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 BOSON SEARCHES WITH THE ATLAS EXPERIMENT

Author
Maxence DRAGUET

Supervisor
Daniela BORTOLETTO

trinity
MMXXIV

M'ILLUMINO D'IMMENSO

UNGARETTI

ARTISTIC REPRESENTATION OF THE HIGGS POTENTIAL BY CERN [1]

ABSTRACT

Identifying the flavour of jets plays an essential role in many ATLAS analyses. The outcome of the hadronisation of quarks and gluons, jets leave a rich signature of numerous particles emitted in a main direction from the initially decaying one. This subject is extensively discussed in this thesis, with a complete review of the algorithmic developments carried out by the ATLAS Collaboration from 2020 to early 2024. Increasingly sophisticated machine learning models called taggers have been developed for this specific purpose. The classical approach relies on a hierarchical construction combining low-level physically-motivated taggers with a Deep Set or a Recurrent Neural Network as inputs to a high-level network predicting the flavour. Recently, 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 a Graph Attention Network or a Transformer Encoder unit. Expert knowledge is passed to the model by optimising multiple tasks, with different physics input types analysed in a multimodal framework. The design and training of these taggers are reviewed, with a study of the hyperparameter optimisation for large networks using techniques from the ML literature on Large Language Models.

Following the 2012 discovery of the Higgs boson by the ATLAS and CMS Collaborations, 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 now all been measured. Attention is shifting to the second-generation fermions, such as the c-quark, and on precision 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 the decay modes allows for a coherent joint analysis strategy improving the constraining of the shared backgrounds. Flavour taggers are used to identify candidate b- and c-jets to reconstruct the Higgs. The full p_T spectrum is covered, with the two candidate jets resolved at low momentum and a single merged boosted signature at high momentum. Three leptonic channels are defined based on the number of electrons and muons found in the final state. A fine categorisation is deployed with dedicated Boosted Decision Trees signal discriminants to increase the sensitivity. The analysis is blinded with an expected 95% CL$_s$ upper limit for the $VH(H \to c\bar{c})$ signal strength of 11.1 $\times$ the Standard Model prediction. The $VH(H \to b\bar{b})$ expected 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 template measurement is performed in stage 1.2 for $VH(H \to b\bar{b})$, in bins of p_T and number of additional jets.

Personal Contributions

The work presented in this thesis is inherently collaborative, having been carried out as a member of the ATLAS Collaboration. This thesis concentrates on the two subjects to which I mostly contributed: the development of new heavy flavour taggers and the combined $VH(H \to b\bar{b}/c\bar{c})$ analysis using the full Run 2. While I produced some "ATLAS" labelled figures, others were taken from public results produced by other members of the Collaboration. Plots without label have been personally produced, with some exceptions in the analysis chapter. This section highlights 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 the new ATLAS software release (R22) for Run 3.
- Modifications to the preprocessing to implement importance sampling when harmonising distributions of different flavour samples with the full simulation statistics.
- Modifications to the training software to include taus, flexible input variables definition, and general debugging.
- Modifications to the postprocessing to implement new visualisation and graphics.
- Retraining DL1r on the new R22 release and comparing it to the previous versions as well as pre-training studies on samples from older software releases.
- First PFlow and variable-radius (VR) training of DL1d with the DIPS sub-tagger.
- Hyperparameter and input optimisation of DL1d.
- Training the DIPS sub-tagger with VR jets.

- Hyperparameter optimisation studies of GN1 and GN2, with modifications to 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 led me to significantly participate in the development of the Umami [2] and Salt [3] software used to train the networks. My contributions have been part of different ATLAS publications, such as Ref. [4] with a DL1r model I trained, Ref. [5] with a DL1d model I trained, Ref. [6] for which I produced the DL1d input to the X_{bb}, as well as an upcoming ATLAS publication on GN2. I led the effort on hyperparameter optimisation of GN2, producing the public results in Ref. [7] presented at the 6^{th} Inter-Experimental Machine Learning Workshop [8] and the Cloud-Native AI Day KubeCon conference [9].

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

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

- Comparing the X_{bb} tagger to DL1r for the boosted $VH(H \to b\bar{b})$ by studying the impact on the signal sensitivity with dedicated analysis MVA trainings.
- 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, with studies leading to the final approach presented in this thesis. Additional study on the Higgs candidate reconstruction strategy in this control region.
- Derivation and harmonisation of the p_T^V-dependent ΔR_{cc} cuts in $VH(H \to c\bar{c})$.
- Training and deployment of the CARL models 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 software 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 modelling 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 as integrating an update to the output results visualisation.

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

Contents

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 Set
DL Deep Learning
DL1 Deep Learner 1 Model
DL1d DL1 with DIPS
DL1r DL1 with RNNIP
DNN Deep Neural Network
DT Decision Tree
ECAL Electromagnetic Calorimeter
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
HCAL Hadronic Calorimeter
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
PCA Principal Component Analysis
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 Vertexer 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 aptly named *Standard Model (SM)* [10, 11]. In the edifice of science, the SM is our current best understanding of the foundation of Nature at its smallest scale. The elegance of its structure resides in its mingling of mathematical concepts such as symmetries and gauge invariance to describe and predict the fundamental 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 [12, 13]. The successes of the SM have been continuosuly experimentally confirmed by countless measurements, particularly with the ATLAS and CMS observation of the theorised Higgs boson in 2012 [14, 15]. The SM is not however a complete theory of the elementary components of the Universe. Gravity is not included, neutrino masses are observed but not accounted for, and the SM does not explain astronomical observations of the existence of dark matter. The mass of fermions is introduced somewhat arbitrarily through Yukawa interactions coupling these particles to the Higgs boson. The origin of the strict mass hierarchy between the different fermionic generations is therefore left unexplained. The High Energy Physics (HEP) community finds itself in the unsual situation of having a remarkably accurate and inerrable yet incomplete model. Searches are actively ongoing to test all predictions of the SM, with the hope to uncover some discrepancies shedding some light on the way forward to correct the theory and include the currently unexplained properties of the Universe. Concerning the Higgs boson, while the leading production modes and the decay modes to third-generation fermions and vector bosons have been observed to agree with the SM, the couplings 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) [16–22].

This thesis presents, 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 for the first time performed jointly to leverage a shared strategy and better constrain the common backgrounds. The vector boson V (W or Z) associated production VH is used, with the V leptonically decaying to 0, 1, or 2 electrons, muons, or neutrinos. This latter requirement reduces the otherwise significant multi-jet background and provides an effective signature to select data to save online through the triggers. The search is performed with the 140 fb^{-1} of proton-proton collision data collected by ATLAS at a centre-of-mass energy of 13 TeV during the Run 2 of the Large Hadron Collider (LHC), from 2015 to 2018.

An essential component in the analysis is to reliably identify b- and c-quarks from the complex reconstructed structure of their decay called a *jet*. This is a challenging task due to the rich structure of jets and the large event rate leading to overlapping signatures. The ATLAS experiment is helped in this mission by the tremendous amount of real and simulated data accessible, leading to the effective deployment of state-of-the-art Machine Learning (ML) techniques. As such, the second central theme of this thesis is the elaboration of evermore sophisticated neural network model called *taggers* to classify the flavour of jets. The recent efforts of the ATLAS Collaboration to improve these methods are extensively described, with a complete account of the algorithmic developments from 2020 to early 2024. First presented is the development and training of DL1d, a hierarchical tagger relying on the DIPS sub-tagger built on a Deep Set network to replace the DL1r tagger using the RNNIP sub-tagger. More recentely, a breakthrough in performance has been obtained by adopting a single complex neural network with a powerful Graph Attention Network (GAT) or a transformer encoder unit at its core, in models respectively called GN1 and GN2. These development, the training, and the performance of these revolutionary new taggers are reported here in detail, as well as the latest 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 strives to present a coherent and connected narrative leading to the combined $VH(H \to b\bar{b}/c\bar{c})$ analysis presented last in Chapter 6. First, the SM, the Higgs mechanism, and Yukawa interactions are reviewed 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, and perhaps even more so in particle physics, Chapter 4 is entirely dedicated to a overview of the field relevant to 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 reported 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 components of the Universe. Nature at this microscopic scale is best represented by an intricate connexion between elementary particles, the undividable constituents of matter, and their interactions, also represented by particles. This framework is encapsulated into the mathematical foundation of Quantum Field Theory (QFT). A major scientific achievement of the second half of the XXth century is the elaboration of the so-called Standard Model (SM) of Particle Physics, a unified patchwork of theories describing all known elementary particles and three of the four fundamental interactions affecting them. This theory has stood the tests of countless experiments and grown with General Relativity into one of the two pillars of modern physics. Among its main achievements, it correctly predicted the existence of the Higgs boson, the W and Z bosons, the gluons, and the top and charm quarks. This chapter reviews relevant elements of the theories on which high energy particle physics rests to contextualise the work presented in this thesis and the significance of the analysis presented in Chapter 6.

2.1 The Standard Model of Particle Physics

To date, the SM is the most successful theory to describe the constituents and the dynamics of matter [11]. It stands at the centre of theoretical particle physics, and was elaborated by combining the theories of quantum mechanics and special relativity in the second part of the XX$^{\text{th}}$ century. Its numerous exploits have been rewarded by a total of 55 Nobel Prizes, and the SM is often hailed as the most successful theory of science due to its unique ability to predict properties of the Universe to a staggering degree of precision: most famously, the anomalous magnetic dipole moment prediction is in agreement with measurements up to 10 significant decimals [23]. The SM is expressed in the language of the dynamics of quantised fields of Quantum

Field Theory (QFT). These fields play two roles: describing matter itself through *fermions*, such as the electron, and the different interactions through *bosons*, such as the W and the Higgs boson H. The latters govern how matter interacts under the electromagnetic, weak, and strong interactions and with the Higgs field. Particles are the results of local excitations of quantised fields that are defined as operator-values distributions over Poincaré-Minkowski spacetime. Figure 2.1 displays the fundamentals particles of the SM.

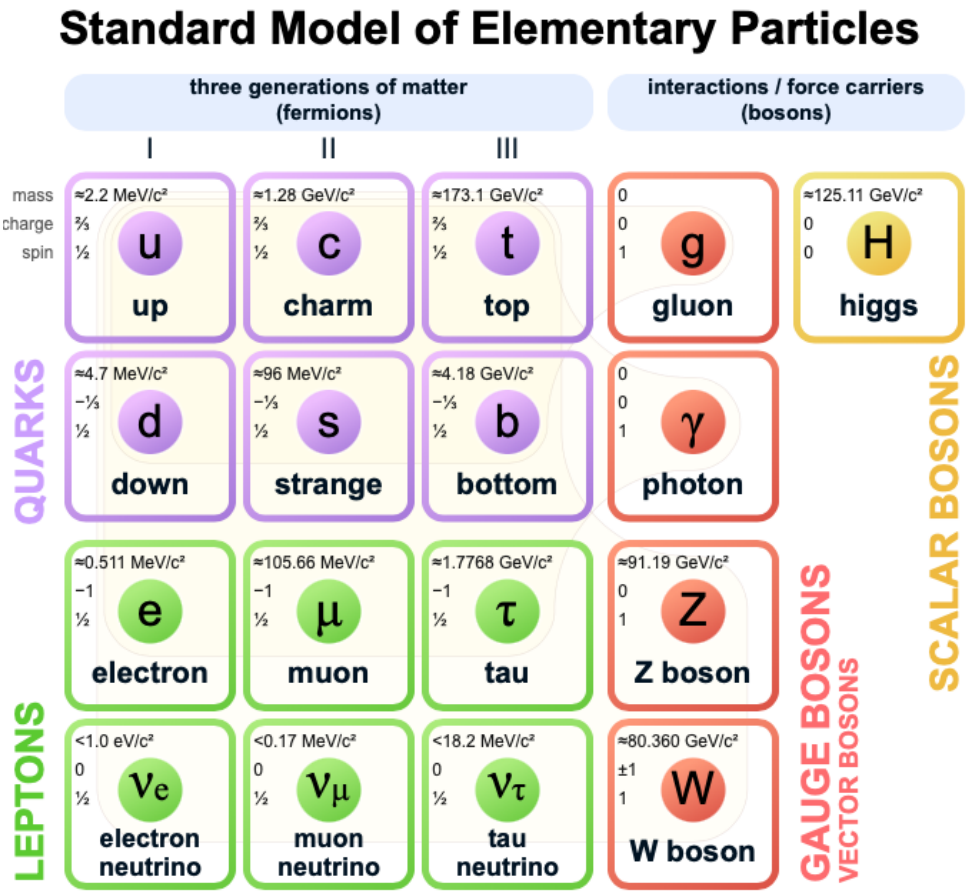


Figure 2.1: Elementary particles of the Standard Model [24]. Elementary fermions (quarks and leptons) are listed in the three left columns representing the three generations, and elementary bosons in the two right ones, with the gauge bosons in the first column and the Higgs scalar boson in the last 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 constituting the fermions, and integer-spin particles obeying the Bose-Einstein statistics establishing the bosons. The elementary fermions are evenly split into 6 quarks and 6 leptons each paired into three generations, with only the first generation being stable. The distinction between quarks and leptons 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 -1p, in units of the proton charge, or are neutral. The charged leptons include the electron e^-, the muon μ^-, and the tau τ^-. The neutral leptons are called neutrinos, with one neutrino ν_ℓ associated per charged lepton ℓ, e.g., the electron-neutrino ν_e for the electron e. In the SM, the numbers of leptons of each generation

is a conserved number. For the quarks, the electromagnetic charge is fractional, dividing them evenly between *up*-type quarks with charge $+2/3$ consisting of the up u, charm c, and top t flavours, and the *down*-type quarks with charge -1/3 and the flavours down d, strange s, and bottom b. To every particle corresponds an *anti*particle, with some quantum numbers changed such as the electrical charge that takes the opposite sign: e.g., the antiparticle of the electron e^- is the positron e^+.

The kinematics and dynamics of the fields representing the particles in the theory are expressed through a Lagrangian density $\mathcal{L}$, a spacetime discretised 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 is 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 symmetries required by 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 and constitute the backbone of Yang-Mills theories [25]. 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 dimensions in the 4-vector formalism. The full SM Lagrangian $\mathcal{L}_{\text{SM}}$ can be decomposed into 4 terms:

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

Each term encodes a different fundamental property within the unified framework of the SM. Three of the four known interactions of nature are encapsulated in the SM: the strong, the electromagnetic, and the weak forces. The gravitational interaction is set aside due to the weakness of its influence at subatomic scales. The mediators of the three included interactions are the gauge bosons, which are spin 1 particles with different properties arising from the nature of the interaction they symbolise. The electromagnetic and weak forces have been successfully unified as a single Electroweak (EW) interaction, while the theory of the strong force is Quantum Chromodynamics (QCD). One essential element in the SM is the so-called *Brout-Englert-Higgs interaction*, a special force through which some particles acquire mass. This interaction, summarised as *Higgs*, is underpinned by the eponymous Higgs field, an excitation of which is called a Higgs boson H. Yukawa interactions between the Higgs field and quarks are introduced to assign masses to the latter through Yukawa couplings. The analysis presented in Chapter 6 of this thesis is dedicated to a measurement of these couplings for the b- and c-quarks. The different interactions and their respective theoretical modelisations 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 carrier is the photon γ. 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-1/2 fermion with

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

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

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

where the dependency on the spacetime coordinate x is now omitted for clarity. Such a Lagrangian models the free dynamics of any fermion, such as the electron e^- or the c-quark. The electrical charge q of particles is conserved by every known interaction. This conservation must in turn be the result of a symmetry, leading the Dirac Lagrangian to be made invariant under a local gauge $U(1)$ transformation

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

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 covariant derivative* D_μ

$$D_\mu = \partial_\mu + iqA_\mu, \tag{2.5}$$

where a new vector field A_μ 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 gauge field A_μ its own dynamics, 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 in $F_{\mu\nu}F^{\mu\nu}$ introduces a kinetic term for the gauge field. The interaction between the fermion ψ and the gauge field A is represented by the term $q\bar{\psi}\not{A}\psi$ combining them with the coupling q. The strength of electromagnetism is defined in scale by the charge e of an electron. In the present case, the electrical charge q is the conserved quantity of the gauge symmetry, which requires the introduction of a gauge field A_μ that is interpreted as the photon field. The Lagrangian is adapted to include the full dynamic of the electromagnetic interaction through the strength tensor $F_{\mu\nu}$. Fermionic fields are introduced with this approach for the different known fermions, ψ_e, ψ_μ, ψ_u, ψ_c, etc. Their interaction with A_μ defines each time a unique conserved electromagnetic charge q_e, q_μ, q_u, q_c, etc. This procedure is general: the gauge invariance of a Lagrangian introduces a spin-1 gauge boson. Interestingly, the required $U(1)$ invariance forbids the presence of mass terms of the form $m^2 A^\mu A_\mu$ in the Lagrangian, seemingly condemning all gauge bosons to be massless.

2.1.2 Electroweak Sector

The weak force is propagated by two massive gauge vector bosons: the $W^{\pm}$ of mass $m_W \approx 80.36$ GeV[1] and the Z^0 of mass $m_Z \approx 91.19$ GeV [26]. This apparent contradiction with the massless requirements of a $U(1)$ symmetry is elegantly solved by the Brout-Englert-Higgs mechanism [12, 13]. This mechanism, described in the next section, is applied to a unified expression of the electromagnetic and weak interactions known as Electroweak (EW) theory in the Glashow-Weinberg-Salam (GSM) model [27–29]. 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. The local $SU(2)$ transformation acts as

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

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. Since they do not commute, the EW sector is built on a non-Abelian group and is therefore a Yang-Mills theory with self-interacting gauge mediators [25]. The gauge field linked to $SU(2)_L$ leads to a covariant derivative, similarly to QED, to ensure the invariance of the Lagrangian under the symmetry. It is expressed as

$$D_\mu = \partial_\mu + igT_aW_\mu^a, \tag{2.8}$$

with three gauge fields W_μ^1, W_μ^2, W_μ^3 and a unique 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 particle 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 $I_W^3 = \pm 1/2$. For the right-handed part, $I_W = 0$ with $I_W^3 = 0$, decoupling it from the gauge bosons W_μ^a. Physically, the observed weak charged current interactions correspond to the $W^{\pm}$ bosons resulting from a linear combination of the first two gauge fields

$$W_\mu^{\pm} = \frac{1}{\sqrt{2}}\left(W_\mu^1 \mp W_\mu^2\right). \tag{2.9}$$

The $W^{\pm}$-bosons only couple to left-handed particles, but an experimentally observed electrically-neutral Z boson couples to both left- and right-handed particles. This is 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 set as $Y = 2(Q - I_W^3)$, so that the electromagnetic charge Q matches observations. The total covariant derivative of the electroweak sector of the SM is therefore expressed in the GSM model as

$$D_\mu = \partial_\mu + ig\frac{\sigma_a}{2}W_\mu^a + ig'\frac{Y}{2}B_\mu \tag{2.10}$$

[1]The unit system adopted throughout 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 adopt SI units and divide by c^2.

where W_μ^a and B_μ are respectively the $SU(2)_L$ and $U(1)_Y$ gauge bosons. The full EW Lagrangian built with this covariant derivative is

$$\begin{aligned}\mathcal{L}_{\text{EW}} = -\frac{1}{4}W_a^{\mu\nu}W_{\mu\nu}^a - \frac{1}{4}B_a^{\mu\nu}B_{\mu\nu}^a + \sum_j \left[\bar{\ell}_{L_j} i\not{D}\ell_{L_j} + \bar{e}_{R_j} i\not{D} e_{r_j}\right] \\ + \sum_j \left[\bar{Q}_{L_j} i\not{D} Q_{L_j} + \bar{u}_{R_j} i\not{D} u_{R_j} + \bar{d}_{R_j} i\not{D} d_{R_j}\right],\end{aligned} \tag{2.11}$$

where the electroweak field strength tensors are the $W_a^{\mu\nu}$ and $B_a^{\mu\nu}$ matrices. The sum over j represents the three fermionic generations, each introduced separately for the lepton fields, with the left-handed doublet ℓ and the right-handed singlet e for the charged lepton, and the quark fields, with Q representing the left-handed doublet and u_R and d_R the right-handed up- and down-type singlets. Neutrinos, that are in the ℓ doublet, can only interact through the weak force and there are therefore no right-handed neutrinos in the SM.

The linear combination of Equation 2.9 is required to represent the physical charged fields $W^\pm$. Similarly, the physically observed electromagnetic photonic field A_μ and the Z boson field Z_μ are the result of a linear combination of the neutral B_μ and W_μ^3 in the GSM model. This combination depends on a fundamental parameter of the SM called the *weak mixing angle* θ_W such that

$$\cos\theta_W = \frac{g}{\sqrt{g^2 + g'^2}} \tag{2.12}$$

thereby establishing a connexion between the coupling strengths of the weak interaction and the electromagnetic interaction, with coupling e the electrical charge of a positron, as

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

The intrinsic strength of the weak interactions is of similar order to that of the electromagnetic interaction, but is weak in appearance due to the large mass of its vector gauge bosons. The weak force is the only known fundamental interaction to violate symmetry under parity transformation. A significant achievement of modern particle physics is the unification of interactions that are perceived as different at low energies. The problem of the mass of the vector gauge bosons however remains. Additionally, the split of fermionic fields into left- and right-handed components leads fermionic mass terms in the QED Lagragian of Equation 2.6 to violate the gauge invariance. Both issues are resolved by considering an additional scalar field introducing the mechanism of the Brout-Englert-Higgs and Yukawa interactions, as explained in the following sections.

2.1.3 The Brout-Englert-Higgs Mechanism

The Brout-Englert-Higgs mechanism, abbreviated *BEH* henceforth, offers an elegant solution to introduce mass terms for the gauge fields $W_\mu^\pm$ and Z_μ [12, 13]. It postulates the existence of an additional 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 a complex scalar field with 4 degrees of freedom:

$$\phi = \begin{pmatrix} \phi^+ \\ \phi^0 \end{pmatrix} = \frac{1}{2} \begin{pmatrix} \phi_1 + i\phi_2 \\ \phi_3 + i\phi_4 \end{pmatrix}. \tag{2.13}$$

This complex scalar field is made to interact with the electroweak gauge fields through their covariant derivative of Equation 2.10 as

$$\mathcal{L}_{\text{Higgs}} = (D_\mu \phi)^\dagger (D^\mu \phi) - V(\phi), \tag{2.14}$$

where the first term describes the kinetic energy of the ϕ field and the second term is the Higgs potential energy

$$V(\phi) = -\mu^2 \phi^\dagger \phi + \lambda (\phi^\dagger \phi)^2. \tag{2.15}$$

The expression of this potential is constrained by the need for the theory to be renormalisable. Two scalar constants govern the Higgs potential: μ and λ describing, respectively, the quadratic and quartic interactions 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 non-null field values such that

$$\phi^\dagger \phi = \frac{1}{2}(\phi_1^2 + \phi_2^2 + \phi_3^2 + \phi_4^2) = \frac{\mu^2}{2\lambda} = \frac{v^2}{2} \tag{2.16}$$

introducing in the last equality the so-called *vacuum expectation value* $v = \sqrt{\frac{\mu^2}{\lambda}}$ of the field ϕ. The infinite degeneracy of the Higgs potential vaccuum 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 expectation is simply

$$\langle 0|\phi|0\rangle = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v \end{pmatrix}. \tag{2.17}$$

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. To model the full field dynamics around the chosen vacuum state, a particular gauge choice is adopted to absorb unphysical Goldstone bosonic fields into the electroweak vector fields called *unitarity gauge* [30], simplifying the expansion to

$$\phi(x) = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v + h(x) \end{pmatrix}, \tag{2.18}$$

where h is the real neutral Higgs scalar field. Introducing this expression into the Higgs Lagrangian of Equation 2.14 gives

$$\begin{aligned} \mathcal{L}_{\text{Higgs}} = & \frac{1}{2}(\partial_\mu h)(\partial^\mu h) + \frac{\mu^2}{2}(v+h)^2 - \frac{\lambda}{16}(v+h)^4 \\ & + v^2 \frac{g^2}{4}(W_\mu^+ W^{\mu-})(1 + \frac{h}{v})^2 + v^2 \frac{g^2 + g'^2}{8}(Z_\mu Z^\mu)(1 + \frac{h}{v})^2 \end{aligned} \tag{2.19}$$

where mass terms appear 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 masses of the gauge vector bosons are

$$m_W = \frac{v}{2}g, \quad m_Z = \frac{v}{2}\sqrt{g^2 + g'^2} \tag{2.20}$$

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

$$m_W = m_Z \cos\theta_W.$$

The Higgs field is massive with mass

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

The BEH mechanism elegantly assigns mass to the gauge vector bosons while leaving the photon massless. It requires the addition of the scalar Higgs boson H as a massive spin-1 elementary particle. Furthermore, the introduction of the related Higgs field permits the expression of mass terms for the fermions in the SM, 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 a conserved quantum number called *colour*. The only particles having a colour charge in the SM are quarks and the gauge mediators of the strong interaction: the gluons g. There are three colour charges typically labelled *red*, *blue*, and *green*, each coming into its direct or anticolour e.g., red-antired. 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

$$D_\mu = \partial_\mu + ig_s\frac{\lambda_a}{2}G_\mu^a \tag{2.21}$$

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 built with the set of λ_a Gell-Mann matrices. The gauge fields introduced here are the G_μ^a corresponding to the gluonic mediators of the strong field. Gluons carry 2 colour charges, leading to the 8 Gell-Mann matrices λ^a and 8 gauge vector fields G_μ^a, indexed by a. The generators of the $SU(3)_C$ group do not commutate since

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

where f^{abc} are the $SU(3)_C$ structure constants. The non-commutation of the $SU(3)_C$ generators means the $SU(3)_C$ part of the SM is a non-Abelian group, and therefore a case of a Yang-Mills theory with self-interacting gauge fields [25]. Gluonic strength tensors are built similarly to the electromagnetic strength tensor as

$$G_{\mu\nu}^a = \partial_\mu G_\nu^a - \partial_\nu G_\mu^a - g_s f^{abc} G_\mu^b G_\nu^c,$$

with the structure constants generating self-interactions. The full QCD Lagrangian is

$$\mathcal{L}_{\text{QCD}} = -\frac{1}{4}G^a_{\mu\nu}G^{\mu\nu}_a + \sum_f \bar{\psi}_f(i\not{D} - m_f)\psi_f, \tag{2.22}$$

where ψ_f are the six quarks fields, one per flavour f, transforming as an $SU(3)_C$ triplet with one component per colour quantum number.

Like every coupling constant, α_s varies with energy. At low energies, the interaction is so strong that perturbative calculations break and the behaviour of *colour confinement* is observed: any attempt to isolate a quark requires such a large amount of energy that a quark-antiquark pair is spontaneously produced. The shear strength of this interaction explains its short propagation distance despite the fact its gluonic mediators are massless. At higher energies, asymptotic freedom and perturbative calculations are possible thanks to the reduced coupling strength. This typically requires higher-order corrections for the calculation to converge, with some terms, such as quark self-energy loops, diverging to infinity. These so-called *ultraviolet divergence* are removed by renormalising fields and parameters so that the infinities 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 [31]. The former is introduced to deal with the 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, and enters the parton distribution and fragmentation functions introduced later in this chapter.

Quarks combine to form colourless aggregate of matters called *hadrons*, with either a di-quark system combining a quark and an antiquark into a *meson*, or a tri-quark system forming a *baryon* of which the proton p (uud, $q_p = +1$) and the neutron n (udd, $q_n = 0$) are prime examples. The particles constituting the hadrons are called partons. The process leading to the neutralisation of the colour charge of an asymptotically free quark is called *hadronisation*.

