a function of jet p_T for the b-tagging 70% working point per bin on the $t\bar{t}$ (left) and the 30% working point per bin on Z' (right). The flavour fraction is set at $f_c^b = 0.018$ for DL1r and DL1d, 0.05 for GN1, and 0.1 for GN2.

Figure 5.37: Comparing the different models light-rejection as a function of jet p_T for the b-tagging 70% working point per bin on the $t\bar{t}$ (left) and the 30% working point per bin on Z' (right). The flavour fraction is set at $f_c^b = 0.018$ for DL1r and DL1d, 0.05 for GN1, and 0.1 for GN2.

Figure 5.38: Comparing the different models b-rejection as a function of jet p_T for the c-tagging 30% working point per bin on the $t\bar{t}$ (left) and Z' (right). The flavour fraction is set at $f_b^c = 0.2$ for all taggers.

Figure 5.39: Comparing the different models light-rejection as a function of jet p_T for the c-tagging 30% working point per bin on the $t\bar{t}$ (left) and Z' (right). The flavour fraction is set at $f_b^c = 0.2$ for all taggers.

can effectively grow to a larger amount of parameters processing ever larger datasets, with no significant overtraining occurring. The story of modern flavour tagging is a story of refining and ever more expressive machine learning. RNNIP and DIPS required 50-60k parameters, which when introduced in the high-level algorithm to form DL1r and DL1d give rise to models with ∼130k parameters. GN1 revolutionises the approach by adopting a single powerful architecture with a total of ∼800k parameters. GN2 modifies this radical new design to adopt a highly efficient, regularised, and parallelisable model that easily scales the number of parameters to ∼1200k, being the first flavour tagger to cross the threshold of having a million parameters. The latest design of GN2 uses 2.6M parameters, and further tests raised this number to ∼70M parameters. Expert knowledge is passed to the latest generation of models using supervised attention, framing the physics intuition as learnable tasks enforced during training instead of as sub-techniques that need to be manually optimised and maintained.

5.3.3 GN2 Hyperparameters Optimisation

The state-of-the-art flavour tagger at ATLAS is, at the time of writing, built on the GN2 architecture. Naturally, fine-tuning the model is required to further push the performance higher. Many studies are ongoing to deliver yet a stronger tagger than the GN2 version presented in this thesis. A non-exhaustive list of ongoing research directions includes:

- Optimising the track selection and the jet reconstruction type. Moving towards yet a looser selection and letting the network sift through a larger set of background tracks could deliver further performance. Assessing the effect of modelling uncertainties is however of particular importance for these modifications.

- The inclusion of neutral constituent information by using jets defined as Unified Flow Object (UFO). Tracks are reconstructed from hits in semiconductor-based detectors. Such hits are only recorded for charged particles flying through the active regions of the sensors. This approach entirely misses neutral particles, such as neutrons, neutral pions and kaons, and neutrinos. All but the latter leave energy in the calorimeters that is measurable and accessible. The UFO jet definition combines track information with calorimeter topocluster objects. Studies are ongoing to add this information to the set of tracks.

- The inclusion of leptonic information. 40% of b-hadrons include either an e or a μ in the jet cone [30]. As seen with GN1, the inclusion of leptonic information in the set of tracks leads to a significant performance increase. Studies are ongoing to build a finer lepton-information analyser within GN2.

- Hadronic decays of τ are a major source of background for analyses focusing on c-jet tagging, due to their similar signatures. Including these leptons in the classification objective has been seen to deliver promising results in initial studies.

- Finer output classes categorisation. Currently, the simple labelling scheme deployed combines topologies with significant differences. For example, purely hadronic and semi-leptonic decays of b-jets are both labelled b-jets. Adopting greater flexibility in the definition of classes allows the model to fully utilise the unique signature of each process.

- Integrating further expert information into the design is known to deliver a great boost to performance. Studies are ongoing to upgrade the set of auxiliary tasks, in particular for secondary vertex fitting and reconstruction. A GN2 model able to reliably reconstruct

this information would have a use case in the ATLAS experiment beyond heavy-flavour jet tagging while benefitting from improved performance for this essential task.

These design considerations are paramount to producing a more efficient tagger. An equally essential endeavour is to fine-tune the architecture to extract the best performance from a chosen strategy. This section focuses on some initial studies to perform Hyperparameter Optimisation (HPO) and network architecture search for GN2. The essential challenge is that a test of a change to the hyperparameters or the model architecture requires fully retraining a GN2 model from scratch. This is a costly process, as a single epoch of GN2 training takes roughly ∼28 min for 2 NVIDIA A100 GPUs each fed data by 20 CPU on a 30 million jets dataset with batchsize 2000 evenly split on the GPUs. GN2 has many hyperparameters that should be optimised to deliver optimal performance, among which the most relevant are: initial lr, maximal lr, end lr, the weights of the 2 auxiliary tasks, the amount of weight decay, the batchsize, and the floating numbers precision. Important architecture-level elements to be optimised are the embedding dimension (output of the initialiser and as input and output of each transformer encoder), the depth of the initialiser, the number of layers and heads in the transformer encoder, the size of the transformer output, the auxiliary tasks DNN, the activation functions, and the specific loss functions and their class-weights used.

Unfortunately, access to GPUs is, at the time of writing, limited for members of the Collaboration. Most of the computing power leveraged to train advanced ML models such as GN2 is accessed on the high-performance cluster of institutes to which members belong. In this respect, a promising area of development is being pursued by CERN, with the introduction of a KubeFlow-backed server hosted on *ml.cern.ch* [157]. KubeFlow was created by Google and is now backed by the Cloud Native Computing Foundation. It is an open-source framework built on Kubernetes to perform machine learning operations such as training, inference, deployment, and hyperparameter optimisation. The project aims to centralise some GPU resources into a single cluster with datastorage, efficient I/O reading capabilities, and dedicated GPU nodes. Katib, KubeFlow's dedicated HPO workload, is a promising approach to perform effective hyperparameter optimisation with state-of-the-art autoML techniques that automate and refine the strategy to test and converge on the best hyperparameters [158]. At the time of writing, the server is still in a beta phase with little hardware accessible, thereby removing it from consideration as a possible solution to carry out the full HPO of GN2. However, the SALT framework used to train GN2 was adapted to run on any KubeFlow platform, with initial tests showing promising

possibilities for the Collaboration. Being accessible to any member of ATLAS, this project would "democratise" access to computing-intensive studies for institutes lacking an advanced High Performance Cluster (HPC).

Large NN such as large language models that are being developed at ATLAS will require clusters designed for machine learning, with many GPUs accessible on dedicated nodes. This paradigm of computing is markedly different from the typical grid-based distributed computing deployed in particle physics experiments. While MC-based samples and sub-sampled datasets can be effectively processed by autonomous parallel jobs, ML requires communication between the different jobs to keep the weights of the model trained synchronised on the different GPUs. A fast connexion between these GPUs is essential, as is having fast read access to the full dataset due to the need to loop over the whole data for each epoch during training. Distributing the computation across different HPC that are geographically distant, as is common with the current CERN computing grid, is not effective for this purpose. The CERN KubeFlow server is a promising area of development for the future computational needs of ATLAS. Furthermore, having a framework compatible with KubeFlow allows operating on multiple platforms, giving the flexibility to scale resource access for computationally demanding tasks, such as HPO. Most private and public cloud providers, such as Google Cloud, Amazon Web Service, and Microsoft's Azure, are KubeFlow-compatible and host a larger amount of state-of-the-art GPUs. Salt can be effectively deployed on the infrastructure of one of these cloud providers or the CERN's KubeFlow server with no noticeable distinctions for the user.

While leveraging a large amount of computing power is a natural solution to the challenging task of HPO of a "large" neural network by ATLAS standards, a more refined technique can be exploited in the present case. Recent works from the ML community suggest that the optimal hyperparameters of a nominal model can be estimated from a smaller model [159]. Here smaller refers to either the depth - the number of layers - or the width - the number of neurons per layer and, in the case of a transformer, also the number of heads in the multihead attention - of the neural network. Ref. [160] establishes the mathematical foundation backing this surprising behaviour of deep neural network: the Maximal Update Parametrisation (μP). The rest of this section is dedicated to introducing and defining the maximal update parametrisation before establishing its relevance for HPO.

Maximal Update Parametrisation

The maximal update parametrisation is first and foremost a *parametrisation*. In this context, the parametrisation of a neural network refers to the definition of the weights of each neuron, the way they are initialised, and how they are updated from a given optimisation algorithm, such as Adam or SGD [112]. The default or *standard* parametrisation (SP) follows the so-called LeCun parametrisation [161]. This parametrisation, routinely deployed in ML frameworks such as PyTorch [91], initialises the weights by sampling them from a Gaussian or Uniform distribution with mean 0 and standard deviation given by the inverse of the input dimension of the layer the weight belongs to. For both Adam and SGD, a single master learning rate (LR) η is used for all weights. For μP, some subtle differences are introduced, as summarised in Table 5.9. Mainly, the output layer weights are sampled from a Gaussian with a standard deviation being the inverse of the input dimension **squared** of the output layer. Concerning the learning rates, the hidden and output layers are scaled down by their respective input dimension for Adam. For SGD, the output layer LR is scaled similarly, but the input and the bias LR are scaled up by the output dimension of these layers.

	Initialisation Distribution		Adam LR		SGD LR	
	SP	μP	SP	μP	SP	μP
$w^{L_{\text{inp}}}$	$\sim \mathcal{N}\left(0, \frac{1}{d^{\text{in}}_{L_{\text{inp}}}}\right)$	$\sim \mathcal{N}\left(0, \frac{1}{d^{\text{in}}_{L_{\text{inp}}}}\right)$	η	η	η	$\eta \times d^{\text{out}}_{L_{\text{inp}}}$
$w^{L_{\text{hid}}}$	$\sim \mathcal{N}\left(0, \frac{1}{d^{\text{in}}_{L_{\text{hid}}}}\right)$	$\sim \mathcal{N}\left(0, \frac{1}{d^{\text{in}}_{L_{\text{hid}}}}\right)$	η	$\eta / d^{\text{in}}_{L_{\text{hid}}}$	η	η
$w^{L_{\text{out}}}$	$\sim \mathcal{N}\left(0, \frac{1}{d^{\text{in}}_{L_{\text{out}}}}\right)$	$\sim \mathcal{N}\left(0, \frac{1}{d^{\text{in}}_{L_{\text{out}}} \times d^{\text{in}}_{L_{\text{out}}}}\right)$	η	$\eta / d^{\text{in}}_{L_{\text{out}}}$	η	$\eta / d^{\text{in}}_{L_{\text{out}}}$
$b^L \ \forall L$	0	0	η	η	η	$\eta \times d^{\text{out}}_{L}$

Table 5.9: Comparing the Standard Parametrisation (SP) to the Maximal Update Parametrisation (μP), as defined in Ref. [159] based on the work of Ref. [160].

This particular derivation of μP, taken from Ref. [159], is equivalent to the original μP derivation introduced in Ref. [160]. μP turns out to be the unique parametrisation that maximally updates the weights of a neural network. The updates are "*maximal*" in the sense that they are as large as they could be for a given LR to avoid any instabilities. For the specific case of the attention mechanism computed by the multi-head attention of transformers, the scaling has to be modified from $\sqrt{d_k} \rightarrow d_k$ to properly scale with width [159]. Figure 5.40 shows a comparison of the size of the pre-activation of a GN2 model with μP parametrisation to a standardly parametrised GN2, referred to as the SP model, at different training steps. Each curve displays, for different embedding widths in the transformer and the track initialiser, the sum of the abso-

lute values of the weights before the activation ($L_1(\text{layer}) = \sum_{w_i \in \text{layer}} |w_i|$) for the initialiser and transformer models only. Three timesteps are displayed for each model, the initialisation ($t = 1$) and after 1 ($t = 2$) and 2 ($t = 2$) training steps. The interesting behaviour highlighted in this figure is that for the SP model, the pre-activation weights blow up with width during training as shown by the exponential rise of the sum of pre-activations. For μP however, the L_1 of each layer stays flat with width even during training, proving the correct parametrisation of the model and the "width-independent" scaling. This unstable behaviour of the SP parametrisation is easily highlighted thanks to the use of a large and fixed learning rate (here $lr = 10^{-2}$).

Figure 5.40: The sum of the absolute value of the pre-activation weights for the different layers in the initialiser and transformer parts of a GN2-like model in standard parametrisation (SP - top) and in μP parametrisation (bottom), at three timesteps: initialisation ($t = 1$ - left), after one training step with $lr = 10^{-2}$ ($t = 2$ - centre), and a second training step ($t = 3$) [6]. The models displayed are labelled GN2-like as they lack auxiliary tasks.

Theoretically, a μP model should deliver equal to better performance to an equivalent SP model when both have optimal hyperparameters. This behaviour is due to the maximal updating of the former, leading to optimal in-depth updates of all layers. The standard parametrisation does not implement this correct updating, with outer layers closer to the loss function having an opacity effect on the propagation of the update for the input layers proportionally to their widths. Scaling down the learning rate is not a sufficient modification to correct the SP: as displayed in Figure 5.40, not all layers update incorrectly with some pre-activation sum staying flat across the widths. By updating all activation maximally independently of the width, μP outperforms SP for a tuned learning rate [160]. A significant advantage of this parametrisation is that the optimal learning rate for a μP architecture becomes width-independent. This leads to the μTransfer algorithm for HPO, where the best hyperparameters for a μP model are found on

a version with fewer neurons per layer (smaller width) and the found optimal ones are transferred to the full-size model at no extra cost (0-shot transferred) [159]. The benefits of adopting the maximal update parametrisation are:

1. Better performance of a μP model compared to an SP model for a tuned learning rate.
2. Improved hyperparameter optimisation with the μTransfer algorithm: performing the HPO scan on a smaller and easier-to-train model to 0-shot transfer the best set of hyperparameters to the full-size models.
3. Better hardware usage for HPO: a smaller model can be trained on a single GPU. This is of particular interest for the ATLAS Collaboration, as most of the GPU resources accessible are scattered through geographically distant computing sites.
4. Simplified architecture: with μP, a wider model outperforms a smaller model if no overtraining occurs. Therefore, the best learning rate hyperparameter has to be found once for all GN2 models of varying widths and the widths are chosen based on the desired computational complexity.

Hyperparameters that can be optimised with the μTransfer algorithms are said to be μTransferable. They consist of [159]:

- Learning rate and parameters of a learning rate scheduler.
- Optimiser parameters, such as the momentum, and the Adam α and β.
- Initialisation parameters, such as the initial per-layer variances.
- Multiplicative constants.

Unfortunately, many parameters do not μTranfer as they combine aspects of the model and the data, and must be studied on the full-size model directly. For example, the regularisation parameters (dropout, weight decay, normalisation, ...) do not scale, as a particular model size will overfit depending on the data. Finally, the last important family of hyperparameters are those defining the scale of the problem. These parameters are not found from μTransfer but rather "μTransfered along". They consist of the width[4], the depth, and the batchsize. Only the scaling along width is theoretically proven thanks to μP, while the others are empirically observed to hold [159].

[4]Number of neurons per layer, number of attention heads in a transformer, ...

Studies of the μP parametrisation and the μTransfer algorithm have been performed for the GN2 flavour tagger. In this architecture, the most relevant dimensions are the width and the depth of the transformer part, tasks with building a conditional representation of the tracks from the embedded tracks processed by the initialiser network. These two dimensions are keys as most of the parameters of the GN2 model are in the transformer and the initialiser, with only a few parameters set in the networks of the primary and auxiliary tasks. As such, the chosen dimension to scale with μTransfer is the embedding width. The number of parameters in the transformer associated with the embedding width scales quadratically with this parameter, making it the most sensitive dimension to define the complexity of GN2.

Figure 5.41: Comparison of a maximal learning rate value scan at an initial learning rate value of 10^{-5} for an SP (left) and a μP GN2 models (right) for three different embedding widths: 64 (yellow), 128 (red), and 256 (purple). The y-axis displays the validation loss attained. Taken from [6].

To demonstrate the effect of μP on GN2, a learning rate hyperparameters optimisation campaign targeting the initial and maximal value of the learning rate[5] is performed using the standard and maximal update parameterisation (SP vs μP). Three embedding widths are considered: the nominal 256 embedding width, defining a GN2 model with 2.3M parameters, a mid-size 128 embedding width (0.72M parameters), and a small 64 embedding width model with 0.23M parameters. Interestingly, this smaller model with an embedding width of 1/4 of the full model only has a 10th of the parameters. Furthermore, the small model was found to be trainable on a single GPU while the full and mid-size models required two GPUs to be trained in a reasonable amount of time. All models are full GN2 models trained on 30M PFlow

[5]The final value - LR end - was not modified and is kept at 10^{-5} for all test due to limited compute.

jets[6] for 40 epochs with batchsize 1024. Parameters not mentioned are kept similar between embedding widths and parametrisation, and the epoch giving the lowest validation loss is chosen for each run. Figure 5.41 displays the main result from this campaign, displaying the various LR max considered at the best LR initial found (10^{-5}). Three main observations are drawn from analysing the result:

1. With μP, the wider GN2 models - larger embedding width - always outperform the smaller versions.

2. Wider models do not always outperform smaller models with SP. In particular, at large LR max, the wider model becomes unstable and its performance in terms of validation loss significantly decreases.

3. The optimal LR max (and LR init as shown in Figure 5.42) are shared across widths with μP, while no such behaviour is guaranteed for SP - but is observed in the present case.

Figure 5.42: Scan of the maximal learning rate (x-axis) versus initial learning rate (individual column) as measured by the validation loss (y-axis) of SP models (top) and the μP model (bottom) with three different embedding widths: 64 (yellow), 128 (red), and 256 (purple). Taken from [6]. The scan at LR initial = 10^{-5} benefitted from more tests to capture the sudden rise in validation loss at larger LR max for SP.

The full LR init vs LR max scans can be found in Figure 5.42 for SP and μP. Changing the LR init has little effect on the reached performance, due to the LR scheduler quickly moving away from the initial value and the common LR end value of 10^{-5} shared by all models at the

[6]Composed of 60% $t\bar{t}$ and 40% Z'.

end of training. The LR max however is a significant hyperparameter having a large impact on performance. All SP models with 256 embedding widths are found to become unstable at large values of max LR. Note that the scan at LR initial $= 10^{-5}$ benefitted from more tests to capture the sudden rise in validation loss at larger LR max. As expected from the previous discussion, all μP models stay stable, even at larger values of the learning rate. On the contrary, SP models become unstable with large LR max. μP models share the same optimal LR parameters, although some variance impacts the precision of the method on the smallest model. Due to the limited computing power available, only one seed was run per test, introducing some unmeasured statistical variance in the output. An essential conclusion in this respect is the computing gain from performing the HPO on the smaller width model than the full-width one:

- The full-width model (embedding size 256) has 2.3M parameters, taking ∼39 min per epoch on 2 A100 GPUs each fed data by 20 CPUs.
- The small-width model (embedding size 64) has 0.23M parameters, taking ∼20 min per epoch on 1 A100 GPU fed data by 20 CPUs.

Essentially, a single full-width model hyperparameter test is in computing terms equivalent to running 4 individual tests on the smaller model. Given a fixed computing budget, one can therefore have a far better coverage of the hyperparameter search space with μTransfer.