2.1.5 Yukawa Interactions

In the QCD Lagrangian of Equation 2.22, the introduction of mass terms for the quarks breaks the gauge invariance of the theory under $SU(2)_L \otimes U(1)_Y$ and must be therefore nullified $m_f = 0$. The masses of all fermions are included in the SM by introducing Yukawa interactions between the Higgs and fermionic fields [32]. These terms are expressed by the following Lagrangian

$$\mathcal{L}_{\text{Yukawa}} = -\frac{1}{\sqrt{2}}\sum_f y_f \bar{\psi}_f(v + h)\psi_f, \tag{2.23}$$

where ψ_f are the fermionic fields and the fundamental *Yukawa couplings* $y_f = \sqrt{2}m_f/v$ for each flavour of electrically charged fermion f are introduced as coupling strengths. Picking the v components in the sum in parentheses give mass terms m_f to the fermions, with the h terms leading to Higgs-fermion interactions proportional to the fermion mass. The vacuum expectation value plays the role of a general mass scale of the theory, with Yukawa couplings refining the specific mass of each fermionic species. For the quark sector, a further 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* [26]

$$V_{CKM} = \begin{pmatrix} V_{ud} & V_{us} & V_{ub} \\ V_{cd} & V_{cs} & V_{cb} \\ V_{td} & V_{ts} & V_{tb} \end{pmatrix}, \tag{2.24}$$

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 current interactions allow for flavour-changing processes. The matrix is almost diagonal in magnitude, 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 modes. Protons are composite particles and, at high energies, the main *hard-scattering* interaction is between components of the protons called the *partons*. These partons consist of the *valence* quarks (*uud* for a proton) but there are also contributions from *sea* quarks, as well as gluons and photons, present within the hadron due to quantum fluctuations. In a pp collision, two interacting partons a and b from each proton undergo the main event $ab \to X$, with the activity from the rest of the protons assigned to the *Underlying Event (UE)*. The cross-section for the global process $pp \to X$ is expressed using the factorisation theorem [31] as

$$\sigma_{pp\to X} = \sum_{a,b} \int_0^1 dx_a \int_0^1 dx_b f_a(x_a, \mu_F) f_b(x_b, \mu_F) \int d\sigma_{ab\to X}\left(x_a P_a, x_b P_b, \mu_R, \mu_F\right), \tag{2.25}$$

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 $p_i = x_i P_i$ taken as a fraction x_i of the proton momentum P_i, and μ_F is the previously introduced factorisation scale symbolising the dependency of the PDF on the energy scale of the underlying process $ab \to X$. The interaction is effectively factorised into two terms: the first picking up the interacting partons and their fraction of momentum and the second considering the main $ab \to X$ event.

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

- *Gluon-gluon fusion (ggF)*: two partonic gluons fuse into a quark loop with a radiated Higgs boson as $pp \to H$. The quarks in the loop couple to the Higgs, hence the massive 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 [33]. This process is favoured thanks to the large contributions of gluons to the protonic PDFs at the energies considered.

- *Vector boson fusion (VBF)*: two off-shell vector bosons V (W or Z) radiated from partonic quarks fuse to form a Higgs as $pp \to qqH$, with cross-section $\sigma_{VBF} = 3.77 \pm 0.09$ pb [33]. The quarks leave the characteristic signature of a forward jets pair in the event.

- *Associated production with a vector boson (VH)*: the Higgs boson is produced in association with a vector boson V as $pp \to VH$. This process is studied in detail in Chapter 6, dedicated to an analysis of Higgs bosons decaying to pairs of b- or c-quarks in this production mode. It has a cross-section of $\sigma_{VH} = 2.24 \pm 0.14$ pb [33]. 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 as $pp \to q\bar{q}H$. 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 $b\bar{b}H$ [33].

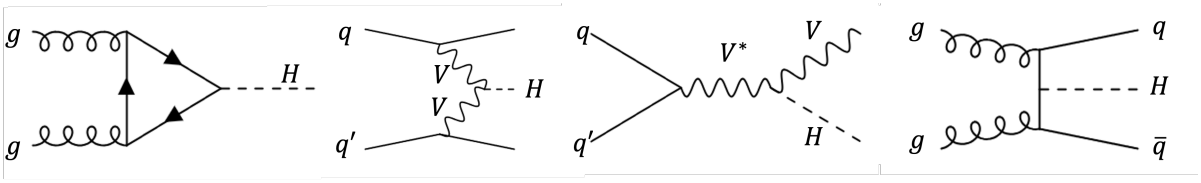


Figure 2.2: The leading order Feynman diagrams for Higgs production at the LHC, from left to right: gluon-gluon fusion, vector boson V fusion, vector boson associated production, and $q\bar{q}$ associated production.

The dependency of the Higgs boson production modes 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 with $m_H = 125$ GeV is $\Gamma_H = 4$ MeV, implying a short lifetime of $\tau_H \sim 10^{-22}$ s and restricting measurement to the decay products. The branching ratios at $\sqrt{s} = 13$ TeV are displayed on the right of Figure 2.3, 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

$$BR(H \to X) = \frac{\Gamma(H \to X)}{\Gamma_H}, \tag{2.26}$$

where the total Higgs width Γ_H is the sum of all partial decay width $\Gamma(H \to X)$, for all possible X. The most likely Higgs decay mode is to a pair $b\bar{b}$ (~ 58 %), followed by the decay to a WW pair (~ 21 %). The $c\bar{c}$ decay branching ratio is ~ 2.9 %.

The WW and ZZ decays can only be achieved via virtual off-shell Higgs bosons, reducing their contributions despite their large mass. Fermionic decays are challenging to observe in hadron colliders due to the large multi-jet background. The vector bosons and di-photon leptonic decays benefit from advantageous experimental conditions, being easier to identify thanks to the

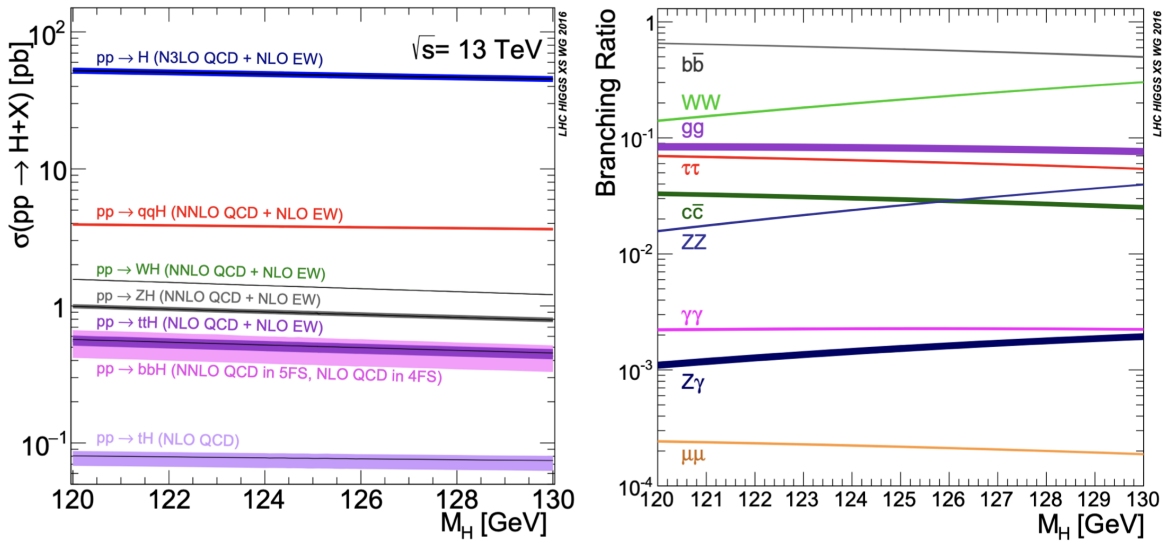


Figure 2.3: The Standard Model production cross-sections from proton-proton collisions (left) and decay branching ratio (right) of the Higgs boson as a function of m_H at $\sqrt{s} = 13$ TeV [33].

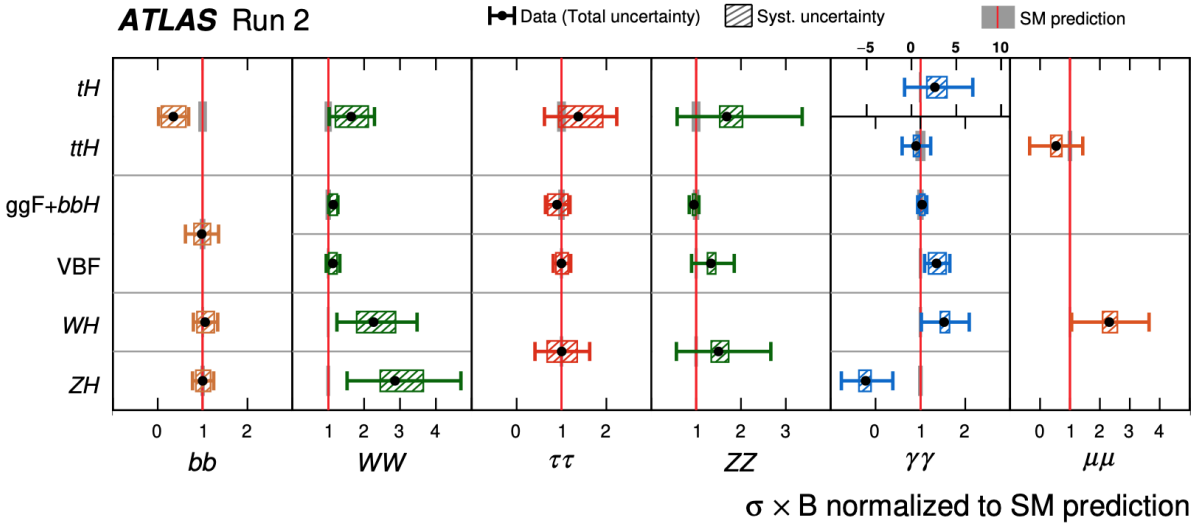


Figure 2.4: Ratio of observed signal strengths to the SM predictions for different combinations of Higgs boson production and decay modes [34]. Horizontal bars denote the 68% confidence interval, with grey bands showing theory uncertainties on the SM cross-section × branching ratio predictions.

presence of leptons and suffering from less background contamination. For these reasons, the ATLAS and CMS Collaborations observed in 2012 a particle of mass $m_H = 125$ GeV with the properties of the Higgs boson by combining searches for $H \to \gamma\gamma$, $H \to ZZ \to \ell^+\ell^-\ell'^+\ell'^-$, and $H \to W^+W^- \to \ell^+\ell^-\nu\bar{\nu}$ [14, 15]. This paved the way to many additional Higgs measurements, summarised in Figure 2.4. Decay modes to the electroweak gauge bosons and third-generation fermions (t, b, τ) have all been observed, and the sensitivity to the second-generation (μ) is approaching evidence-level. The observed Higgs boson is remarkably consistent with the SM predictions [34].

CHAPTER 3

The LHC & ATLAS Experiment

Modern particle physics explores the frontier of the technological reach of science. To discover the Higgs boson, a remarkably complex infrastructure is necessary to probe physics at the required scale. The Centre Européen pour la Recherche Nucléaire (CERN) hoest the largest and most powerful particle accelerator ever built: the Large Hadron Collider. It has held this title since its construction concluded in 2008 [35], and easily ranks as one of the most intricate machines ever created. Protons are accelerated to up to 99.9999991% of the speed of light in its giant 27 km long ring-shaped beamline, buried 100 m below the surface of the French-Swiss border in the suburb of 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, composed of bunches of particles, are collided at four precise interaction points where large detectors are built and operated by dedicated collaborations: ATLAS [36], CMS [37], ALICE [38], and LHCb [39]. The first two are multipurpose experiments with overlapping physics programs, while ALICE and LHCb respectively study heavy ion and heavy flavour physics. This chapter describes the experimental setup of the LHC and the ATLAS experiment, focusing on proton-proton collisions and introducing relevant elements for the work presented in this thesis.

3.1 The Large Hadron Collider

The last machine in the complex multi-stage accelerator complex of CERN depicted in Figure 3.1, the Large Hadron Collider (LHC) is capable of frontally colliding proton or heavy ion beams packed into bunches. The beams collide at four interaction points, where dedicated experiments such as ATLAS measure the resulting physics signatures in large detectors designed for their specific physics programmes. The life of a proton beam starts innocuously in a bottle of ionised

hydrogen H^- gas, the content of which is passed through a linear accelerator called LINAC 4[1] to reach energies of 160 MeV [40]. After stripping the ionised hydrogen atoms of their two electrons to leave bare protons, the next acceleration stage happens in the Proton Synchrotron Booster (BOOSTER), bringing the beam energy to 2 GeV [41]. The protons are then handed to increasingly larger synchrotrons: the Proton Synchrotron (PS) to reach energies of 26 GeV [42], followed by the Super Proton Synchrotron (SPS) to reach energies of 450 GeV [43]. The beam is finally injected into the LHC in two different beamlines circulating the proton in opposite directions [44]. Superconducting dipole magnets generating a 8.33 T field steer the highly energetic beams, while complex geometries of magnets such as quadrupoles and sextupoles refine the bunch shape through focusing effects. Powerful radiofrequency cavities accelerate the protons to their final energy of 6.5 TeV, giving a total pp collision centre-of-mass energy of $\sqrt{s} = 13$ TeV in Run 2.

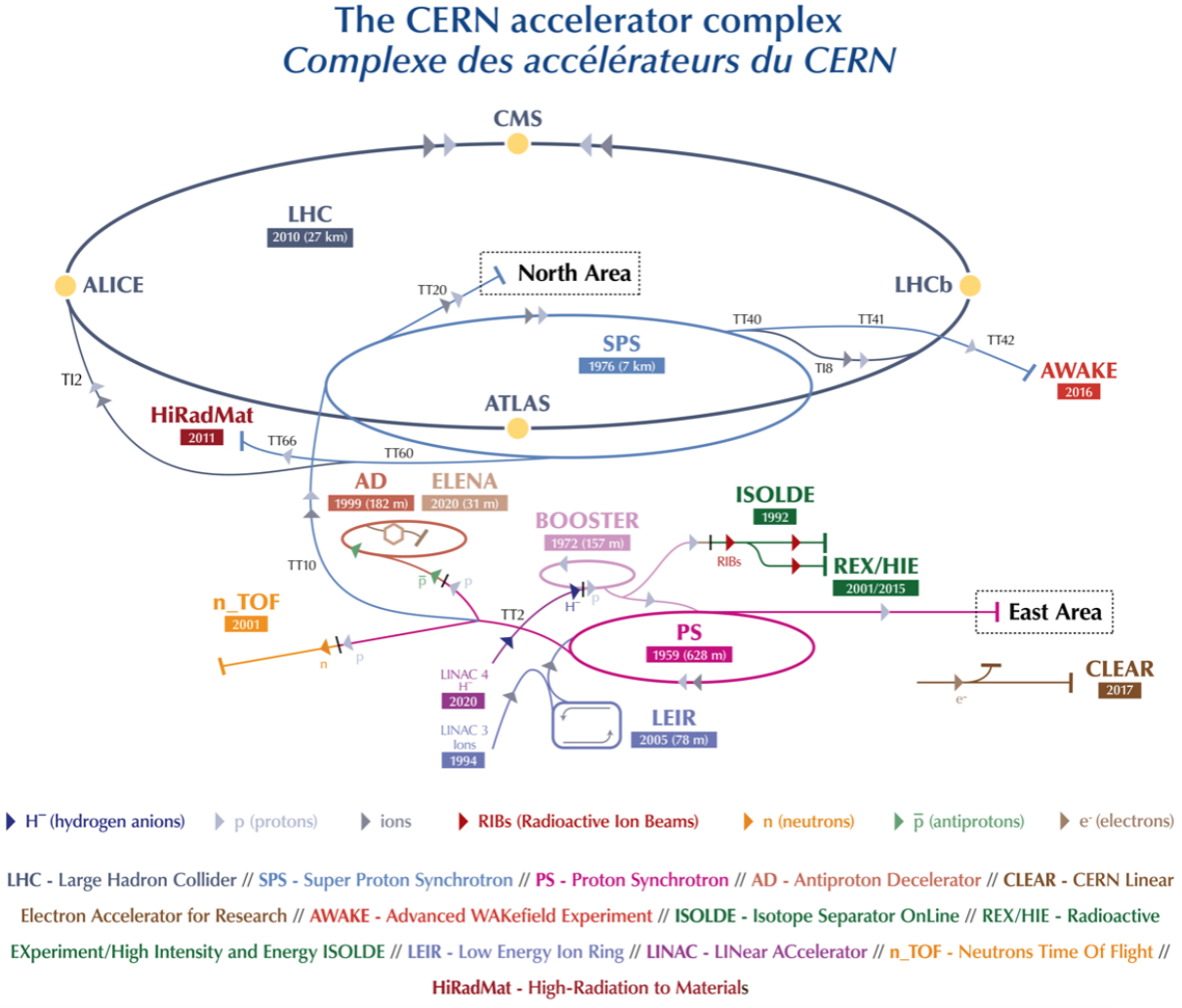


Figure 3.1: The complete accelerator complex of CERN for Run 3 [45].

The operation of the LHC is split into dedicated *runs* of data taking separated by *shutdowns* to maintain or upgrade the infrastructure. Key metrics about these runs from the point of view of the ATLAS experiment are displayed in Table 3.1. Run 2 operated at a larger centre-of-mass energy ($\sqrt{s}$) and higher average instantaneous luminosity $\mathcal{L}$ than Run 1, with the ongoing Run 3 yet again pushing up the limit.

[1] LINAC 2 before 2020.

	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 accelerator performance of the LHC in the different runs of data taking. The reported values correspond to those recorded by the ATLAS experiment [46–48]. 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 $\langle\mu\rangle$.

The average instantaneous luminosity $\mathcal{L}$ measures the quantity of data collected from the relation

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

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 geometrical extensions of the beam density distribution in the x- and y-direction. The integrated luminosity $\int \mathcal{L}\, dt$ measures the number of events 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 by ATLAS corresponds to 140.1 ± 1.2 fb^{-1}, with a small uncertainty of 0.83 % [47] thanks to a complex measurement involving luminosity-dedicated detectors such as LUCID-2 [49]. Figure 3.2 shows the cumulative distribution of the integrated luminosity during Run 2, jointly with another important machine parameter: the average number of interactions per bunch crossing $\langle\mu\rangle$.

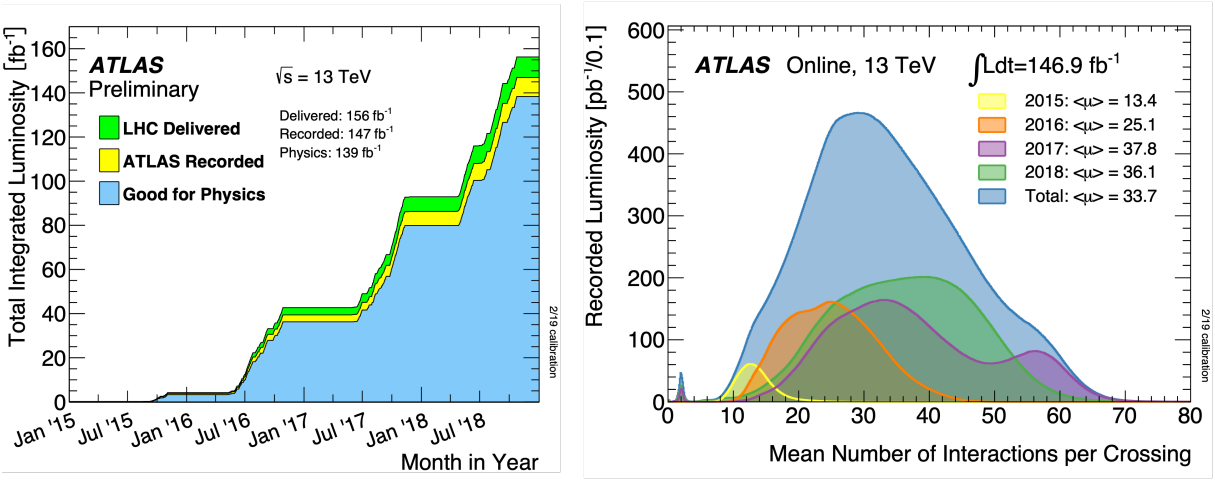


Figure 3.2: The ATLAS cumulative integrated luminosity delivered, recorded, and useful for physics (left) and the average luminosity-weighted pile-up distribution (right) during Run 2 [50]. The luminosities listed correspond to an early calculation that was corrected in Ref. [47].

The main event during the collisions of the proton bunches is the inelastic hard scattering where most of the energy transfer occurs. Other protons in the bunches can have softer interactions leading to background activity referred to as *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 those involved in the hard scattering, and *out-of-time PU* if the protons are from lateral bunches. The LHC separates bunches by a 25 ns delay, corresponding to a machine frequency of 40 MHz. To control the luminosity, the angle of attack of the beams are tweaked so that their geometrical overlap measured by σ_x and σ_y at the point of impact is tunable. 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 maintains and operates the eponymous cylindrically shaped multi-layered detector, lying 100 m underground with a length of 45 m and a 26 m diameter [36], as presented in Figure 3.3. The experiment is designed to probe a broad 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 endcaps closing the geometry at its extremities. Essential requirements in the technical design had to be met to manage the extreme event rate by requiring a fast response from radiation-hard sensors with state-of-the-art readout electronics in combination with good spatial and temporal resolution to disentangle the effect of pile-up.

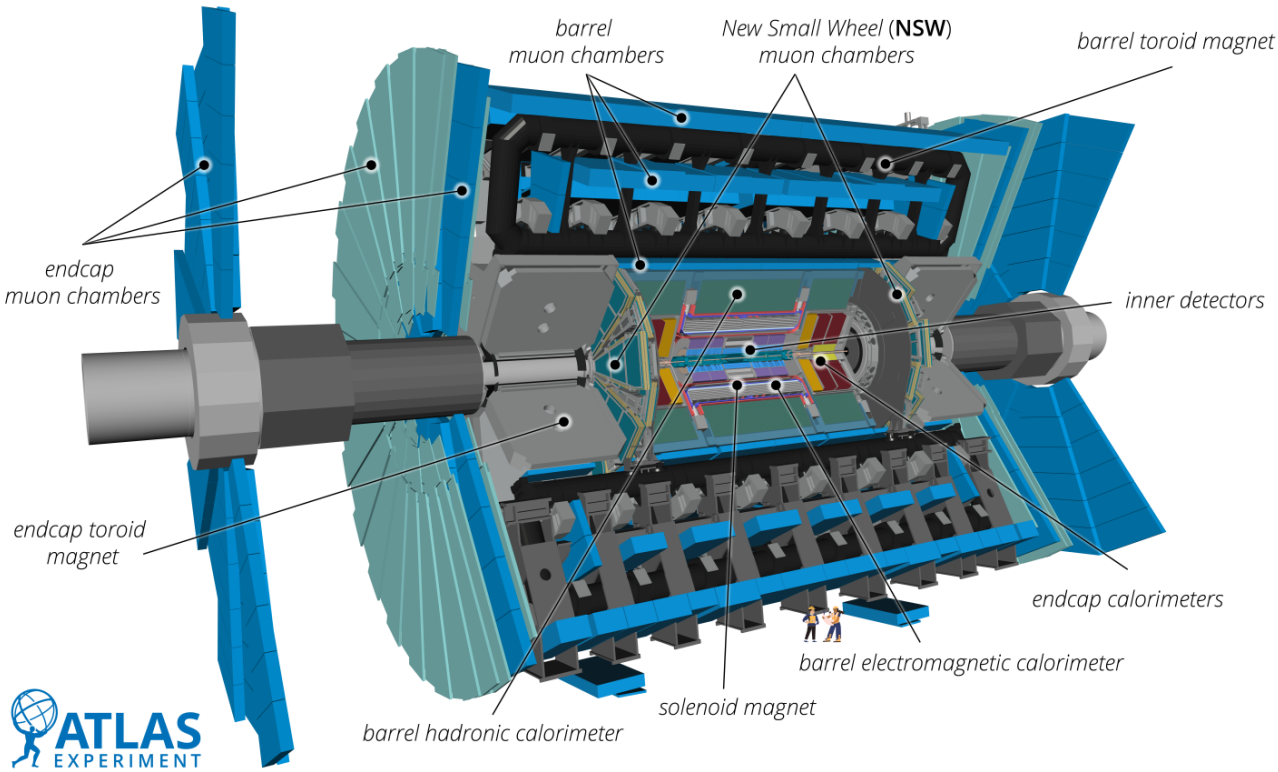


Figure 3.3: Cut-away view of the ATLAS detector [51].

The coordinate system adopted in ATLAS is described in Figure 3.4: the x-axis points to the centre of the LHC ring, the y-axis points upwards, and the z-axis is in the longitudinal

direction along the beamline, anti-clockwise when viewed from atop. The azimuthal angle ϕ is defined in the transverse plane $x - y$, and the polar angle θ is measured upwards from the beam-axis. The transverse momentum p_T of a particle is obtained from its momentum vector $\boldsymbol{p} = (p_x, p_y, p_z)$ of magnitude p as $p_T = p \sin\theta = \sqrt{p_x^2 + p_y^2}$. This projection plays a crucial role as the momentum's longitudinal component p_z is not fully resolvable due to the openings for the beamline and the interacting partons carrying only a fraction of the original proton momenta. Only the transverse momentum can therefore be reliably measured. Since the partons are mostly longitudinally boosted, the transverse momenta in an event are approximately balanced. The rapidity y of a particle, a crucial invariant in 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_z the energy and longitudinal momentum of the particle. In the ultrarelativistic limit, when $p >> m$, the rest mass is negligible and $E \approx p$. In this case, the rapidity y is well approximated by the experimentally reconstructable pseudo-rapidity η:

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

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

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

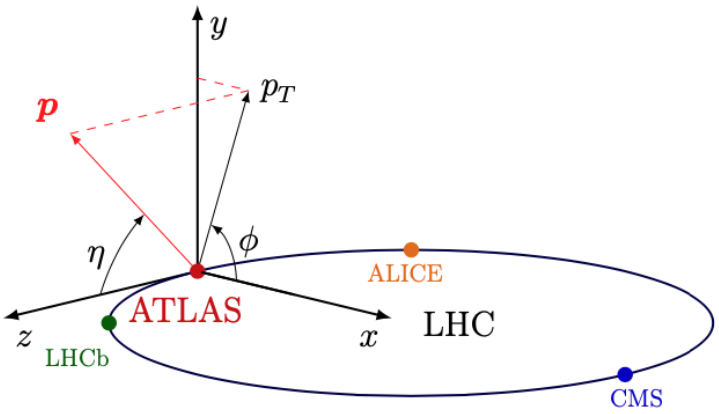


Figure 3.4: The ATLAS coordinate system [52].

As depicted in Figure 3.3, ATLAS combines different systems into a single precision machine. These subdetectors measure information in the range $|\eta| < 2.5$, with some specialised subdetectors such as the calorimeters extending further. An essential component of the detector system is its two types of superconducting magnets. Four central solenoid magnets enclose the point of interaction to 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 particle of momentum p is deflected by a magnetic field B due to the Lorentz force, leading to a relation

between the radius of curvature R of the trajectory and the momentum p such that

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

where $p_\perp$ is the magnitude of the momentum perpendicular to the magnetic field $\boldsymbol{B}$, and q is expressed in the unit of proton charge. Therefore, by measuring the curvature, the component of the momentum transversal to the magnitude B can be inferred. Higher magnetic fields induce 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) [53]. This tracker covering the range $|\eta| < 2.5$ in a radius of 3 cm to 1 m is designed to record hits in silicon semiconductors or straw tubes from the passage of charged particles so that their trajectory or *track* can be reconstructed from the combined signature. The powerful 2 T magnetic field of the central solenoids enables this detector to measure both the charge and the momentum of charged particles. The ID combines three subsystems, represented in Figure 3.5.

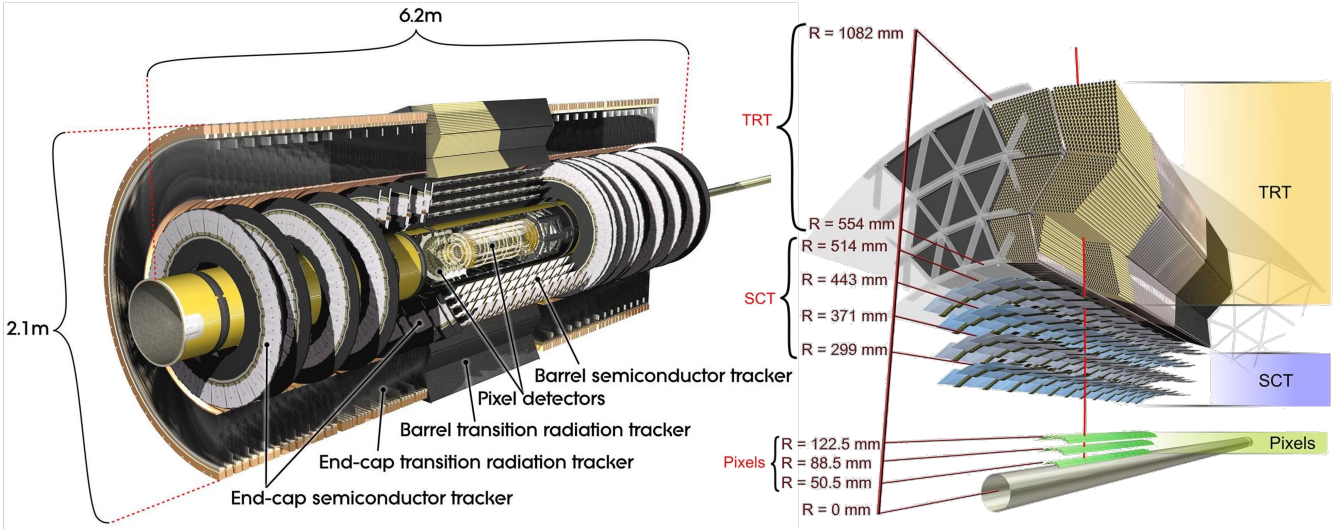


Figure 3.5: The Inner Detector of ATLAS [51].

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 pixels [53, 54]. During Run 2, an additional *Insertable B-Layer (IBL)* was added at a 33 mm radius from the centre, with 12 million pixels [55]. This detector gives robust and precise tracking performance and plays a major role in flavour tagging, as described in Chapter 5. Pixels measure 50 × 400 μm^2 and are organised along $R\phi \times z$, with a smaller 50 × 250 μm^2 for the IBL. The geometrical position resolution delivered is of 10 μm (67 μm) in the transverse $R\phi$ plane (z-direction) [56, 57].

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 [58, 59]. The resolution is typically of 17 μm in $R\phi$ and 580 μm in z [60].

The final system is the *Transition Radiation Tracker (TRT)*, a gas-based straw-tube tracker aiding track reconstruction by delivering numerous hits [61]. 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 is fitted a conducting wire at its centre and the surface is electrically charged, so that the passage of a charged particle ionises the gas leading to a measurable discharge. Polyethylene is placed between the tubes to encourage the emission of transition radiations from relativistic particles proportionally to their Lorentz boost $\gamma \sim E/m$. Consequently, the TRT is used for both 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 130 μm in the $R\phi$ plane for the barrel and the $z-\phi$ plane in the endcaps [62].

Altogether, the track inverse transverse momentum resolution of the ATLAS ID is

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

where $\oplus$ denotes a sum in quadrature [36]. This corresponds to a relative error of about 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 the muons that are largely left unscathed. The system is composed of an Electromagnetic Calorimeter (ECAL) and a Hadronic Calorimeter (HCAL) both covering the $|\eta| < 3.2$ region and a forward calorimeter for the $3.2 < |\eta| < 4.9$ region [36], as displayed in Figure 3.6. Each calorimeter interlays specific layers of active and passive materials. The passive material has a large atomic number to induce a cascade of particles called *shower*. The active material, typically liquid argon (LAr) in ATLAS, collects the energy from these showers through ionisation or scintillation light.

The ECAL is designed to collect the energy of electrons and photons and contributes to the measurement of the energy of jets. The active material is LAr, with absorbing plates of lead used as passive material to encourage *bremsstrahlung* $e \to e\gamma$ and pair production from photons $\gamma \to e^+e^-$. The ECAL has a depth of 22 X_0, where the unit of *radiation length* X_0 tracks the distance for an electron to retain only $1/e$ of its original energy. The energy resolution is parametrised into three terms representing the sampling term, the electric noise, and a constant contribution for miscalibrations, summed in quadrature as [63]

$$\frac{\sigma_E^{\text{ECAL}}(E)}{E} = \frac{10\%}{\sqrt{E}} \oplus \frac{0.17[\text{GeV}]}{E} \oplus 0.7\%, \tag{3.8}$$

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

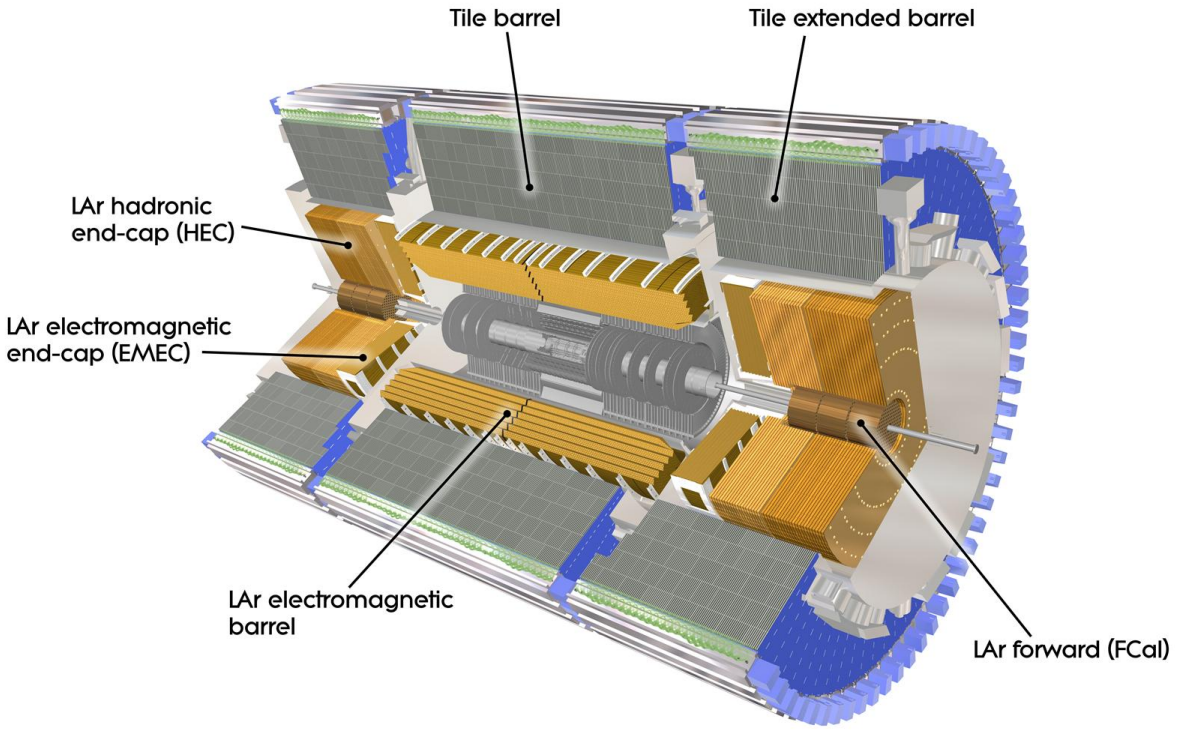


Figure 3.6: The calorimeter systems of ATLAS [51].

The HCAL is designed to capture the energy of hadronic showers, 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 approximately $10X$, 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 of hadrons, with an HCAL resolution expressed as [63]

$$\frac{\sigma_E^{\text{HCAL}}(E)}{E} = \frac{52.9\%}{\sqrt{E}} \oplus 5.7\%, \tag{3.9}$$

because the electrical noise is negligible. This translates into a resolution of ∼17% (∼6%) at energies of ∼10 GeV (∼100 GeV).

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 energies studied means they barely interact with the previously described system, only leaving a few hits in the inner detectors and flying largely undisturbed through the calorimeters. For this reason, the outmost subdetectors of ATLAS are specially designed to be sensitive to muons. The Muon Spectrometer (MS), shown in Figure 3.7, is a dedicated muon tracking system that also provides an effective triggering hardware, as described later in this chapter. The muon tracker is composed of drift tubes split between the barrel region for $|\eta| < 1.2$ and the endcaps $1.2 < |\eta| < 2.7$, with cathode strip

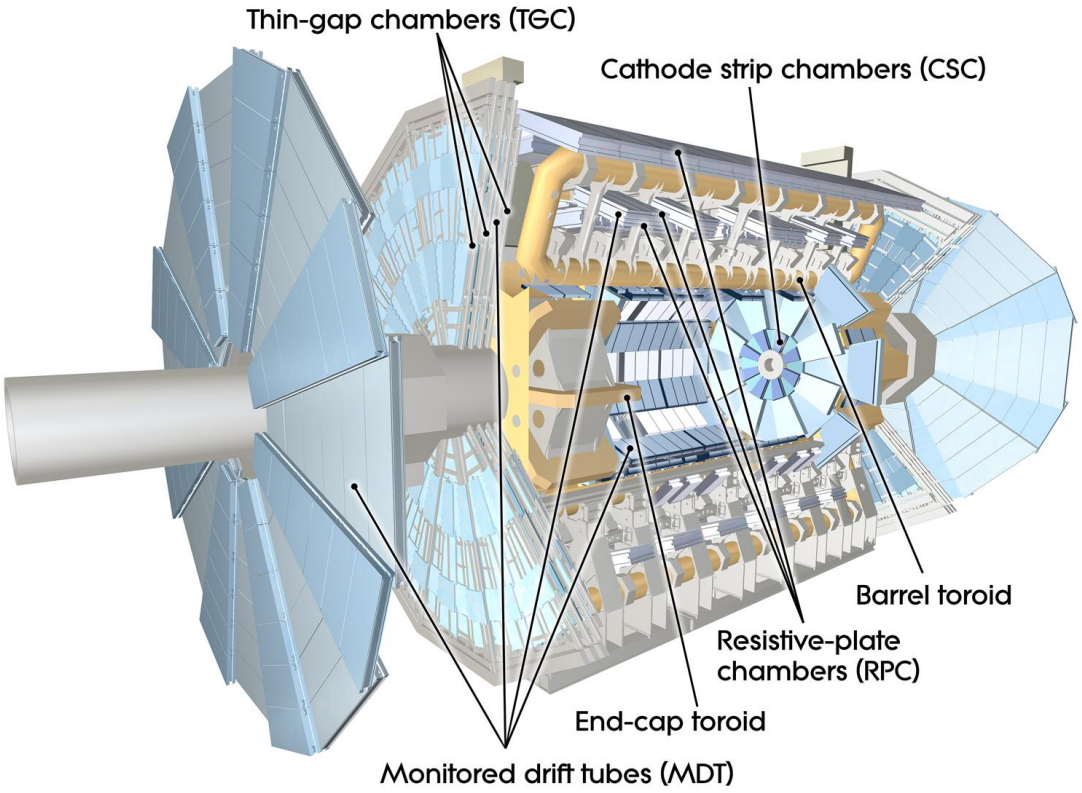


Figure 3.7: The muon detectors of ATLAS [51].

chambers in the inner layers of the endcaps to resist the larger activity. The trigger system relies on resistive-plate chambers in the barrel and thin-gap chambers in the endcaps. To improve momentum and charge measurements from the reconstructed 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 [64].

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 40 MHz, due to the 25 ns bunch-crossing. This unfortunately leads to a data generation rate that is too high for the electronics and computing resources available, requiring the Collaboration to design specific approaches to reduce the rate to a manageable level [65]. This is the task of the trigger system, which is described first in this section. Events that pass the trigger thresholds are stored and must be further analysed 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 for object types relevant to the presented work. This latter step is performed thanks to the extensive ATLAS software [66, 67], exploiting the specific signatures of the different detected particles as schematised in Figure 3.8.

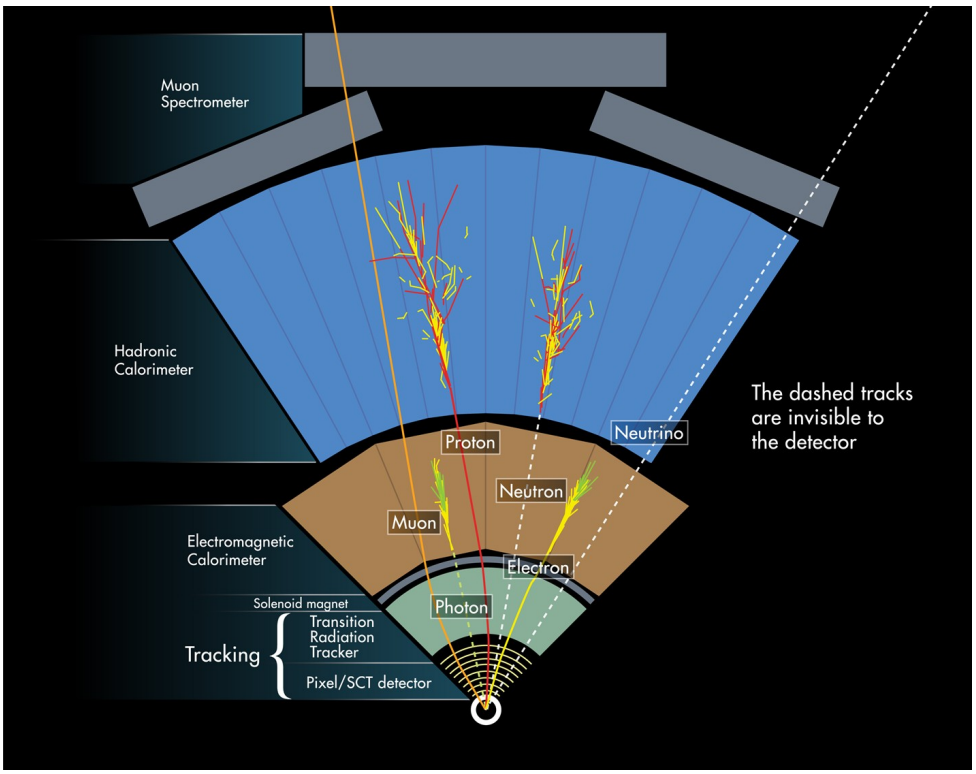


Figure 3.8: Schematics of different particles signatures in the ATLAS detector [68].

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 $\sim$ 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 suitable for data storage [69]. 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 periods for each setup. Analyses can then select data based on the specific signatures sought.

3.3.2 Low-Level Signatures: Tracks, Vertices, and Clusters

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 the collected localised energy deposits called *hits*. With denser pile-up activity, the number of hits in a single event becomes significant, making track reconstruction a challenging computational problem [70, 71]. The trajectories are curved thanks to the previously described superconducting magnets. From a set of hits, tracks are fitted inside-out [71]: clusters of three hits in the Pixel or SCT detectors are first identified as *seeds*, with additional hits associated by a com-

binatorial Kalman Filter [72] based on compatibility criteria with 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 while favouring high p_T tracks. The process is then extended to the TRT from the outside-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 Pixel 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 Parameters (IPs), respectively z_0 and d_0, measuring the distance from the Primary Vertex (PV) to the perigee[2] of the track in their respective plane.

Tracks are drawn by charged particles that emanate from the main event, the decay of particles, or are radiated through different physics processes. If a reconstructed charged particle is produced in the hard scattering event, its trajectory leads back to this special location called the *Primary Vertex (PV)* [73]. If the particle is produced in a subsequent decay, the point of emanation can sometimes be distinguished and is labelled *Secondary Vertex (SV)* [74]. Reconstructing the vertices is crucial for the physics programme of the Collaboration. The primary vertex is identified from a seed vertex first on the set of all well-reconstructed tracks [75]. The vertex position is then iteratively refined by removing tracks incompatible with the reconstructed vertex and refitting, until some quality criteria are met. Discarded tracks are then used to identify secondary and 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 cells with energy deposits matching specific criteria, either from the *sliding window* or *topocluster* algorithms [76]. The former generates fixed-size rectangular clusters by translating a window to maximise the transverse energy E_T measured. The latter clusters neighbouring cells based on a signal-to-noise criterion. As the sliding window method is easier to calibrate, it is used in electron, photon, and hadronic-τ reconstruction. Topoclusters 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, mainly in the ID and ECAL. In the central region $|\eta| < 2.5$, electrons are identified and reconstructed with both subdetectors. The forward region $2.5 < |\eta| < 4.9$ is only covered by the calorimeters, and the shape of the shower is used to identify electrons. Here, only centrally produced *prompt*-electron reconstruction is 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 jets can be mistaken as electrons, with identification and isolation criteria derived to provide high-purity and effective selections for analyses. The reconstruction relies on calorimeter clusters and track information. When relevant, tracks are matched to clusters with the expected energy loss taken into consideration. The track is extrapolated to ensure compatibility with the cluster barycentre, and

[2]The point of closest approach of the track with respect to the PV.

the process is run again with more stringent conditions after refitting the matched tracks. A prompt-electron is required to have a track matched to the PV. The absence of precision hits or a matched track leads to considering the calorimeter clusters as a photon deposit. Photons can however be mistaken as electrons due to the photon-conversion process, where $\gamma \to e^+e^-$.

To further distinguish prompt-electrons from non-prompt electrons and photons, a likelihood-based identification algorithm built on a Multivariate Analysis (MVA) discriminant is deployed [77]. Features exploited include the number of hits in each tracker layer, the track IPs, and some calorimeter cluster parameters. Several operating points that are progressively more selective are defined on the Multivariate Analysis (MVA) discriminant, from *Very Loose*, *Loose*, *Medium*, to *Tight*. Prompt-electron candidates are required to be isolated from other tracks and energy deposits, with specific isolation criteria that are either ID- or calorimeter-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 deposits in a calorimeter cone around the electron cluster. As further described in Chapter 6.7, the efficiencies of the electron reconstruction, including identification and isolation, are estimated 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, without a matched for unconverted photons. Converted photons are allowed hits in the outer layers of 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 faint track in the ID and some energy deposits in the calorimeters. The fragmented signature in the different subdetectors is combined to identify muon candidates. In the MS, tracks are constructed from a fit of the successive 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 operating points are defined to identify muons as *Loose*, *Medium*, *Tight*, and *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 $Z \to \mu^+\mu^-$ and $J/\psi \to \mu^+\mu^-$ samples.

3.3.5 Jets

Quarks and gluons are the most commonly produced particles in a hadron collider. As described in Chapter 2, these particles carry colour charges and therefore undergo hadronisation when produced to neutralise their free colour. This complex phenomenological process leaves 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 the energy deposited in the hadron calorimeters. These aggregated objects are constructed by applying a clustering algorithm on tracks and/or calorimeter clusters, depending on the jet definition.

The most notorious clustering method is the anti-k_T algorithm, thanks to the robustness of

the defined jets to collinear splitting and additional soft emissions [78]. The algorithm starts by considering high-momentum objects, after which softer objects are considered and potentially added to grow the jets or start a new jet. Two objects are considered at a specific step of the algorithm: the seed object i, either the highest momentum object or the jet in construction, and the currently unassigned highest transverse momentum object j. Two distances are evaluated when considering whether to cluster these objects

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

The first distance, d_{ij}, combines the angular apperture $\Delta R_{ij} = \sqrt{(\eta_i - \eta_j)^2 + (\phi_i - \phi_j)^2}$ between i and j with the transverse momentum k_T of the two objects and a fixed *radius* parameter R. This distance defines a radius limiting the size of the jet cone. It is compared to the second distance, d_{iB}, assessing the size of the already formed jet i. If $d_{ij} < d_{iB}$, j is clustered with i into a larger jet i, otherwise i is identified as a jet and removed from consideration. The algorithm proceeds after updating the distances until all constituents are assigned. Typical radii 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 to identify heavy object 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 properties of the particle they represent. The following types are all reconstructed with the anti-k_T algorithm but clustering different objects:

- PFlow jets combine particle-flow objects [79] with a radius $R = 0.4$. These objects combine tracking information from the ID with the calorimeter clusters, leading to a better energy resolution at low p_T and lower pile-up contamination after calibration [80].
- EMTopo jets are constructed from denoised topological calorimeter clusters called *topoclusters*, based on the per cell energy significance $S_{\text{cell}} = E_{\text{cell}}/\sigma_{\text{cell}}$, where E_{cell} is the energy and σ_{cell} the expected noise level in the cell [81]. 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 trimmed to remove the contributions from soft contamination, which is mainly due to PU and UE activity, leading to an improved mass resolution [82].
- Track-jets are constructed with a Variable Radius (VR) depending on the jet p_T, such that the wide cone used at low p_T ($R \sim 0.4$) becomes narrower at high p_T ($R \sim 0.02$). They are typically identified as sub-jets of a large-R jet, to give access to the single-jet flavour tagging techniques described in Chapter 5.

PFlow and VR 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 are assigned a flavour based on the presence 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 found within the jet, as described in detail in Chapter 5.

Jets benefit from an extensive calibration to correct their reconstructed properties such as the mass, the energy, and the jet axis. 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 calorimeters and leakage out of the active regions. The *Jet Energy Scale (JES)* calibration implements these corrections in successive steps [83]:

- *Origin*: the jet axis, initially constructed from the centre of ATLAS, is corrected to point from the PV, and the reconstructed p_T is updated.
- *Pile-up*: both in-time and out-of-time pile-up leave additional energy deposits in the calorimeters. This is subtracted from the jet, first from an overall estimation based on the average PU and then from the actual number of interactions and vertices in the event.
- *Absolute*: absolute energy corrections dependent on the energy E and η are derived to match the data energy scale to the particle-level energy scale with dedicated simulation samples.
- *Eta inter-calibration*: the detector is not homogenous and the forward region measurements are typically less accurate. Corrections are applied to forward jets based on central jets ($|\eta| < 1.4$).
- *Global sequential calibration*: energy leakage in the calorimeters is accounted for with a set of momentum corrections based on five different observables tracking the jets shape and energy.
- *In-situ calibration*: corrects any possible differences due to an incorrect description of the detector in the simulations by performing a fit to data in the dedicated measurement of a well-reconstructed object. Events from the following processes are used at increasing p_T scales: Z+jet events with the leptonic Z decays, γ+jet, and QCD multi-jet events.

The JES is parametrised by p_T, and uncertainties are derived for analyses to include in their modelling. The *Jet Energy Resolution (JER)* is then defined as σ_{p_T}/p_T, and also calibrated with uncertainties derived from a fit to di-jet events [83]. Despite the JES correction procedure, PU jets can still be significant and a Jet Vertex Tagger (JVT) is used to reject this background [84]. This implements a 2D likelihood method built from track variables. From this discriminant, different selection criteria are derived as operating points with specific PU jet rejections.

3.3.6 Taus

Taus are the heaviest generation of charged lepton, with a mass of 1.8 GeV slightly higher than that of c-quarks [26]. Their lifetime is so short that they mostly decay within the beampipe without directly leaving any signature. They leptonically decay 35% of the time to neutrinos and an e or a μ, hence their hadronic decays are more frequent. The leptonic decays are hard to disentangle from prompt electrons and muons. Hadronic decays however leave a discernable signature reconstructed as a small-R jet identified by a Recurrent Neural Network (RNN), to disentangle them from PU and QCD jets [85]. Different operating points are derived at specific efficiencies.

3.3.7 Missing Transverse Energy

Some physics objects do not leave a signature in the detector, such as neutrinos. Their presence is not directly detectable but can be inferred thanks to the negligible initial p_T of the two interacting partons. Requiring the transverse energy E_T^{miss} and momentum to be balanced, the missing transverse energy is calculated as the negative vectorial sum of the transverse momenta of objects, as

$$\boldsymbol{E}_T^{\text{miss}} = -\sum_{\text{hard}} \boldsymbol{p}_T - \sum_{\text{soft}} \boldsymbol{p}_T, \tag{3.11}$$

where the sum is decomposed into a *hard* term englobing all high-level physics objects and a soft term including good-quality ID tracks associated with the primary vertex but not to one of the high-level physics objects [86]. 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, a definition of the often confonded terminology is presented. This is followed by presenting 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.
- *Creating*, the ability to generate new content or information.
- *Acting*, the ability to perform a task or to modify the direct environment.

Artificial Intelligence (AI) research is a large field of investigation that studies these various aspects through numerous subjects such as robotics, Natural Language Processing (NLP), computer vision, generative modelling, and Reinforcement Learning (RL). Artificial intelligence is broadly separated into three levels according to the performance of the underlying system:

1. *Narrow Intelligence* represents artificial intelligence capabilities on a unique task, for which the software is specifically 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 to make 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.

2. *General Intelligence* refers to an artificial intelligence capable of matching human problem-solving skills in multiple environments. In particular, this hypothetical setting require machine to learn new tasks on their own and extrapolate from 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 the automated learning process to any task.

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 type is accessible and routinely deployed, with the second one the focus of ambitious research for state-of-the-art research laboratories. The inception of reactive AI, the initial approach attempted, comes from the research into games in the 50s and 60s. This paradigm saw the rise of algorithms capable of searching for optimal moves 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 the worthiness of an action: e.g., a queen is typically 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. State space exploration of many realistic problems however scales asymptotically with the dimension of the input, quickly rendering reactive-based approaches impractical. Combined with the need for human-encoded insights 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 tractable, such as autonomous driving, speech recognition, seamless robotics, etc. For such problems, complete programmatic problem formulations do not exist, prohibiting a reactive-based approach. The revolutionary paradigm of limited-memory AI has 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 introduced a paradigm shift to the field, moving away from human-declared logic-based rules written in a specific syntax, the wholemark of reactive AI. The lattter involves the execution of statements such as

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

for an input x and an output y. With limited-memory AI, the state representation and decision steps are encoded in the mathematical models of 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 specific 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 $f(x) = \hat{y} \to y$ under training convergence. The model f is trained to become an accurate estimator of the label y.

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, Support Vector Machine (SVM), logistic regression, kernel methods, k-Nearest Neighbours, etc.
- *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, constituting a specialised approach to building models from the core artificial neuron 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. ML can be deployed on various different tasks:

- *Classification:* assigning a discrete variable called *label*, to a datapoint: e.g., identifying b-jet. The general case is multiclass, with n possible labels, and a particularly common case is binary classification ($n = 2$).
- *Regression:* predicting a continuous variable for a datapoint: e.g., momentum reconstruction.

- *Features extraction:* given a dataset with specific internal features, construct new features, e.g., reconstructing the secondary vertex from a set of tracks. A special subcase of this category is embedding data points into a different hyperspace. The dimension of this final space can be larger, when embedding the data into a richer space, or smaller in the case of dimensionality reduction. A common example of the latter is Principal Component Analysis (PCA), which projects the data to a subspace spanned by the principal eigenvectors, those associated to the largest eigenvalues.

- *Generation:* sampling new data from a distribution matching the training dataset distribution, e.g., sampling new $t\bar{t}$ events from a learnt 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 according to the amount of human intervention [87]:

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

- *Unsupervised learning:* the data is not endowed with the extra information the model must predict but rather has underlying features that to be extracted. The model is therefore trained with an objective to optimise without explicit targets and should discover patterns and insights without 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 in *active learning*. A prime example is to combine an unsupervised clustering task with a classification of the clusters. This is particularly fruitful when the cost of labelling the data is expensive, as is the case with real-world data.

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

- *Reinforcement Learning:* this paradigm of ML is dedicated to the setting of a game-theoretic environment. An agent explores and interacts with an environment by choosing actions from a learnable policy by estimating its current situation and expected reward. In RL, the agent learns to construct the best policy to satisfy a reward function and obtain the best outcome.

These different settings are particularly explored in deep learning, which is widely recognised as the most performant technique thanks to its ease of scaling in complexity.

4.1.3 Deep Learning

Deep Learning (DL) refers to family of methods predominantly derived in the 1980s that have quickly grown in popularity in the last decade, with widely advertised results on competitive

benchmark tasks in pattern recognition, such as the super-human performance of the *DanNet* model [88] based on Convolutional Neural Networks (CNNs) [89]. The basis of any deep learning method is the artificial neuron, a logical unit inspired by the design of a human neuron. Several such units are combined into layers of any number of neurons defining the width of the layer, and the layers 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 depth-wise between layers or width-wise between neurons. 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) simulation programs of both the physics process and the detector reconstruction, the simulations 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 rich 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, aligning with the main data representations studied by 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 most relevant approaches for HEP.

4.2 Machine Learning Methods for Physics

High-energy physicists enjoy a special relationship with machine learning. 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 HEP analysis can be described as five steps process:

1. Data collection: real data is collected from a detector exposed to the underlying physics desired.
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 required by modern particle physics experiment are composed of a complex set of subdetectors sensitive to different physical phenomena, as described in Chapter 3. The low-level information collected by different devices must be processed and recombined to generate *objects*, aggregated information that hold physical meaning. 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 that is most sensitive to the studied *signal* or process. The sensitivity aspect underlies the need to take into account limitations in the knowledge of theoretical physics, the precision of the apparata, and the statistics of both simulated and measured events. 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 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 in Chapter 3.3.1, can be upgraded to use sophisticated DL models running online thanks to a hardware backbone built on Field-Programmable Gate Arrays (FPGAs) or Graphics Processing Units (GPUs).