This optimisation study was carried out to demonstrate the benefits of μP on GN2. Interestingly, the optimal value found for both the μP and SP models is at an LR max $= 5 \times 10^{-4}$ and LR initial $= 10^{-5}$. The default values used in the prior training of GN2 were, by luck, the same LR max but a larger LR init of 10^{-7}. To quantify the effect on performance, the b-efficiency versus c- and light-rejection on $t\bar{t}$ and Z' of two μP models are displayed in Figure 5.43, with the suboptimal one being the worst performing full-width model (LR max $= 5 \times 10^{-5}$, LR init $= 10^{-7}$) and the optimal one the best performing one (LR max $= 5 \times 10^{-5}$, LR init $= 10^{-5}$). While the optimal and suboptimal models had close validation loss, respectively 0.601 and 0.591, a significant difference in background rejection at all efficiencies is observed. At a b-tagging working point of 70%, the suboptimal GN2 model underperforms the optimal one on $t\bar{t}$ by 18% (14%) on c-rejection (light-rejection) and the disparity is even higher on Z', rising to 24% (26%) at a b-tagging working point of 30% - which is equivalent to the 30% WP on $t\bar{t}$.

Figure 5.43: The c- and light-rejections as a function of the b-jet tagging efficiency in the $t\bar{t}$ (left) and Z' (right) test samples, from [6]. Models compared are the optimal μP GN2 (LR max = 5×10^{-5}, LR init = 10^{-5}) and the suboptimal μP GN2 (LR max = 5×10^{-5}, LR init = 10^{-7}), all with 256 embedding width. Shaded regions represent the binominal error band.

Additional tests of μP performed with GN2 showed a similar correct scaling across depth with similar optimal hyperparameters being transferrable, as expected from empirical results [159]. Due to the limited computing power available, the study of SP versus μP only encompassed two hyperparameters: the initial and maximal learning rate. The validity of the method has been confirmed and future studies optimising all the learning rate scheduler hyperparameters (including the warm-up and the learning rate at the end) will be carried out. Other hyperparameters that can best optimised with μTransfer are the initialisation variances of the different layers and the auxiliary objectives individual weights of Equation 5.10. To summarise this section on HPO, the present work introduces two approaches that are combined to deliver an improved hyperparameter optimisation:

- Executing the HPO on KubeFlow with the Katib workload to benefit from state-of-the-art autoML algorithm.
- Leveraging the μP parametrisation to increase the performance of the tuned GN2 and benefit from the factor 4 boost in hyperparameter test coverage from μTransfer.

The full optimisation of GN2 is, at the time of writing, an ongoing effort of the ATLAS Collaboration.

5.3.4 GN2X: GN2 Variant for Boosted Higgs to Heavy Flavours Decays

A final aspect of the GN2 model presented in this thesis is an application of the architecture to a specialised objective: identifying boosted Higgs boson decaying into a pair of b- or c-quarks. Having an effective tagger to identify these boosted decays can significantly help analyses studying the decay of Higgs particles to a $c\bar{c}$ pair [162], for the precise measurement of the Higgs boson p_T spectrum [163], and for beyond the SM measurements [164]. To perform this task, a new algorithm labelled GN2X is introduced based on the design of GN2 [5]. Its main task is to discriminate jets from boosted Higgs boson decaying into a $b\bar{b}$ or a $c\bar{c}$ pair from those originating from the fully-hadronic top-quark decay and the multi-jet processes. While other taggers presented in this chapter relied on small-radius ($R = 0.4$) PFlow jets or VR jets, GN2X is trained on jet reconstructed with a large-radius ($R = 1.0$) with UFO objects to capture the majority of the decay products [165]. UFO combines PFlow [166] and Track-Calo clusters objects [167], thereby including neutral and charged components in the reconstruction. UFO large-R jets are reconstructed with the anti-k_T algorithm with a radius $R = 1.0$ [69].

To train the algorithm, Higgs produced in association with a Z-boson and decaying to a pair of heavy flavour quarks ($b\bar{b}$ or a $c\bar{c}$) are simulated. To not bias the result towards a specific p_T, η, and mass distributions of the jets, the simulations are resampled to have an approximately flat distribution of jet mass in the training set, while the validation set follows the SM ZH production for a Higgs boson H of a mass equal to 125 GeV. Similarly, the top-quark decay with subsequent hadronic decay of the W boson in the $t \to bW$ chain is simulated for the training samples using a hypothetical Z'-boson of 4 TeV mass decaying as $Z' \to t\bar{(}t)$ with approximately flat jet p_T distribution. The evaluation sample uses the SM $t\bar{t}$ decay with filters on the scalar sum of the objects p_T in the event. Finally, the multi-jet process is simulated in slices of particle-level jet p_T to have the same spectrum. More details on the simulated samples used can be found in Ref. [5]. After resampling the samples to enforce the same p_T, η, and mass distributions, there are 62 million jets split between 15 million $H_{b\bar{b}}$, 15 million $H_{c\bar{c}}$, 10 million top, and 22 million multi-jets.

The previous algorithm for this task that now serves as benchmark in this study is the X_{bb} tagger, a feed-forward network combining the flavour tagging discriminants of DL1r or DL1d for up to three VR sub-jets associated to the large-R jet [168, 169]. The track selection is similar to that of the GN-models (Section 5.3), and the inputs of the model are equivalent to those of Table 5.6, with the jet variables defined on the large-R jet with the addition of the mass of the

large-R jet. At most 100 tracks associated with a jet are supplied to the network, as sorted by the decreasing transverse impact parameter significance S_{d_0}. The same auxiliary tasks as in GN2 are used with the same respective weights and neural network designs. The initialiser has a 192 embedding dimension and the transformer encoder combines 6 layers with 4 attention heads. The global representation is again obtained from an attention-weighted sum over the conditional tracks, with learnable attention weights. GN2X contains in total 1.5 million parameters and is trained on 4 A100 GPUs for 40 epochs (∼1 hour per epoch) with a batchsize of 1000.

Figure 5.44: The ROC curves for $H(b\bar{b})$ (left) and $H(c\bar{c})$ tagging (right) on an SM simulated test samples, from [5]. The respective tagging efficiency is displayed versus the top and multi-jet rejections, for jets with a $p_T > 250$ GeV and a mass $50 < m_J < 200$ GeV. Models compared are the baseline X_{bb} tagger, using the variable-radius DL1r of at most 3 identified sub-jets in the large-R jet, the tag obtained by combining the tag on two variable-radius jets within the large-R jet with the single-jet GN2 tagger, and the GN2X model. The former is only available for $H(b\bar{b})$ tagging, and the $H(b\bar{b})$ rejection is displayed for $H(c\bar{c})$ tagging. The $H(c\bar{c})$ background is negligible for $H(b\bar{b})$ tagging. Shaded regions represent the binominal error band.

The model outputs four probabilities $p_{H_{b\bar{b}}}$, $p_{H_{c\bar{c}}}$, p_{top}, and p_{QCD} that are combined in a discriminant score equivalent to Equations 5.1 and 5.2:

$$D_{H_{b\bar{b}}} = \log \frac{p_{H_{b\bar{b}}}}{f_{H_{c\bar{c}}} \cdot p_{H_{c\bar{c}}} + f_{\text{top}} \cdot p_{\text{top}} + (1 - f_{H_{c\bar{c}}} - f_{\text{top}}) \cdot p_{\text{QCD}}}, \tag{5.11}$$

where the flavour fractions were chosen from dedicated performance studies to be $f_{H_{c\bar{c}}} = 0.02$

and $f_{\text{top}} = 0.25$. A discriminant for $H_{c\bar{c}}$ is similarly defined:

$$D_{H_{c\bar{c}}} = \log \frac{p_{H_{c\bar{c}}}}{f_{H_{b\bar{b}}} \cdot p_{H_{b\bar{b}}} + f_{\text{top}} \cdot p_{\text{top}} + (1 - f_{H_{b\bar{b}}} - f_{\text{top}}) \cdot p_{\text{QCD}}}, \tag{5.12}$$

with $f_{H_{b\bar{b}}} = 0.3$ and $f_{\text{top}} = 0.25$. The performance of GN2X can be assessed from the ROC curves presented in Figure 5.44. An additional performance to X_{bb} and GN2X is presented, where two individual VR sub-jets are b- or c-tagged by a VR-trained GN2 model. The jets used are the leading VR sub-jets associated with the large-R jet. Note that X_{bb} was not retrained on the specific samples but uses the VR-trained DL1d previously introduced. A clear performance gained is delivered by the GN2X method above both the X_{bb} tagger and the combination of two individual tags with GN2. The latter approach does not access correlations between the sub-jets, explaining its lower performance at higher $H(b\bar{b})$ and $H(c\bar{c})$ efficiencies than the GN2X and X_{bb} model. At a 50% $H(b\bar{b})$ WP, GN2X improves the top rejection (multi-jet rejection) on X_{bb} by a factor 1.6 (2.5) [5]. For $H(b\bar{b})$ tagging, the $H(c\bar{c})$ background is negligible. GN2X also improves the performance for $H(c\bar{c})$ tagging over the approach combining two individual VR tagged-jets: at a 50% WP, GN2X improves the top rejection by a factor 3, the multi-jet rejection by a factor 5, and the $H(b\bar{b})$ rejection by a factor 6. This novel approach to perform boosted object tagging is the first of its kind in ATLAS and is now integrated into the ATLAS software.

5.4 Calibration

All flavour taggers presented in this chapter are trained on MC-simulated events, as described in Section 5.1.3. As such, they depend on and acquire specific features of the simulated data that might not be present in the real data collected by the ATLAS experiment. While the Collaboration aims to generate the highest-fidelity simulations possible thanks to advanced software built on GEANT4 [142] and many other specialised frameworks, inherent and unavoidable differences are left. To quantify the effect of using a simulation-trained network on real data, the ATLAS Collaboration performs Data-Monte Carlo agreement and calibration studies. These are performed in two steps:

- Data-MC SFs are derived, comparing the output of the tagger on a simulated and real dataset using the same selection [118, 170–172]. The efficiencies ϵ^f for each flavour $f \in b, c, \text{light}$ are measured, both on the simulated and real dataset, with

$$\epsilon^f(p_T) = \frac{N^f_{\text{tagged}}(p_T)}{N^f_{\text{all}}(p_T)},$$

where $N^f_{\text{tagged}}(p_T)$ is the number of jet of flavour f in the bin of p_T that are b-tagged and N^f_{all} the total number of jet of flavour f in the same bin. Scale factors to apply to simulations are then derived for each flavour f as

$$\text{SF}^f_{\text{Data-MC}}(p_T) = \frac{\epsilon^f_{\text{Data}}(p_T)}{\epsilon^f_{\text{MC}}(p_T)},$$

giving the ratio of the measured efficiency in data over simulation. To include dynamics-dependent effects of the tagger, the efficiencies ϵ^f and SF are derived in bins of jet p_T. Such calibration factors correct the efficiencies of tagging and mis-tagging and are applied to all analyses using the flavour tagger. This calibration is performed independently for each output flavour of the tagger, as it relies on selecting a portion of the ATLAS data with a dominating proportion of the specific flavour. The b-tagging efficiency is derived from a sample of $t\bar{t}$ with two charged leptons in the final state, as described in Ref. [118]. The SF for c-jet misstagging is calibrated on a $t\bar{t}$ sample decaying to exactly one charged lepton and several jets [172]. Finally, the SF for light-jets is derived in a sample of Z-bosons produced in association with jets (Z+jets) [173]. Due to the extreme rejection power of modern flavour taggers, a special technique called *flip tagger* is used for this last SF, in which a tagger is modified to have a reduced light-rejection.

- MC-MC SFs are derived between the chosen nominal Monte Carlo simulator used for training and other simulators or by changing the tuning [174]. This dependency is measured by applying the same tagger to samples simulated with different generators, mainly PYTHIA [135], HERWIG [175], and SHERPA [176] for variation to the parton shower and hadronisation and MADGRAPH for variation to the matrix element [177]. The decay chains of b- and c-hadrons in ATLAS are further simulated with the EVTGEN package [140]. These effects are measured into SFs using the same technique as the data-MC scale factors. For an alternative generator, the SFs of flavour f is derived by composing the Data-MC SFs with the nominal sample and the MC-MC SFs as

$$\text{SF}^f_{\text{Alternative}}(p_T) = \frac{\epsilon^f_{\text{Data}}(p_T)}{\epsilon^f_{\text{Nominal MC}}(p_T)} \times \frac{\epsilon^f_{\text{Nominal MC}}(p_T)}{\epsilon^f_{\text{Alternative MC}}(p_T)} = \frac{SF^f_{\text{data-MC}}(p_T)}{SF^f_{\text{MC-MC}}(p_T)}.$$

These scale factors are applied in physics analyses as a per-jet weight to the discriminant. Some early studies of both scale factor types have been performed in Ref. [4], showing good agreement between the data and simulated performance of DL1d and GN1. Variations due to the change of generator are also found to be at most 8% with respect to the nominal choice.

5.5 Conclusion

This chapter introduces the main machine learning models developed for heavy-flavour jet identification in ATLAS during the period covering 2020 to 2024. Work carried out in and presented in this thesis includes the first training of the DL1d model, including the DIPS sub-tagger of the first time in the ATLAS software. DL1d is found to have improved background rejections at a fixed working point for both b- and c-tagging compared to the at-the-time released tagger: DL1r. Significant changes in addition to the development of this new tagger were made to the preprocessing pipeline of the Umami framework [1] and the architecture as well as the list of features used. Finally, the new family of taggers based on graph neural network for GN1 and transformers for GN2 is presented, with the architecture adopted fully described and the performance of the different taggers compared. The hyperparameter optimisation of GN2 is also discussed, introducing the possibilities of using the new infrastructure of the KubeFlow server managed by CERN as well as the relevance of the maximal update parametrisation for improving the search for optimal hyperparameters of GN2. Significant contributions were made to the development of the Umami and Salt frameworks used to train the DL1 and GN families of models were made to support these studies [1, 2].

CHAPTER 6

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

Perhaps the most important raison d'être of the Large Hadron Collider (LHC) was to discover the Brout-Englert-Higgs boson (Higgs - H), a feat achieved by the ATLAS and CMS Experiments in July 2012 [13, 14]. Theorised in 1964 by two independent papers introducing the mechanism of spontaneous symmetry breaking to give mass to the gauge bosons [11, 12], its discovery almost fifty years later marked one of the greatest achievements of the particle physics community. The Higgs boson is an essential part of the Standard Model (SM). It is tied to the mechanism through which particles acquire mass without breaking the electroweak gauge invariance, as described in Chapter 2. While the gauge bosons W and Z gain mass through symmetry breaking, in the SM the fermions - quarks and leptons - acquire theirs by interacting with the Higgs fields - a scalar field carried by the Higgs boson H, as described by the Yukawa mechanism [178].

6.1 Introduction

The Higgs boson H [11, 12, 179, 180] was discovered in 2012 by the ATlAS and CMS Collaborations using the data of the LHC Run 1 [13, 14]. This triggered a race by both experiments to study the specific properties of the discovered particle, and in particular to observe its different

production processes and decay channels. The initial decay channels studied for the ground-breaking discovery where the bosonic decays of the Higgs to final states of photons and leptons: $H \to \gamma\gamma$, $H \to ZZ$, and $H \to WW$. These channels benefit from clean experimental conditions, reliable measurements, and limited backgrounds. The new particle is now being studied in ever finer detail, confirming its coupling to many massive particles of the SM and showing remarkable agreement with the properties dictated by the theory. During the LHC Run 2, corresponding to data taken from 2015 to 2018, the $t\bar{t}H$ production mechanism was observed for the first time, providing the first measurement of the top Yukawa coupling [181, 182]. Additionally, the decay of Higgs bosons to a pair of τ-lepton is now well established and different cross-section measurements have been performed [183, 184]. Importantly, the decay channel of the Higgs boson to a $b\bar{b}$ pair was observed by both ATLAS and CMS [185, 186]. This last decay channel is of particular significance since it has the largest predicted branching ratio of 58% for $m_H = 125$ GeV in the SM.

Concerning the second generation of fermions, there now is evidence of the decay to a $\mu^-\mu^+$ pair by CMS [187] and a 2 σ excess over the background-only hypothesis by ATLAS [188]. Furthermore, constraints on the branching ratio of the H to another second-generation fermion, the c-quark, have been set by both collaborations studying the $H \to c\bar{c}$ decay mode [121]. This decay mode is the most common Higgs decay mode that has yet to be observed. It is indeed particularly challenging due to the small predicted branching ratio of 2.9% [189], the large background rates, and the experimental difficulties in identifying c-jets. It is a fertile ground for new physics Beyond the Standard Model (BSM) due to the smallness of the predicted Yukawa coupling $y_c^{\mathrm{SM}} \approx 3.99 \times 10^{-3}$ [190] for the c-quark as well as an important test of the validity of the model [15–21]. The fermion Yukawa couplings in the SM are indeed largely added ad-hoc and do not explain the distinct mass hierarchy between the three generations of quarks. This question can be probed by studying the coupling strengths of the quarks to the Higgs boson. The $VH(H \to b\bar{b}/c\bar{c})$ analysis, to which this chapter is dedicated, probes the hierarchy of mass between the b- and c-quark, respectively a 3rd and 2nd generation quark.

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

While $H \to b\bar{b}$ enjoys the largest branching ratio at the observed Higgs mass, the large multi-jet background in a hadron collider like the LHC makes this decay mode very challenging. The measurements for both the $b\bar{b}$ and $c\bar{c}$ decay modes are therefore performed in a so-called *associated production mode*, where the H is produced in addition to an extra vector boson V (W or

Z) decaying leptonically, to electrons (e), muons (μ), neutrinos (ν), or a combination $e\nu$ or $\mu\nu$. Despite the relatively small cross-section of the VH production mode ($\sigma_{VH} = 2.25$ pb compared to the total H production $\sigma_H \approx 51$ pb), the process benefits from experimentally favourable conditions thanks to the presence of leptons in the event signature: these allow for efficient triggering and greatly reduce the contribution of the multi-jet background. Other analyses relying on full-hadronic final states in the associated or other production modes are also performed by the ATLAS Collaboration, but they are less sensitive to the Higgs coupling to heavy-flavour quarks. In the ATLAS Collaboration, the $VH(H \to b\bar{b})$ and $VH(H \to c\bar{c})$ analyses adopt very similar strategies. The main ingredient is the ability to reliably tag the flavour of jets produced in an event to reconstruct the heavy quark pair produced in the H decay, using the tools described in Chapter 5.

Using the full Run 2 dataset with a total integrated luminosity of 140 fb^{-1}, the published $VH(H \to c\bar{c})$ ATLAS analysis obtained the following upper limits on the signal strength of the $VH(H \to c\bar{c})$ as predicted by the SM: an observed (expected) upper limit of 26 $\times$ SM (31 $\times$ SM) [122]. The measurement also provided the first constraint on the Higgs-charm coupling modified $|\kappa_c| < 8.5$. For comparison, CMS reported an observed (expected) upper limit of 14.4 $\times$ SM (7.6 $\times$ SM) and a constraint of $1.1 < \kappa_c < 5.5$ [123].

For the $VH(H \to b\bar{b})$, thanks to a larger expected signal, the ATLAS analysis reaches a sensitivity of 6.7 standard deviations [191]. Having reached the observation level, the focus of this analysis has shifted towards a precision differential measurement of the fiducial cross-sections as a function of momentum in the reduced Simplified Template Cross-Section (STXS) scheme. To probe larger p_T ranges, the analysis is now split into the *resolved* [191] and the *boosted* [192] analyses, with the latter restricting to values of the transverse momentum of the associated vector boson p_T^V above 250 GeV - a variable highly correlated to the p_T of the Higgs p_T^H. The name of these analyses comes from the approach to reconstruct the Higgs boson candidate. At low p_T^V, the two b-jets from the H-boson decay can be independently resolved into two distinct small cone radius (small-R) jets. At high p_T^V, the H-boson is highly Lorentz-boosted requiring a change of strategy: the candidate H-boson is efficiently reconstructed as a single large-radius ($R = 1$) jet merging the two b-jets. The measured signal strengths, the ratio of the measured yield to the SM predictions, are:

- For the resolved analysis in Run 2: a signal strength of $1.02^{+0.18}_{-0.17}$ corresponding to an

observed (expected) significance of 6.7 (6.7) standard deviations [191]. Due to the good sensitivity of the analysis, the result is further detailed into the WH and ZH production processes with observed (expected) significances of, respectively, 4.0 (4.1) and 5.3 (5.1) standard deviations. Furthermore, the VH cross-section times the $H \to b\bar{b}$ and $V \to$ leptons branchings fractions ($\sigma \times BR$) are reported in the reduced Simplified Template Cross-Section (STXS) scheme. Finally, limits are set on the coefficients of effective Lagrangian operators which can affect the VH production and the $H \to b\bar{b}$ decay.

- For the boosted analysis: a signal strength of $0.72^{+0.39}_{-0.36}$ corresponding to an observed (expected) significance of 2.1 (2.7) standard deviations [192].

Some preliminary studies aiming at combining the different analyses have already been performed, with the resolved $VH(H \to b\bar{b})$ and $VH(H \to c\bar{c})$ analyses combined in Ref [122] and the resolved and boosted $VH(H \to b\bar{b})$ combined[1] in Ref [194]. These combinations require careful studies to remove the overlap between the analyses, such as by introducing a switch in p_T^V at 400 GeV between the resolved and boosted strategies. However, they rely on the published analyses and are therefore not optimised. The objective of the Combined Analysis presented here is to define a common analysis strategy, correlating as much as possible the experimental and modelling uncertainties for both Higgs decay modes and p_T^V regimes, thereby improving the measurements of $VH(H \to b\bar{b})$ and $VH(H \to c\bar{c})$ simultaneously. This new combined measurement has several additional benefits:

- The Higgs-charm and -beauty coupling modifiers, κ_c and κ_b, can be measured directly, as well as their ratio κ_c/κ_b.
- The auxiliary measurements of background processes are shared, leading to a better knowledge of background processes that contribute to both phase spaces such as the V+jets and top-quark processes.
- The combined analysis benefits from improved signal selection thanks to upgraded physics objects and event reconstruction techniques. In particular, new machine learning-based techniques are integrated for both the event selection and flavour tagging.

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

[1] CMS published an analogous combination in Ref [193].

modifications to the analysis are expected in the soon-to-be-published final result, in particular to the modelling strategy and the categories entering the fit. The work presented here is largely based on the internal documentation of the experimental team and personal work carried out during the duration of the DPhil project.

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

The Combined Analysis is performed with the full ATLAS Run 2 proton-proton collision data, collected from 2015 to 2018, for a total integrated luminosity of 140 fb^{-1} at a centre of mass energy $\sqrt{s} = 13$ TeV. The regions and boundaries between the different regimes of the analysis are illustrated in Figure 6.1. The $VH(H \to b\bar{b})$ and $VH(H \to c\bar{c})$ parts are separated by the required presence of two b-tagged jets or a c-tagged jets respectively[2]. The p_T^V cut marks the difference between the Higgs candidate reconstruction scheme of the resolved and boosted $VH(H \to b\bar{b})$: two small radius (R = 0.4) jets for $p_T^V < 400$ GeV and, above, one large radius (R = 1) jets with two b-tagged Variable Radius (VR) track-jets associated to the large-R jet.

Figure 6.1: The analysis regimes considered in the combined $VH(H \to b\bar{b}/c\bar{c})$ analysis.

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

[2]More precisely a tight c-tagged jet, as described in Section 6.5

with good statistical significance, thereby offering a suitable test to verify the validity of the strategy deployed. The main backgrounds are the production of a vector boson with additional jets(V+jets, mostly Z+jets in 0L and 2L, W+jets in 1L) and the top-quark processes (*Top*, predominantly the top-quark pair production $t\bar{t}$, with one of the t decaying leptonically, and a sub-leading contribution from single top-quark production with an extra W boson, both in 0L and 1L). Minor backgrounds are the QCD multi-jet[3], the single-top process (without an extra associated W boson) and non-signal diboson pair productions (VV). The processes are further described in the Section 6.4.

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

All backgrounds are simulated using Monte Carlo (MC) simulation packages, except for the multi-jet and the top background in 2-leptonic channel which are estimated from a data-driven method. The multi-jet is only included in the 1-lepton channel, as it is negligible in the other channels. The chapter is separated into different sections introducing the datasets and MC simulations (Section 6.4), describing the objects reconstruction techniques (Section 6.5.1), the analysis selection and regimes (Section 6.5.2), the event categorisation (Section 6.5.3), the experimental and process modelling (Section 6.8 and 6.9), the fit framework (Section 6.10.1), and finally the main results (Section 6.10.2).

6.4 Data and Simulated Samples

The combined analysis is performed on data collected during Run 2 of the LHC, with proton-proton collisions recorded between 2015 and 2018 at a $\sqrt{s} = 13$ TeV for an integrated luminosity of 140 fb^{-1} [40]. Data events passing some quality requirement are selected, ensuring for example

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

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

Table 6.1: The nominal Monte Carlo samples used in the $VH(H \to b\bar{b}/c\bar{c})$ analysis, and the corresponding process cross-sections at $\sqrt{s} = 13$ TeV. The PDF sets mentioned in the table are used for the matrix element.

that all the sub-detectors were correctly operating. The analysis requires extensive and accurate Monte Carlo (MC)-based modelling of the signal and the background processes, except for the QCD multi-jet and the $t\bar{t}$ background in the 2-lepton channel which have data-driven estimations. All MC samples are simulated with ATLAS detector effects [141] using GEANT4 [142]. The nominal samples are produced using the prescriptions described in Table 6.1, detailing the Matrix Element (ME) generators, Parton Shower (PS), and Parton Distribution Function (PDF) releases used as well as the cross-sections. Samples are normalised either to the best theoretical cross-section predictions or the generator cross-sections.

Both simulated samples and data are reconstructed with the offline reconstruction software of ATLAS. The EVTGEN 1.6.0 program is used to simulate the properties of b- and c-hadrons decays[4] [140]. Pile-up is included in the simulation, both from multiple interactions in the same and adjacent bunch crossing. This is performed by overlaying events with minimum bias simulated using PYTHIA 8 with A3 tune and interfaced with the NNPDF 2.3 PDFs [135]. The rest of this section gives more details about the simulation of the different processes. When relevant, alternative samples generated from a different setup to the nominal samples are introduced. These alternative samples are used to assess process modelling uncertainties in Section 6.9.1, as summarised in Figure 6.14.

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

6.4.1 Signal Processes

The analysis targets the $VH(H \to b\bar{b})$ and $VH(H \to c\bar{c})$ processes, here called *signals*. The Leading Order (LO) Feynman diagrams contributing to the associated production VH are the qq-initiated modes depicted in Figure 6.2. A gluon-initiated production of ZH is also possible at Next-to-Leading Order (NLO) with a quark loop (mostly top-quark), as depicted in Figure 6.3. The ME calculations are based on the POWHEG-BOX V2 generator [132, 133]. The qq-initiated VH samples are simulated with the POWHEG generator with the multiscale improved NLO (MiNLO) procedure [195], with one-loop amplitudes computed with the GoSam automated software [196]. The qq-initiated samples simulate Parton Shower (PS), UE, and multiple parton interactions with PYTHIA 8.245, while the gg-initiated use PYTHIA 8.307 [135]. Both use the AZNLO tune [197] with PDFs based on the NNPDF3.0NLO for matrix elements [134].

Figure 6.2: Leading order Feynman diagrams for $VH(H \to b\bar{b}/c\bar{c})$, qq-initiated with V decaying into leptons.

Figure 6.3: Feynman diagrams of the gg-initiated contributions to $ZH(H \to b\bar{b}/c\bar{c})$.

The inclusive cross-sections for WH and ZH are calculated at NNLO in QCD [198] and NLO in Electroweak (EW) [134]. The gg-initiaed ZH contribution relies on the LO prediction from POWHEG instantiated with PYTHIA 8.

Alternative samples: are simulated with POWHEG+MiNLO+HERWIG 7.0, with the same simulation stack as the nominal but replacing PYTHIA 8 by HERWIG 7.0 [175] for the simulation of the PS, hadronisation, UE, and multiple parton interactions.

6.4.2 Background Processes

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

V+jets

The production of a gauge vector boson V in association with jets is the largest background in the analysis. Some leading contributing Feynman diagrams to this process are presented in Figure 6.4. Both the Z+jets and W+jets are simulated with SHERPA 2.2.11 [199], which delivers NLO precision on ME computation for up to 2 jets and LO accuracy for between 3 and 5 jets. PS and hadronisation are treated by the default SHERPA generator, with the NNLO PDFs based on NNPDF3.0nnlo [134]. Uncertainties from missing higher orders are evaluated by varying the QCD renormalisation and factorisation scales, μ_R and μ_F, in the matrix elements by respective factors 0.5 and 2. Flavour filtering is applied to generate samples enriched with heavy flavour quarks.

Figure 6.4: Examples of Feynman diagrams of the V+jets process. The left diagram gives jet pairs of the same flavour due to the gluon splitting, while the right one can give mixed flavours.

Alternative samples Two sets of alternative samples are available:

- MADGRAPH FXFX samples are produced for the modelling studies, using the MADGRAPH5_AMC@NLO 2.6.5 program [177]. This generates events with V-boson and up to three additional partons in the final state at NLO accuracy. The scales μ_R and μ_F are set to 1/2 the transverse mass of all final-state partons + leptons. PYTHIA 8.240 is interfaced for showering and hadronisation, with the A14 tune and the NNPDF2.3lo PDF set with $\alpha_s = 0.13$.

- SHERPA 2.2.1 [176] samples are used as alternative generator. They give different p_T^V distributions to 2.2.11 [200], an important modification given the observed data-MC dis-

agreements in the p_T^V distributions when using Sherpa 2.2.11. These samples are similar to those used in the standalone $VH(H \to c\bar{c})$ [122].

Top-pair Production

The $t\bar{t}$ process is the second most important background in the analysis. The leading order Feynman diagram for this process is shown in Figure 6.5. The nominal samples are generated for the 0L and 1L channels with Powheg at NLO calculation of the matrix element [130, 131]. It is interfaced with Pythia 8.230 with the NNPDF3.0NLO PDFs using the A14 tune for PS, hadronisation, and UE description. Filtering is applied while simulating to enhance statistics. Cross-sections are calculated to NNLO in QCD, with resummation of the Next-to-Next-to-Leading Logarithmic (NNLL) soft-gluon terms [201].

Figure 6.5: Feynman diagrams of the $t\bar{t}$ production and decay, where each W can decay leptonically or hadronically.

Alternative samples Several alternatives are simulated for modelling studies:

- Replacing Pythia by Herwing 7.0 with H7UE tune [202] for parton shower, hadronisation, and UE modelling while keeping the same nominal Powheg setup. This sample is used to systematically assess variations to the parton shower model.
- Replacing Powheg by MadGraph5_aMC@NLO [177] for NLO hard-scattering matrix element modelling with the nominal Pythia for PS, hadronisation, and the UE simulation. This sample is used to systematically assess variation in the matrix element prediction.
- Weights variations tuning the initial and final state radiations (Initial State Radiation (ISR) and Final State Radiation (FSR)) contributions relative to the nominal setup. There are 4 such variations, based on the nominal Powheg + Pythia 8.230:
  - A high- and low-variations of ISR, where the μ_R and μ_F scales are doubled and halved.

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

Single-top Production

The so-called single-top process combines different minor channels, with the leading Feynman diagrams depicted in Figure 6.6. The dominant contribution is the associated top-production Wt channel, with the $t \to Wb$. Two other contributions are the t- and s-channel, with the former having a significantly increased cross-section over the s-channel. These processes are simulated similarly to the $t\bar{t}$, with the cross-sections calculated for a top-quark mass of $m_t = 172.5$ GeV at NLO in QCD for the s- and t-channel [203, 204] and with approximate NNLO accuracy from NNLL soft-gluon resummation for the fiducial Wt production cross-section [205, 206].

Figure 6.6: Feynman diagrams of the Wt-production (top) and the single top production (bottom) in the t-channel (left) and the s-channel (right).

The Wt production channel has diagrams overlapping with the $t\bar{t}$ production at NLO in QCD. In the analysis, a diagram subtraction (DS) scheme is applied to remove the overlap with $t\bar{t}$ by locally cancelling the $t\bar{t}$ contribution in the NLO Wt cross-section calculation [207].

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

- The 2 alternative generators and the 4 changes to the ISR and FSR used for the alternatives of $t\bar{t}$ are also applied to the Wt- and t-channels.
- For Wt only, a sample using an alternative overlap removal procedure is produced with the alternative diagram removal (DR) scheme [207] to systematically model the overlap

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

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

Diboson Process

The diboson processes WW, WZ, and ZZ enter the analysis both as a background, with a hadronically decaying V-boson mistaken for the Higgs, and as a cross-check signal when decaying into a $b\bar{b}$ or $c\bar{c}$ pair. Some leading qq-initated Feynman diagrams are depicted in Figure 6.7, with gluon-initiated diagrams also possible via quark-loops. The qq-initiated diboson are simulated similarly to the V+jets, using Sherpa 2.2.11 [199]. The gg-initiated processes are simulated with the older Sherpa 2.2.2 version. In both cases, the cross-sections are computed at NLO precision, with the NNLO PDFs based on NNPDF3.0nnlo [134] for both the matrix element and parton shower.

Figure 6.7: Feynman diagrams of the diboson production in the t- (left) and s-channel (centre & right). The t-channel can lead to any combination of W and Z depending on the initial quark-pair.

Alternative samples:

- Powheg v2 interfaced with Pythia 8 samples are produced to systematically assess ME and PS variations.
- Sherpa 2.2.1 samples are produced to systematically model the impact of varying the fragmentation function.

QCD Multi-jet

This process is estimated from data instead of simulations because of the difficulty in generating sufficient statistics samples due to the low selection efficiency, despite having a much larger production cross-section than the Higgs. QCD multi-jet events can be selected when heavy flavour hadrons decay semi-leptonically or jets are misidentified as leptons. Such leptons are

normally not isolated, and only a small fraction passes the lepton requirements. The multi-jet is negligible in the 0-lepton and 2-lepton channels thanks to the strict selections available. In the 1-lepton resolved channel, the remaining contribution is assessed from data-driven templates for $VH(H \to b\bar{b})$ or as a side control region for $VH(H \to c\bar{c})$. In both cases, a region enriched in multi-jet is defined by inverting the lepton isolation requirements. The residual multi-jet is mostly present at low p_T^V values and is therefore ignored in the boosted regime.

6.5 Selection and Categorisation

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

6.5.1 Object Selection

As introduced in Chapter 3, the ATLAS software supports different object reconstruction techniques. The reconstruction strategy of the relevant objects to the $VH(H \to b\bar{b}/c\bar{c})$ analysis is presented in this section.

Primary Vertex: all events considered in the analysis are required to have at least one primary vertex reconstructed from tracks in the ID [208].

Electrons: are reconstructed by matching a deposit in the electromagnetic calorimeter with a track in the ID [68, 209]. Electrons are required to have $p_T > 7$ GeV and $|\eta| < 2.47$. They are identified with a *loose* working point of a likelihood discriminant, matching the calorimeter shower shape to an associated track. The e candidates must satisfy p_T-dependent isolation criteria in both the ID and calorimeter. In the 1L channel, the *tight* likelihood criterion is used with stricter calorimeter isolation requirements to better reject the multi-jet background. Additional requirements on the electron selection are lepton channel-dependent and summarised

in Table 6.2. VH-Loose electrons require a loose likelihood identification and are applied in all channels. Additionally, the WH-Signal and ZH-Signal criteria are respectively applied in the 1L and 2L channels, with a tighter p_T due to the trigger threshold. The 1L likelihood identification and isolation selections are tighter to suppress the multi-jet background.

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

Table 6.2: Electron Selection requirements.

Muons: are reconstructed by matching an energy deposit in the muon detector with information from the ID and Muon Spectrometer (MS) [210]. They are required to have $p_T > 7$ GeV, $|\eta| < 2.7$, to satisfy a *loose* identification criteria, and be isolated in the ID according to p_T-dependant criteria. These requirements are summarised in Table 6.3 and vary depending on the lepton channel similarly to the electron requirements. The VH-Loose requirements are applied to muons in all channels. The WH-Signal and ZH-Signal are additionally applied to the 1L and 2L channels respectively, with a stricter track-based isolation used in 1L to suppress the multi-jet background.

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

Table 6.3: Muon Selection requirements.

Taus: hadronically decaying τ-leptons are identified and vetoed in 1L using an RNN-based tagger [76]. Taus are required to have a $p_T > 20$ GeV, $|\eta| < 2.5$, and to have 1 or 3 associated tracks. In 0L and 2L, if the jet passes a *loose* working point requirement for hadronically decaying τ-leptons, it is no longer considered a jet and cannot be considered as a candidate for the reconstruction of the Higgs boson.

Missing Transverse Energy: neutrinos are not detectable in ATLAS and their presence is inferred from momentum imbalance in the transverse plane to the beam axis. E_T^{miss}, also called MET, is the negative vectorial sum of the transverse momentum of physics objects, namely

electrons, muons, hadronic τ, and jets. An additional track-based *soft term* is added, to include a contribution from good quality tracks associated with the Primary Vertex (PV) but not to any reconstructed physics object [77].

Jets Three types of jets are used by the analysis, all reconstructed with the anti-k_t algorithm [69]:

1. Small-R jets: are reconstructed from topological clusters of energy deposit in the hadronic calorimeter based on the reconstructed PFlow objects with a radius $R = 0.4$. A jet is considered as *central* if $|\eta| < 2.5$ and $p_T > 20$ GeV, and as *forward* if $2.5 \leq |\eta| < 4.5$ and $p_T > 30$ GeV. Central (forward) jets with a $p_T < 60$ GeV ($p_T < 120$ GeV) are required to originate for the primary vertex as identified by the jet vertex tagger (JVT) to limit the pile-up background [211]. *Tight* jet cleaning criteria are applied to suppress non-collision background. Central jets are used in the resolved regime to reconstruct the Higgs candidate, based on flavour tagging.

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

3. Variable-R (VR) track-jets: are reconstructed with a p_T-dependent radius, optimised for double b-tagging efficiency of the boosted $H \to b\bar{b}$ decays [212]. They are required to have $p_T > 10$ GeV and $|\eta| < 2.5$. These track-jets are used to reconstruct the b-tagged objects inside the large-R jet and to define a boosted control region rich in the top background.