2. Simulating a dataset through Monte Carlo is a computationally intensive task. Each event must pass through a selection of probabilistic steps, with only a simulated data point satisfying all requirements reaching the final sample. While This process can be optimised significantly with refined MC methods, 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 current limitation of these approaches is the difficulty to fully incorporate the sophisticated theoretical model required to simulate the data, as any discrepancy or non-closure introduces levels of disagreements that are counter-productive in the final the physics analysis.

3. ML is particularly well-suited for object reconstruction. Broadly, machine learning offers scalable, efficient, and accurate techniques for this essential task. Important examples in ATLAS are particle identification (e.g., τ identification), E_T^{miss} reconstruction, and heavy-flavour jets classification, as demonstrated in Chapter 5.

4. Historically, physicists have relied on a cut-base approach to selecting data: the relevance of different variables is analysed for the physics problem at hand, to identify the best features to use to select events through manually defined restrictions. For example, in a leptonically decaying Z boson measurement to two charged leptons $\ell^+\ell^-$, restricting the invariant mass of the lepton pair $m_{\ell^+\ell^-}$ to lie close the Z boson rest mass is beneficial to

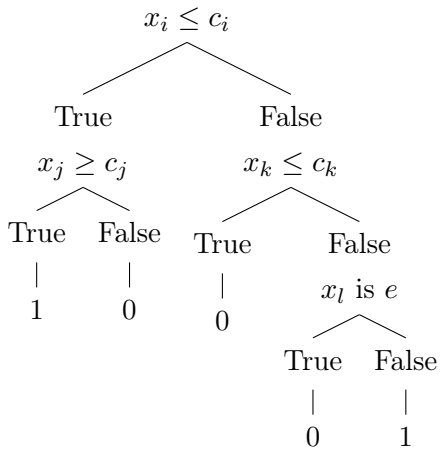
select this process. Machine learning entirely bypasses this process, learning directly from an appropriate set of signal and background samples a transformation of the input features optimising the separation of signal from background.

5. The likelihood function of the constructed statistical test, quantifying 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. Additionally, anomaly detection, such as the search for unknown resonances, can be automated with unsupervised machine learning.

Contributing to step 3 in the aforementioned list is of the main focus of this thesis: developing DL-tools for improved jet classification. The analysis presented in the latter part of the text also introduces some classical ML techniques of data selection as suggested in step 4.

4.2.1 Decision Trees

Decision Trees (DTs), 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, and are interpretable. The model relies on the recursive partitioning of the input space [87]. Each partition step is a *node* from which a tree structure emerges, where an initial *root* state is subsequently partitioned along different branches with one *leaf* per final output region. The splits are performed on the features of the input data, with the method accepting both discrete categorical values (e.g., the label of a lepton as e, μ, τ) and continuous values (e.g., $m_{\ell^+\ell^-}$). The following is a simple example of a classification tree outputting the predicted class as 0 or 1:



At each node, a condition is learnt 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.

Finding the optimal set of partitions of a dataset is an NP-complete problem and therefore intractable for large datasets. Instead, a greedy approach is adopted to build a tree, relying on a heuristic to find a satisfying solution. The most common approach is to successively choose the

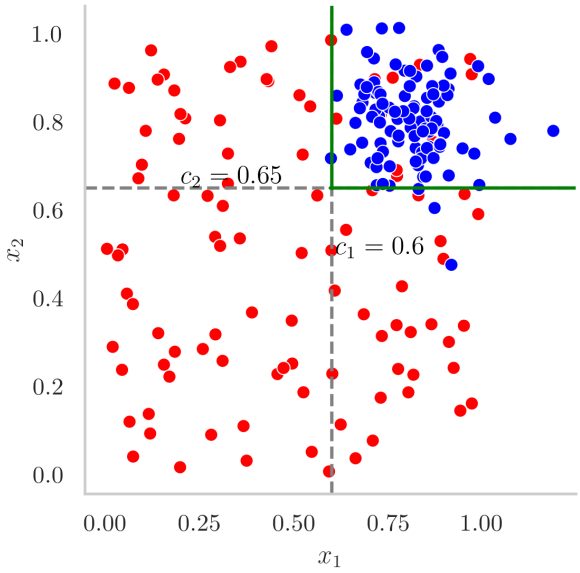


Figure 4.1: A binary classification problem with two features. A decision tree applies two successive cuts c_1 and c_2 to isolate most of the blue class from the red.

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

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

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 depends 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 is often 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 operator defined as $\mathbb{I}(x) = 1$ 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 as

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

  with *boldsymbol*π a vector $(\pi_1, \pi_2, ..., \pi_C)$ of the class-condition probabilities.

- *Information Gain:* an equivalent formulation to the entropy, measuring the gain in information from the change in entropy induced by adding a selection on feature X_j to the current selection

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

- **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.

Algorithm 1 Recursive Procedure to Train a Decision Tree [87].

```
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
```

DTs can overfit a dataset, when the model tunes itself to specific features of the training set that do not generalise. Regularisation serves as an important step to contain this mostly undesirable behaviour. For trees, a common procedure to avoid overtraining is to interrupt the growth of the tree when it is no longer worth doing so or to *prune* the tree by removing 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 random subsets of the data and aggregate the results into a single prediction. For example, taking as regressive output the average over several N_l base learners

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

over base learner predictions $y_i(x)$ for a common input x. For classification, the predicted class can be decided with majority voting. This statistical technique of combining different predictors is referred to as *bagging* or *ensembling*. The different predictors can be built on subsets of the input features and training datapoints to further decorrelate them, thereby forming a *random forest*.

4.2.2 Boosted Decision Trees

A popular extension to the simple DT approach is to introduce the concept of *boosting*, leading to a technique referred to as *Boosted Decision Trees (BDT)* or *Multivariate Analysis (MVA)*. Boosting is a greedy algorithm leveraging a weak learner and applying it sequentially to weighted versions of the data, with a larger weight given to misclassified datapoints. This method is hugely popular in data science, having earned the title *"best off-the-shelf classifier in the world"* [90]. Two particularly useful approaches are adaptive boosting (AdaBoost) [91] and gradient boosting

[92], 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 $y_i \in \mathbb{R}^d$ 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 combinations of the weak learners f_i with weights $\alpha_i \in \mathbb{R}$. 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 $L(y_i, F_m(x_i))$. AdaBoost relies on the exponential loss

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

that the new weak learner $\alpha_m f_m$ added at step m has to minimise. The typical case for AdaBoost is binary classification with $y_i \in \{-1, 1\}$, but the algorithm is generalisable to multi-class [87]. 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 $w_{i,m}$ [87], where

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

For the first step $m = 1$, the weights are initialised to $1/N$. They 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 input in 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

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

that combines the new weak learners f_m with optimal weight α_m to the current strong learner F_{m-1}. The AdaBoost algorithm is summarised in Algorithm 2.

Algorithm 2 Adaboost for Binary Classification with Exponential Loss [87]

Initialise weights: $w_{i,1} = 1/N$, 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 renormalise $\sum_i w_{i,m+1} = 1$.
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, is not restricted to a specific loss function. The objective to minimise is the empirical risk, the expected value of the loss function L on the training set

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

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 is the gradient of each datapoint 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, an undesirable feature that is remedied by using a weak learner to approximate the negative gradients. 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 region indxed by $j = 1, ..., J_m$:

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

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

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

with α_m selected by minimising the empirical risk

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

The full algorithm for gradient boosting is presented in Algorithm 3, where the update rule is added a *learning rate* hyperparameter lr to introduce regularisation and reduce the risk of overfitting. By keeping $0 < lr \leq 1$, 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 a higher computational complexity. Further regularisation techniques are bootstrap aggregation - training each weak learner on a random subset of the data -, limiting the number of leaves, penalising models of larger complexity, and pruning branches that do not sufficiently reduce the loss.

Algorithm 3 Gradient Boosting [87]

Initialise $f_0(x) = \arg\min_\alpha \sum_{i=1}^{N} L(y_i, \alpha)$
for $m = 1$ to M **do**
  Compute the gradient residuals 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 $f_m(x) = f_{m-1}(x) + lr \times \alpha_m h_m(x)$
end for
return $f(x) = f_M(x)$

Boosted Decision Treess (BDTs) resist better to overtraining thanks to the regularisation effect. 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. 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 [93, 94].

Pros:

- *High Accuracy:* BDTs easily achieve high accuracy in many tasks.
- *Ease of Deployment:* BDTs typically perform reasonably well out-of-the-box and are easy to train with few hyperparameters to optimise.
- *Adaptability to Different Distributions:* boosting algorithms adapt to different types of data distributions and capture non-linear relationships.
- *Ensemble Learning:* combining multiple weak learners to create a strong learner improves the overall model performance.

- *Robustness to Overfitting:* boosting mitigates overfitting, enhancing the generalisation of the model to unseen data.

Cons:

- *Sensitivity to Noisy Data:* BDTs are sensitive to noisy data and outliers.
- *Computational Complexity:* Training multiple weak learners sequentially can be computationally expensive, especially on large datasets.
- *Parameter Tuning:* BDTs require some fine-tuning of the hyperparameters for optimal performance.
- *Black Box Nature:* The ensemble nature of BDTs make them somewhat of a black box, sacrificing the interpretability of DTs 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 [95], is the logical unit at the core of modern deep learning. Notably, the Multilayer Perceptron (MLP) or Deep Neural Network (DNN) are obtained by stacking layers of artificial neurons. Inspired by biological principles, the perceptron, 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 are modified during training to correct the output of the perceptron. Artificial neurons are a direct generalisation on this principle, with the output no longer being thresholded but applied a chosen function f after being added a learnable bias term b.

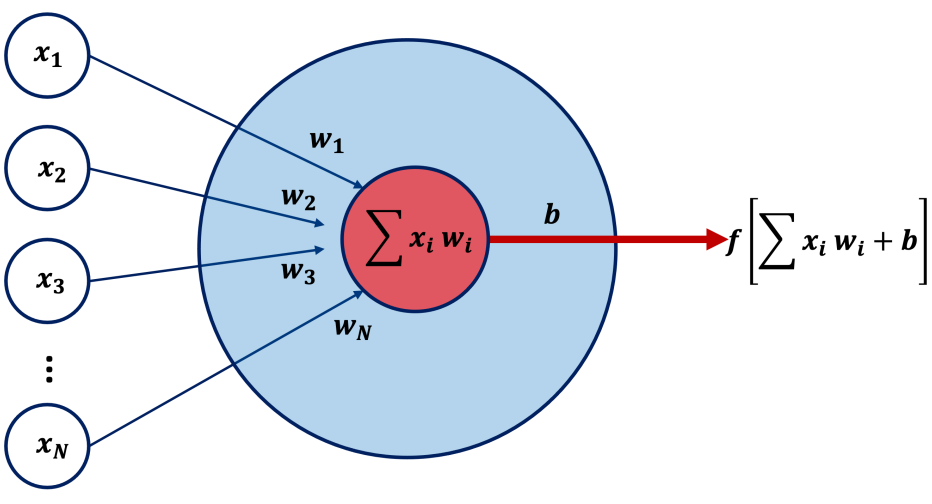


Figure 4.2: An artificial neuron: the inputs x_i $(i = 1, ..., N)$ are multiplied by learnable weights w_i, summed and added a learnable bias b and passed to an activation function f.

The interest in artificial neurons stems from a significant theoretical result: stacks of artificial neurons are *universal function approximator* [96, 97], as shown in the next section. This theoretical result is built on a mathematically advantageous function chosen for f: the sigmoid σ,

illustrated in Figure 4.3 and defined as

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

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 probability distribution mappings. An essential mathematical property of the sigmoid, particularly relevant for DL, is the ease to compute its derivative:

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

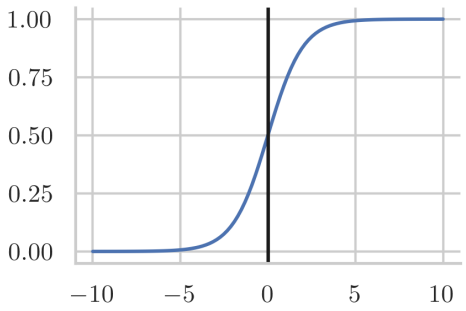


Figure 4.3: The sigmoid function σ.

The power of artificial neurons stems from the ability to be efficiently combined them into ordered structure with powerful representation powers. 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 leads to a mathematical model that can approximate any continuous function.

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 only Neural Network (NN) - is constructed by stacking layers of artificial neurons as shown in Figure 4.4. Each neuron in a layer receives as input the output of the neurons of the previous layer, and connects to the neurons 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 flowing from the input x to the output y. Mathematically, a single layers at depth i with m units given as input the previous n-neurons layer at depth $i-1$ computes an affine transformation

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

where $W^i \in \mathbb{R}^{m \times n}$ is the matrix of learnable weight of layer i - one row per unit of layer i, one column per unit of layer $i-1$, $b_i \in \mathbb{R}^m$ the vector of learnable biases, f^i is the activation function of layer i, and $a^{i-1} \in \mathbb{R}^n$ are the n activated outputs of the previous layer. The activations can differ for the units of the same layer but are often kept similar to vectorise the mathematical operations.

Neural networks implement a recursive system of computation based on Equation 4.8. A powerful theoretical property of neural networks is their capacity to be *Universal Function Ap-*

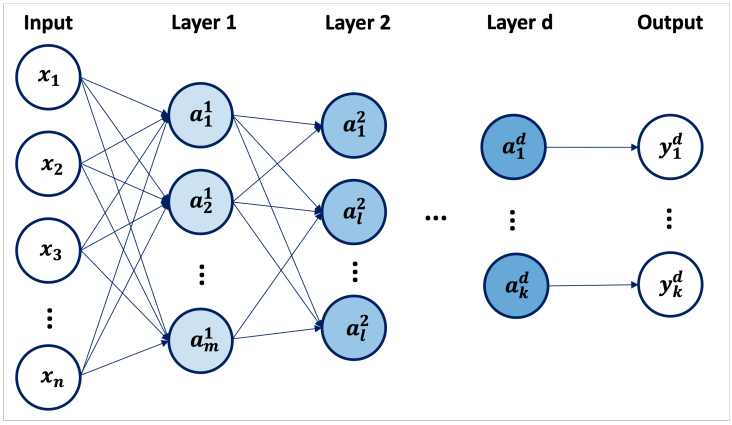


Figure 4.4: A deep neural network with d layers of width m, l ..., k. Each artificial neuron, represented by a ball of darkening blue along the depth, computes an affine transformation of the input of the layer followed by an activation function. The input of the DNN is x and the output is y.

proximators. This family of theorems, defined for various types of activations and network, demonstrates 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 [96, 97]:

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

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

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 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 would indeed collapse into a linear regression model. Such a non-linear 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 deep learning [98]. Som e common activation functions, shown in Figure 4.5, are:

- The sigmoid function of Equation 4.7.
- The hyperbolic tangent function $\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$.
- The Rectified Linear Units (ReLU) function[99]

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

The non-linearity here is strictly between positive and negative inputs, making this activation function the simplest that can be leveraged in DNN. A generalisation of ReLU called leakyReLU $= \max(\alpha x, x)$ introduces a linear function in the negative range, with $\alpha \in [0, 1[$.

- The Exponential Linear Unit (ELU) function modifies the leakyReLU in the negative domain while keeping the saturation property as

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

  with the hyperparameter $\alpha > 0$.

- The softmax function which, for an $x \in \mathbb{R}^n$, return a vector $\text{softmax}(x) = [..., \frac{e^{x_i}}{Z}, ...]$ with $Z = \sum_i e^{x_i}$. For a 2-dimensional x, the softmax is equivalent to the sigmoid. 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.

The sigmoid is no longer the choice of reference, due to its tendency to quickly saturate - meaning its gradient for large positive or negative values *vanishes* by tending to 0. The hyperbolic tangent offers larger gradients thanks to its [-1, 1] range with steeper curvature, making it the activation of choice for autoregressive architecture such as the Recurrent Neural Network (RNN). The ReLU function is the most widely used activation function as its derivative is trivial 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 some neurons can be stuck in *off*-mode. For this reason, it is important to correctly initialise the weights of a DNN. A workaround for this limitation is the introduction of leakage for the negative inputs with leakyReLU or Elu activations.

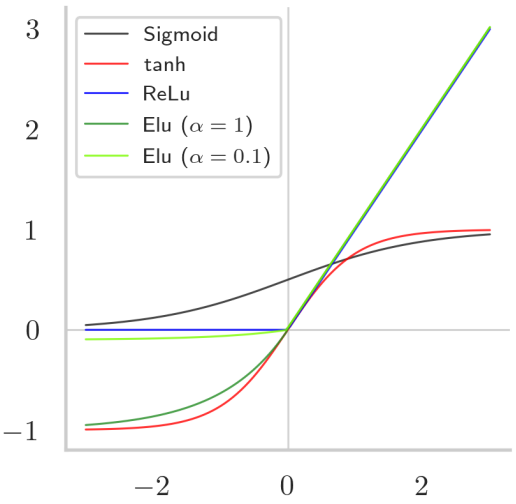


Figure 4.5: The most common non-linear activations used in deep learning.

While the Universal Function Approximator theorem is a powerful endorsement of neural netowrks, it does not state *how* to derive the best network. The task of choosing the right architecture depthwise and widthwise and the correct weights and biases is approximated by a learning strategy updating the parameters to reduce the error between the empirical risk. What sets apart neural networks from other Universal Function Approximators is the simplicity of the procedure to update their weights: with a suitable computational structure and activation function choice, the NNs are *differentiable*. Gradients can therefore be computed from a loss

function measuring the quality of the output $\hat{y}$ and *backpropagated* across the neurons to update the weights and biases. 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 [100] and TensorFlow [101], 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 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 implemented by neural networks requires a large dataset to correctly assign values to the parameters without suffering from overtraining. The *backpropagation* algorithm of Algorithm 4 circumvent these problems [102].

Algorithm 4 Backpropagation Algorithm

function UPDATE(x, y, N, $\mathcal{L}$)
  Forward step: propagate input x through network N to get prediction $\hat{y}$
  Loss: compute the loss or reward of N as $\mathcal{L}(y, \hat{y})$
  while $\exists$ a layer without local gradients **do**
    Take the right-most layer required a gradient
    Take the gradient at the input of the subsequent layer
    With the chain rule, propagate the gradient of the next layer to 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

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

where θ encapsulates all parameters of the model, with a per datapoint x_i loss of l_i. The backpropagation algorithm starts with a forward pass through the network before computing the gradient of each layer by starting from the output and applying the chain rule of calculus. Once all the local gradients are available, the parameters are updated to reduce the loss by taking a step in the direction opposite the gradients, giving the largest reduction in loss. For example, for a specific parameter w_{ij} at training step $t+1$:

$$w_{ij}^{T=t+1} \leftarrow w_{ij}^{T=t} - lr \times \text{grad}\left[w_{ij}^{T=t}\right], \tag{4.12}$$

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 ($\to 0$) or exploding ($\to \infty$) gradients. This requires some care in the architecture choice and can motivate the use of a specific activation function over another. Concerning the loss function $\mathcal{L}$, some typical choices are:

- The cross-entropy loss function is based on the definition of entropy in Equation 4.2. It is

commonly used to assign probabilities in a classification problem 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 Squarred Error (MSE) and Mean Absolute Error (MAE)

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

To regularise the model, *regularisers* are added to the loss function to restrict the size of the weights, either with the 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 results in sparse networks with unnecessary weights set to 0. The amount of regularisation is controlled by the hyperparameter λ. Further regularisation can be obtained by randomly dropping out some connexions in the network during training, a technique called *dropout* and controlled by the dropout probability p to include a neuron.

Pros:

- *Universal Function Approximators:* neural networks are universal function approximators. With enough parameters, they can approximate any continuous function to arbitrary precision.
- *Flexible Architecture:* the architecture of NNs is flexible, allowing for extensive customisation of the number of layers, number of neurons, and the activation function. This flexibility makes them suitable for various tasks.
- *Feature Learning:* neural networks automatically learn hierarchical representations of features from the input data. They can therefore capture complex non-linear features.
- *Availability of Optimisation Techniques:* optimisation techniques built on gradient descent are well-suited for training neural networks.

Cons:

- *Limited Modeling of Sequential and Geometrical Data:* these networks are not naturally designed for handling sequential data and temporal dependencies nor images.

- *Fixed Input Size:* they require a fixed input size. While techniques like padding or resizing can be employed, these might not be suitable for handling inputs of varying lengths.
- *Lack of Memory:* neural networks do not have an inherent memory mechanism.
- *Vanishing and Exploding Gradients:* training DNNs can be challenging due to vanishing and exploding gradients.
- *Need for Sufficient Labelled Data:* DNNs require a large amount of data for effective training. In domains where data is scarce, the performance may be limited.

This section introduces deep neural networks, the fundamental architecture constituted of artificial neurons stacked into layers. There are many refinements to this base architecture, and the most important ones are next explored.

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 NLP. The main adaption 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 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 dynamically adapted to variable length of the input. 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

$$y_t = W(h_t) = W(V(x_t) + U(h_{t-1})), \tag{4.13}$$

where U, V, and W are three different DNN mappings 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 exploding or vanishing gradients, the *tanh* function is often used as activation in RNN thanks to its smooth distribution and limitation to the range $[-1, 1]$. As the network relies on repetitive multiplications of numbers in the range $[-1, 1]$, the effect of much earlier timestep ($h_{t<<T}$) is lost when processing later input at T. This process is referred to as *memory loss*. This undesired property is remedied with architectural modifications to RNN that improve their operational memory, such as the Long-Short Term Memory (LSTM).

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 [103]. 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

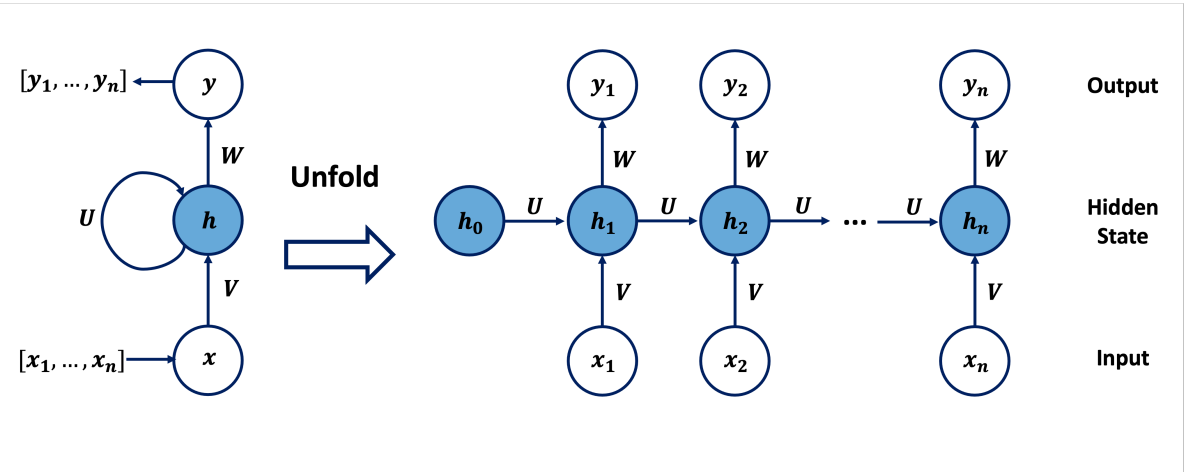


Figure 4.6: A recurrent neural network, using 3 feed-forward neural networks (DNN) U, V, W, to map the input sequence $x = [x_1, x_2, ..., x_N]$ to the output $y = [y_1, y_2, ..., y_N]$ using the internal hidden state h^t evolving for each timestep t. h_0 would typically be obtained by a mapping of the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

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.

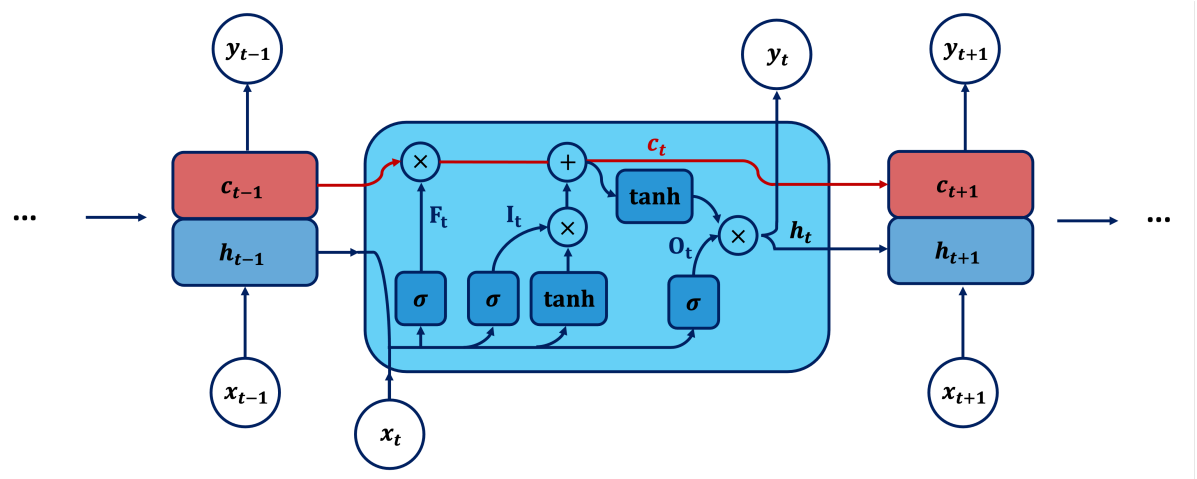


Figure 4.7: An LSTM cell. Arrows and lines that merge imply concatenation of the inputs, the $\times$, $+$, and tanh are element-wise operations and the σ are different layered transformations (1-layer feed-forward network). F_t is the forget gate of the memory cell c, I_t the input gate, and O_t the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

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. 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 flexibility of Universal Turing Machine [104]. A significant drawback however is the impossibility to fully parallelise the processing of sequence due to the strict ordering, making RNN expensive models

to train. 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.
- *Limited Short-Term Memory:* Traditional RNNs struggle to capture long-range dependencies due to their limited short-term memory.
- *Complexity:* While the LSTM architecture can mitigate the two points above, the cost is a more complex architecture that is harder to train.
- *Interpretability:* The internal workings of a RNN is challenging to interpret.

4.2.6 Convolutional Neural Networks

Convolutional Neural Networks (CNN) [89, 105] 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 consists of convolutional layers - implementing the fundamental convolution operation-, pooling layers, and DNNs. This architecture, presented in Figure 4.8, 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.

The CNN leverage 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 is parallelisable for an image with multiple channels. For classification and regression, a CNN typically stacks 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 $\mathbb{R}^{x \times y}$.

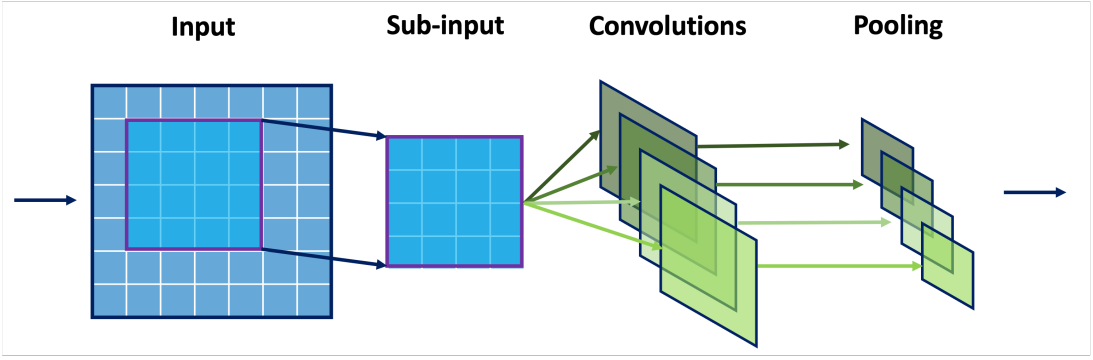


Figure 4.8: A layer of a convolutional neural network, implementing a convolution with 4 kernels followed by a pooling operation. This design can be stacked to create deep architecture, and combined with a feed-forward neural network, after flattening the output at some depth, before reaching the final loss function layer.

CNN-based models, such as AlexNet [105] and ResNet [106], 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. For a given image:

- A DNN given the flatten the image requires $x \times y$ neurons.
- A CNN with k kernels of size $\alpha \times \beta$ requires $k \times \alpha \times \beta$ artificial neurons.

For example, for an image of size 100 $\times$ 100, a DNN requires 10,000 neurons while a CNN can process the image with only 25 units if a single kernel of size 5 $\times$ 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.
- *Spatial Hierarchies:* convolutional and pooling layers enable the model to capture spatial hierarchies in the input data while respecting the properties of images.
- *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 is computationally intensive.
- *Large Datasets:* CNN often require large datasets for effective training.
- *Interpretability:* The internal workings of CNNs are challenging to interpret.

4.2.7 Graph Neural Networks

Recentely, Graph Neural Networks (GNNs) have garned attention for their ability to model and analyse complex relationships within graph-structured data [107]. 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 both local and global structures defined by the graph. 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 directional weighted edges [108]. A particular feature arising from this property is that graphs are permutation equivariant: the order of nodes can be rearranged without impact.

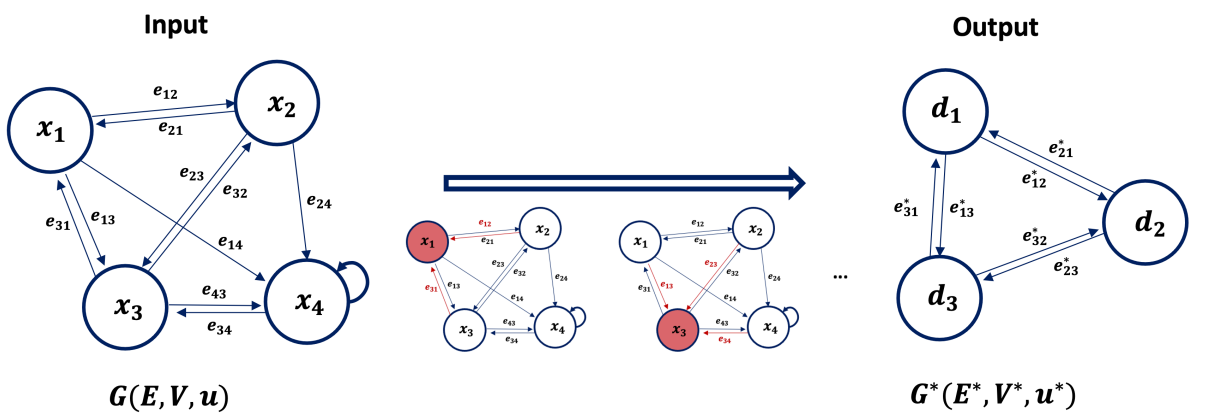


Figure 4.9: A graph neural network update on a directed graph $G(V, E, u)$ with a global representation u, four initial nodes $\in V$ and edges $e_{ij} \in E$ connecting nodes $i \to j$. By analysing the neighbours of each node, the graph is updated to a new graph $G^*(V^*, E^*, u^*)$.

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 4.9. Different architectures implement different update processes for the graph, with popular GNN architectures being Graph Convolutional Networks (GCNs) [109] and Graph Attention Network [110]. 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 $u \in \mathbb{R}^u$.

The most general graph update algorithm to describe an update stage of a full GNN block

Algorithm 5 Steps of Computation in a Full Graph Network Block [108]

function GraphNetworkUpdate(E, V, u)
  for $k \in \{1 \dots N^e\}$ **do**
    $e_k^* \leftarrow \phi^e(e_k, v_{r_k}, v_{s_k}, u)$
  end for
  for $i \in \{1 \dots N^n\}$ **do**
    Let $E_i^* = \{(e_k^*, r_k, s_k)\}$ for $k = 1 : N^e$ where $r_k = i$
    $\overline{e_i^*} \leftarrow \rho^{e \to v}(E_i^*)$
    $v_i^* \leftarrow \phi^v(\overline{e_i^*}, v_i, u)$
  end for
  Let $V^* = \{v^*\}_{i=1}^{N^v}$
  Let $E^* = \{(e_k^*, r_k, s_k)\}_{k=1}^{N^e}$
  $\overline{e^*} \leftarrow \rho_{e \to u}(E^*)$
  $\overline{v^*} \leftarrow \rho_{v \to u}(V^*)$
  $u^* \leftarrow \phi_u(\overline{e^*}, \overline{v^*}, u)$
  return (E^*, V^*, u^*)
end function

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. The aggregation function must 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 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 $\overline{e_i^*}$ connected 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 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 4.10. A notable example for this thesis is the Deep Set architecture [111], designed to run specifically on

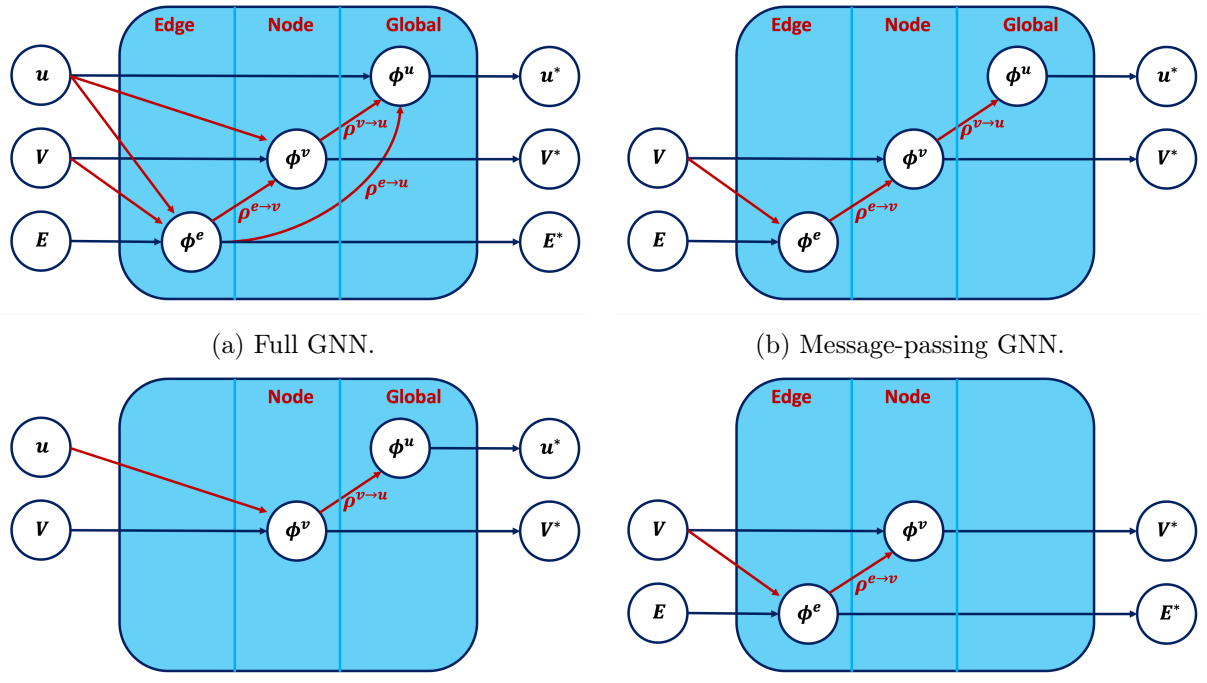
sets where the ordering does not matter. It essentially simplifies the graph network by dropping altogether the edges and considering instead a fully connected graph with static edges, with an update of the global representation only based on pooled node information:

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

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

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

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 [112].



(a) Full GNN.

(b) Message-passing GNN.

(c) Deep Set.

(d) Non-local NN.

Figure 4.10: Different types of GNN update rules, defining different GNN architectures [108].

A different approach introduced in [113] 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 a weighted version of the inputs, with weights standing for the degree of attention to be given to the different parts. Attention is not restricted to GNN but is easily encapsulated in its formalism. As will be shown in the next section, the transformer is a special case of a non-local NN. In this section, only the Graph Attention Network (GAT) is introduced for the sake of conciseness [110]. GATs implement 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 the nodes $v_i, v_j \in \mathbb{R}^d$ are connected 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 operator $[,]$ stands for matrix concatenation. These scores are then combined for each node i over its neighbours $\{j\}$ to give α_{ij} attention scores

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

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

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

where the sum over j is taken over neighbouring nodes of i, σ is an 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 abritrary relationships between entities.
- *State-of-the-Art Performance:* GNNs have achieved state-of-the-art results in various graph-related tasks.

Cons:

- *Computational Complexity:* training GNNs can be computationally expensive.
- *Limited Global Context:* some GNN architectures may struggle to capture long-range dependencies in graphs, limiting their ability to consider global context.
- *Interpretability:* GNNs are not directly interpretable.

4.2.8 The Rise of the Transformers

The transformer architecture, introduced in 2017 [114], has become a foundational and ubiquitous design across machine learning. It has significantly impacted the field, enabling the development of state-of-the-art models such as BERT [115] and GPT [116]. 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) [117].

The transformer architecture is based on the mechanism of self-attention introduced in the previous section. It removes the sequential processing of RNN by favouring a fully parallelisable approach, allowing for efficient computation on dedicated hardware. The key components in the transformer are the self-attention mechanism and position-wise feed-forward networks. Self-attention allows the model to weigh the importance of different words or tokens in a sequence for each token. This mechanism enables the model to capture long-range dependencies in the input data without the added complexity of LSTM. Strictly speaking, the input of a transformer is an

unordered sequence that has no strict order. For NLP, the ordering is built into the model with position-wise embedding, giving the model a handle to determine the index of the token in the sequence. In computer vision, 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 classical transformer.

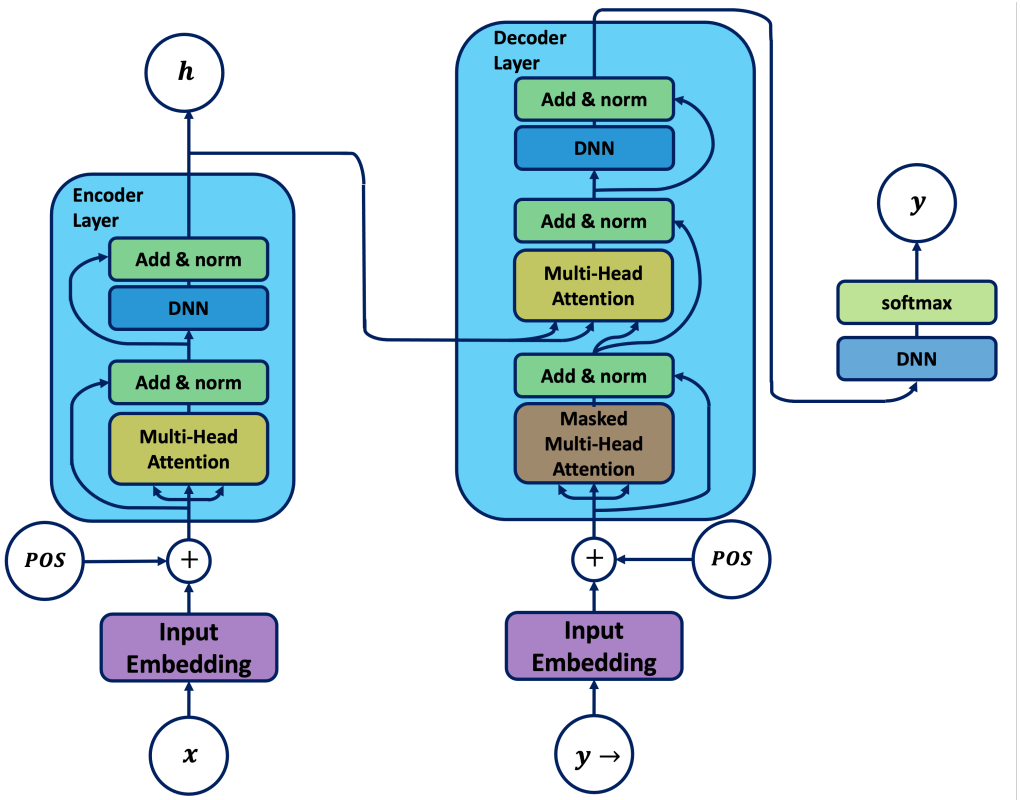


Figure 4.11: The full transformer architecture, combining an encoder and a decoder each made of an arbitrary number of layers. The input x is first embedded with a dedicated mapping which can, when order matters, be supplemented with positional embedding. The encoder generates an internal representation 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.

The general transformer architecture, presented in Figure 4.11, consists of an encoder and a decoder. 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 feed-forward 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 [118]. 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 4.14 and schematised in Figure 4.12a, 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 $\mathbb{R}^{d_q \times d_v}$ of values reweighted by the attention scores. This combines d_q different queries of the d_k keys mapping to d_v values. Therefore, Equation 4.14 implements for each query a weighted sum of the values, based on a compatibility function established by comparing the queries and keys:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{Q^T K}{\sqrt{d_k}}\right) V, \tag{4.14}$$

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 subsequent 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-consuming than using a DNN mapping to compute the attention - a technique referred to as *additive attention* [119]. As shown in Figure 4.12b, 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:

$$Q_i = QW_i^Q, \quad K_i = KW_i^K, \quad V_i = VW_i^V,$$

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

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

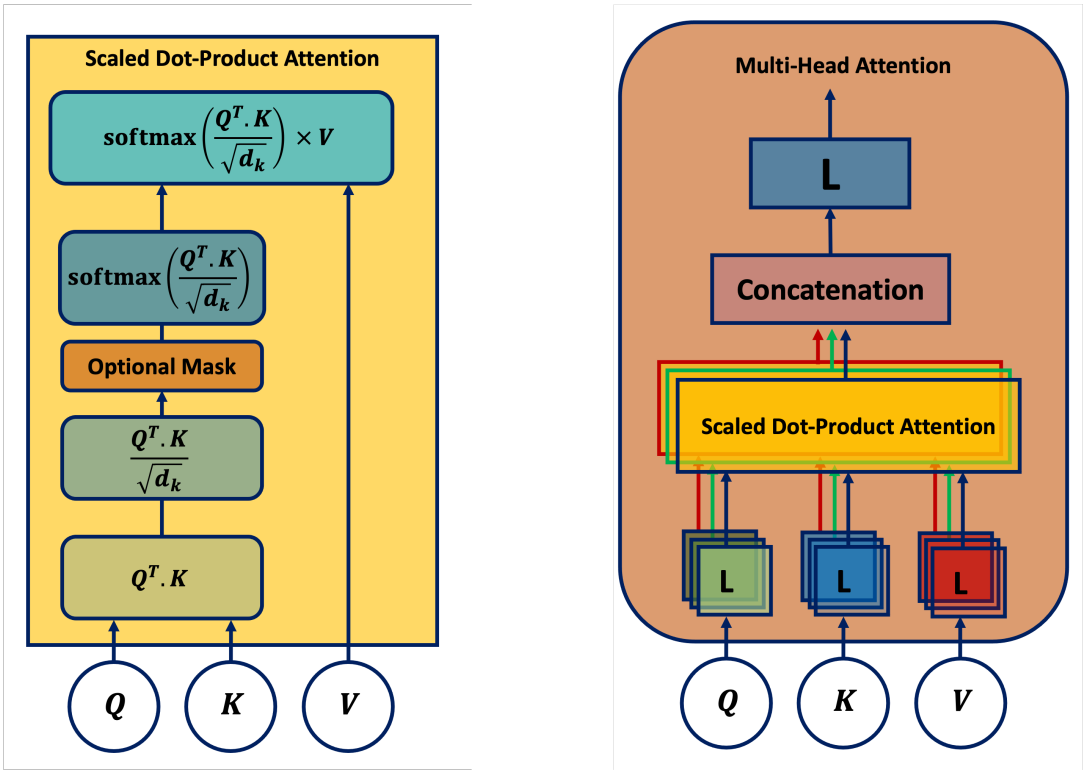
$$\text{Multi-Head Attention}(Q, K, V) = [H_1, ..., H_h]\, W^o.$$

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:

$$Q = xW_i^Q, \quad K = xW_i^K, \quad V = xW_i^V.$$

Pros:

- *Versatility:* the transformer architecture has been successfully applied to various NLP and computer vision tasks.
- *Parallelisation:* the architecture is efficiently parallelised, speeding up the computations.
- *Capturing Dependencies:* the self-attention mechanism enables the model to capture complex long-range dependencies.
- *Stability:* transformers benefit from inbuilt regularisation effects due to residual connexions



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

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

Figure 4.12: 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.

and normalisation layers. Models based on this structure can therefore simply scale in complexity and resist well to overtraining.

Cons:

- *Computational Complexity:* large transformer are computationally expensive to train.
- *Interpretability:* the attention mechanism is challenging to interpret.
- *Data Dependency:* transformer models require large amounts of data for effective training, limiting their application in domains with limited statistics.

4.3 Training and Optimising Deep Learning Models

Training and optimising neural network involve a combination of selecting appropriate architectures, fine-tuning the hyperparameters that are not learnt by backpropagation, and employing acceleration techniques to improve efficiency and convergence. In this section, key aspects of the training process are explored.

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 gradient descent rule of Equation 4.12, and each method has different advantages. The two main approaches are:

- Stochastic Gradient Descent (SGD): instead of deriving the gradients on the whole dataset (full-batch), the expectated gradient 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 the gradient of the parameter w computed for a single data point. A common observation is that for sub-batches - from now on referred to as just *batches* - of sufficient size b, the statistical estimator of the gradient based on the batch is unbiased. This greatly reduces the time needed 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 per-batch gradient has a larger variance than a full-batch, making it harder for the model to overtrain but slowing the convergence to an optimum.

- Adam is an algorithm published in 2014 leveraging an adaptive moment estimation approach as well as batch processing [120]. 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 steeper slope that can be quickly traversed so that any slowing-down due to a changing curvature of the objective function landscape can be mitigated from 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

$$m_t \leftarrow \beta_1 m_w^{t-1} + (1-\beta_1)\nabla w_s^t,$$

  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 $(\nabla w_s^t)^2$. 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

$$v_t \leftarrow \beta_2 v_w^{t-1} + (1-\beta_2)(\nabla w_s^t)^2,$$

  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

$$\hat{m}_w^t \leftarrow \frac{m_w^t}{1-\beta_1} \quad \text{and} \quad \hat{v}_w^t \leftarrow \frac{v_w^t}{1-\beta_2}.$$

The two contributions are then combined into a single gradient descent step as

$$w^t \leftarrow w^{t-1} - lr \times \frac{\hat{m}_w^t}{\sqrt{\hat{v}_w^t} + \epsilon}, \tag{4.15}$$

where a very small ϵ is added for numerical stability.

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*, modifying the parameter throughout training to resolve different parts of the loss function landscape. Initially, a relatively large 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 reduces the learning rate so that smaller steps can be taken later to approach the chosen optimum. At the beginning, the rate is typically not set to its maximum to move the gradient updates to a valley of interest. An equivalent choice is to modify the batch size while keeping the lr fixed [121]. 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 an effective way to improve the training performance. Some methods, such as Adam, have other specific hyperparameters that should be optimised.

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 and L2 (L1) regularisations that 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, 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:

- *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, transformer,

...), the choice of activation functions (ReLU, tanh, ...), and the number of layers, nodes, and connexions between the units.

- *Hyperparameter Tuning:* optimising the hyperparameters - parameters that are not optimised through backpropagation of the loss - is essential for achieving the best possible performance. Key hyperparameters include optimiser-related parameters such as the learning rate, the batch size, regularisation parameters, and initialisation of the weights and biases.

The optimisation process for both hyperparameter tuning and architecture selection is expensive: the model must be trained with different combinations of hyperparameters or 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 trial and error, with the ML literature offering little insight 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:* can significantly speed up the training process. Instead of preparing a single batch, multiple batches can be loaded by different processing units in parallel to avoid this bottleneck.

- *Early Stopping:* prevents overfitting and save computation time by interrupting the training when the performance saturates.

- *Hardware Accelerators:* specialised hardware accelerators can accelerate the training. The Graphics Processing Unit (GPU) is specially designed to perform matrix operations in parallel, and therefore well suited for ML. Utilising GPUs for training and inference can give substantial speedups compared to CPU-based computing.

- *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 pre-train the foundational model giving the latter a headstart. Fine-tuning these models on specific tasks or connecting additional modules can significantly reduce the required training time and data compared to training 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 modified for the new task. Such large foundational models are already available in NLP (e.g., the 7 billion parameters transformer Mistral 7B [122]) and computer vision (e.g., the 500 million parameters Florence-2 [123]).

Training and optimising 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 depends on available resources and trade-offs between training time, model performance, and interpretability.

CHAPTER 5

Flavour Tagging

This chapter is focused an essential task for the ATLAS experiment: identifying particles passing 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 quark flavour should be associated with an observed signal is called flavour tagging. Free quarks and gluons hadronise as per the rules of QCD, forming many particles that can themselves further decay. Such a dynamic results in an ensemble of particles radiated within a cone centred around the initial coloured 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 hyperparameter 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 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 colour-neutral 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 chosen in this chapter is to label the jets based on the species of hadrons found: 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 analyses (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 scattering 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 [124]. 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 $\Lambda_b^0 = udb$ [125]. For b-quarks, the hadronisation process is hard and most ($\sim 75\%$) of the quark momentum is passed to the b-hadron [124]. 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 CKM matrix $|V_{bc}|$ 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 [26]. 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 increases with rising jet p_T, and at $p_T \sim 250$ GeV even overpasses the first detector layer of the IBL located at a radius of 33 mm from the centre of ATLAS, as shown in Figure 5.2. The location of the hadron decay, called the *SV*, can often be reconstructed with the ATLAS detector [126]. Some important variables describing the decay of hadrons are the IPs d_0 and z_0 of the tracks left by charged particles emanating from the SV. As shown in Figure 5.1, d_0 and z_0 are the transversal and longitudinal distances from the primary vertex to the perigee 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 to 4 or 5 charged stable particles [125]. Another characteristic of b-jets is the likely presence of leptons in the jet cone, as $\sim 40\%$ of the b and c-hadrons decays are semi-leptonic [26].

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

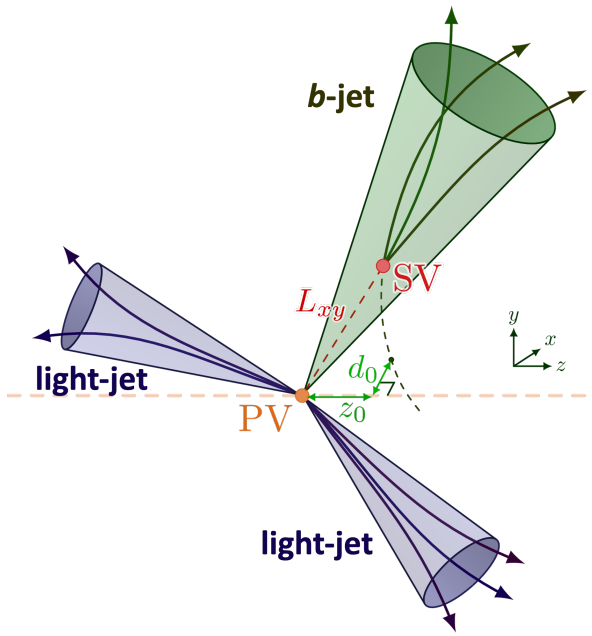


Figure 5.1: Representation of a b-jet [127].

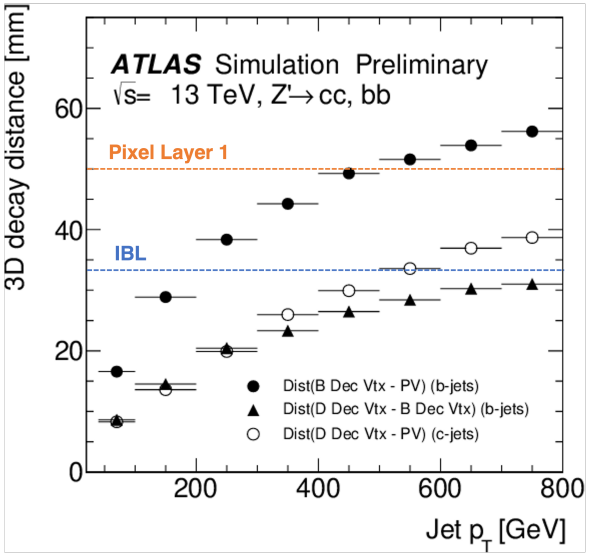


Figure 5.2: b-hadron and c-hadron (labelled D-hadron) decay radius as a function of jet p_T reconstructed for two b-jets or c-jets in Z' events, adapted from [128]. The IBL and first Pixel layer transverse distances are indicated as blue and orange dashed lines.

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 [26], is smaller than for b-hadrons and is harder to resolve with the currently deployed tracker, as highlighted in Figure 5.2. 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 to be easily mistaken for c-jets, as these leptons can hadronically decay into a similar number of particles and leaving a jet signature in the detector. Tagging c-jets is however becoming particularly important due to the focus on the challenging $H \to c\bar{c}$ search [129–131], 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 deep learning methods, given their vastly superior effectiveness. 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 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 [132]. 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) [133], leveraging the Recurrent Neural Network Impact Parameter (RNNIP) sub-tagger [134], and a Deep Set approaches for DL1 with DIPS (DL1d), leveraging the Deep Impact Parameter Set (DIPS) sub-tagger [135]. This last tagger is,

at the moment of writing this thesis, the state-of-the-art calibrated tagger in the ATLAS software [66]. Algorithms from this family were mainly developed for the end of Run 2 of the ATLAS experiment [136], with DL1d developed just before the start of Run 3.

2. The GN family of taggers are built on more advanced deep learning 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) [4, 5, 137]. 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, to be used in Run 3 analyses [66].

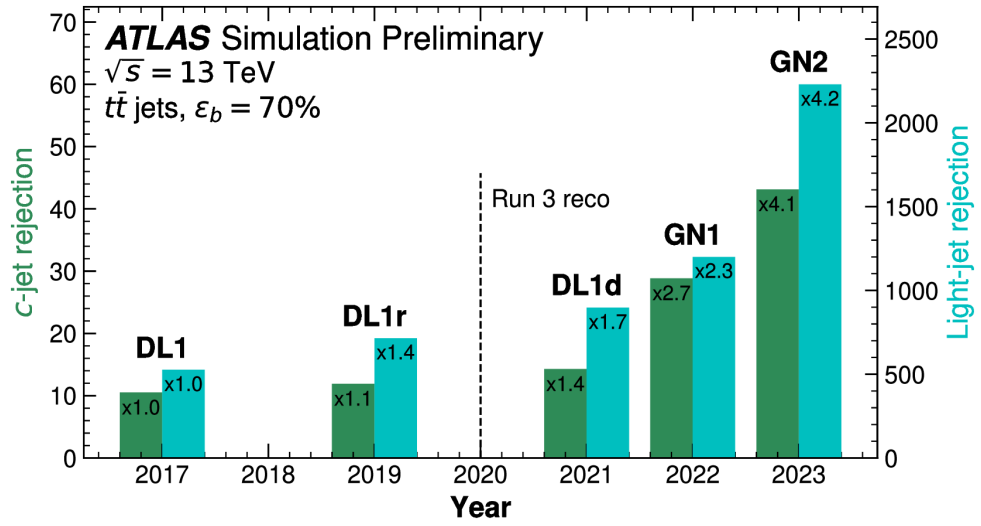


Figure 5.3: Comparison of the performance of flavour tagging models introduced through the years [5]. Light and c-jet rejections (inverse of the mistag efficiency) are plotted for different taggers at a fixed b-jet tagging efficiency of 70% on a $t\bar{t}$ evaluation set. The multiplicative factors in 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 5.3, showing a remarkable and consistent increase in light- and c-jet rejections at a fixed b-tagging efficiency of 70% evaluated on 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 [66].

5.1.3 Datasets

ATLAS analyses scan a p_T spectrum that covers a wide range of energies due to the fractional momentum of the partons. 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 Beyond the Standard Model (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 (out-of-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 [138–141]. The hard scattering matrix element is computed for proton-proton collision with the NNPDF3.0NNLO set of parton distribution functions (PDF) [142], and the simulated hard scattering events are interfaced with PYTHIA 8.230 [143] using the A14 parameter tune [144] and the NNPDF2.3LO PDFs for the parton shower, hadronisation, and underlying event simulations [145]. Studies in Refs. [146, 147] 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 [148].

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 [149] built on GEANT4 [150]. 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 operating point for jets with $p_T < 60$ GeV and $|\eta| < 2.4$ [84]. Tracks are associated with a jet 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 an ensemble of 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 deep learning, with some instead directly implementing physics-inspired algorithms. They consist of [136, 151]:

- *IP Likelihood Discriminants*: IP2D and IP3D (summarised IPxD) are likelihood-ratio templates in 2D and 3D to assign flavour-discriminating weights based on the transversal and

global impact parameters significances[1] S_{d_0} (35 bins) and S_{z_0} (20 bins) of the tracks, and 14 bins of tracks categorisation in addition for IP3D [133]. For the three main 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 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) [132].