Flavour Tagging: Jet flavour tagging is perhaps the most important part of the event reconstruction. The latest available DL1r tagger from Run 2 is used in the analysis for both b- and c-tagging in the resolved and boosted regime [128]. The methodology differs slightly between the two regimes of the analysis due to the different flavour tagging needs:

- In the resolved $VH(H \to b\bar{b}/c\bar{c})$, DL1r is used to tag both b- and c-jets. The so-called Pseudo-Continuous Flavour Tagging (PCFT) scheme, illustrated in Figure 6.8, is deployed to allow for a coherent joint definition and simultaneous calibration of b- and c-tagged jets, adopting the technique first introduced for 2D c-tagging in the $VH(H \to c\bar{c})$ analysis [122]. The DL1r tagger assigns a b-score[6] and a c-score[7] to every selected jets. To tag

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

a jet, the associated score must be higher than a specific cutoff value, defined to give a specific selection efficiency in simulated data, also known as a Working Point (WP). From this score, the jet is assigned one of 4 possible labels, based on 2 b-tagging WP and 2 c-tagging WP. These WP are tested in strict successive order, with first a 60% tight b-tagging working point (bin 4) followed by a looser 70% b-WP (bin 3). A jet passing these selections is labelled B[8]. Otherwise, it is considered for c-tagging with first a *tight* working point at 20% efficiency (bin 2), followed by a *loose* WP at an exclusive efficiency of 20% (bin 1) on the remaining jets - so that 40% of the c-jets are effectively selected in the combined tight and loose bins. A jet selected by the tight c-tagging WP is labelled T, and L if it only passes the loose WP. A jet failing to pass all WPs is not tagged and labelled N (bin 0). The b-tagging WPs correspond to official ATLAS ones for DL1r [128], while those for c-tagging are optimised for the purpose of the analysis. The tagging efficiency of each bin is displayed in Table 6.4, shown for the main flavours as well as τ-leptons reconstructed as jets. The calibration of all five bins of Figure 6.8 is performed simultaneously for the

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

Table 6.4: Jet tagging efficiency for b-jet, c-jet, light-jet and τ-jet in the Pseudo-Continuous Flavour Tagging (PCFT) bins, measured from a POWHEG+PYTHIA 8 simulated sample of semi-leptonic $t\bar{t}$ events.

analysis following the methodology described in Ref [128], with some results presented in Appendix **??**.

- The boosted regime only targets b-jets, with the single-jet DL1r tagger used. As such, the standard pseudo-continuous b-tagging method is used [128]. The track-jets associated with the leading large-R jet are given a b-score based on the per-flavour probabilities outputted by DL1r. The 85% working point is adopted to maximise the signal yield, due to the important statistical limitations in the boosted regime. Track-jets passing this working point are B-tagged, otherwise, they are untagged N. Studies showed that the very loose DL1r WP gives a better expected statistical significance than the then-available X_{bb} tagger. The official calibration from Ref [128] is used and extended to higher p_T with uncertainty

[8]The difference between these b-tagged is used in the discriminant MVA of the analysis

Figure 6.8: The Pseudo-Continuous Flavour Tagging (PCFT) scheme used to simultaneously define 2 b-tagged, a tight c-tagged, a loose c-tagged, and a non-tagged bins.

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

While the analysis was underway, the superior single-jet GNN taggers and the boosted objects $GN2X$ tagger introduced in Chapter 5 were not yet available as their calibration was an ongoing effort. Furthermore, switching to these new taggers was not feasible from a practical point of view in the timing of the analysis. They represent, however, an exciting avenue for progress in future iterations of this search, in both the resolved and boosted regimes.

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

6.5.2 Event Selection

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

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

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

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

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

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

large-R jet $p_T > 250$ GeV.

In all regimes, the number of reconstructed charged lepton in the final state defines three channels as the 0-lepton (0L), 1-lepton (1L), and 2-lepton (2L). The objective of this leptonic selection is to reconstruct the associated V-boson. The selection of events in the resolved regime is presented in Table 6.6 and Table 6.5 for the boosted regime. Additional channel-specific requirements are also introduced to limit background contamination and reviewed in this section.

Selection specific to the 0-lepton channel

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

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

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

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

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

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

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

Selection specific to the 1-lepton channel

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

Selection specific to the 2-lepton channel

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

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

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

Table 6.5: Summary of the event selection in the 0-, 1- and 2-lepton channels of the boosted $VH(H \to b\bar{b})$ regime.

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

Table 6.6: Summary of the event selection in the 0-, 1- and 2-lepton channels of the resolved $VH(H \to b\bar{b}/c\bar{c})$ regimes. The resolved 1L & 2L Top CR tagging definition ignores the candidate jet tagging requirements For $VH(H \to c\bar{c})$, an extra CR for V+lf changes the candidates tagging to one L-tag + no-tag (LN).

6.5.3 Event Categorisation

Selected events are further categorised following a successive decomposition into regions of defined tag, vector boson V transverse momentum p_T^V, and number of jets N_{jet}. The full categorisation gives rise to signal and control regions that enter the statistical analysis defined in the fit framework of Section 6.10. The control regions are defined to help constrain the modelling of specific backgrounds, such as by correcting their yields. The definition of the regions depends on the analysis regime and the targeted Higgs decay, with Figure 6.17 providing a condensed global overview of the different regions defined.

Resolved Regime Categorisation

In the resolved regime, the number of central + forward jets in an event defines different N_{jet} categories, separated to maximise the signal sensitivity. A $p_T > 30$ GeV cut is considered for non-Higgs candidate jets just to determine this number of jets categorisation. This was found to limit the signal migration across STXS bins in $VH(H \to b\bar{b})$ while having almost no impact on the $VH(H \to c\bar{c})$ sensitivity[12]. All distributions of the resolved regime regions with processes normalised to their posfit expectations from the fit described in Section 6.10 are presented in Appendix **??**. The plots presented in this section show the prefit and blinded distributions in the different regions, and simulated processes are therefore not normalised to data. The variables displayed correspond to those chosen for the fit, as detailed in Section 6.7. The precise definitions of the analysis regions are reviewed in this section.

Resolved $VH(H \to b\bar{b})$ SRs: requires exactly 2 b-tagged jets (BB), with no extra B- nor any T-tags, and events are separated into different categories based on N_{jet}. All lepton channels have a 2-jet and a 3-jet categories. The 0L channel has an additional 4-jet category, and the 2L an extra 4 or more jets (4p or $\geq$4) category[13]. They are included to improve the STXS measurements sensitivity in bins with at least one additional jet. All regions are further split into different bins of p_T^V as [75, 150] GeV (except for 0L[14]), [150, 250] GeV, and [250, 400] GeV. Selected $VH(H \to b\bar{b})$ signal regions in the analysis are presented in Figure 6.9.

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

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

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

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

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

Figure 6.9: BB-tagged 2-jet $150 < p_T^V < 250$ GeV signal regions in all lepton channels.

region is then split based on the remaining candidate tag into a 2 c-tags region and a 1 c-tag region: the former requiring an extra loose (LT) or tight c-tag (TT)[15], the latter an additional non-tagged jet N (NT). The p_T^V bins are similar to $VH(H \to b\bar{b})$, except for the highest p_T^V one that is relaxed to ≥ 250 GeV given the limited impact of the overlap with the boosted $VH(H \to b\bar{b})$. Adding the p_T^V region above 400 GeV was found to increase the total $VH(H \to c\bar{c})$ sensitivity by 10%. The jet multiplicity N_{jet} defines a 2 and a 3 jets categories, with the latter being 3 or more jets (3p or $\geq$3) only in 2L thanks to a reduced $t\bar{t}$ background. A selection of 2 c-tagged signal regions is presented in Figure 6.10, with Figure 6.11 presenting some 1 c-tagged signal regions. The 1 c-tag Signal Regions (SRs) in the 75 GeV $< p_T^V <$ 150 GeV range are not included in the fit to their significant background yield.

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

Figure 6.10: Selection of 2 c-tagged ($TT + LT$) $150 < p_T^V < 250$ GeV signal regions.

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

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

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

Figure 6.11: Selection of 1 c-tagged $250 < p_T^V$ signal regions.

$\Delta R(j_1, j_2)$[16] (shortened as ΔR in this document) between the Higgs-candidate jets. This split is governed by a p_T^V-dependent cut on the ΔR that is derived to give a specific signal purity in the SR: keeping 95% (85%) of the signal yield in the 2-jet (3 or more jets) SRs. The mathematical expression of the cuts is presented in Table 6.7, and illustrated in Figure 6.12, with more details given in Appendix **??**. Events with a ΔR below the cutting line enter the signal region, while those above go in a High ΔR CR, also called *CRHigh*. To avoid some mis-modelling effect at high ΔR and keep the High ΔR CR kinematically close to the SR, an uppercut of $\Delta R \leq \pi$ is applied to all events. This effectively removes $\sim 40\%$ of events in the High ΔR CR, with a negligible impact on the signal region. For $VH(H \to c\bar{c})$, CRHighs are considered for every 1 and 2 c-tagged SRs[17], with the TT- and LT-tagged events separated in the CRHigh to respectively constrain the $V{+}hf$ and $V{+}mf$instead of being merged as in the SR. In $VH(H \to b\bar{b})$, the CRHighs are used to extract the normalisation of the backgrounds while in $VH(H \to c\bar{c})$ the shapes of the $m_{c\bar{c}}$ and p_T^V spectrum are also used, as detailed in Section 6.7. Some High ΔR CRs are shown in Figure 6.13.

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

Table 6.7: Cuts defining the High ΔR (centre) and Low ΔR (right) control regions, CRHigh & CRLow. The inequalities are set to enter the control regions, with p_T^V expressed in GeV.

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

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

Figure 6.12: The p_T^V-$\Delta R_{c\bar{c}}$ 2D signal yield map of the 1-lepton $VH(H \to c\bar{c})$, for the 2-jet (left) and 3-jet (right) regions. The lines are the results of fitting the high and low $\Delta R_{c\bar{c}}(p_T^V)$ cuts for various signal tags, with the yellow curve showing the $VH(H \to b\bar{b})$ $\Delta R_{b\bar{b}}$ cut used in the analysis, with the CRHigh above the top yellow line, and the SR below. A Low ΔR CR can be defined by the bottom lines, splitting an extra region from the signal region for the $VH(H \to b\bar{b})$ only.

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

Figure 6.13: Some High ΔR CRs (CRHigh) with 2 jets and $150 < p_T^V < 250$ GeV.

The Low ΔR Control Regions: low-ΔR control regions ($CRLow$) are defined in $VH(H \to b\bar{b})$ 1L to better constrain the W+hf contribution. They are based on p_T^V-dependant cuts defined similarly to the High ΔR ones, separating 10% of the diboson events from the signal regions, as displayed in the bottom parts of the plots in Figure 6.12. The cuts used are defined in the right of Table 6.7, where events with a ΔR above the cutting line enter the signal region while those below go to the CRLow. In $VH(H \to c\bar{c})$ and the 0L and 1L $VH(H \to b\bar{b})$, the CRLow is not separated from the signal region as it has little impact on the sensitivity of the fit. One of the CRLow regions is presented on the left of Figure 6.15.

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

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

and $VH(H \to c\bar{c})$, with similar p_T^V and jet multiplicity categorisation as the SRs. In 0L and 1L, they are defined by requiring events to have at least one B-tag and at least one tight c-tag T, making them orthogonal to the $VH(H \to b\bar{b})$ signal regions. The Higgs candidate is reconstructed from the leading B jet among B-tagged jet and leading T jet among T-tagged jet, for kinematic similarity to the SRs. The Top(bb) component, which is significant for $VH(H \to b\bar{b})$ due to the required tag, is controlled from the previously defined CRHighs in 0L and 1L, thanks to the large ΔR between the produced b jets in a $t\bar{t}$, as shown in Figure 6.13a. Two Top BT control regions are presented on the left of Figure 6.14.

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

Figure 6.14: Two Top CR BT-tagged (left & centre) and a Top $e\mu$ CR (right), all with 3 jets and $150 < p_T^V < 250$ GeV.

Top Control Regions in 2L: there, the Top background is mostly made of di-leptonic $t\bar{t}$ decays, with both subsequent W decaying leptonically. High purity Top CRs are derived for the 2-lepton channels by requiring leptons of different flavours ($e\mu$ or μe) instead of the same flavour (ee or $\mu\mu$). This mix of flavours is possible as the leptons are produced in distinct W-boson decays. These so-called *Top $e\mu$ CRs* are used to derive a $t\bar{t}$ background template in a data-driven way for the 2-lepton SRs in $VH(H \to b\bar{b})$. For $VH(H \to c\bar{c})$, the $t\bar{t}$ is a less significant component due to the flavour tagging requirements, and the Top $e\mu$ CRs contribute to the fit as single-bin 2-lepton CRs per p_T^V and jet multiplicity, with at least one T-tag jet. An example of such a CR is presented in Figure 6.14c.

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

[19]The V+lf corresponds to a grouping of the V+jets components with light-flavour jets, as introduced in Section 6.9.3.

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

Figure 6.15: A BB-tagged Low ΔR CR (left) and two LN-tagged $V+l$ CRs (centre & right), both with 2 jets and $150 < p_T^V < 250$ GeV.

B-tagged jet in the event. The selection is otherwise similar to that of the 1 c-tagged signal regions[20], with the candidate pair now tagged as LN, where N is the leading untagged central jet. The 1L $V+l$ CRs are 60% pure in W+lf, while the 2L $V+l$ CRs reach a 70% Z+lf purity. An example of the former is shown in Figure 6.15b, while a 2L $V+l$ CR is shown in Figure 6.15c.

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

Figure 6.16: Some boosted BB-tagged with $400 < p_T^V < 600$ GeV regions signal regions (left & centre) and boosted Top CR (right).

Boosted Regime Categorisation

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

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

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

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

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

are moved to the boosted Top control regions in the 0L and 1L channels. The $t\bar{t}$ process is the main background in these lepton channels, where a t-quark decay is captured as a single large-R jet merging the produced b and a hadronically decaying W. The boosted Top CRs effectively capture this signature by identifying the b-quark from the other decaying t-quark in the $t\bar{t}$ pair, with the same 85% b-tagging WP. An example of such a region is displayed in Figure 6.16.

Figure 6.17: The $VH(H \to b\bar{b}/c\bar{c})$ combined analysis regions, showing the Signal Regions (SR), High and Low ΔR control regions (CRHigh & LowCR), the Top BT CR, the Top $e\mu$ CR, the $V + l$ LN-tagged CR, and the boosted Top CR. $VH(H \to c\bar{c})$ is in blue, the shared Top BT CR in the resolved regime in yellow, and $VH(H \to b\bar{b})$ in green and red for the resolved and boosted regimes respectively. Regions used in the fit as single-bin distributions to derive an absolute normalisation are indicated in italics.

6.6 Tagged-jets corrections

Several corrections to the energy are applied to tagged jets from the previously introduced selection. The objective is to improve the energy resolution of the pair of jets selected to form the Higgs candidate. All jets benefit from a default jet energy calibration called the *Global Sequential Calibration (GSC)* [71], as introduced in Section **??**. This global correction is not optimal for b- and c-jets that benefit from special features, motivating the use of additional dedicated corrections for such jets. Table 6.8 summarises the additional corrections presented in this section.

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

Table 6.8: The different H-candidate jet energy correction.

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

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

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

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

c-jets.

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

FSR recovery is deployed for events with 3 or 4 jets in the 2L resolved regime, to further improve the resolution of the m_{bb} or m_{cc} peak after the kinematic fit correction. Such events are likely to have jets emanating as Final State Radiation (FSR), whereby a quark or a gluon is emitted by a final state particle. A continuous cut on the sum $\Delta R_{j,j_1} + \Delta R_{j,j_2}$ of angular separations between a third or fourth jet (j) to the Higgs-candidate jets (j_1 and j_2) is applied as a function of p_T^V. Any additional jet below the cut is considered as a radiation and is added to the closest candidate jet. This effectively corrects the reconstructed mass of Higgs bosons as well as the jet multiplicity, leading to an expected 7% improvement in $VH(H \to b\bar{b})$ STXS sensitivity by reducing the migration between measurement bins. Due to the possible increased acceptance of the $t\bar{t}$ background in the sensitive region from the reduction of additional jets, this correction is not applied to 0L or 1L.

The effects of the different reconstruction techniques are illustrated in Figure 6.18 for some selected 2-lepton resolved and boosted distributions.

6.7 Discriminant Variables

The analysis leverages a varied set of reconstructed physical variables in the fit to constrain the different processes and control mis-modelling effects. Figure 6.19 displays the variables used for each control region in the fit in the resolved regime. The reconstructed Higgs mass directly offers some separation power of the signals from their major backgrounds, hence some control regions such as the Top BT CR and CRHigh are modelled with these relevant variables: m_{bb}, m_{cc}, and m_J, depending on the targeted decay and the regime. Some CRs passed to the fit are modelled with the p_T^V distribution, such as the CRHigh in the 2L-channel NT and BB tagged-regions and the $V + l$ CR (LN-tag). Directly fitting this distribution helps constrain a Monte-Carlo

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

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

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

Figure 6.18: Performance of the energy corrections on simulated samples of different analysis schemes in the 2-lepton channels, inclusive in p_T^V and number of jets.

mis-modelling in the p_T^V distributions of the SHERPA 2.2.11 V+jets samples, as detailed in Section 6.9. To optimise signal and background separation in the statistical analysis, dedicated BDT, also called MVA, are trained with the TMVA Root framework [213] in the signal regions of the combined analysis. Simple one-dimensional discriminants are built from the outputs of fine-tuned BDTs trained on specific sets of event-level input variables, as described in greater detail in this section.

6.7.1 Multivariate Analysis

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

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

Figure 6.19: Illustration of the discriminant variables used per control regions of the resolved regime in the fit. Norm-only indicates a region to extract a global normalisation and not binned by a variable.

The wide adoption of BDT-discriminants in all regimes of the analysis marks a significant improvement over the standalone $VH(H \to c\bar{c})$ and boosted $VH(H \to b\bar{b})$ analyses, generalising the successful approach first introduced in the resolved $VH(H \to b\bar{b})$ [191]. Before training, the object and event selections of Section 6.5 and the jet energy corrections of Section 6.6 are applied. To limit the number of training runs and the risk of overtraining from the low statistics of some kinematic regions, the BDTs are trained on inclusive regions combining the SRs and the ΔR-based CRs (CRHigh and CRLow). The BDTs are trained to discriminate the respective signal of the different targeted decays [21] from background samples, including V+jets, $t\bar{t}$, single-top, and diboson. BDTs were specifically trained in the following specific categories, covering the fine analysis categorisation to guarantee sufficient statistics and avoid overtraining:

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

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

  The low p_T^V bin is separated from the higher $p_T^V > 150$ GeV due to its large statistics and different background compositions.

- **Boosted $VH(H \to b\bar{b})$:** one BDT is trained per lepton channel in an inclusive bin of $p_T^V \geq 400$ GeV.

For training, the full MC samples statistics in all analysis regimes are used thanks to the so-called *GNN truth tagging*. Instead of filtering down the simulated samples by cutting away events failing to pass the flavour tagging requirements - the standard application of the selection called *direct tagging* -, this technique applies a weight to each event to represent its probability of passing the tagging selection. The result is a weighted distribution possessing the statistical precision of the full MC-samples but shaped as the directly tagged distributions. The weights in this truth tagging procedure are predicted by a GNN-based neural network passed event-level information, with more details given in Appendix **??**. The truth tagging procedure is applied to BB for $VH(H \to b\bar{b})$, and to TT, TL, and NT for $VH(H \to c\bar{c})$. The variables used for each lepton channel in both regimes are listed in Table 6.9, with more precise definitions of each variable given in Appendix **??**. Features with long tails are clipped to contain 99% of the centred distributions, and given a specifically chosen default value when they are not defined for an event. The sets of features used are the result of hyperparameter optimisation campaigns, with many other variables tested but eventually not included due to their negligible impact on the performance.

The architectures of the different BDTs are optimised, with the gradient boosting technique of Section 4.2.2 deployed in the resolved regime to improve performance and to capture effects outside the bulk of the distributions. In the boosted regime, due to the lower statistics available

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

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

Table 6.9: The variables used for the 0-, 1- and 2L channels MVA's in the resolved and boosted regimes for the $VH(H \to b\bar{b}/c\bar{c})$combined analysis. The variables are further described in Appendix **??**.

and large tails in the distributions, the AdaBoost method - introduced in Section 4.2.2 - is adopted to help stabilise the training [82]. Tables 6.10 and 6.11 list the architectures used for the $VH(H \to b\bar{b})$ and $VH(H \to c\bar{c})$ BDTs respectively, with the main and diboson BDTs sharing the same hyperparameters. For $VH(H \to c\bar{c})$, the hyperparameters are further tuned to avoid overtraining from the smaller available statistics in the 2L channel and the diboson cross-check.