- *Track Collection Analyser*: either with RNNIP [134] or DIPS [135]. These are deep learning 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 fully described later in this chapter.

- *Secondary Vertexer 1 (SV1)*: combining a secondary vertex finder and a tagger to offer flavour discrimination information [136]. The former, based on the VKALVRT vertex reconstruction package [74], 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 $\to B \to D$, with the assumption that the vertices of the weakly decaying B- and D-hadrons tend to align with the PV [128, 136].

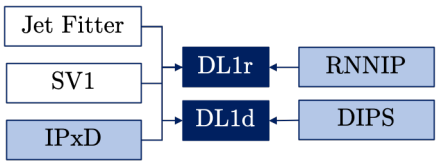


Figure 5.4: The algorithms for flavour tagging in the DL1 family. High-level taggers are in dark blue, track-based taggers in light blue, and vertex taggers in white.

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 made of 44-45 features. This high-level tagger outputs three scores p_X for the analysed jet corresponding to the b-, c-, or light-flavour (indicated with the letter u) probabilities such that $p_b + p_c + p_u = 1$. A b-tagging discriminant D_b is then derived by computing a scaled log-likelihood ratio

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

where f^b_c is the c-fraction, a parameter that can be modified to tweak the relative importance of the rejected flavours. The analogous c-tagging discriminant D_c relying on the f^c_b b-fraction parameter is

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

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 (mistag) efficiencies.

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

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 mistag 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 tagger. 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, ordered by the absolute transverse IP significance $|S_{d_0}|$, to extract tagging information from correlations between tracks [134]. The vector of track features, described 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 [151]. RNNIP outputs a probability p_X for the jet to belong to flavour $X \in [b, c, \text{light}, \tau]$. The architecture of RNNIP is an RNN-based model leveraging an LSTM core, as depicted in Figure 5.5. 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 a 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, delivering superior performances to the IP-based approaches with only ~ 40 % of the parameters - 11,636 trainable parameters for RNNIP [151].

5.2.2 DIPS

The DIPS tagger based on the Deep Set architecture [111], as depicted in Figure 5.6, is an alternative to RNNIP to model the correlations between an arbitrary number of tracks [135]. As introduced in Section 4.2.7, such a model is composed of two fully-connected feed-forward neural

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) [132].

Table 5.1: Track variables passed to the initial version of the RNNIP model [134]. Later versions removed the category embedding and added the per track hit information shown for DIPS in Table 5.2.

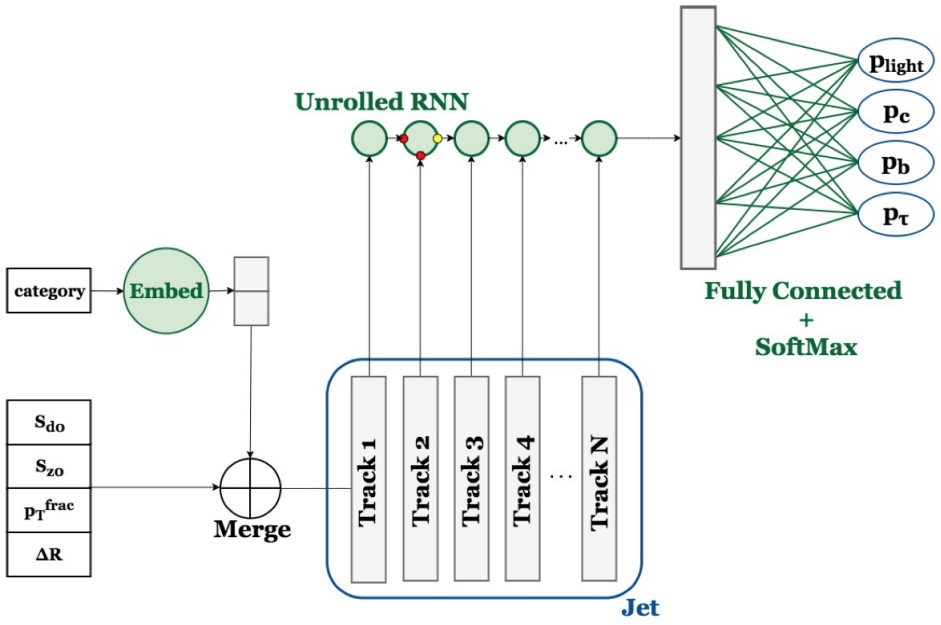


Figure 5.5: Diagram of the RNNIP tagger [151]. The input consists of track features augmented with an embedding of track categories. Tracks are then ordered by absolute transverse IP significance and fed through an LSTM core. The unrolled sequence from this LSTM is padded to a fixed size and processed by a DNN to output the per flavour probabilities.

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 Section 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.

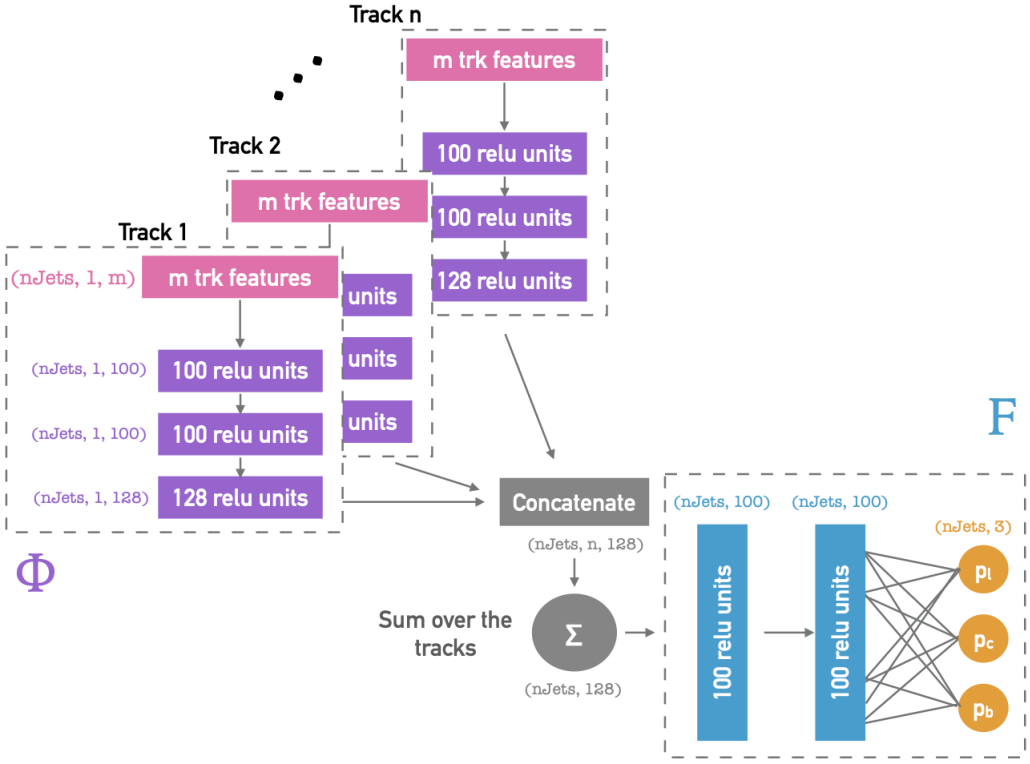
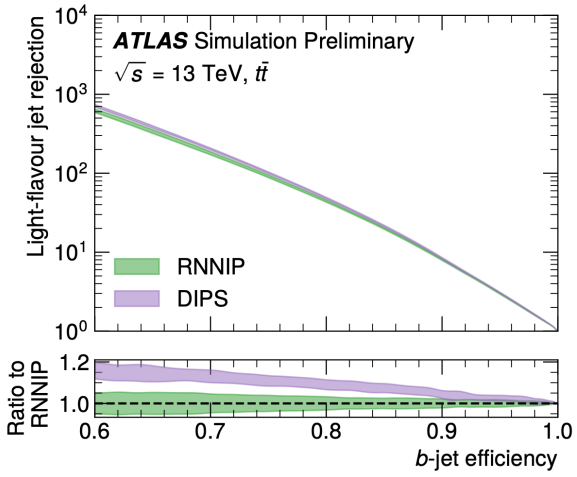


Figure 5.6: Diagram of the DIPS tagger for flavour tagging [135]. The input consists of a set of N tracks, each represented by a feature vector. Each track is embedded by a DNN track network Φ into a fixed-dimension vector. All embedded track vectors are then pooled by summation to a fixed-size vector. 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.

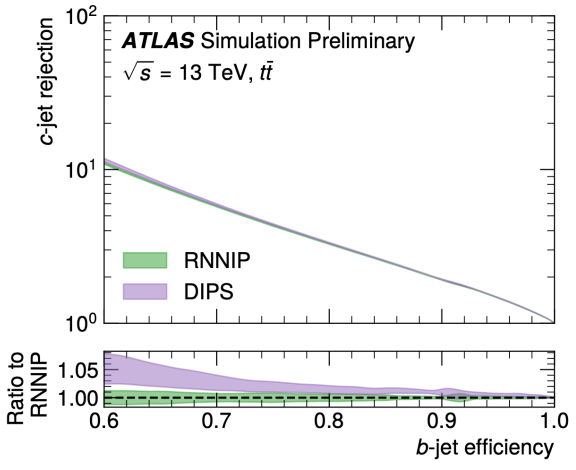
This approach has several advantages over RNNIP, mainly the physically motivated permutation-invariance of the input and the improved training and evaluation time thanks to a more parallelisable architecture, as the track embedding performed by Φ can be massively parallelised on GPUs. These motivations are translated in an appreciable performance delivered by DIPS, which globally outperforms RNNIP while operating at a reduced computational cost [135]. The performance can be assessed from Figure 5.7, 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 $t\bar{t}$ evaluation sample.

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 $\log p_T^{\text{frac}} = \log p_T^{\text{track}}/p_T^{\text{jet}}$.
$\log \Delta R$(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 [135]. 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 [71].



(a) Light-rejection.



(b) c-rejection.

Figure 5.7: Light- (left) and c-rejection (right) as a function of b-jet tagging efficiency for RNNIP (green) and DIPS (purple), taken from [135]. The curves and error bands show the mean and standard deviation of the rejections for 5 trainings per algorithm. The bottom panel shows the ratio to RNNIP.

The training times on the same GPU hardware for a 48k parameters DIPS model is 78 ± 4 seconds per epoch, while a 47k parameters RNNIP requires roughly thrice as much, 241 ± 14 seconds per epoch [135]. The faster training time allowed the Collaboration to focus on

optimisation studies of the hyperparameters. An important observation was that loosening the track selection criteria led to a performance improvement. For RNNIP, IP2D, and IP3D, the selected tracks must pass the following quality selection: ≥ 8 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 [135]. Loosening the selection and keeping the top 25 tracks as ranked by decreasing S_{d_0} to capture more tracks from heavy-flavour decays gives 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. Some studies on interpreting the performance of DIPS are summarised in Appendix A.1.

5.2.3 Training 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. 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 larger 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 semi-leptonic 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 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. The angular separation ΔR between the products of a decaying particle X of large mass m_X scales inversely to the transverse momentum [82]:

$$\Delta R \approx \frac{2m_X}{p_T^X}. \tag{5.5}$$

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 [152]. 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

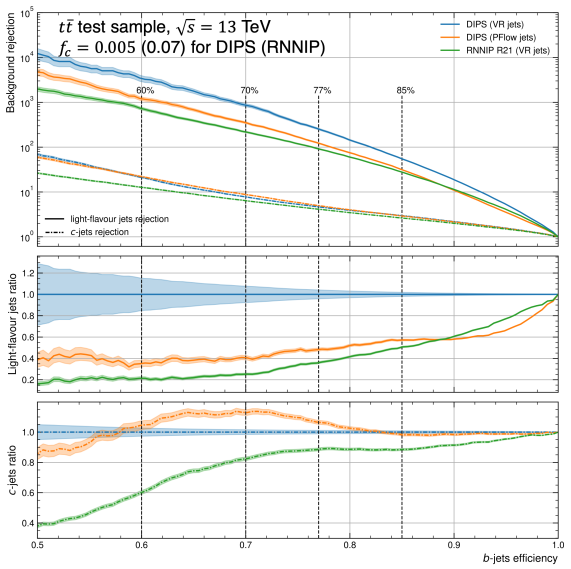
with radius $R = 1.0$. Using such a fixed large radius overestimates the size of boosted jets which are easily contaminated by the PU, as well as the underlying event and initial-state radiations.

Another approach to reconstruct jets from boosted objects decay is the VR jet algorithm [153], 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 5.5. 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:

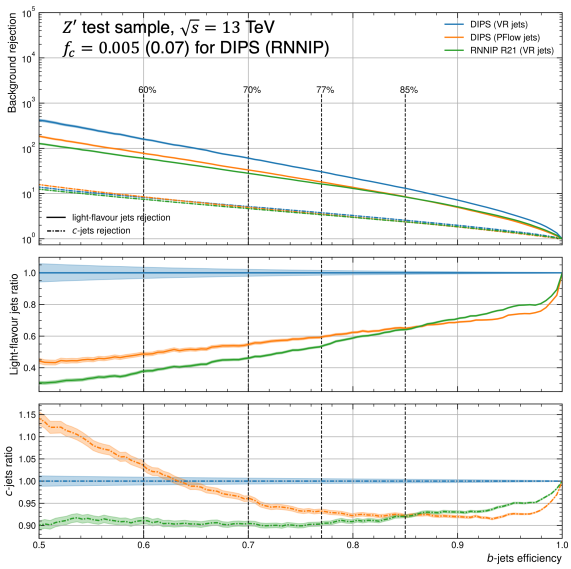
- 85 % of jets are sampled from the $t\bar{t}$ with a maximal p_T of 400 GeV. At least one of the W boson from the t-quark is required to decay leptonically.
- 7.5% are sampled from Z' events, where an exotic boson Z' decays as $Z' \to q\bar{q}$, with a variable Z' mass to generate a flat p_T spectrum 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 A.2 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 that bin have to be oversampled, while values lower than 1 indicate they should be downsampled. Jets are iteratively resampled until the distribution of each flavour matches the target distribution. As displayed in Appendix Figure A.2a 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.

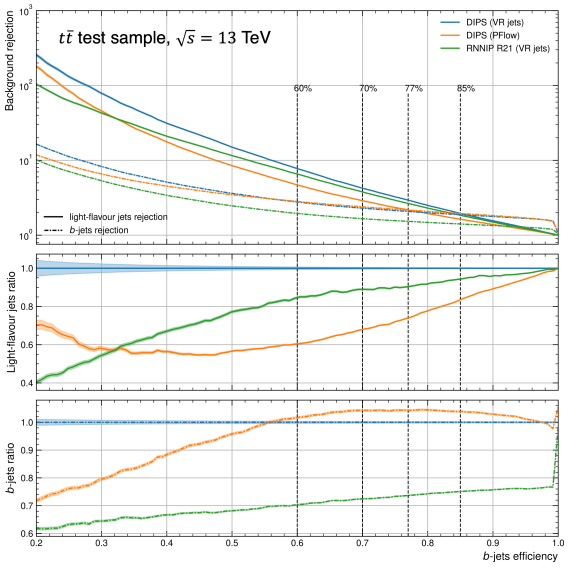
The optimised DIPS model with 62,167 learnable parameters from the previous section was trained for 200 epochs on 4 Quadro RTX 8000 GPUs. The learning rate started at 0.001 and was reduced by a factor of 0.8 on plateaus of 3 epochs, with a batch size of 15k jets, batch normalisation, and a dropout rate 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 validation set of 300k jets as well as the best light- and c-rejections at a fixed 77% b-tagging efficiency is selected.



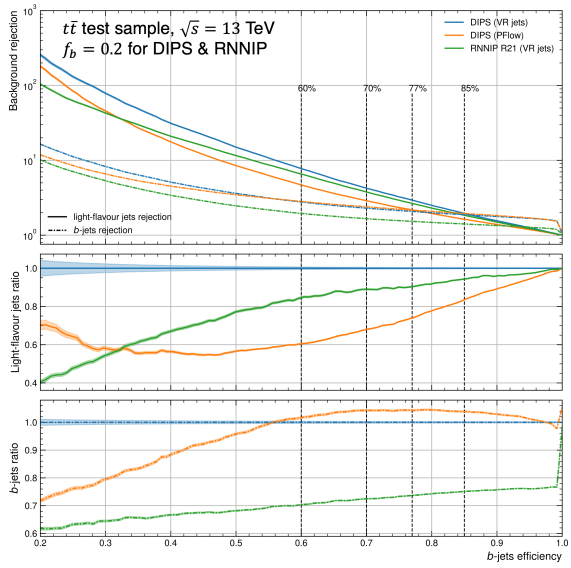
(a) b-tagging on $t\bar{t}$.



(b) b-tagging on Z'.



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

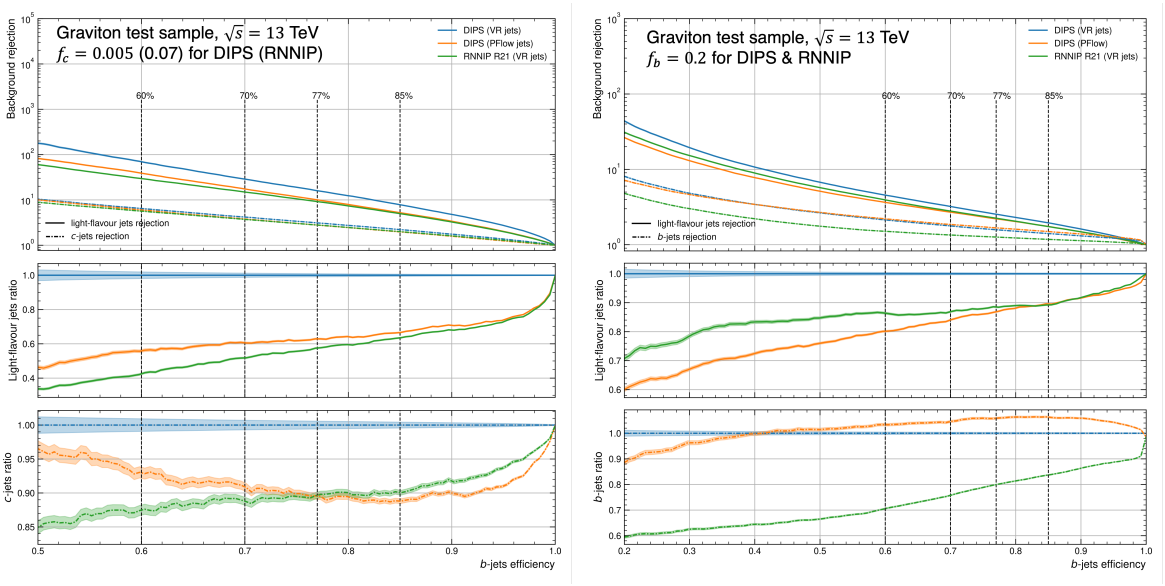


(d) c-tagging on Z'.

Figure 5.8: ROC curves for b-tagging and c-tagging on 300k jets test samples of $t\bar{t}$ (left) and Z' (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.

Figures 5.8 and 5.9 show 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 (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 $Y \in [c$, 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



(a) b-tagging. (b) c-tagging.

Figure 5.9: ROC curves for b- and c-tagging on 300k jets of the graviton samples, similar to Figure 5.8.

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 is reserved for the final model as presented in Section 5.2.5. Figures 5.8c, 5.8d, and 5.9b lead to similar conclusions for c-tagging.

5.2.4 Training DL1d and DL1r with PFlow Jets for Run 3

This work presents the first study of the retraining of DL1r on a new ATLAS software release for the Run 3 of the LHC, and the first training of DL1d including the DIPS sub-tagger in a high-level flavour tagging tool. Other important novelties of this work are the possible inclusion of τ-jets in the 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', are simulated in proton-proton collisions at $\sqrt{s} = 13$ 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 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.

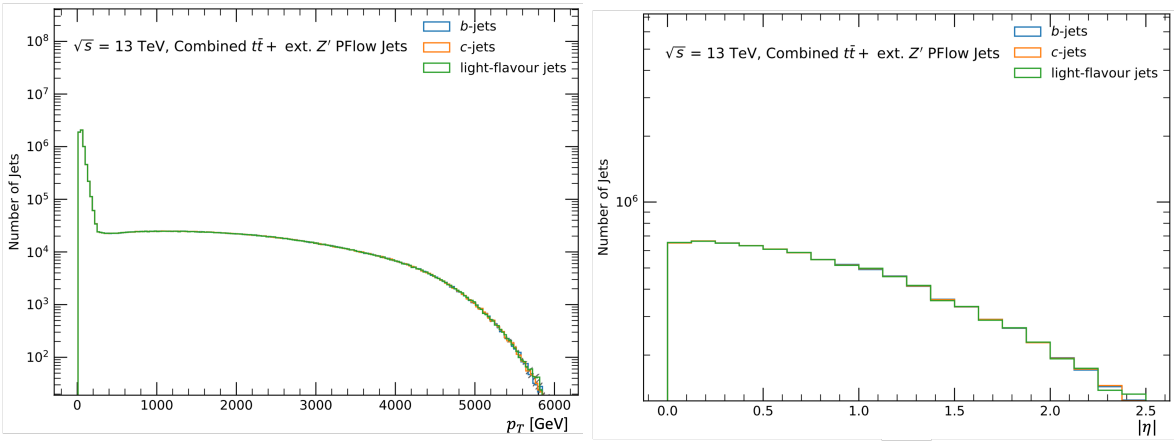


Figure 5.10: The p_T (left - in MeV) and $|\eta|$ distributions of the resampled b-, c-, and light-jets in, respectively, blue, orange, and green. The three sets are resampled to have the same $p_T - |\eta|$ 2D distributions. The flat p_T spectrum extending up to several TeV is due to the exotic Z' process 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 software [2] based on TensorFlow [101] for 300 epochs with a variable learning rate schedule and the default network structure adopted in the previously released DL1r: 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 conclusion driven by the lack of improvements from these attempts is that models built on this simple DNN structure with such a large dataset are already likely saturating their performance. 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 software release 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.

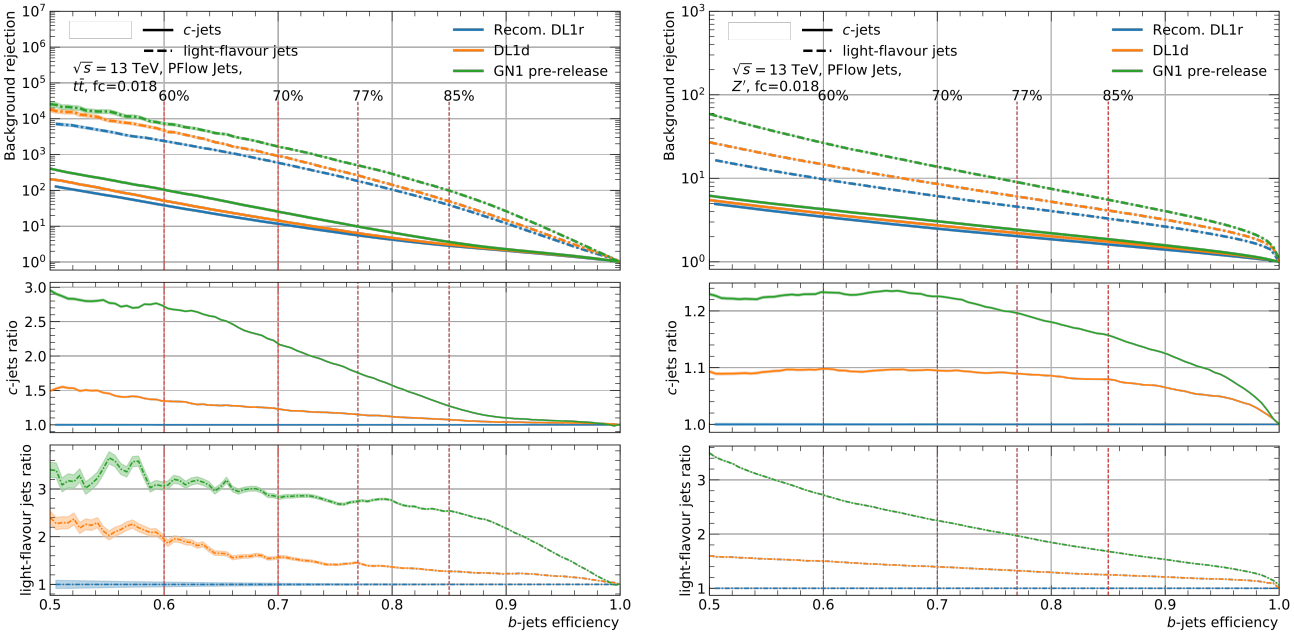


Figure 5.11: Performance for b-tagging with a flavour fraction of $f_c^b = 0.018$. Left: $t\bar{t}$; right: Z'. Top: ROC curves; centre: ratio of c-jets rejection from b-jets relative to DL1r; bottom: same ratio for light-jets rejection. The recommended DL1r from the previous release is in blue. The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

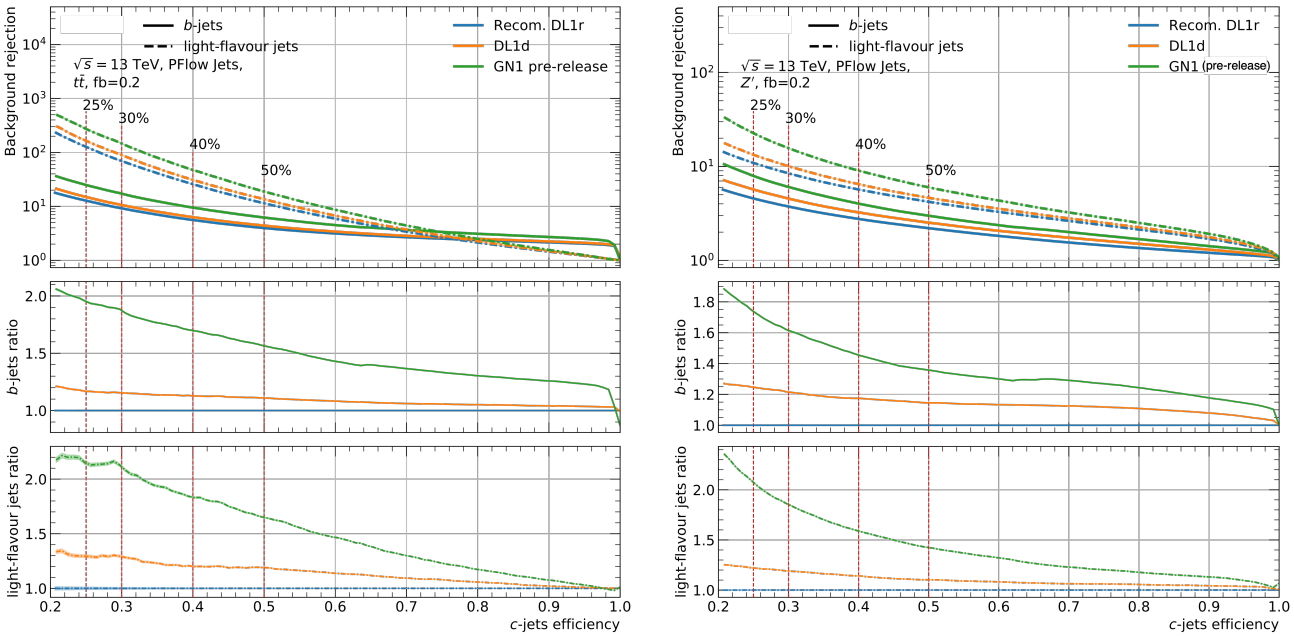


Figure 5.12: Performance for c-tagging with a flavour fraction of $f_b^c = 0.2$. Left: $t\bar{t}$; right: Z'. Top: ROC curves; centre: ratio of b-jets rejection from c-jets relative to DL1r; bottom: same ratio for light-jets rejection. The recommended DL1r from the previous release is in blue. The new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

Figure 5.11 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 5.8. 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, 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 sub-tagger 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 noisier set of 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 [4]. 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.

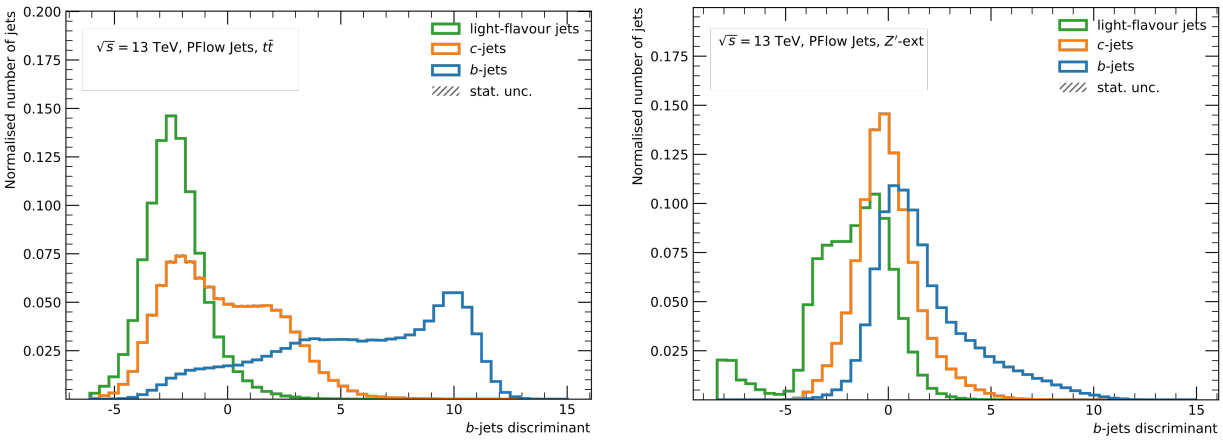


Figure 5.13: Distribution of DL1d b-tagging discriminant with $f_c = 0.018$ for the different jet flavours, evaluated on $t\bar{t}$ (left) and Z' (right).

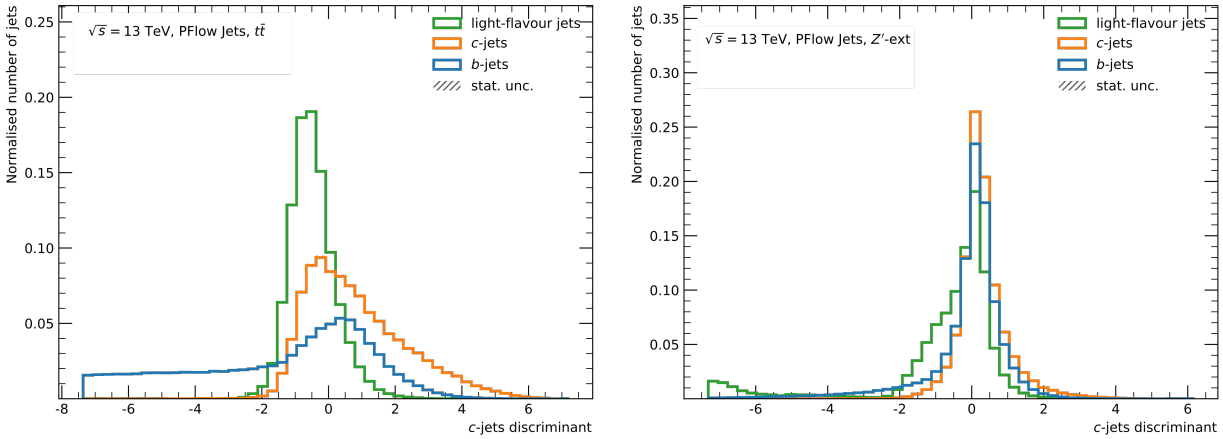


Figure 5.14: Distribution of DL1d c-tagging discriminant with $f_b = 0.2$ for the different jet flavours, evaluated on $t\bar{t}$ (left) and Z' (right).

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.

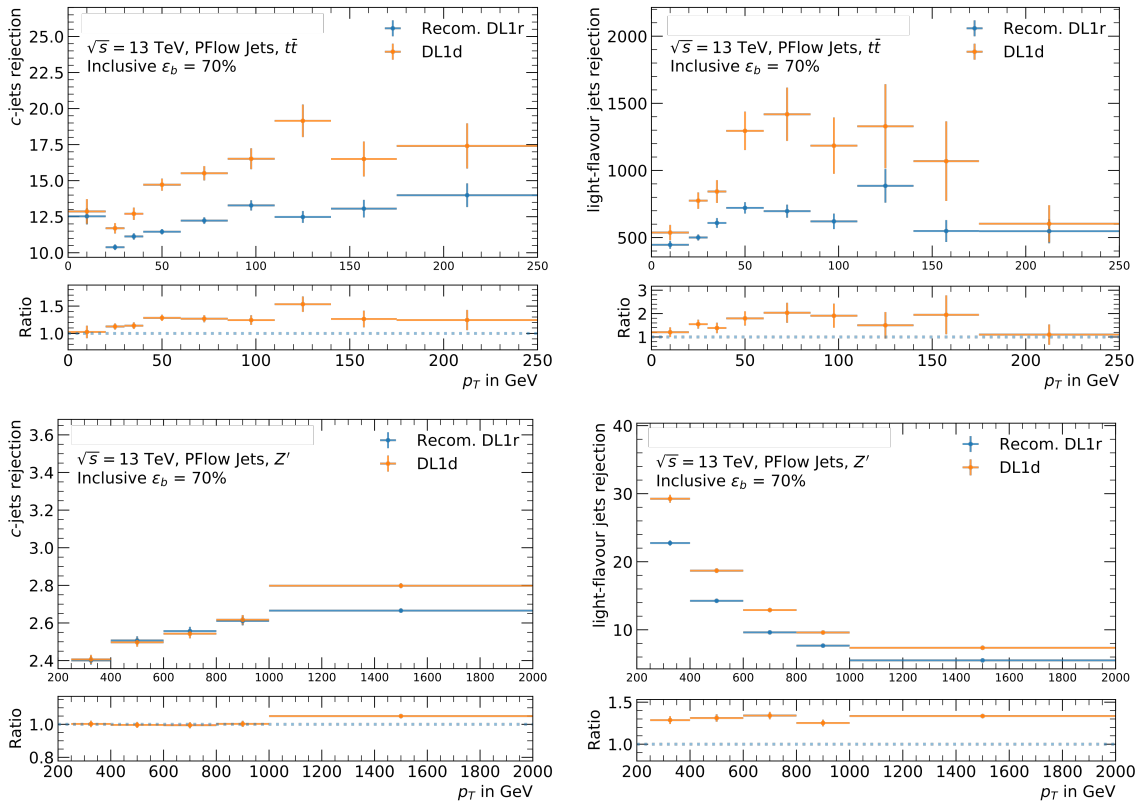


Figure 5.15: Background flavour rejections at a fixed b-tagging efficiency of 70% (per region shown) for the various taggers. Top: $t\bar{t}$; bottom: Z'; left: c-rejection; right: light-rejection. For each plot, the bottom panel presents the ratio to the recommended DL1r.

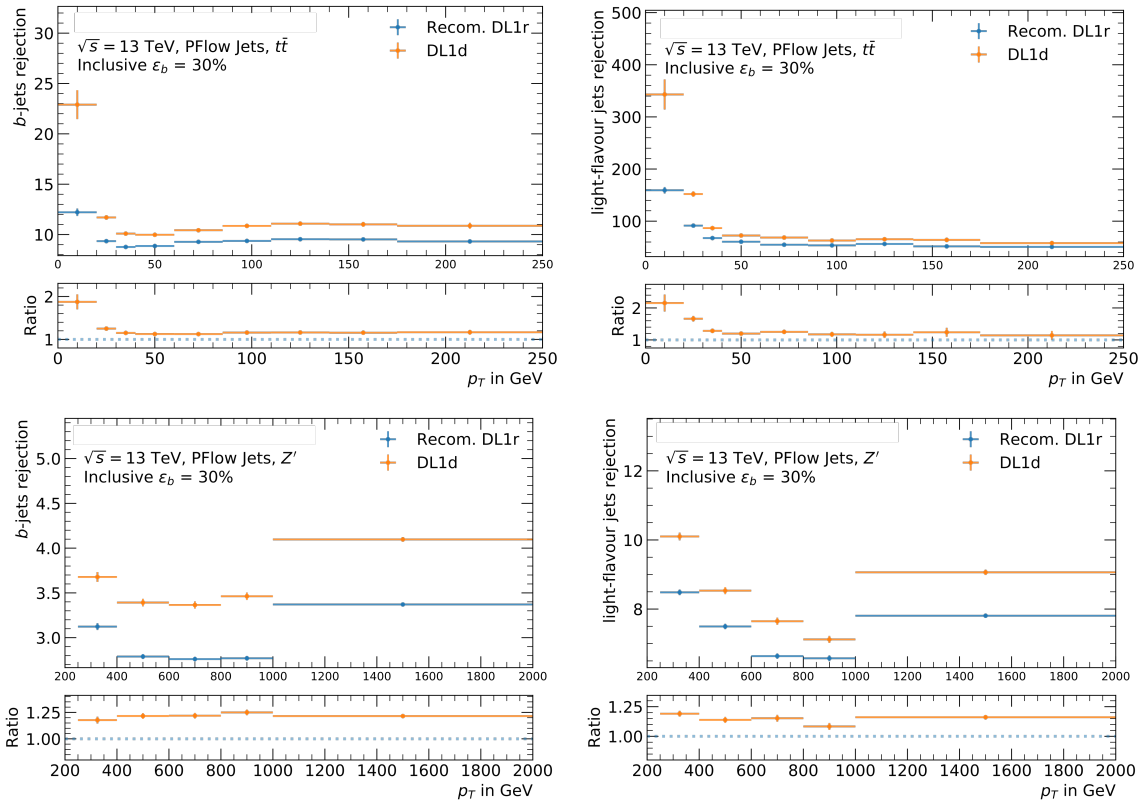
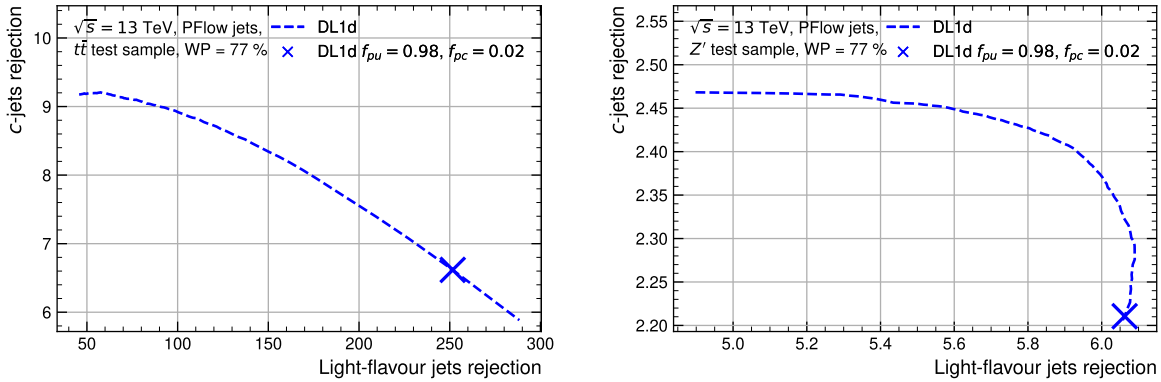
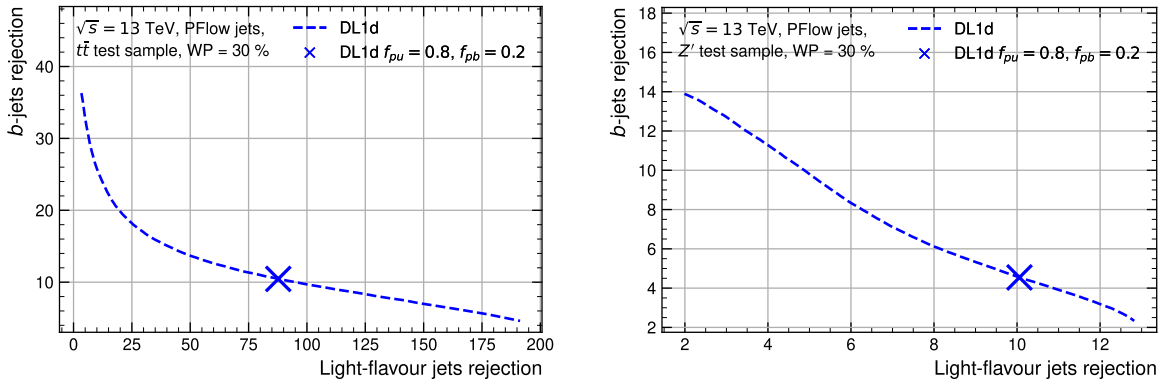


Figure 5.16: Background flavour rejections at a fixed c-tagging efficiency of 30% (per region shown) for the various taggers. Top: $t\bar{t}$; bottom: Z'; left: b-rejection; right: light-rejection. For each plot, the bottom panel presents the ratio to the recommended DL1r.

The DL1d model was quickly integrated into the ATLAS software thanks to its similarities with DL1r. Its fast calibration led to its rapid introduction to the Collaboration and deployment in early Run 3 analyses [154]. 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).



(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 operating point of 77% (33%). Increasing f_c or f_b shifts the marker upwards along the curves.

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 [94]. Figures 5.18 and 5.19 present the outcome of applying this framework, as proposed in Ref. [155] 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 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 - 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-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 explained by the distinct kinematic properties of 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 because the information is better provided by DIPS. This anti-correlation behaviour of sub-algorithms to the D_c discriminant is expected, these methods having been primarily designed to help b-tagging.

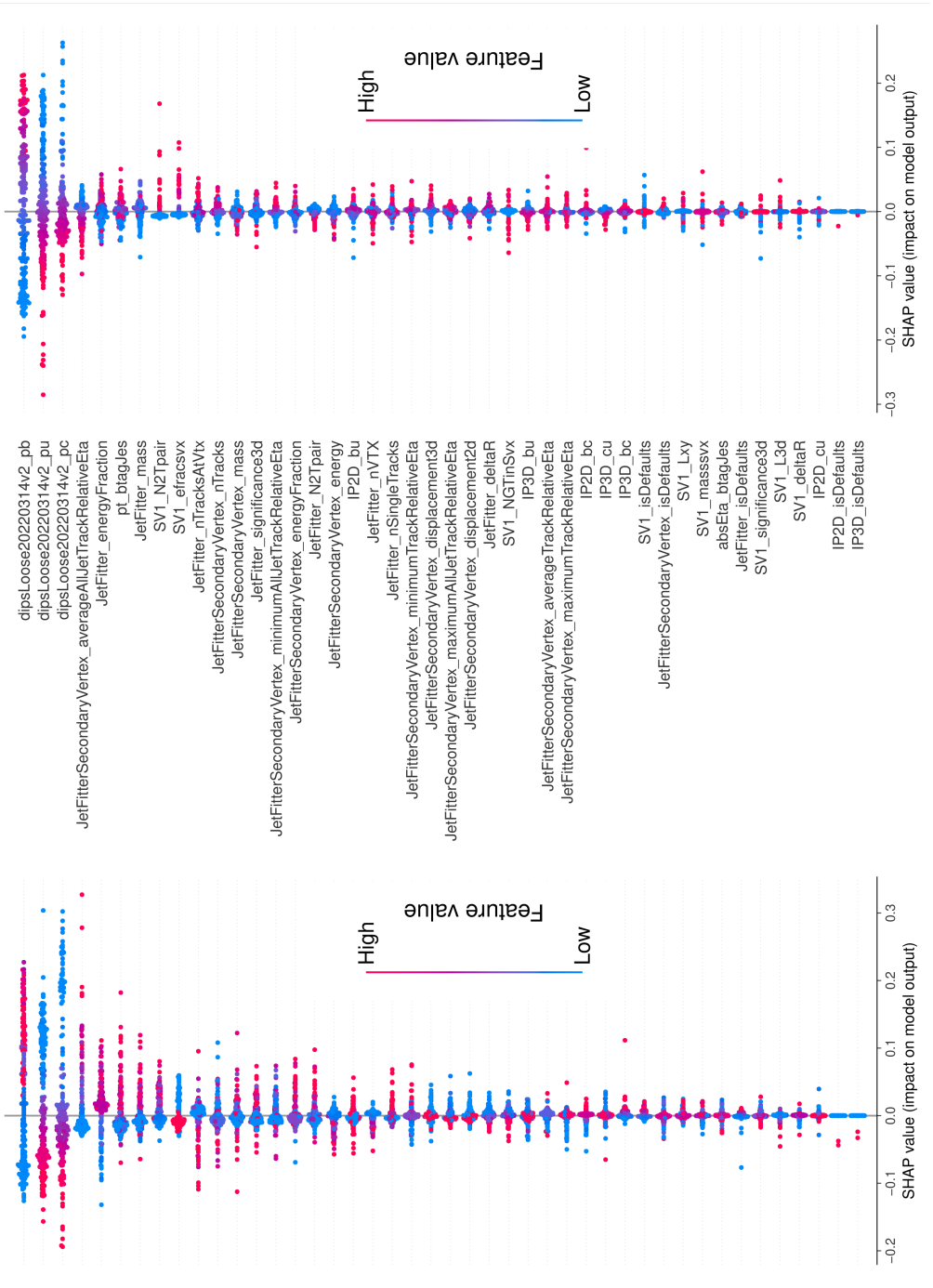


Figure 5.18: Shapley values of the different inputs variables of DL1d for b-tagging, $t\bar{t}$ on the left and Z' on the right. High feature values are marked as red dots, while low feature values are blue.

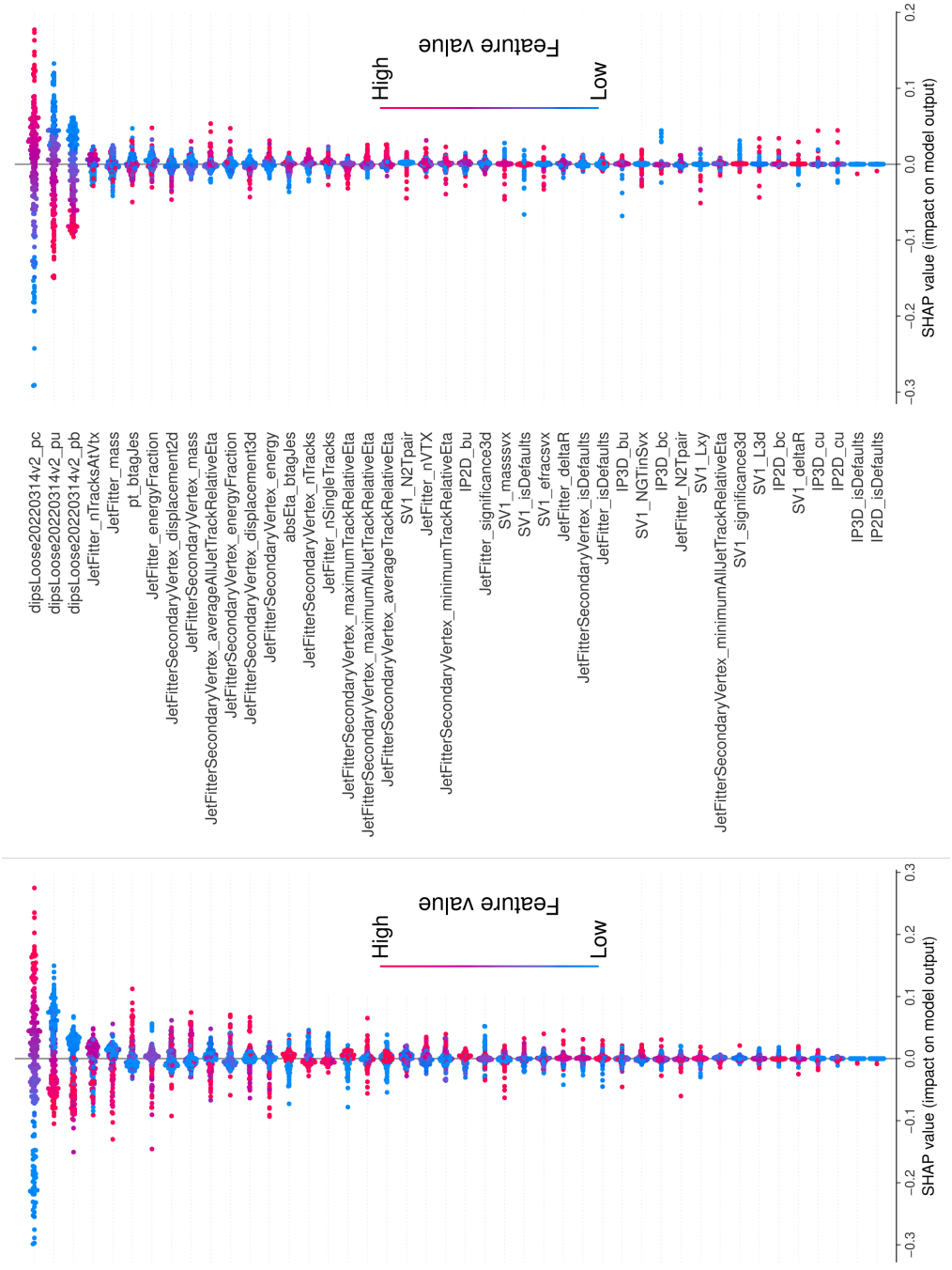
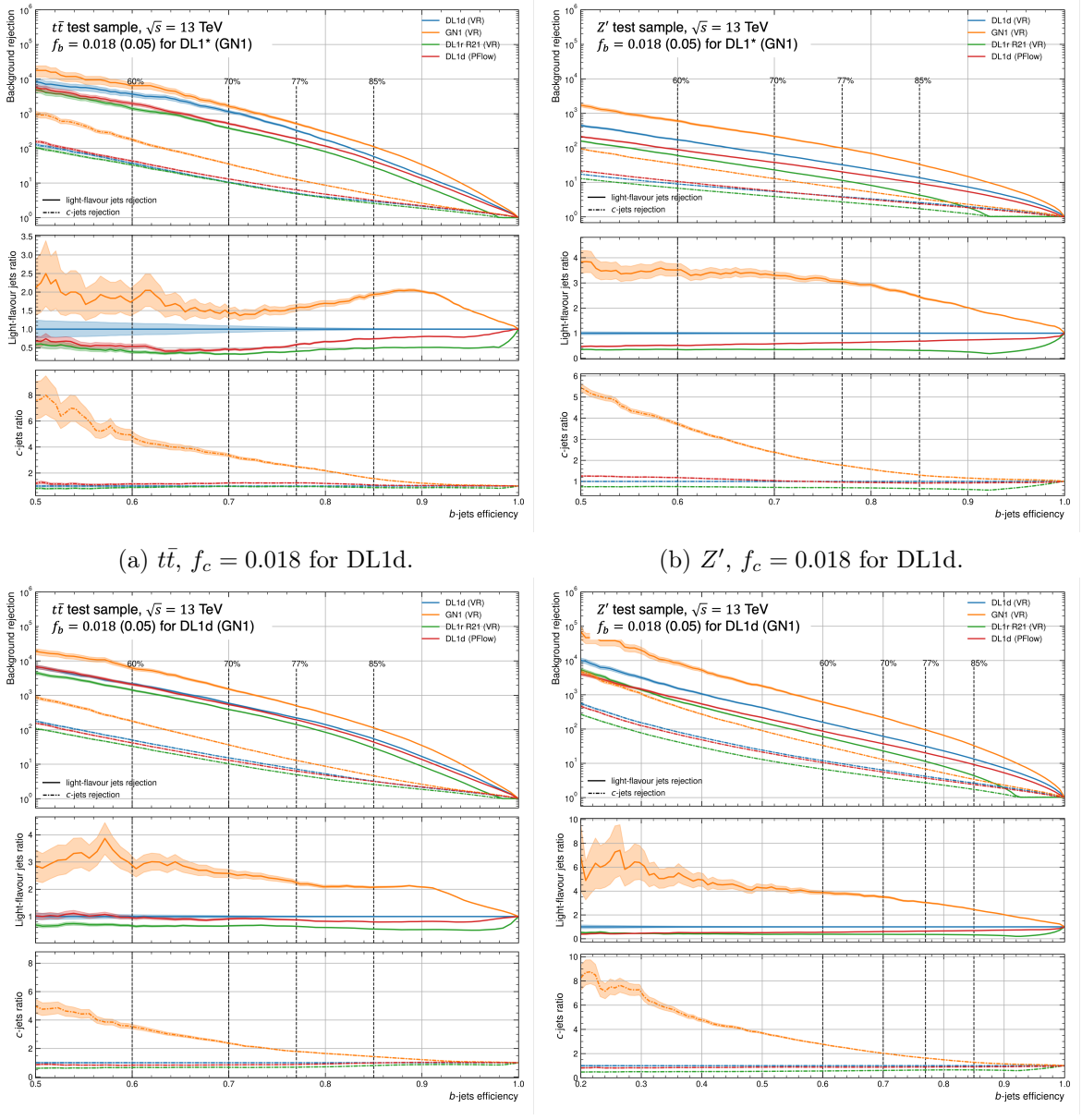


Figure 5.19: Shapley values of the different inputs variables of DL1d for c-tagging, $t\bar{t}$ on the left and Z' on the right. High feature values are marked as red dots, while low feature values are blue.

5.2.5 Training DL1d with 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 a DL1r version trained on VR jets and a pre-release GN1 trained on 20 million VR jets.



(a) $t\bar{t}$, $f_c = 0.018$ for DL1d.

(b) Z', $f_c = 0.018$ for DL1d.

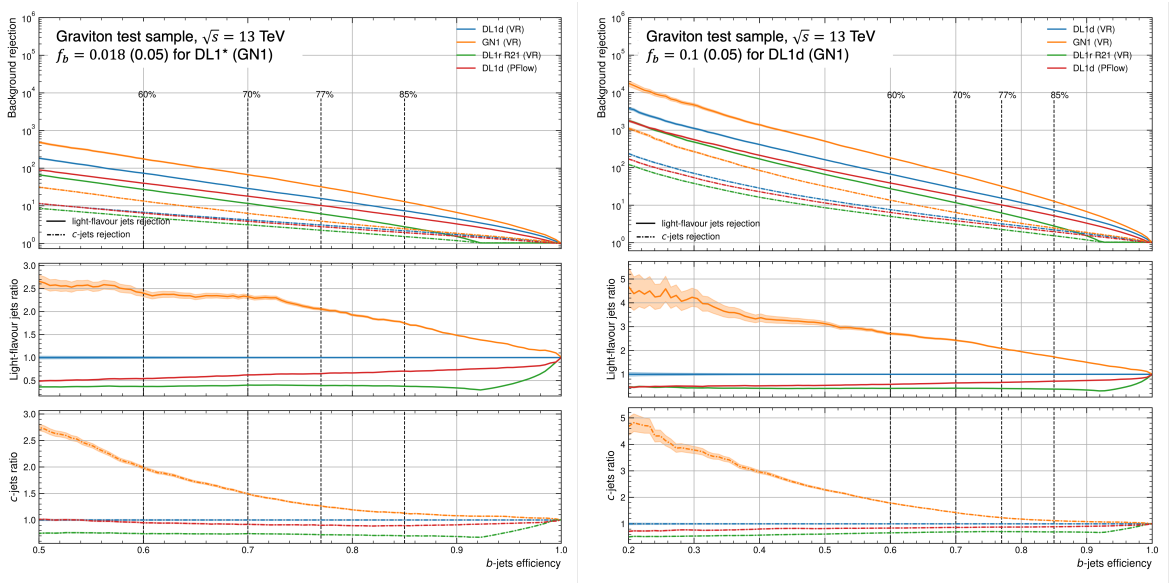
(c) $t\bar{t}$, $f_c = 0.1$ for DL1d.

(d) Z', $f_c = 0.1$ 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, DL1r trained on VR-jets in green, and the PFlow DL1d in red.

A clear benefit from retraining on the dedicated VR jet sets is observed in the ROC curves of Figures 5.20 and 5.21, with the VR-DL1d outperforming the PFlow version for all b- and

c-tagging efficiencies considered. Introducing DIPS in the Deep Learner 1 Model (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) DL1d $f_c = 0.018$. (b) DL1d $f_c = 0.1$.

Figure 5.21: ROC curves for b-tagging. Similar to Figure 5.20 for the graviton process.

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 A.3 of Appendix A.3, displaying flavour fractions scans for b- and c-tagging, this choice of f_b^c is indeed suboptimal for the 30% WP.

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

The new generation of classifiers developed for flavour tagging at ATLAS introduces a fundamental shift in design, moving away from the hierarchical approach. Instead, a single large neural

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 ($f_b^c = 0.2$).

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 5.22, this change to the flavour tagging software stacks greatly simplifies the maintenance and development, with all the attention focused on a single network. A new software called SALT [3] built on PyTorch [100] 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) [110, 156] for GN1 and a transformer encoder for GN2 [114].