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

Table 6.10: Hyperparameters of the BDTs per lepton channel of the $VH(H \to b\bar{b})$ resolved and boosted. All models used the Gini index as separation method, without pruning.

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

Table 6.11: Hyperparameters of the BDTs per lepton channel of $VH(H \to c\bar{c})$. The 2L low p_T^V region mentioned covers 75 GeV $<$ pTV $<$ 150 GeV. All models used the Gini index as separation method, without pruning.

Trainings are performed with the k-fold method, setting $k = 2$, to use the full statistics while assessing the overtraining risk. In other words, each BDT is doubly trained: once on odd events, and once on even events. The performance is assessed on the held-out fold and the final discriminant is the combination of the odd- and even-trained BDTs, thereby exposed to the whole training set. Additional overtraining checks are performed on each fold-training, comparing the trained distribution to a test distribution obtained by applying the BDT on the heldout set for the fold, as presented in Figure 6.20. The BDTs deliver a good discrimination performance, with a typical AUC of the ROC of ~ 0.9 and a large increase on the expected statistical significance of the analysis compared to using the Higgs candidate mass as discriminating variable.

(a) BB-tagged model, test AUC = 0.9.

(b) 2 c-tagged model, test AUC = 0.898.

Figure 6.20: Overtraining checks for the BDTs trained for the resolved $VH(H \to b\bar{b})$ (left) and $VH(H \to c\bar{c})$ (right) in the 0L 2-jet region with $p_T^V \geq 150$ GeV. The binned histograms are the training data (blue) and background (purple) distributions, while the data points are the equivalent test distributions - the bottom plots show the ratio of test/train.

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

6.7.2 Output Variable Transformation

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

$$Z = z_s \frac{n_s}{N_s} + z_b \frac{n_b}{N_b}, \tag{6.1}$$

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

The z_s and z_b parameters are manually tuned for each analysis regime and lepton channel, giving signal regions with a final amount of BDT bins varying from 4 to 15, as displayed in the postfit plots of Appendix **??**. An additional protection is added to avoid bins with too few data or MC statistics, by requiring at least 3 signal + background events per bin after transformation. The specific tunes of the parameters for the different regimes of the combined analysis are presented in Table 6.12.

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

Table 6.12: The optimised tune of the z_s and z_b parameter to rebin the MVAs with the *Transformation D* algorithm in different phase spaces of the combined analysis. XT is the 2 c-tagged region.

6.8 Experimental Uncertainties

While efforts are made to correctly simulate the collection and reconstruction of information with the ATLAS detector, inaccuracies permeate this procedure and must be taken into account in the statistical analysis of Section 6.10. Several types of experimental uncertainties are considered in the combined analysis, to cover the systematics effects due to the detector performance, the reconstruction of objects such as leptons and jets, and the effects of flavour tagging. Table 6.13 summarises the various contributions that are sources of uncertainty, which are detailed in this section.

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

Table 6.13: Summary of all experimental systematic uncertainties.

Luminosity & Pile-up: The measured Run 2 luminosity for ATLAS is 140 fb^{-1} with an uncertainty of 0.83% [40]. The measurement is performed by $x - y$ beam separation scans and is combined with information from dedicated luminosity-sensitive detectors. The pile-up uncertainty for simulated events is obtained by varying the data rescaling factor of the nominal average pile-up $\langle\mu\rangle$. This factor is introduced due to the observation that MC-simulated samples match data at a higher μ than used in their simulation. This rescaling factor is therefore used to reweight the data, matching a simulated-$\mu = 1.0$ to a data-$\mu = 1.09$, a rescaling summarised as 1.0/1.09. A 1σ uncertainty on the average pile-up is measured by varying the factor from 1.0/1.0 to 1.0/1.18.

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

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

- E_T^{miss}: SFs factors are included to account for the direction of the E_T^{miss} as well as the soft-term contributions.
- Electrons: uncertainties on the reconstructed values, the identification efficiency, isolation efficiency, and the energy scale and resolution are included. These are derived by comparing in data and simulations kinematic distributions in $Z \to e^+e^-$, $W \to e\nu$, and $J/\psi \to e^+e^-$ events [68].
- Muons: uncertainties on the reconstruction and identification efficiencies for muons with $p_T > 15$ GeV and $p_T < 15$ are included separately, using respectively samples of $Z \to \mu^+\mu^-$ and $J/\psi \to \mu^+\mu^-$ [210]. Additionally, uncertainties on the isolation efficiency, track-to-vertex association efficiency, momentum scale and resolution as well as charge-dependent misalignment effects are considered.
- Taus: hadronically decaying τ-leptons[23] uncertainties on the reconstruction and RNN-

[23] About 65% of τ decays are hadronic.

based identification efficiencies as well as the electron veto efficiencies are derived from samples of $Z \to \tau^+\tau^-$ and top-quark decays to taus [76, 214, 215].

Jets Jets are calibrated in dedicated analyses, of which two reduced sets of uncertainties are propagated to the combined $VH(H \to b\bar{b}/c\bar{c})$ for small- and large-R jets. For the small-R jets, these uncertainties cover *in-situ* analyses, η-intercalibration, flavour composition, punch-through jets, high-p_T hadrons, and pile-up effects as well as the jet energy scale and resolution measured in data [74, 216]. The reduced set is derived by a principal component analysis to preserve the largest correlations in certain regions of jet kinematics. Large-R jets uncertainties for the energy scale and resolution are also estimated from data [217]. An uncertainty covering the calibration discrepancy between data and MC-simulations is also included.

Flavour Tagging A dedicated calibration is performed to derive flavour tagging scale factors in the resolved regime, as described in Section 6.5, while the common ATLAS uncertainties are used for the boosted regime, as described in 5.4. These flavour tagging calibration SFs are derived by combining data-MC efficiency modelling SFs and MC-MC SFs to account for variations to parton showering and hadronisation. These scale factors are smoothed using a local polynomial kernel estimator to avoid distortions in the kinematic variables [218]. For each jet flavour, there is one uncertainty per p_T bin in the calibration. A τ-jet uncertainty is derived by copying the c-jet values and decorrelating them. Principal component analysis is deployed to reduce the large set of systematics uncertainties to 45 (5) for b-jet, 20 (4)for c-jet, and 20 (4) for light-jets in the resolved (boosted) regime. Additional uncertainties are added to model to extrapolate the performance to high-p_T jets. Truth tagging uncertainties are expected to be covered by these flavour tagging uncertainties, so no dedicated uncertainties are considered.

6.9 Signals & Backgrounds Modelling

Similarly to the experimental process, the simulations of the signals and backgrounds cannot entirely be accurate and mis-modellings are to be expected in the derived samples. These inaccuracies must be taken into account in the fit to avoid introducing bias. The modelling of the signals and backgrounds in the $VH(H \to b\bar{b}/c\bar{c})$ combined analysis is discussed in this section. The composition of the different processes changes depending on the lepton channel and the analysis category, as shown in Figure 6.21. The V+jets backgrounds are the dominant ones in the signal regions of the 0-lepton and 2-lepton channels, while the top processes contribute more

(a) Boosted regime (≥ 400 GeV). (b) 0L, $p_T^V \in [150, 250]$ GeV. (c) 0L, $p_T^V \in [250, 400]$ GeV.

(d) 1L, $p_T^V \in [75, 150]$ GeV. (e) 1L, $p_T^V \in [150, 250]$ GeV. (f) 1L, $p_T^V \in [250, 400]$ GeV.

(g) 2L, $p_T^V \in [75, 150]$ GeV. (h) 2L, $p_T^V \in [150, 250]$ GeV. (i) 2L, $p_T^V \in [250, 400]$ GeV.

Figure 6.21: The background composition of the different analysis regimes and lepton channels, with the data - Monte Carlo prefit agreement displayed in the bottom panels.

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

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

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

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

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

6.9.1 General Modelling Strategy

The combined analysis adopts some common strategies to model the backgrounds and signals that are described in this section before reviewing the specificities adopted for each process. A guideline for the modelling is to treat backgrounds coherently across analysis regimes and correlate uncertainties between the $VH(H \to b\bar{b})$ and $VH(H \to c\bar{c})$ sides when possible. The normalisations of the major backgrounds, the V+jets and Top, are free to float in the fit, with Floating Normalisation (FN) split by p_T^V and jet multiplicity when the statistics allow. Minor backgrounds are fixed at MC predictions with a normalisation uncertainty. To account for MC-generator modelling uncertainty, comparisons of the nominal samples to alternative samples introduced in section 6.4 and summarised in Table 6.14 are performed. For each process, the uncertainties are split into normalisation, relative acceptance, and shape uncertainties.

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

Table 6.14: Summary of nominal and alternative samples in the analysis. Alternative samples include different generator and systematics effects from modification to the nominal setup.

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

Acceptance uncertainties: relative acceptance uncertainties for each process cover possible changes in the distribution of events of the specific process across the different regions of the analysis phase space. They account for the migration of events between these regions and are assessed by measuring the change in the ratio of events between regions when switching to differently generated samples (indexed by i here). The priors on these uncertainties are calculated with the double ratio of Equation 6.2:

$$\text{Acceptance Unc}_i = \frac{\text{Acceptance}[\text{Cat.}^B(\text{alternative}_i\ \text{MC})]}{\text{Acceptance}[\text{Cat.}^A(\text{alternative}_i\ \text{MC})]} \Bigg/ \frac{\text{Acceptance}[\text{Cat.}^B(\text{nominal MC})]}{\text{Acceptance}[\text{Cat.}^A(\text{nominal MC})]}, \tag{6.2}$$

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

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

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

Shape uncertainties: the BDTs, m_{bb}, m_{cc}, and p_T^V shapes of the processes in the different regions are given some flexibility in the fit by introducing shape uncertainties derived from a comparison of the nominal to the alternative samples. The combined analysis introduces the novel Calibrated Likelihood Ratio Estimator (CARL) technique to derive a reweighted shape uncertainty using a neural network [219]. A DNN is trained to discriminate nominal events from alternative ones, with the process repeated for each alternative sample. The output of the CARL network is a score representing the probability for an event to belong to the alternative sample. This is used to reweight the nominal distribution into the alternative distribution, similarly to

the process of truth tagging. The advantage of this technique is that the reweighted nominal distributions benefit from much larger statistics than the alternative ones, thus smoothing out bin fluctuations and reducing the MC statistics uncertainties. Examples of such derived CARL shape uncertainties modelling the PS with MADGRAPH5_AMC@NLO for the single-top Wt process in 1-lepton are presented in Figure 6.22. Additional shape uncertainties from EW corrections, QCD scales, V+jets and diboson p_T^V modelling with SHERPA 2.2.1, parton shower alternative for the signal samples, and uncertainties for the single-top Wt DS / DR shapes are directly derived by comparing samples.

(a) SR BDT distribution. (b) High ΔR CR m_{bb} distribution.

Figure 6.22: CARL closure plots, between the nominal POWHEGPYTHIA8 (*PwPy8*, with the DR scheme) and the alternative MADGRAPH5_AMC@NLO (*aMCatNLO*), for the single-top Wt production in $VH(H \to b\bar{b})$, 1-lepton, 75 GeV $< p_T^V <$ 400 GeV, and 2 jets. The CARL interpolation (orange) of the nominal (blue) into the alternative (red) is smoother and with lower MC-stats. uncertainty. The top plots show the distributions, the middle plots the ratios, and the bottom plots the residuals.

An overview of the signals and backgrounds modelling systematics considered is presented in Figure 6.15, with the full details listed in Appendix **??**. All uncertainties presented here are further processed before entering the fit. To remove large statistical fluctuations potentially present in shape systematics, these shapes are smoothed by iteratively rebinning the distribution until the statistical uncertainty in each merged bin of the nominal distribution is smaller than 5%. When a systematic has a negligible impact on the distributions entering the fit, it is pruned away to ease convergence and reduce the fit complexity. This is applied to systematics causing a normalisation effect smaller than 0.5% or when both the up- and down-variations have the same sign. Shape uncertainties are pruned if no bin in the distribution has a deviation above 0.5% after the overall normalisation, or if only one of the up- or down-variation is non-zero. For very

Uncertainties	Resolved	Boosted
Signal		
$qqWH$ / $qqZH$ / $ggZH$ normalisations / acceptance	Values from previous analyses [122, 191, 192]	
$H \to bb$ Branching Ratio	1.61%	
$H \to cc$ Branching Ratio	From +5.53% to -1.99%	
Z+jets		
Z+hf normalisations	Floating	
Z+mf normalisation	Floating	35%
Z+lf normalisation	Floating	35%
Z+hf flavour composition ratios	8% - 12%	6% - 9%
Z+mf flavour composition ratios	4% - 10%	6% - 9%
High-ΔR CR-SR ratios	5% - 30%	-
Top CR-SR extrapolation ratios	-	15% - 25%
2L to 0L acceptance ratios	2% - 10%	3%
p_T^V extrapolation ratios	-	15%
W+jets		
W+hf normalisations	Floating	
W+mf normalisation	Floating	36%
W+lf normalisation	Floating	38%
W+hf flavour composition ratios	4% - 25%	11%
W+mf flavour composition ratios	14% - 29%	9% - 15%
W+lf flavour composition ratios	9%	-
High / Low ΔR CR-SR extrapolation ratios	2% - 63%	-
Top CR-SR extrapolation ratios	-	16% - 27%
1L to 0L acceptance ratios	3% - 30%	20%
p_T^V extrapolation ratios	-	3%
N_{jet} extrapolation ratios	12% - 20%	-
Top ($t\bar{t}$ + single-top Wt) 0L & 1L resolved		
Top(bb) normalisations	Floating	-
Top(bq/qq) normalisations	Floating	-
Flavour acceptance ratios	5% - 10%	-
1L to 0L acceptance ratios	2% - 8%	-
High / Low ΔR CR-SR extrapolation ratios	2% - 10%	-
Wt / $t\bar{t}$ ratios	12% - 48%	-
Top ($t\bar{t}$ + single-top Wt) 2L resolved		
Normalisations in $VH(H \to c\bar{c})$	Floating	-
Normalisation in $VH(H \to b\bar{b})$	0.08%	-
Single-top (t-channel) 0L & 1L resolved		
Normalisations s - t	4.6% - 17%	-
High / Low ΔR CR-SR extrapolation ratios	3% - 17%	-
p_T^V extrapolation ratios	7% - 15%	-
N_{jet} acceptance ratios	15%	-
1L to 0L acceptance ratio	6%	-
$t\bar{t}$ and single-top boosted		
$t\bar{t}$ normalisations	-	Floating
single-top s, t, Wt, normalisations	-	4.6% - 10% - 25%
$t\bar{t}$ 1L to 0L acceptance ratios	-	6% - 20%
$t\bar{t}$ Top CR-SR acceptance ratios	-	10%
Wt p_T^V extrapolation ratio	-	20%
Wt 1L to 0L acceptance ratios	-	20% - 40%
Diboson		
WW / ZZ / WZ normalisations	16% / 17% / 19%	16% / 17% / 27%
$ggVV$ normalisation	30%	
Lepton channel acceptance	2% - 23%	7%
N_{jet} acceptance	10% - 30%	-
p_T^V acceptance	3% - 16%	8% - 40%
SR / CR acceptance	6% - 16%	-
STXS binning acceptance	-	1.2% - 42.2%
Multi-jet (1L)		
Normalisations	20% - 100%	-

Table 6.15: Summary of the modelling systematic uncertainties in the analysis. The values given refer to the size of the uncertainty affecting the yield of each background. Uncertainties in the shapes of the distributions are not shown but taken into account for all backgrounds.

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

6.9.2 Signal Modelling

The three main signal productions $qq \to WH$, $q\bar{q} \to ZH$, and $gg \to ZH$ are modelled separately, with uncertainties addressing the production and the decay mode of the Higgs into $b\bar{b}$ or $c\bar{c}$. The goal of the analysis is to measure the fiducial cross-sections of the $VH(H \to b\bar{b})$ and the signal strength of the $VH(H \to c\bar{c})$. This first objective is approached with the adoption of the Simplified Template Cross-Section (STXS) in the reduced scheme of stage 1.2 [220, 221], depicted in Figure 6.23. The bins are defined in successive regions of transverse momentum of the vector boson p_T^V, from truth information in the simulated samples, and the number of additional jets N_{jet} in the event, at 0 or more than 1 additional jet.

Figure 6.23: The Standard Template Cross-Section scheme in the reduced stage 1.2 used for the combined $VH(H \to b\bar{b}/c\bar{c})$ analysis.

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

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

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

- *QCD scale uncertainties:* obtained by varying the renormalisation and factorisation scales μ_R and μ_F. These variations are the most impactful in the theoretical prediction of the VH production cross-sections. They are considered as shape uncertainties, implemented to cover modifications to the inclusive cross-sections and to parametrise possible migrations across p_T^V and additional jet multiplicity bins, following Ref [222]. The quark- and gluon-initiated signal processes have cross-section modifications parametrised separately.

- *PDF + α_s uncertainties*: alternative parton distributions from the PDF4LHC15_30 modifying the VH cross-sections in STXS bins are considered [223]. The VH cross-sections in each STXS bin are systematically modified by comparing the nominal PDF to 30 alternatives. Furthermore, the α_s estimated at the Z-mass is varied for the nominal setup within its uncertainties. These uncertainties are separately calculated for qq-initiaded WH and ZH, and for gg-initiated ZH. Shape effects on the resolved regime p_T^V distributions are considered, while variations to the boosted large-R mass m_J and the invariant mass m_{bb} or m_{cc} are negligible.

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

- *Branching ratio*: a theoretical uncertainty of 1.61% on the $H \to b\bar{b}$ branching ratio and an uncertainty covering the range from -1.99% to +5.53% for the $H \to c\bar{c}$ branching ratio are used [31]. The ZH (WH) cross-sections cover 96.52% to 104.11% (97.95% to 101.98%) of their values thanks to additional uncertainties.

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

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

6.9.3 V+jets Modelling

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

Z+jets

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

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

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

- *Channel extrapolation 2L → 0L uncertainties*: for the Z+*hf*, Z+*mf*, and Z+*lf* respectively.
- *Flavour composition uncertainties*: accounting for the variation on the yields of different flavours in the combinations with the double ratio of Equation 6.2. These include a ratio of *cc* to *bb* for Z+*hf*, and of *bc* and *bl* to *cl* for Z+*mf*. They are decorrelated in p_T^V and jet multiplicity N_{jet} bins and cover the $VH(H \to b\bar{b})$ and $VH(H \to c\bar{c})$ sides of the resolved regime.
- *Region extrapolation uncertainties*: are included to model the acceptance of different regions, and derived with the double ratio Equation 6.2 from a high purity region to a lower purity as:

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

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

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

The values of the acceptance uncertainties are presented in Table **??** of Appendix **??**, with a summary mentioned in Table 6.14. In addition, 4 different types of shape uncertainty are considered:

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

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

W+jets

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

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

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

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

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

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

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

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

6.9.4 Top Modelling

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

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

(a) (b)

Figure 6.24: The non-rebinned MVA distributions of the top background components (direct tagged) in the $VH(H \to c\bar{c})$ signal regions (TN-tagged on the left, TL-tagged on the right) with 150 GeV $< p_T^V <$ 250 GeV and 3 jets. Top(bb) in black, top(bc) in red, top(bl) in blue, and top(qq) in green. The bottom panels show the normalised distributions.

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

These groupings are based on the shared kinematics of the components, where the selected jets are either both b-jets and thus likely to directly come from the top-decays (bb), 1 b-jet likely from a top decay and 1 non b-jet from a subsequent hadronic W-decay or a radiated jet (bc and bl, summarised bq), or neither directly from the top-decay (cc, cl, and ll, summarised qq). The bc and bl are combined into a single top(bq) component as they indeed share the same kinematics, as illustrated in Figures 6.24 in the signal regions of $VH(H \to c\bar{c})$.This top(bq) background is particularly significant in the $VH(H \to c\bar{c})$ analysis as it peaks at the signal mass (having a mass $\sim (m_{\text{top}} + m_W)/2 \approx m_H$) and therefore exhibits signal-like properties such as reaching high MVA scores, as shown in Figure 6.24. Due to the small contribution of the top(qq) component, it is merged with the top(bq) into a single top(bq/qq) component, with the different sub-components shapes modelled by flavour composition uncertainties. The rest of this section details the modelling of the top backgrounds in the analysis regimes for 0L and 1L, followed by the single-top t- and s-channels in resolved, and finally the modelling adopted for the boosted regime.

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

There are three main elements in the top background modelling scheme in the 0L and 1L resolved regime: floating normalisation, acceptance uncertainties, and shape uncertainties. On the first point, free-floating normalisations are applied for the top(bb) and the top(bq/qq) components, constrained respectively by the BB-tagged High ΔR CR and the BT-tagged Top CR. These FNs are separated in jet multiplicity N_{jet} (2-jet, 3-jet, and only for the 0L channel 4-jet) as well as p_T^V, for a total of 16 FNs. Concerning the second point, several types of acceptance uncertainties are applied, as summarised in Table 6.14 and detailed in the Appendix Table **??**:

- *Channel extrapolation 1L → 0L uncertainties*: the Top is dominant in 1L, hence the FNs derivation is driven by the 1-lepton channel and applied to the 0-lepton one. This uncertainty is split in p_T^V: 2% in [150, 250] GeV and 8% in [250, 400] GeV.
- *Flavour composition uncertainties*: the top(bq/qq) includes differently shaped sub-components. Uncertainties are derived from the alternative samples with the double ratio Equation 6.2, comparing bl and qq to bc (of 5% and 10% respectively).
- *Region extrapolation uncertainties*: the top(bb) is dominant in the CRHigh while the top(bq/qq) leads in the Top BT CR, hence the extrapolations differ for the components. They are all derived from the double ratios with alternative samples.

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

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

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

- CARL shapes: modelling the difference between the nominal samples (Powheg+Pythia 8) and the alternative modelling of the parton shower (Powheg+Herwig 7) and matrix element (MadGraph5_aMC@NLO+Pythia 8). These CARL models are trained separately for $t\bar{t}$ and Wt and per lepton channel, inclusively in flavour compositions and N_{jet}. The DR scheme is used as nominal for these training of Wt because the alternative samples use the same $t\bar{t}$ overlap removal scheme.

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

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

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

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

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

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

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

Resolved Regime Top Backgrounds in 2L

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

Boosted Regime Top Backgrounds

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

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

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

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

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

6.9.5 Diboson Modelling

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

In the resolved regime, the diboson processes are a small background in the analysis, so only normalisations uncertainties are used for ZZ (17%), WW (16%), and WZ (19%) for the qq-initiated, and $ggVV$ (30%) for the gg-initiated. All uncertainties are correlated between $VH(H \to b\bar{b})$ and $VH(H \to c\bar{c})$. The $VZ(\to b\bar{b})$ and $VZ(\to c\bar{c})$ are considered as signals of the cross-check analysis, and denoted as $VZbb$ and $VZcc$ in the diboson modelling. The rest of the WW, WZ, and VZ are classified as background components, denoted as VVbkg. Acceptance uncertainties, summarised in Table 6.14 and detailed in the Appendix Table **??**. For the signal components, the uncertainties are split between the ZZ and WZ components and include:

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

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

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

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

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

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

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

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

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

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

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

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

- CARL shape uncertainties comparing the nominal Sherpa 2.2.11 samples to the two alternative samples Powheg+Pythia8 and Sherpa 2.2.1. The former accounts for differences to the matrix element and parton shower, while the latter accounts for the mis-modelled p_T^V shape. These uncertainties are applied to all regions.

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

- QCD scale shape uncertainties are included to model changes to the scales μ_R and μ_F, similarly to the V+jets.
- PDF shape uncertainties modelling variation to α_s are considered.
- EW shape uncertainties are considered, similarly to V+jets.

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

6.9.6 Multi-jet Modelling

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

6.10 Statistical Analysis

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

- $VH(H \to b\bar{b})$:
  - Inclusive signal strength $\mu_{VH_{bb}}$ and significance: 1 POI.
  - Signal strengths for $WH(\to b\bar{b})$ and $ZH(\to b\bar{b})$: 2 POIs.
  - Fiducial STXS measurements in the reduced stage 1.2, described in Section 6.9.2 and Figure 6.23: 15 POIs (8 for ZH, 7 for WH).

  - Constraints on Yukawa Higgs-bottom coupling modifier κ_b.

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

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

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

6.10.1 Likelihood Function Definition

All parameters of interest are estimated by comparing theory-based expectations baked into MC-simulated samples to real collected data in a fit. This fit is performed by maximising the binned-likelihood function in all the analysis regions simultaneously, as a function of the signal strengths and statistical and systematic uncertainties. The full binned-likelihood function is composed of three terms, accounting, respectively, for the number of events per bin $\mathcal{L}_{\text{Events}}$, the impact of systematics $\mathcal{L}_{\text{Systematics}}$, and finally the impact of the limited statistics of the simulated sample $\mathcal{L}_{\text{MC-stats}}$. These are combined into the likelihood function of Equation 6.3:

$$\mathcal{L} = \mathcal{L}_{\text{Events}} \times \mathcal{L}_{\text{Systematics}} \times \mathcal{L}_{\text{MC-stats}}. \tag{6.3}$$

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

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

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

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

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

The nominal value is by convention set at $\theta_0 = 0$, with $\theta = \pm 1$ representing a $\pm 1\ \sigma$ variation. The effect of each NP is determined in auxiliary measurements, following the prescriptions introduced in the modelling Sections 6.8 and 6.9. For example, if an NP tracking the normalisation of a background with a 10% prior is moved upwards by 1 standard deviation in the fit, the yield of the background is increased by 10%. After the fit, the central values of the NPs can be moved upwards or downwards, with a deviation from the initial central value called a *pull* and defined as

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

where the prefit value $\theta_0 = 0$ and $\sigma_{\theta_0} = 1$ for all NPs. The *constraint* indicates the change in certainty on the NP after the fit, estimated by the variance $\sigma_{\hat{\theta}}$ measured in the inverse Hesse matrix at the maximal likelihood point. For the normalisation of the major backgrounds, special unconstrained NPs are included with no likelihood penalty and said to be *free-floating* (FNs). They are free to vary and determined from data in control regions with an enhanced purity of the processes they normalised. These special NPs and have prefit values θ_0 set at 1.

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

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

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

Bringing every component together, the full likelihood function of Equation 6.3 is thus defined

as

$$\mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\theta}, \boldsymbol{\gamma}) = \prod_{i \in \text{bins}} \mathcal{P}(N_i \,|\, \boldsymbol{\mu} s_i(\boldsymbol{\theta}) + b_i(\boldsymbol{\theta}, \boldsymbol{\gamma})) \times \prod_{\theta \in \boldsymbol{\theta}} \mathcal{N}(\theta \,|\, 0, 1) \times \prod_{i \in \text{bins}} \mathcal{N}(\beta_i \,|\, \gamma_i \beta_i, \sqrt{\gamma_i \beta_i}). \quad (6.4)$$

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

$$\lambda(\boldsymbol{\mu}) = \frac{\mathcal{L}\left(\boldsymbol{\mu}, \hat{\hat{\boldsymbol{\theta}}}, \hat{\hat{\boldsymbol{\gamma}}}\right)}{\mathcal{L}\left(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\gamma}}\right)}. \quad (6.5)$$

The λ ratio is bound to the range [0, 1], with higher values implying a good agreement between the data and the chosen $\boldsymbol{\mu}$ while lower values are signs of disagreements. This pattern permits the construction of a likelihood ratio test statistics $t_{\boldsymbol{\mu}}$, defined as [224]

$$t_{\boldsymbol{\mu}} = \begin{cases} -2 \ln \lambda(\boldsymbol{\mu}) & \hat{\boldsymbol{\mu}} \geq \boldsymbol{\mu} \\ 0 & \hat{\boldsymbol{\mu}} < \boldsymbol{\mu} \end{cases}, \quad (6.6)$$

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

$$p_{\boldsymbol{\mu}} = \int_{t_{\boldsymbol{\mu},\text{obs}}}^{\infty} f(t_{\boldsymbol{\mu}} \,|\, \mathbf{0}) dt_{\boldsymbol{\mu}}, \quad (6.7)$$

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

$$Z = \Phi^{-1}(1 - p_{\boldsymbol{\mu}}), \quad (6.8)$$

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

To determine a 95% upper limit CL on a signal strength, a modified frequentist CL_s method is deployed [224, 225], based on the test statistics $\tilde{t}$ defined as:

$$\tilde{t} = -2\ln\frac{\mathcal{L}_{s+b}}{\mathcal{L}_b} = -2\ln\frac{\mathcal{L}\left(\mu=1,\hat{\hat{\boldsymbol{\theta}}}(\mu=1),\hat{\hat{\boldsymbol{\gamma}}}(\mu=1)\right)}{\mathcal{L}\left(\mu=0,\hat{\hat{\boldsymbol{\theta}}}(\mu=0),\hat{\hat{\boldsymbol{\gamma}}}(\mu=0)\right)}, \tag{6.9}$$

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

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

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

There are 15 POIs for the $VH(H \to b\bar{b})$ side and 1 POI for $VH(H \to c\bar{c})$. The binning used and regions included as well as the variables defining the underlying distributions entering the fits are detailed in Sections 6.5 and 6.7. A dense summary of the full categorisation is presented in Figure 6.17, underscoring the complexity of an analysis spanning 164 different regions, 84 of which are in $VH(H \to c\bar{c})$ (30 SRs, 6 Top $e\mu$ CRs, 10 $V+l$ CRs, 48 CRHigh), 48 in the resolved $VH(H \to b\bar{b})$ (21 SRs, 6 CRLow, 21 CRHigh), 12 BT-tagged Top CRs shared in the resolved regime, and 10 in the boosted regime (6 SRs, 4 boosted Top CRs). Experimental and modelling uncertainties are introduced to account for any mis-modelling and avoid biasing the fit, as described in Sections 6.8 and 6.9. The analysis described in this thesis is not yet concluded, with modifications to the modelling under active investigation at the time of writing. Consequently, the fit is still blinded, with the data in signal regions bins most sensitive to the signal hidden. For m_{bb} or m_{cc}

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

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

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

Concerning the $VH(H \to c\bar{c})$ signal strength measurement, the 95% CL_s expected upper limits are shown for the different lepton channels and combined in Figure 6.25a for the postfit Asimov and Figure 6.25b for the prefit Asimov.

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

Figure 6.25: The 95% CL_s upper limit on the $VH(H \to c\bar{c})$ signal strength from the combined analysis posfit (left) and prefit (right).

Significant improvements are expected for all lepton channels. The combination of all lepton channels leads to a remarkable improvement on the 95% CL_s upper limit on μ_{VHcc} from 31 × SM in the latest published ATLAS $VH(H \to c\bar{c})$ analysis [122] to 11.1 × SM (10.2 × SM) in the postfit (prefit) Asimov combined one, a factor 2.8 improvements in sensitivity. Gains are expected to be made in all lepton channels, which now have very similar sensitivity thanks to modifications to the analysis strategy. Compared to the published analysis, the 0-lepton channel upper limit is reduced from 40 × SM → 17.5 × SM, the lepton from 60 × SM → 18.8 × SM, and the 2-lepton from 51 × SM → 18.8 × SM [122]. These correspond to relative sensitivity improvement factors of 2.3, 3.2, and 2.7. Most of the gains are made in the 1- and 2-lepton channels, although the 0-lepton channel remains the most sensitive one.

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

On the $VH(H \to b\bar{b})$ side, combining the resolved and boosted regime, the postfit expected significance on the $VH(H \to b\bar{b})$ signal strength is 7.9 σ over the background-only prediction, corresponding to a 23% improvement over the published expected significance at 6.3 σ [194]. This is achieved thanks to a postfit expected significance of 4.7 σ in the 0-lepton channel (15% improvement to published result), 5.3 σ in 1-lepton (30% improvement), and 4.4 σ in 2-lepton (3% improvement). The most sensitive channel is distinctively the 1-lepton channel in this case.

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

Finally, adopting the fine splitting of the STXS stage 1.2 with 15 bins defined by p_T^V and additional jet multiplicity N_{jet}, with 8 bins in ZH and 7 bins in WH, the $VH(H \to b\bar{b})$ analysis reaches the per bin sensitivities listed in Table 6.16, with evidence-level only attained for the 150 GeV $< p_T^V <$ 250 GeV with no additional jet ZH bin. The impact of systematics and statistical uncertainties on the signal strengths of the different bins is shown in Figure 6.26.

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

Table 6.16: The expected prefit significance in the different STXS bins of the combined analysis.

The Diboson Cross-Check

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

Figure 6.26: The constraints on the prefit STXS signal strength.

$VZ(\to b\bar{b})$ part, the postfit expected significance reaches a large value of 15.1 σ when combining lepton channels. The 0-lepton channel reaches 11.2 σ, 1-lepton 6.2 σ, and 2-lepton 9 σ. On the $VZ(\to c\bar{c})$ side, the combined analysis expects to reach observation level for the first time, with a combined postfit expected significance of 5.1 σ. This represents a significant improvement of a factor 2.3 from the 2.2 σ published expected result [122]. The combined analysis reaches a postfit expected significance of 3.9 σ in 0-lepton, 2.6 σ in 1-lepton, and 3.1 σ in 2-lepton.

Additional Fit Results

In addition to the main results highlighted above, some further insights into the output of the fits are given before concluding this chapter. To verify the MC-samples correctly reproduce the data after the fit, some posfit plots are presented in Figure 6.27 for selected signal regions and control regions, with all of them listed in Appendix **??**. Interestingly, good agreement between the data and posfit MC samples is also observed for validation distributions not directly constrained in the fit. Figure 6.28 displays posfit distributions for a BL-tagged region analoguous to the used BT-tagged Top CR, an LL-tagged region similar to a c-tagged signal region, and the p_T^V spectrum of the whole 2L BB-tagged 2-jet signal region. The good agreement observed between data and MC-samples in all regions is a sign of the correct constraining by the fit.

The breakdown of the uncertainties, presented in Table 6.17, is a measure of the different contributions of the uncertainties to the $VH(H \to b\bar{b}/c\bar{c})$ analysis. The NPs are grouped based on their origin, and their impact on the signal strengths of $VH(H \to b\bar{b})$ and $VH(H \to c\bar{c})$ is assessed by iteratively re-running fits with successive groups of NPs fixed at their postfit values. The notation adopted is to label the signal strengths (one for $VH(H \to b\bar{b})$ and one for

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

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

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

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

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

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

Figure 6.27: Selected posfit signal regions (top row) and control regions (bottom row), for the 0L (left), 1L (centre), and 2L (right).

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

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

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

Figure 6.28: Posfit distributions in a BL-tagged Top CR-like (left) and LL-tagged SR-like (centre) validations regions and the 2L p_T^V spectrum in the 2-jet BB-tagged SR.

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

$$\text{Impact} = \sqrt{\sigma_{\hat{\mu}}^2 - \sigma_{\hat{\mu}'}^2}. \tag{6.10}$$

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

Table 6.17: Breakdown of the different systematics and statistical uncertainties.

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

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

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

A second technique to assess the importance of different nuisance parameters on the signal strengths is to change their NP values upwards and downwards by their postfit uncertainties σ_θ and re-run the fit with the modified NP fixed. For each NP, this requires running two fits in addition to the nominal fit from which $\hat{\theta}$ and $\sigma_{\hat{\theta}}$ are measured: one with the NP fixed at $\hat{\theta} + \sigma_{\hat{\theta}}$ and one with $\hat{\theta} - \sigma_{\hat{\theta}}$. NPs are ranked by the difference in the signal strengths between these new fits and the nominal one, as shown in Figure 6.29. In these plots, the central values of NPs are set at 0 (1 for FNs and γ-factor) as the dataset is the postfit Asimov set. For $VH(H \to b\bar{b})$, the W+hf extrapolations have a significant impact on the predicted signal strength, with several of these systematics highly ranked. Shape uncertainties associated with the diboson process and Higgs modelling uncertainties as well as the Wt DS-DR shape uncertainty and a b-jet tagging uncertainty also contribute meaningfully. The floating normalisation of W+hf in the boosted region is the only FN to make the ranking, due to its significant pull as is shown below in Figure 6.30. For $VH(H \to c\bar{c})$, the Top process CARL shapes are the leading nuisance parameters,

with the $Z + cc$ shape and the W+jets cc/bb acceptance ratio. The Z+lf and, to a lesser extent, W+lf floating normalisations have a large impact on the predicted signal strength, despite the constraints offered by the $V + l$ CR. The light- and c-jets uncertainties from flavour tagging are the biggest contributors in this category. Finally, the multi-jet process normalisation enters the ranking as this process contributes more in $VH(H \to c\bar{c})$. The γ-factor listed corresponds to the last unblinded bin in the 1L high p_T^V 2-jet SR, shown in Figure 6.27b, where a large amount of signal is expected, and the effect of this NP should be reduced once the signal is no longer constrained to its SM expectations in a conditional data fit.

Figure 6.29: The 15 most highly ranked Asimov postfit nuisance parameters for the $VH(H \to b\bar{b})$ (left) and $VH(H \to c\bar{c})$ (right) signal strengths. Modelling NPs are written in black, experimental NPs in blue, and floating normalisation (and γ-factor) in green, with values indicated by the bottom axis showing $\Delta\theta = \hat{\theta} - \theta_0$. Black points are nuisance parameters with their central value at 0 showing the pull (γ-factor with central value at 1), and red points are floating normalisation with central values at 1. The error bars on the point show the 1 σ uncertainty of the NPs. The effect of changing the NP by +1 σ (-1 σ) induces the change in signal strength $\Delta\mu$ shown by the hashes (empty) blue rectangle, with respect to the top axis.

In the combined analysis, the major backgrounds have free-floating normalisations decorrelated across the different p_T^V and jet multiplicity bins. The values set by a conditional likelihood fit to data, where the $VH(H \to b\bar{b}/c\bar{c})$ signal strengths are set to the SM expectations, are presented in Figure 6.30. They are compared to the same FNs obtained in the cross-check analysis with $VZ(\to b\bar{b}/c\bar{c})$ as signals. Good agreement is observed between the two sets of floating

Figure 6.30: The floating normalisations of the major background in the combined analysis targeting the $VH(H \to b\bar{b}/c\bar{c})$ in blue, versus the cross-check analysis $VZ(\to b\bar{b}/c\bar{c})$ in green.

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

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

The correlations between the different floating normalisations are displayed as a heat map in Figure 6.31. A rich structure of dependencies emerges from such a plot. As expected, FNs related to each process are highly correlated with the other FN of the same process, in different analysis p_T^V and N_{jet} categories. Some striking exceptions are visible: the boosted $t\bar{t}$ displays a very small uncorrelation with the resolved top(bb) and top(bq/qq). Concerning correlations across processes, the Top(bb) and Top(bq/qq) are respectively seen to have large correlations with the Z+*hf* (and the W+*hf* in lesser extend) and the W+*mf* and Z+*mf*, as expected from the presence of the Z+jets and W+jets in the CRHigh and the 0L and 1L BT-tagged Top CR. The W+*hf* normalisations are slightly anti-correlated to the V+*lf*and slightly correlated to the Z+*hf* in the low N_{jet}Ṡimilarly, the V+*lf* are strongly correlated between the W and Z.

6.11 Conclusion

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

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

Figure 6.31: The correlations between the floating normalisations of the major background in the combined $VH(H \to b\bar{b}/c\bar{c})$ analysis.

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

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

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

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

CHAPTER 7

Conclusion and Outlook

This thesis follows a logical succession, with first the theory underpinning the modern edifice of particle physics presented in Chapter 2. The SM has been extensively validated by many experiments across the world, and in particular the ATLAS Collaboration using data collided by the LHC as presented in Chapter 3. Many properties of the discovered particle in 2012 have been confirmed to correspond to those of the predicted SM Higgs boson. Nevertheless, the ATLAS Collaboration continues to systematically study the new particle in evermore challenging measurements, searching for any possible discrepancy between observations and theoretical predictions. These require state-of-the-art detectors and reconstruction software. At its core, a particle physics analysis is a data challenge that is well adressed by modern machine learning and artificial intelligence, as reviewed in Chapter 4. The recent progress in this field provides an exiciting avenue of development for ATLAS, helping the Collaboration propels the performance of its software to new heights by designing effective network-based models. One such promising area of development concerns jet flavour tagging, which has continuously benefitted from the adoption and development of advanced ML in recent years, as outlined in Chapter 5. GN2, the newest generation of tagger, relies on a single mutlimodal network exploiting a Transformer Encoder at its core, with multiple tasks targeted to distill expert knowlegde in the network. The excellent performance obtained promises more refined measurements from the numerous analyses targetting heavy-flavour quarks in their final state during the ongoing Run 3 of the LHC.

Two analyses benefitting starkly from flavour tagging are the $VH(H \to b\bar{b})$ and $VH(H \to c\bar{c})$. These are now combined into the $VH(H \to b\bar{b}/c\bar{c})$ analysis, described in Chapter 6. At the time of writing, the analysis is in its last phase with final studies on the modelling and the fit framework, hence the results presented here are still blinded. Excitingly, great strives were made in the effort to observe the $H \to c\bar{c}$ decay and measure the c-quark Yukawa coupling. The expected upper limit on the signal strength has been reduced by a factor 2.8 to 11.1 $\times$ SM expectations, compared to the last published ATLAS result [122]. Similarly, great progress was made in the measurement of the $H \to b\bar{b}$, with an expected signal strength sensitivity of 7.9 σ corresponding to a 23% improvement over the last ATLAS published result [194]. For the first time, both production modes are expected to be observed in a $b\bar{b}$ decay mode, with respective significance of 5.5 σ for WH and 6.2 σ for ZH. An Simplified Template Cross-Section (STXS) measurement of the different cross-sections of $VH(H \to b\bar{b})$ is also performed in stage 1.2.

To continue exploring the limit of our understanding in the particle physics realm, algorithmic and machine developments are required throughout the experiment. Collecting large datasets is crucial to analyses searchingfor rare signatures and performing precision measurements. This requires the detector to be operated in more challenging conditions: the price for higher instantaneous luminosity is more pile-up. The subdetectors must be upgraded to deal with this increased activity. Simultaneously to improving the hardware, the software of the ATLAS Collaboration must be upgraded to further push the sensitivity of the detector and deal with the new conditions. In this respect, flavour tagging benefited greatly from adopting advanced new NN architecture such as the Transformer, but also from the multimodal input and multitask paradigms to pass expert knowledge. Future areas of progress primarily consist in continuing this path, adding further low-level input information and defining additional tasks helping the main classification objective. Across science and industry, modern machine learning plays a crucial role in this modernisation effort. This is particularly the case in HEP, where the large databases from collisions and simulations are effectively exploited to create effective and precise models. Such networks are trained for all the uses of the field: from generative AI to effectively produced simulated samples, to deployed on FPGAs triggers, DL to reconstruct physics objects from the rich set of low-level collected data, and finally ML in analyses to improve signal discrimination from background and help constrain the modelling of the different processes.

Bibliography

[1] *Umami Framework*. https://gitlab.cern.ch/atlas-flavor-tagging-tools/algorithms/umami. Accessed: 2023-04-21.

[2] *Salt Framework*. https://gitlab.cern.ch/atlas-flavor-tagging-tools/algorithms/salt. Accessed: 2024-02-20.

[3] *Graph Neural Network Jet Flavour Tagging with the ATLAS Detector*. Tech. rep. Geneva: CERN, 2022. URL: %5Curl%7Bhttps://cds.cern.ch/record/2811135%7D.

[4] ATLAS Collaboration. *Jet Flavour Tagging With GN1 and DL1d. Generator dependence, Run 2 and Run 3 data agreement studies*. Tech. rep. Geneva: CERN, 2023. URL: %5Curl%7Bhttps://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PLOTS/FTAG-2023-01/%7D.

[5] ATLAS Collaboration. *Transformer Neural Networks for Identifying Boosted Higgs Bosons decaying into $b\bar{b}$ and $c\bar{c}$ in ATLAS*. Tech. rep. Geneva: CERN, 2023. URL: %5Curl%7Bhttps://cds.cern.ch/record/2866601%7D.

[6] Maxence Draguet. "Training and optimisation of large transformer models at CERN: an ATLAS case study on Kubeflow". In: *Sixth Inter-Experiment Machine Learning Workshop*. Geneva, 2024. URL: %7Bhttps://indico.cern.ch/event/1297159/contributions/5729198/%7D.

[7] Maxence Draguet. *Training and optimisation of large transformer models at CERN: an ATLAS case study on Kubeflow*. https://indico.cern.ch/event/1297159/contributions/5729198/. Accessed: 2024-04-26.

[8] Maxence Draguet and Ricardo Rocha. *Training and Optimisation of Large Transformer Models: An ATLAS and CERN Use Case*. https://colocatedeventseu2024.sched.com/event/1YFdZ/training-and-optimisation-of-large-transformer-models-an-atlas-and-cern-use-case-ricardo-rocha-cern-maxence-draguet-university-of-oxford-atlas. Accessed: 2024-04-26.

[9] Matthew D. Schwartz. *Quantum Field Theory and the Standard Model*. Cambridge University Press, 2013.

[10] M.K. Gaillard, P.D. Grannis, and F.J. Sciulli. "The standard model of particle physics". In: *Rev. Mod. Phys.* 71 (1999). DOI: 10.1103/RevModPhys.71.S96.

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

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

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

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

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

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

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

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

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

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

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

[22] D. Hanneke, S. Fogwell Hoogerheide, and G. Gabrielse. "Cavity control of a single-electron quantum cyclotron: Measuring the electron magnetic moment". In: *Phys. Rev. A* 83 (5 2011), p. 052122. DOI: 10.1103/PhysRevA.83.052122. URL: https://link.aps.org/doi/10.1103/PhysRevA.83.052122.

[23] *Wikipedia: The Standard Model (of Physics)*. https://simple.wikipedia.org/wiki/Standard_Model. Accessed: 2024-03-15.

[24] Sheldon L. Glashow. "Partial-symmetries of weak interactions". In: *Nuclear Physics* 22.4 (1961), pp. 579–588. ISSN: 0029-5582. DOI: https://doi.org/10.1016/0029-5582(61)90469-2. URL: https://www.sciencedirect.com/science/article/pii/0029558261904692.

[25] Steven Weinberg. "A Model of Leptons". In: *Phys. Rev. Lett.* 19 (21 Nov. 1967), pp. 1264–1266. DOI: 10.1103/PhysRevLett.19.1264. URL: https://link.aps.org/doi/10.1103/PhysRevLett.19.1264.

[26] Abdus Salam. "Weak and Electromagnetic Interactions". In: *Conf. Proc. C* 680519 (1968), pp. 367–377. DOI: 10.1142/9789812795915_0034.

[27] Steven Weinberg. "General Theory of Broken Local Symmetries". In: *Phys. Rev. D* 7 (4 Feb. 1973), pp. 1068–1082. DOI: 10.1103/PhysRevD.7.1068. URL: https://link.aps.org/doi/10.1103/PhysRevD.7.1068.

[28] C. N. Yang and R. L. Mills. "Conservation of Isotopic Spin and Isotopic Gauge Invariance". In: *Phys. Rev.* 96 (1 Oct. 1954), pp. 191–195. DOI: 10.1103/PhysRev.96.191. URL: https://link.aps.org/doi/10.1103/PhysRev.96.191.

[29] John C. Collins, Davison E. Soper, and George Sterman. *Factorization of Hard Processes in QCD*. 2004. arXiv: hep-ph/0409313 [hep-ph].

[30] Particle Data Group et al. "Review of Particle Physics". In: *Progress of Theoretical and Experimental Physics* 2020.8 (Aug. 2020), p. 083C01. ISSN: 2050-3911. DOI: 10.1093/ptep/ptaa104. eprint: https://academic.oup.com/ptep/article-pdf/2020/8/083C01/34673722/ptaa104.pdf. URL: https://doi.org/10.1093/ptep/ptaa104.

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

[32] ATLAS Collaboration. "A detailed map of Higgs boson interactions by the ATLAS experiment ten years after the discovery". In: *Nature* 607.7917 (2022), pp. 52–59. DOI: 10.1038/s41586-022-04893-w. arXiv: 2207.00092 [hep-ex].

[33] Lyndon Evans and Philip Bryant. "LHC Machine". In: *Journal of Instrumentation* 3.08 (2008), S08001. DOI: 10.1088/1748-0221/3/08/S08001. URL: https://dx.doi.org/10.1088/1748-0221/3/08/S08001.

[34] ATLAS Collaboration. "The ATLAS Experiment at the CERN Large Hadron Collider". In: *Journal of Instrumentation* 3.08 (2008), S08003. DOI: 10.1088/1748-0221/3/08/S08003. URL: https://dx.doi.org/10.1088/1748-0221/3/08/S08003.

[35] CMS Collaboration. "The CMS experiment at the CERN LHC". In: *Journal of Instrumentation* 3.08 (2008), S08004. DOI: 10.1088/1748-0221/3/08/S08004. URL: https://dx.doi.org/10.1088/1748-0221/3/08/S08004.

[36] ALICE Collaboration. "The ALICE experiment at the CERN LHC". In: *Journal of Instrumentation* 3.08 (2008), S08002. DOI: 10.1088/1748-0221/3/08/S08002. URL: https://dx.doi.org/10.1088/1748-0221/3/08/S08002.

[37] LHCb Collaboration. "The LHCb Detector at the LHC". In: *Journal of Instrumentation* 3.08 (2008), S08005. DOI: 10.1088/1748-0221/3/08/S08005. URL: https://dx.doi.org/10.1088/1748-0221/3/08/S08005.

[38] *Accelerator Complex of CERN*. https://home.web.cern.ch/science/accelerators. Accessed: 2024-03-21.

[39] ATLAS Collaboration. "Improved luminosity determination in pp collisions at $\sqrt{s} = 7$ TeV using the ATLAS detector at the LHC". In: *The European Physical Journal C* 73.8 (2013), p. 2518. DOI: 10.1140/epjc/s10052-013-2518-3. URL: https://doi.org/10.1140/epjc/s10052-013-2518-3.

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

[41] ATLAS Collaboration. *Preliminary analysis of the luminosity calibration of the ATLAS 13.6 TeV data recorded in 2022*. Tech. rep. Geneva: CERN, 2023. URL: https://cds.cern.ch/record/2853525.

[42] G. Avoni et al. "The new LUCID-2 detector for luminosity measurement and monitoring in ATLAS". In: *Journal of Instrumentation* 13.07 (2018), P07017. DOI: 10.1088/1748-0221/13/07/P07017. URL: https://dx.doi.org/10.1088/1748-0221/13/07/P07017.

[43] *Public ATLAS Luminosity Results for Run-2 of the LHC*. https://twiki.cern.ch/twiki/bin/view/AtlasPublic/LuminosityPublicResultsRun2. Accessed: 2024-03-22.

[44] *ATLAS Schematics*. https://atlas.cern/Resources/Schematics. Accessed: 2024-03-22.

[45] Giles Chatham Strong. "On the impact of selected modern deep-learning techniques to the performance and celerity of classification models in an experimental high-energy physics use case". In: *Mach. Learn. Sci. Tech.* 1 (2020), p. 045006. DOI: 10.1088/2632-2153/ab983a. arXiv: 2002.01427 [physics.data-an].

[46] ATLAS Collaboration. *ATLAS inner detector: Technical Design Report, 1*. Technical design report. ATLAS. Geneva: CERN, 1997. URL: https://cds.cern.ch/record/331063.

[47] Karolos Potamianos. "The upgraded Pixel detector and the commissioning of the Inner Detector tracking of the ATLAS experiment for Run-2 at the Large Hadron Collider". In: *PoS* EPS-HEP2015 (2015), p. 261. arXiv: 1608.07850 [physics.ins-det].

[48] M Capeans et al. *ATLAS Insertable B-Layer Technical Design Report*. Tech. rep. 2010. URL: https://cds.cern.ch/record/1291633.

[49] H. Pernegger. "The Pixel Detector of the ATLAS experiment for LHC Run-2". In: *Journal of Instrumentation* 10.06 (2015), p. C06012. DOI: 10.1088/1748-0221/10/06/C06012. URL: https://dx.doi.org/10.1088/1748-0221/10/06/C06012.

[50] ATLAS Collaboration. *IBL Efficiency and Single Point Resolution in Collision Events*. Tech. rep. Geneva: CERN, 2016. URL: https://cds.cern.ch/record/2203893.

[51] ATLAS Collaboration. "The silicon microstrip sensors of the ATLAS semiconductor tracker". In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 578.1 (2007), pp. 98–118. ISSN: 0168-9002. DOI: https://doi.org/10.1016/j.nima.2007.04.157. URL: https://www.sciencedirect.com/science/article/pii/S0168900207007644.

[52] ATLAS Collaboration. *Technical Design Report for the ATLAS Inner Tracker Strip Detector*. Tech. rep. Geneva: CERN, 2017. URL: https://cds.cern.ch/record/2257755.

[53] ATLAS Collaboration. "Operation and performance of the ATLAS semiconductor tracker". In: *Journal of Instrumentation* 9.08 (2014), P08009. DOI: 10.1088/1748-0221/9/08/P08009. URL: https://dx.doi.org/10.1088/1748-0221/9/08/P08009.

[54] ATLAS TRT collaboration. "The ATLAS Transition Radiation Tracker (TRT) proportional drift tube: design and performance". In: *Journal of Instrumentation* 3.02 (2008), P02013. DOI: 10.1088/1748-0221/3/02/P02013. URL: https://dx.doi.org/10.1088/1748-0221/3/02/P02013.

[55] A Vogel. *ATLAS Transition Radiation Tracker (TRT): Straw Tube Gaseous Detectors at High Rates*. Tech. rep. Geneva: CERN, 2013. URL: https://cds.cern.ch/record/1537991.

[56] Francesca Cavallari. "Performance of calorimeters at the LHC". In: *Journal of Physics: Conference Series* 293.1 (Apr. 2011), p. 012001. DOI: 10.1088/1742-6596/293/1/012001. URL: https://dx.doi.org/10.1088/1742-6596/293/1/012001.

[57] ATLAS Collaboration. "Muon reconstruction performance of the ATLAS detector in proton-proton collision data at $\sqrt{s}$=13 TeV". In: *The European Physical Journal C* 76.5 (2016), p. 292. DOI: 10.1140/epjc/s10052-016-4120-y. URL: https://doi.org/10.1140/epjc/s10052-016-4120-y.

[58] M. zur Nedden. "The LHC Run 2 ATLAS trigger system: design, performance and plans". In: *Journal of Instrumentation* 12.03 (2017), p. C03024. DOI: 10.1088/1748-0221/12/03/C03024. URL: https://dx.doi.org/10.1088/1748-0221/12/03/C03024.

[59] ATLAS Collaboration. *The ATLAS Collaboration Software and Firmware*. Tech. rep. Geneva: CERN, 2021. URL: %5Curl%7Bhttps://cds.cern.ch/record/2767187%7D.

[60] ATLAS Collaboration. *ATLAS Computing Acknowledgements*. Tech. rep. Geneva: CERN, 2020. URL: https://cds.cern.ch/record/2717821.

[61] Joao Pequenao and Paul Schaffner. "How ATLAS detects particles: diagram of particle paths in the detector". 2013. URL: https://cds.cern.ch/record/1505342.

[62] ATLAS Collaboration. "Operation of the ATLAS trigger system in Run 2". In: *Journal of Instrumentation* 15.10 (Oct. 2020), P10004. DOI: 10.1088/1748-0221/15/10/P10004. URL: https://dx.doi.org/10.1088/1748-0221/15/10/P10004.

[63] ATLAS Collaboration. *The Optimization of ATLAS Track Reconstruction in Dense Environments*. Tech. rep. Geneva: CERN, 2015. URL: https://cds.cern.ch/record/2002609.

[64] ATLAS Collaboration. "Performance of the ATLAS track reconstruction algorithms in dense environments in LHC Run 2". In: *The European Physical Journal C* 77.10 (2017), p. 673. DOI: 10.1140/epjc/s10052-017-5225-7. URL: https://doi.org/10.1140/epjc/s10052-017-5225-7.

[65] R. E. Kalman. "A New Approach to Linear Filtering and Prediction Problems". In: *Journal of Basic Engineering* 82.1 (Mar. 1960), pp. 35–45. ISSN: 0021-9223. DOI: 10.1115/1.3662552. eprint: https://asmedigitalcollection.asme.org/fluidsengineering/article-pdf/82/1/35/5518977/35_1.pdf. URL: https://doi.org/10.1115/1.3662552.

[66] ATLAS Collaboration. "Reconstruction of primary vertices at the ATLAS experiment in Run 1 proton–proton collisions at the LHC". In: *Eur. Phys. J. C* 77.5 (2017), p. 332. DOI: 10.1140/epjc/s10052-017-4887-5. arXiv: 1611.10235 [physics.ins-det].

[67] W Lampl et al. *Calorimeter Clustering Algorithms: Description and Performance*. Tech. rep. Geneva: CERN, 2008. URL: https://cds.cern.ch/record/1099735.

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

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

[70] ATLAS Collaboration. "Jet reconstruction and performance using particle flow with the ATLAS Detector". In: *The European Physical Journal C* 77.7 (2017), p. 466. DOI: 10.1140/epjc/s10052-017-5031-2. URL: https://doi.org/10.1140/epjc/s10052-017-5031-2.

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

[72] ATLAS Collaboration. "Topological cell clustering in the ATLAS calorimeters and its performance in LHC Run 1". In: *The European Physical Journal C* 77.7 (2017), p. 490. DOI: 10.1140/epjc/s10052-017-5004-5. URL: https://doi.org/10.1140/epjc/s10052-017-5004-5.

[73] ATLAS Collaboration. "Performance of jet substructure techniques for large-R jets in proton-proton collisions at $\sqrt{s} = 7$ TeV using the ATLAS detector". In: *Journal of High Energy Physics* 2013.9 (2013), p. 76. DOI: 10.1007/JHEP09(2013)076. URL: https://doi.org/10.1007/JHEP09(2013)076.

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

[75] ATLAS Collaboration. *Tagging and suppression of pileup jets with the ATLAS detector*. Tech. rep. Geneva: CERN, 2014. URL: %5Curl%7Bhttps://cds.cern.ch/record/1700870%7D.

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

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

[78] Kevin P. Murphy. *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012. ISBN: 0262018020.

[79] Dan Cireşan, Ueli Meier, and Juergen Schmidhuber. "Multi-column Deep Neural Networks for Image Classification". In: *Proceedings / CVPR, IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Feb. 2012). DOI: 10.1109/CVPR.2012.6248110.

[80] Yann LeCun et al. "Handwritten Digit Recognition with a Back-Propagation Network". In: *Advances in Neural Information Processing Systems*. Ed. by D. Touretzky. Vol. 2. Morgan-Kaufmann, 1989. URL: %5Curl%7Bhttps://proceedings.neurips.cc/paper_files/paper/1989/file/53c3bce66e43be4f209556518c2fcb54-Paper.pdf%7D.

[81] Leo Breiman. "Bagging predictors". In: *Machine Learning* 24.2 (1996), pp. 123–140. DOI: 10.1007/BF00058655. URL: %5Curl%7Bhttps://doi.org/10.1007/BF00058655%7D.

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

[83] Jerome H. Friedman. "Greedy function approximation: A gradient boosting machine." In: *The Annals of Statistics* 29.5 (2001), pp. 1189–1232. DOI: 10.1214/aos/1013203451. URL: https://doi.org/10.1214/aos/1013203451.

[84] Lloyd S Shapley. "A Value for n-Person Games". In: *Contributions to the Theory of Games II*. Ed. by Harold W. Kuhn and Albert W. Tucker. Princeton: Princeton University Press, 1953, pp. 307–317.

[85] Benedek Rozemberczki et al. "The Shapley Value in Machine Learning". In: *ArXiv* abs/2202.05594 (2022). URL: https://api.semanticscholar.org/CorpusID:246822765.

[86] F. Rosenblatt. "The perceptron: A probabilistic model for information storage and organization in the brain." In: *Psychological Review* 65.6 (1958), pp. 386–408. ISSN: 0033-295X. DOI: 10.1037/h0042519. URL: %5Curl%7Bhttp://dx.doi.org/10.1037/h0042519%7D.

[87] G. Cybenko. "Approximation by superpositions of a sigmoidal function". In: *Mathematics of Control, Signals and Systems* 2.4 (1989), pp. 303–314. DOI: 10.1007/BF02551274. URL: %5Curl%7Bhttps://doi.org/10.1007/BF02551274%7D.

[88] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. "Multilayer feedforward networks are universal approximators". In: *Neural Networks* 2.5 (1989), pp. 359–366. ISSN: 0893-6080. DOI: \url{https://doi.org/10.1016/0893-6080(89)90020-8}. URL: https://www.sciencedirect.com/science/article/pii/0893608089900208.

[89] Zhou Lu et al. "The Expressive Power of Neural Networks: A View from the Width". In: NIPS'17. Long Beach, California, USA: Curran Associates Inc., 2017, pp. 6232–6240. ISBN: 9781510860964.

[90] Abien Fred Agarap. *Deep Learning using Rectified Linear Units (ReLU)*. 2019. arXiv: 1803.08375 [cs.NE].

[91] Adam Paszke et al. "PyTorch: An Imperative Style, High-Performance Deep Learning Library". In: *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019, pp. 8024–8035. URL: %5Curl%7Bhttp://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf%7D.

[92] Martín Abadi et al. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. Software available from tensorflow.org. 2015. URL: %5Curl%7Bhttps://www.tensorflow.org/%7D.

[93] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. "Learning representations by back-propagating errors". In: *Nature* 323.6088 (1986), pp. 533–536. DOI: 10.1038/323533a0. URL: %5Curl%7Bhttps://doi.org/10.1038/323533a0%7D.

[94] Sepp Hochreiter and Jürgen Schmidhuber. "Long Short-term Memory". In: *Neural computation* 9 (Dec. 1997), pp. 1735–80. DOI: 10.1162/neco.1997.9.8.1735.

[95] Kyunghyun Cho et al. "Learning phrase representations using RNN encoder-decoder for statistical machine translation". In: *Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*. 2014.

[96] Stephen Chung and Hava Siegelmann. "Turing Completeness of Bounded-Precision Recurrent Neural Networks". In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato et al. Vol. 34. Curran Associates, Inc., 2021, pp. 28431–28441. URL: https://proceedings.neurips.cc/paper_files/paper/2021/file/ef452c63f81d0105dd4486f775adec81-Paper.pdf.

[97] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "ImageNet Classification with Deep Convolutional Neural Networks". In: *Advances in Neural Information Processing Systems*. Ed. by F. Pereira et al. Vol. 25. Curran Associates, Inc., 2012. URL: https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

[98] Kaiming He et al. "Deep Residual Learning for Image Recognition". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 770–778. DOI: 10.1109/CVPR.2016.90.

[99] Franco Scarselli et al. "The Graph Neural Network Model". In: *IEEE Transactions on Neural Networks* 20.1 (2009), pp. 61–80. DOI: 10.1109/TNN.2008.2005605.

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

[101] Thomas N. Kipf and Max Welling. "Semi-Supervised Classification with Graph Convolutional Networks". In: *Proceedings of the 5th International Conference on Learning Representations* (ICLR). ICLR '17. Palais des Congres Neptune, Toulon, France, 2017. URL: https://openreview.net/forum?id=SJU4ayYgl.

[102] Petar Veličković et al. "Graph Attention Networks". In: *International Conference on Learning Representations* (2018). URL: https://openreview.net/forum?id=rJXMpikCZ.

[103] Manzil Zaheer et al. "Deep Sets". In: *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper/2017/file/f22e4747da1aa27e363d86d40ff442fe-Paper.pdf.

[104] Charles R. Qi et al. "PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017.

[105] Xiaolong Wang et al. "Non-local Neural Networks". In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2018, pp. 7794–7803. DOI: 10.1109/CVPR.2018.00813.

[106] Ashish Vaswani et al. "Attention is All you Need". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc., 2017.

[107] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://aclanthology.org/N19-1423.

[108] Alec Radford et al. "Improving language understanding by generative pre-training". In: (2018).

[109] Alexey Dosovitskiy et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale". In: *International Conference on Learning Representations*. 2021. URL: https://openreview.net/forum?id=YicbFdNTTy.

[110] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. *Layer Normalization*. cite arxiv:1607.06450. 2016. URL: http://arxiv.org/abs/1607.06450.

[111] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. "Neural Machine Translation by Jointly Learning to Align and Translate". In: *CoRR* abs/1409.0473 (2014). URL: https://api.semanticscholar.org/CorpusID:11212020.

[112] Diederik P. Kingma and Jimmy Ba. "Adam: A Method for Stochastic Optimization". In: *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. 2015. URL: http://arxiv.org/abs/1412.6980.

[113] Samuel L. Smith, Pieter-Jan Kindermans, and Quoc V. Le. "Don't Decay the Learning Rate, Increase the Batch Size". In: cite arxiv:1711.00489Comment: 11 pages, 7 figures. 2017. URL: https://openreview.net/pdf?id=B1Yy1BxCZ.

[114] Albert Q. Jiang et al. *Mistral 7B*. 2023. arXiv: 2310.06825 [cs.CL].

[115] Bin Xiao et al. *Florence-2: Advancing a Unified Representation for a Variety of Vision Tasks*. 2023. arXiv: 2311.06242 [cs.CV].

[116] B.R. Webber. "Fragmentation and Hadronization". In: *Int. J. Mod. Phys. AS* 15S1 (2000), pp. 577–606. DOI: 10.1142/S0217751X00005334. URL: %5Curl%7Bhttps://cds.cern.ch/record/419784%7D.

[117] ATLAS Collaboration. *Comparison of Monte Carlo generator predictions for bottom and charm hadrons in the decays of top quarks and the fragmentation of high pT jets*. Tech. rep. Geneva: CERN, 2014. URL: %5Curl%7Bhttps://cds.cern.ch/record/1709132%7D.

[118] ATLAS Collaboration. "ATLAS b-jet identification performance and efficiency measurement with $t\bar{t}$ events in pp collisions at $\sqrt{s} = 13$ TeV". In: *Eur. Phys. J. C* 79.11 (2019), p. 970. DOI: 10.1140/epjc/s10052-019-7450-8. arXiv: 1907.05120.

[119] *Representation of a b-jet*. https://tikz.net/jet_btag/. Accessed: 2024-03-20.

[120] Samuel Van Stroud. "Graph Neural Network Flavour Tagging and Boosted Higgs Measurements at the LHC". Presented 2023. UCL, 2023. URL: %5Curl%7Bhttps://cds.cern.ch/record/2869719%7D.

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

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

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

[124] ATLAS Collaboration. *Expected performance of the ATLAS b-tagging algorithms in Run-2*. Tech. rep. Geneva: CERN, 2015. URL: %5Curl%7Bhttps://cds.cern.ch/record/2037697%7D.

[125] ATLAS Collaboration. "Optimisation and performance studies of the ATLAS *b*-tagging algorithms for the 2017-18 LHC run". In: (July 2017).

[126] ATLAS Collaboration. *Identification of Jets Containing b-Hadrons with Recurrent Neural Networks at the ATLAS Experiment*. Tech. rep. ATL-PHYS-PUB-2017-003. CERN, 2017. URL: %5Curl%7Bhttps://cds.cern.ch/record/2255226%7D.

[127] ATLAS Collaboration. *Deep Sets based Neural Networks for Impact Parameter Flavour Tagging in ATLAS*. Tech. rep. ATL-PHYS-PUB-2020-014. CERN, 2020. URL: %5Curl%7Bhttps://cds.cern.ch/record/2718948%7D.

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

[129] Arnaud Duperrin. *Flavour tagging with graph neural networks with the ATLAS detector*. 2023. arXiv: 2306.04415 [hep-ex].

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

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

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

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

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

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

[136] ATLAS Collaboration. *ATLAS Pythia 8 tunes to 7 TeV data*. Tech. rep. Geneva: CERN, 2014. URL: %5Curl%7Bhttps://cds.cern.ch/record/1966419%7D.

[137] Richard D. Ball et al. "Parton distributions with LHC data". In: *Nuclear Physics B* 867.2 (2013), pp. 244–289. ISSN: 0550-3213. DOI: https://doi.org/10.1016/j.nuclphysb.2012.10.003.

[138] ATLAS Collaboration. *Studies on top-quark Monte Carlo modelling for Top2016*. Tech. rep. Geneva: CERN, 2016. URL: %5Curl%7Bhttps://cds.cern.ch/record/2216168%7D.

[139] ATLAS Collaboration. *Study of top-quark pair modelling and uncertainties using ATLAS measurements at $\sqrt{s}$=13 TeV*. Tech. rep. Geneva: CERN, 2020. URL: %5Curl%7Bhttps://cds.cern.ch/record/2730443%7D.

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

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

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

[143] ATLAS Collaboration. *Machine Learning Algorithms for b-Jet Tagging at the ATLAS Experiment*. Tech. rep. ATL-PHYS-PROC-2017-211. CERN, 2017. DOI: 10.1088/1742-6596/1085/4/042031.

[144] V Kostyukhin. *VKalVrt - package for vertex reconstruction in ATLAS*. Tech. rep. ATL-PHYS-2003-031. CERN, 2003. URL: %5Curl%7Bhttps://cds.cern.ch/record/685551%7D.

[145] *Topological b-hadron decay reconstruction and identification of b-jets with the JetFitter package in the ATLAS experiment at the LHC*. Tech. rep. Geneva: CERN, 2018. URL: %5Curl%7Bhttps://cds.cern.ch/record/2645405%7D.

[146] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps". In: *CoRR* abs/1312.6034 (2013). URL: https://api.semanticscholar.org/CorpusID:1450294.

[147] "Boosted Object Tagging with Variable-R Jets in the ATLAS Detector". In: (July 2016). URL: %5Curl%7Bhttps://cds.cern.ch/record/2199360%7D.

[148] David Krohn, Jesse Thaler, and Lian-Tao Wang. "Jets with variable R". In: *Journal of High Energy Physics* 2009.06 (2009), p. 059. DOI: 10.1088/1126-6708/2009/06/059. URL: https://dx.doi.org/10.1088/1126-6708/2009/06/059.

[149] Scott M Lundberg and Su-In Lee. "A Unified Approach to Interpreting Model Predictions". In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon et al. Curran Associates, Inc., 2017, pp. 4765–4774. URL: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

[150] Shaked Brody, Uri Alon, and Eran Yahav. "How Attentive are Graph Attention Networks?" In: *International Conference on Learning Representations*. 2022. URL: https://openreview.net/forum?id=F72ximsx7C1.

[151] Dasol Hwang et al. *Self-supervised Auxiliary Learning with Meta-paths for Heterogeneous Graphs*. 2021. arXiv: 2007.08294 [cs.LG].

[152] Hadar Serviansky et al. *Set2Graph: Learning Graphs From Sets*. 2020. arXiv: 2002.08772 [cs.LG].

[153] Rachel E. C. Smith et al. *Differentiable Vertex Fitting for Jet Flavour Tagging*. 2023. arXiv: 2310.12804 [hep-ex].

[154] Sam Shleifer, Jason Weston, and Myle Ott. *NormFormer: Improved Transformer Pretraining with Extra Normalization*. 2021. arXiv: 2110.09456 [cs.CL].

[155] Leslie N. Smith. *A disciplined approach to neural network hyperparameters: Part 1 – learning rate, batch size, momentum, and weight decay*. 2018. arXiv: 1803.09820 [cs.LG].

[156] Leslie N. Smith and Nicholay Topin. *Super-Convergence: Very Fast Training of Neural Networks Using Large Learning Rates*. 2018. arXiv: 1708.07120 [cs.LG].

[157] *ml.cern.ch*. https://ml.docs.cern.ch. Accessed: 2024-02-16.

[158] Johnu George et al. *A Scalable and Cloud-Native Hyperparameter Tuning System*. 2020. arXiv: 2006.02085 [cs.DC].

[159] Greg Yang et al. "Tuning Large Neural Networks via Zero-Shot Hyperparameter Transfer". In: *Advances in Neural Information Processing Systems*. Ed. by A. Beygelzimer et al. 2021. URL: https://openreview.net/forum?id=Bx6qKuBM2AD.

[160] Greg Yang and Edward J. Hu. "Tensor Programs IV: Feature Learning in Infinite-Width Neural Networks". In: *Proceedings of the 38th International Conference on Machine Learning*. Ed. by Marina Meila and Tong Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, 2021, pp. 11727–11737. URL: https://proceedings.mlr.press/v139/yang21c.html.

[161] Yann A. LeCun et al. "Efficient BackProp". In: *Neural Networks: Tricks of the Trade: Second Edition*. Ed. by Grégoire Montavon, Geneviève B. Orr, and Klaus-Robert Müller. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 9–48. ISBN: 978-3-642-35289-8. DOI: 10.1007/978-3-642-35289-8_3. URL: https://doi.org/10.1007/978-3-642-35289-8_3.

[162] ATLAS Collaboration. "Direct constraint on the Higgs-charm coupling from a search for Higgs boson decays into charm quarks with the ATLAS detector". In: *Eur. Phys. J. C* 82 (2022), p. 717. DOI: 10.1140/epjc/s10052-022-10588-3. arXiv: 2201.11428 [hep-ex].

[163] ATLAS Collaboration. "Constraints on Higgs boson production with large transverse momentum using $H \to b\bar{b}$ decays in the ATLAS detector". In: *Phys. Rev. D* 105 (9 2022), p. 092003. DOI: 10.1103/PhysRevD.105.092003. URL: https://link.aps.org/doi/10.1103/PhysRevD.105.092003.

[164] ATLAS Collaboration. "Anomaly detection search for new resonances decaying into a Higgs boson and a generic new particle X in hadronic final states using $\sqrt{s} = 13$ TeV pp collisions with the ATLAS detector". In: *Phys. Rev. D* 108 (2023), p. 052009. DOI: 10.1103/PhysRevD.108.052009. arXiv: 2306.03637 [hep-ex].

[165] ATLAS Collaboration. "Optimisation of large-radius jet reconstruction for the ATLAS detector in 13 TeV proton–proton collisions". In: *The European Physical Journal C* 81.4 (2021), p. 334. DOI: 10.1140/epjc/s10052-021-09054-3. URL: https://doi.org/10.1140/epjc/s10052-021-09054-3.

[166] ATLAS Collaboration. "Jet reconstruction and performance using particle flow with the ATLAS Detector". In: *The European Physical Journal C* 77.7 (2017), p. 466. DOI: 10.1140/epjc/s10052-017-5031-2. URL: https://doi.org/10.1140/epjc/s10052-017-5031-2.

[167] ATLAS Collaboration. *Improving jet substructure performance in ATLAS using TrackCaloClusters*. Tech. rep. Geneva: CERN, 2017. URL: %5Curl%7Bhttps://cds.cern.ch/record/2275636%7D.

[168] ATLAS Collaboration. *Identification of Boosted Higgs Bosons Decaying Into $b\bar{b}$ With Neural Networks and Variable Radius Subjets in ATLAS*. Tech. rep. Geneva: CERN, 2020. URL: %5Curl%7Bhttps://cds.cern.ch/record/2724739%7D.

[169] ATLAS Collaboration. *Efficiency corrections for a tagger for boosted $H \to b\bar{b}$ decays in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector*. Tech. rep. Geneva: CERN, 2021. URL: %5Curl%7Bhttps://cds.cern.ch/record/2777811%7D.

[170] ATLAS Collaboration. *Calibration of the ATLAS b-tagging algorithm in $t\bar{t}$ semi-leptonic events*. Tech. rep. Geneva: CERN, 2018. URL: %5Curl%7Bhttps://cds.cern.ch/record/2638455%7D.

[171] ATLAS Collaboration. *Calibration of light-flavour b-jet mistagging rates using ATLAS proton-proton collision data at $\sqrt{s}$ = 13 TeV*. Tech. rep. Geneva: CERN, 2018. URL: %5Curl%7Bhttps://cds.cern.ch/record/2314418%7D.

[172] ATLAS Collaboration. "Measurement of the c-jet mistagging efficiency in $t\bar{t}$ events using pp collision data at $\sqrt{s} = 13$ TeV collected with the ATLAS detector". In: *The European Physical Journal C* 82.1 (2022), p. 95. DOI: 10.1140/epjc/s10052-021-09843-w. URL: https://doi.org/10.1140/epjc/s10052-021-09843-w.

[173] ATLAS Collaboration. "Calibration of the light-flavour jet mistagging efficiency of the b-tagging algorithms with Z+jets events using 139 fb^{-1} of ATLAS proton-proton collision data at $\sqrt{s} = 13$ TeV". In: *Eur. Phys. J. C* 83.8 (2023), p. 728. DOI: 10.1140/epjc/s10052-023-11736-z. arXiv: 2301.06319 [hep-ex].

[174] ATLAS Collaboration. *Monte Carlo to Monte Carlo scale factors for flavour tagging efficiency calibration*. Tech. rep. Geneva: CERN, 2020. URL: %5Curl%7Bhttps://cds.cern.ch/record/2718610%7D.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[208] ATLAS Collaboration. *Vertex Reconstruction Performance of the ATLAS Detector at* $\sqrt{s} = 13$ *TeV*. Tech. rep. Geneva: CERN, 2015. URL: https://cds.cern.ch/record/2037717.

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

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

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

[212] ATLAS Collaboration. *Variable Radius, Exclusive-k_T, and Center-of-Mass Subjet Reconstruction for Higgs($\to b\bar{b}$) Tagging in ATLAS*. Tech. rep. Geneva: CERN, 2017. URL: https://cds.cern.ch/record/2268678.

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

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

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

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

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

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

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

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

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

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

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

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

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

Appendices