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 [157].

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.

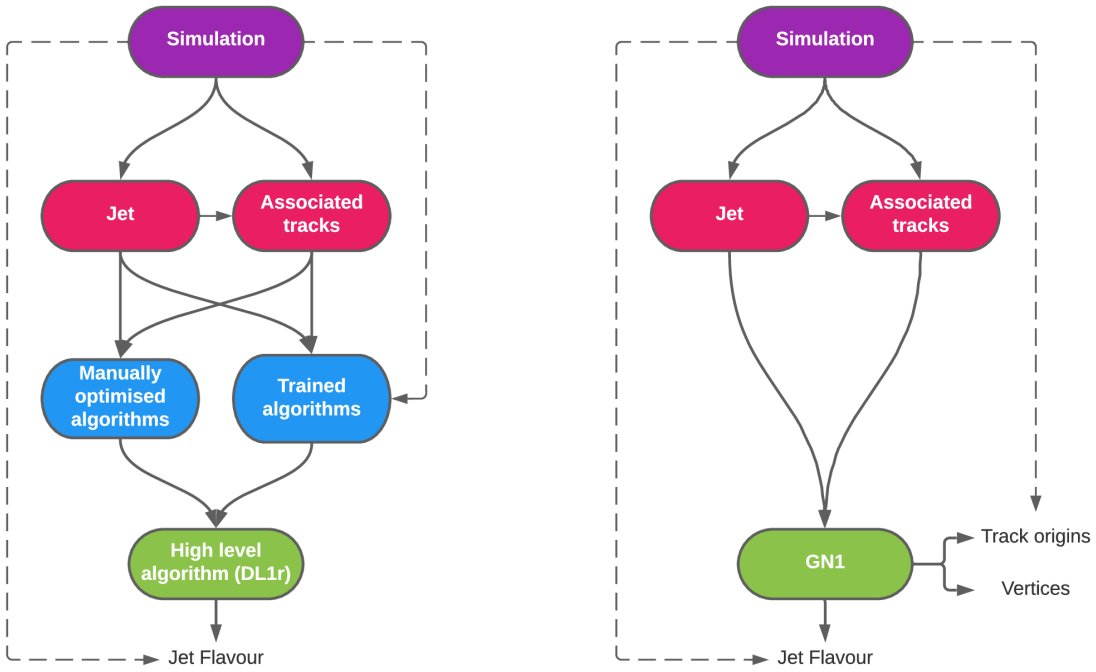


Figure 5.22: Comparison of the tagging scheme between the DL1 family (left) and the GN family (right) [4]. Solid lines represent reconstructed information while dashed lines represent truth information only accessible from the simulations.

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 [4]. Charged particles and tracks are matched using the truth matching probability [71], 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
nSCTHoles	Number of SCT holes

Table 5.6: Input features of the GN family of models [4].

$p_T > 0.5$ 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 5.23, relies on a modified graph attention network [156] specifically designed for graph learning on sets, the so-called *Set2Graph* [158]. 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.

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

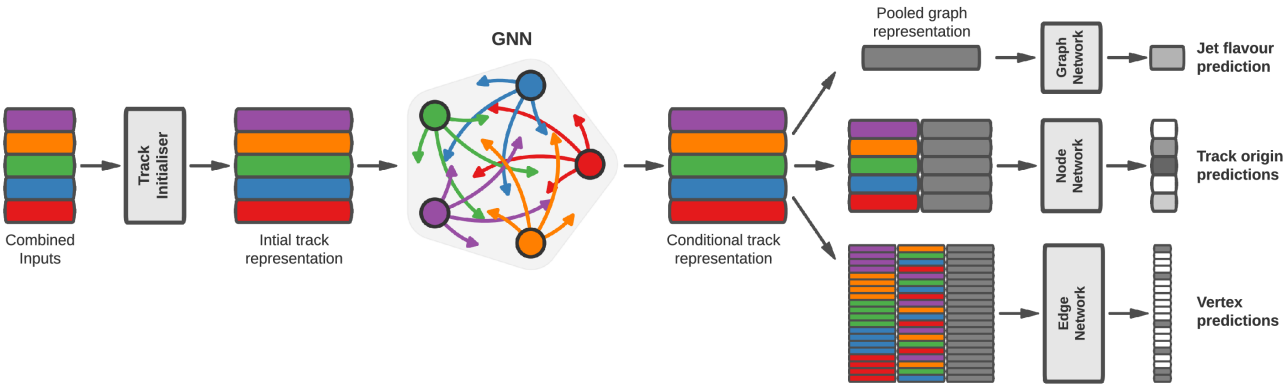


Figure 5.23: The architecture of the GN1 network [4]. The combined input is made of the set of tracks, each of which is given a copy of the two jet variables in addition to the track features, as described in Table 5.6. After a first embedding taking the input to an enriched latent representation, a 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.

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. [156]. 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 [4]:

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

$$e\left(h_i, h_j\right) = V^T \theta\left([Wh_i, Wh_j]\right), \tag{5.6}$$

   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:

$$a_{i,j} = \text{softmax}_j\left(e(h_i, h_j)\right). \tag{5.7}$$

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

$$h'_i = \sum_j a_{i,j}\,.\,Wh_j, \tag{5.8}$$

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 [4]:

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).

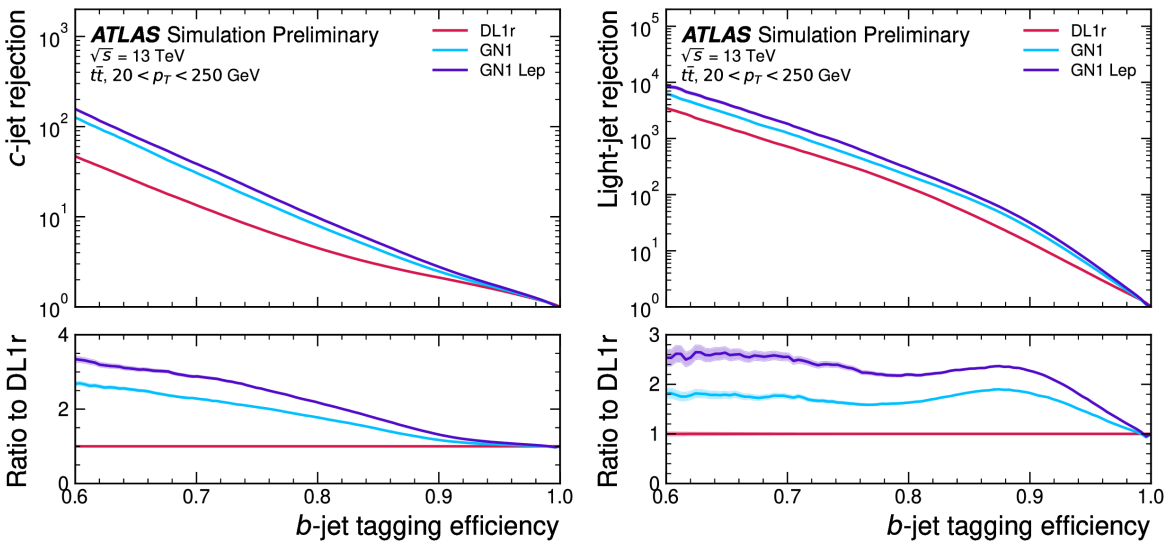
2. *Track origin prediction*: performed by a nodes classifier processing each conditional track representation with the global representation. This network is built with three layers of reducing size 128, 64, and 32 to finish on the output layers of size 7 with softmax, matching to the 7 classes corresponding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

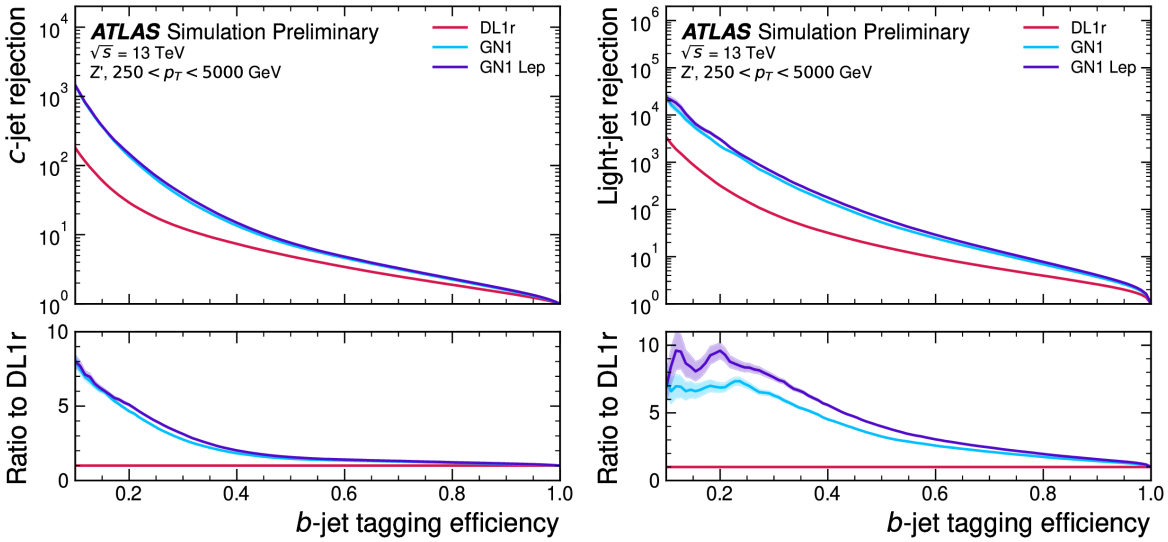
The architecture of GN1 is somewhat similar to an enhanced version of DIPS, with the track initialiser and graph classifiers corresponding to Φ and F. Added elements are the powerful GNN layers and conditional representation pooling layer with attention, as well as the auxiliary objectives. GN1 is trained by minimising the combined loss function $\mathcal{L}_{\text{total}}$ defined as

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{flavour}} + \alpha\,\mathcal{L}_{\text{track}} + \beta\,\mathcal{L}_{\text{vertex}}. \tag{5.9}$$

$\mathcal{L}_{\text{flavour}}$ is the categorical 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 categorical 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. In Equation 5.9, 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 $\beta = 1.5$ [4] 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. [4], 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 [120] with a learning rate of 10^{-3} and a batch size of 4000 jets spread across 4 GPUs.



(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 5.24: 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 [4]. 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.

The results of the training are presented in Figures 5.24 and 5.26 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 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. The performance of DL1d is approximately 20% to 50% above DL1r at the 70% WP, 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 WP 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%) increase in rejection at a 30% WP on Z', which roughly corresponds to applying the 70% WP 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 5.25. The distributions for GN1 move the b-jet distribution to higher values of the discriminants, indicating higher confidence in the associated predicted p_b.

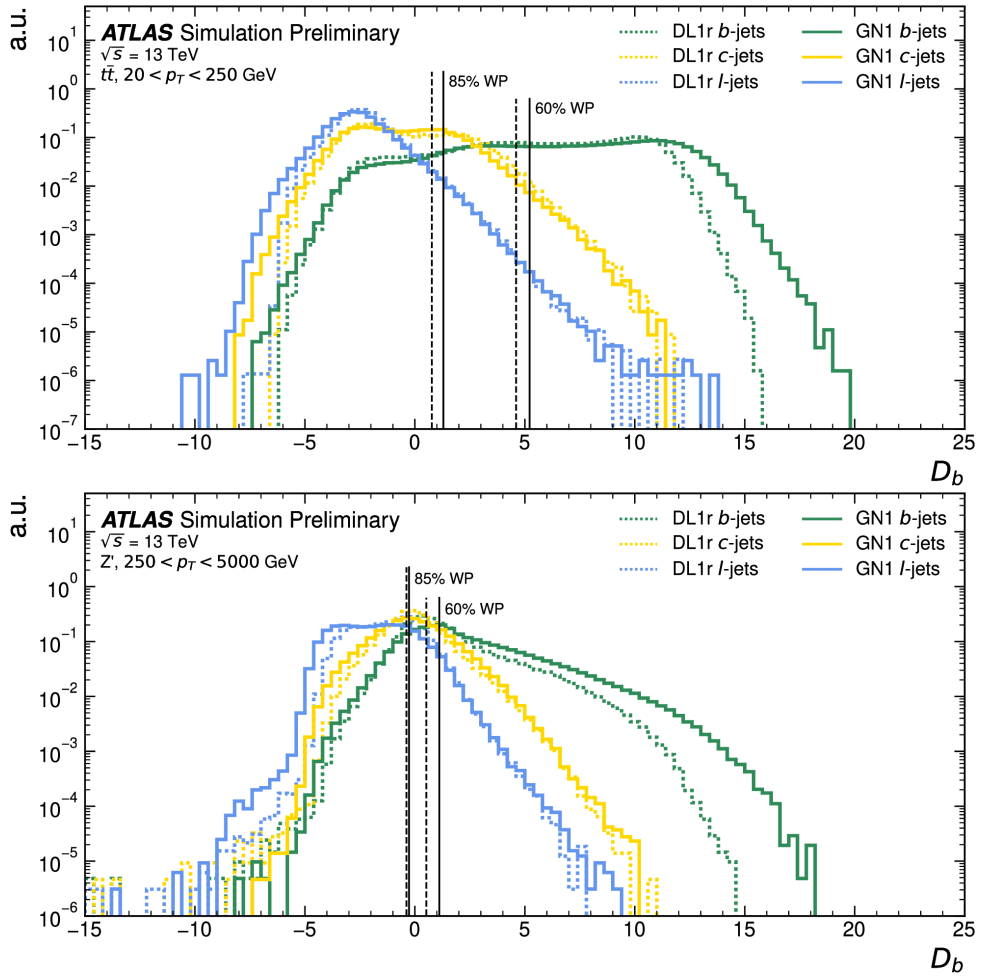
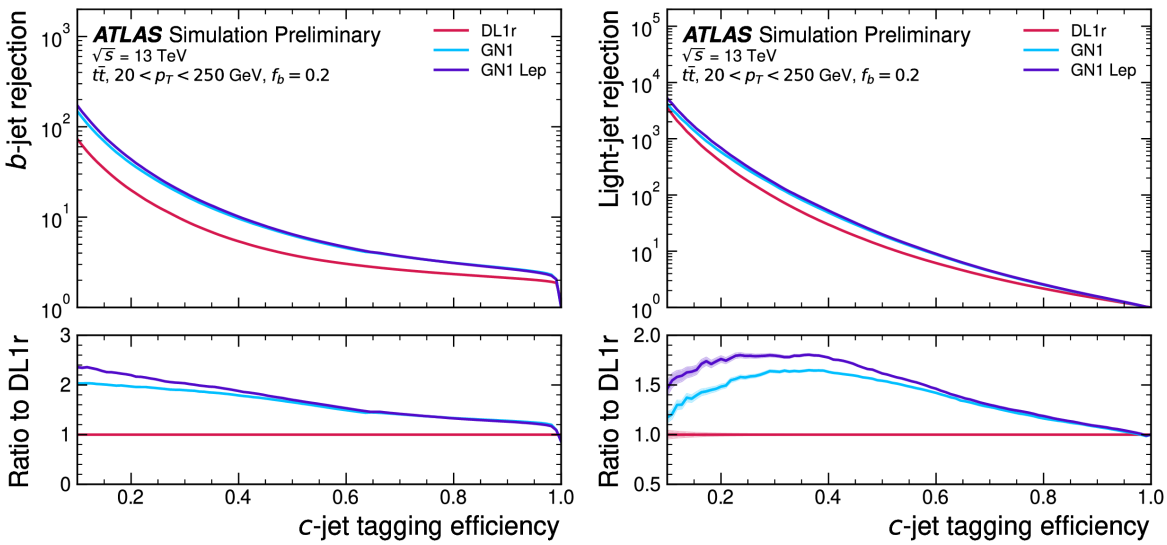
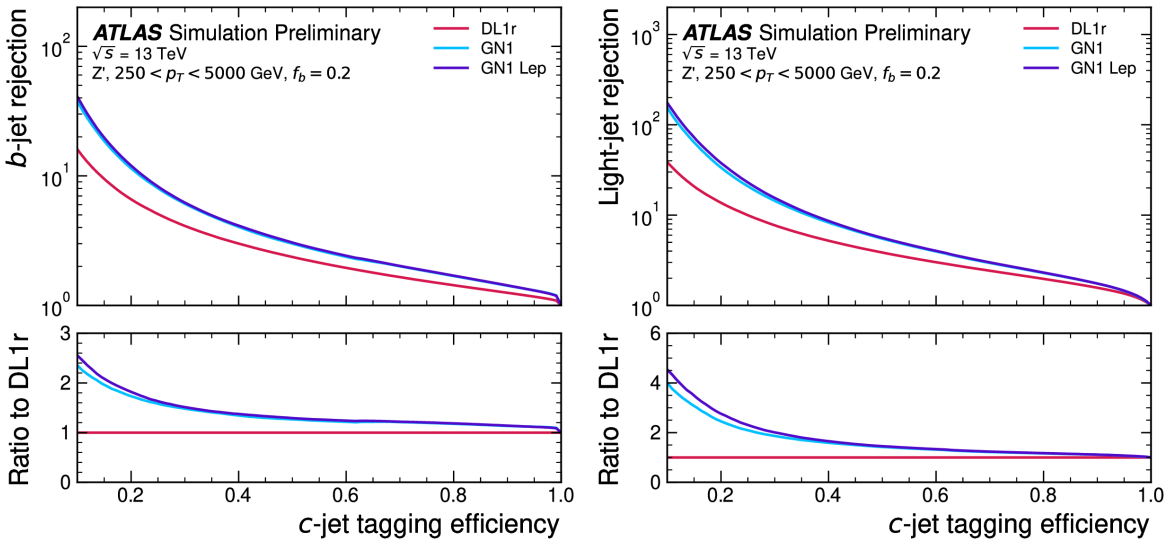


Figure 5.25: Comparing the GN1 and DL1r b-tagging discriminants D_b normalised distributions on the $t\bar{t}$ (top) and Z' (bottom) test samples [4]. Models compared are DL1r in dashed lines and GN1 in continuous lines. Each flavour is indicated by a different colour: green for b-jets, yellow for c-jets, 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 5.26 and 5.27, 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.



(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 5.26: 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 [4]. 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.

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, 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 [71]. This degradation of the track quality impacts the jet tagging performance significantly, as displayed in Figure 5.28 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.

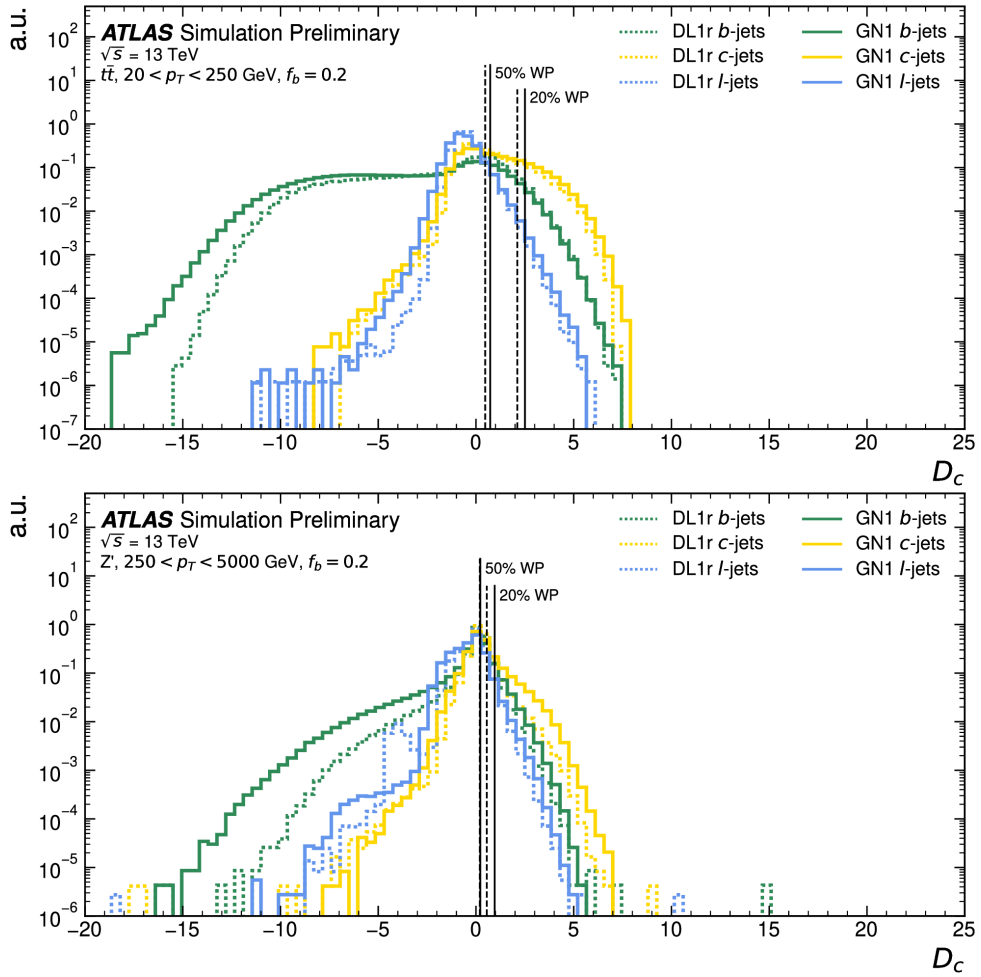


Figure 5.27: Comparing the GN1 and DL1r c-tagging discriminants D_c normalised distributions on the $t\bar{t}$ (top) and Z' (bottom) test samples [4]. Models compared are DL1r in dashed lines and GN1 in continuous lines. Each flavour is indicated by a different colour: green for b-jets, yellow for c-jets, and blue for light-jets. The flavour fraction is set at $f_b^c = 0.2$.

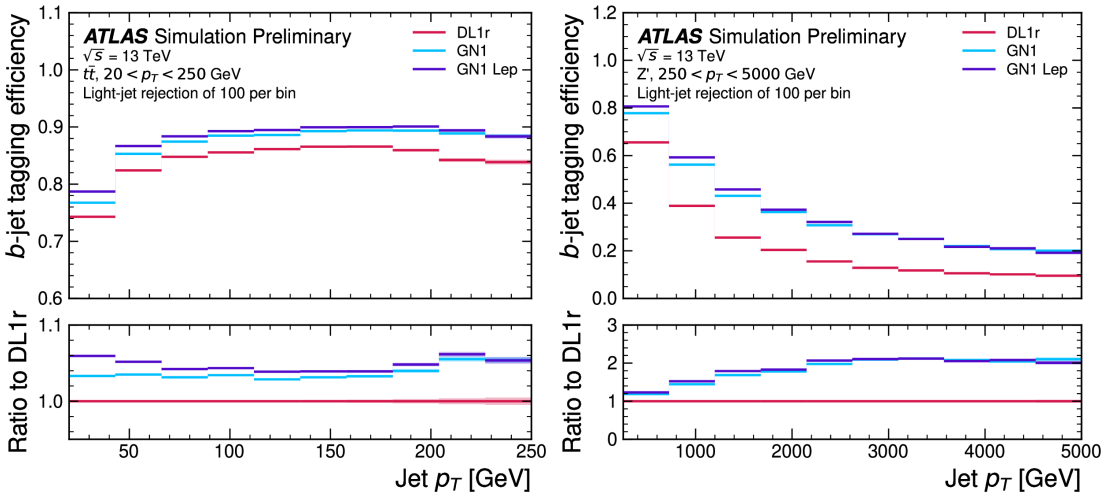


Figure 5.28: Comparing the GN1 and DL1r b-tagging efficiency as a function of jet p_T at a fixed 100 light-jet rejection in each bin on the $t\bar{t}$ (left) and Z' (right) test samples [4]. Models compared are DL1r in dashed lines and GN1 in continuous lines. The flavour fraction is set at $f_c^b = 0.018$ for DL1r and 0.05 for GN1 and GN1 Lep.

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, 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, for the model labelled "GN1 Vert".
- The track classification objective without vertex, referred to as "GN1 TC".

Figure 5.29 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 is not directly relevant as they are only there to distil information to help the primary goal. The track-pairs vertexing performance can be 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 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. [159] is under study to introduce an auxiliary differentiable single vertex fitting task.

Concerning the track origin classification performance, Figure 5.30 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.

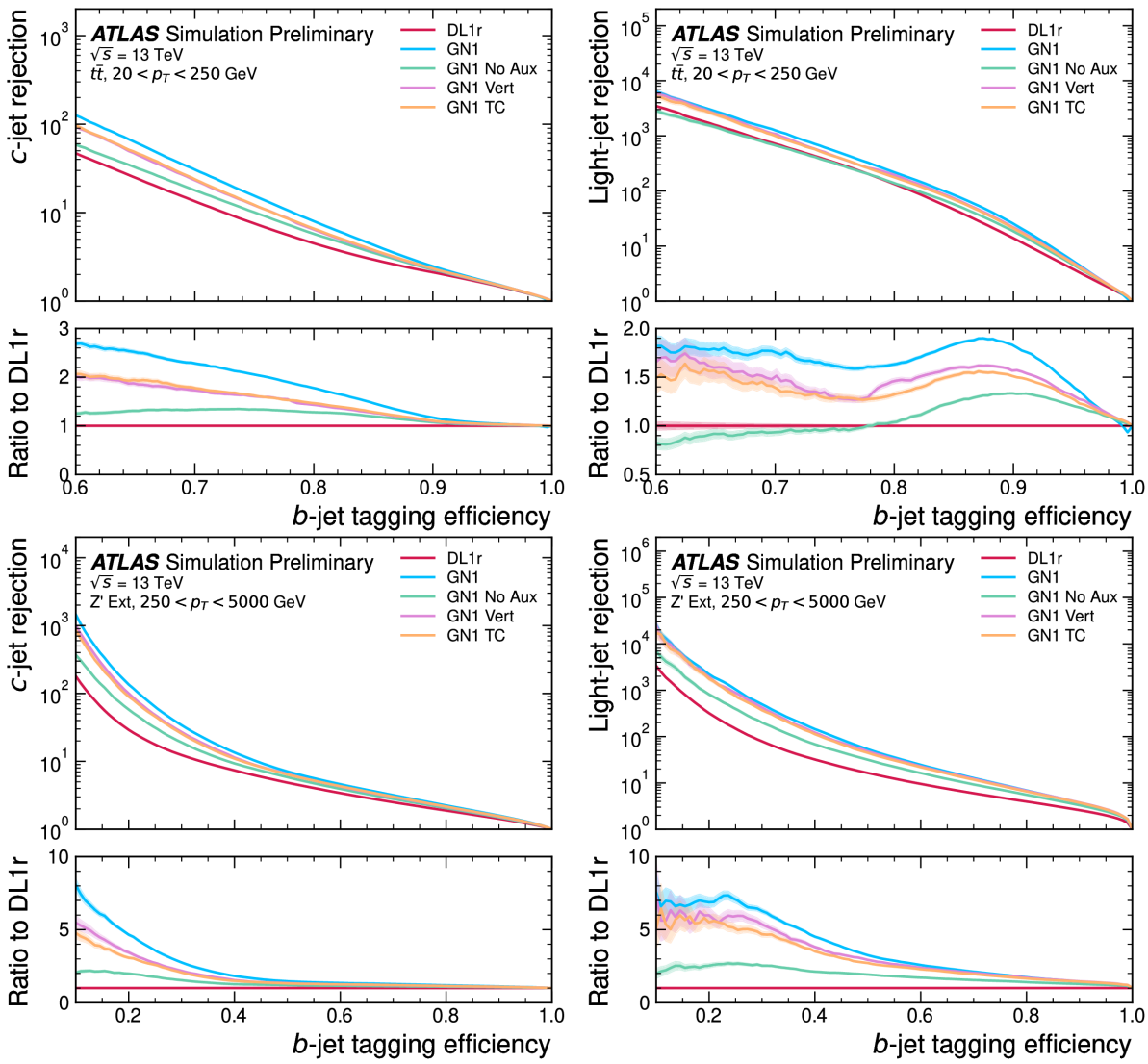


Figure 5.29: ROC curves tracing the b-tagging efficiency versus the c-rejection (left) and light-rejection (right) for the $t\bar{t}$ (top) and Z' (bottom) test samples [4]. Models compared are DL1r in red, GN1 in blue, and versions of GN1 with missing auxiliary tasks. GN1 No Aux in green has none of 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.

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' [4]. 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.

GN1 shows 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 underlines the superiority of deploying a single network built around an

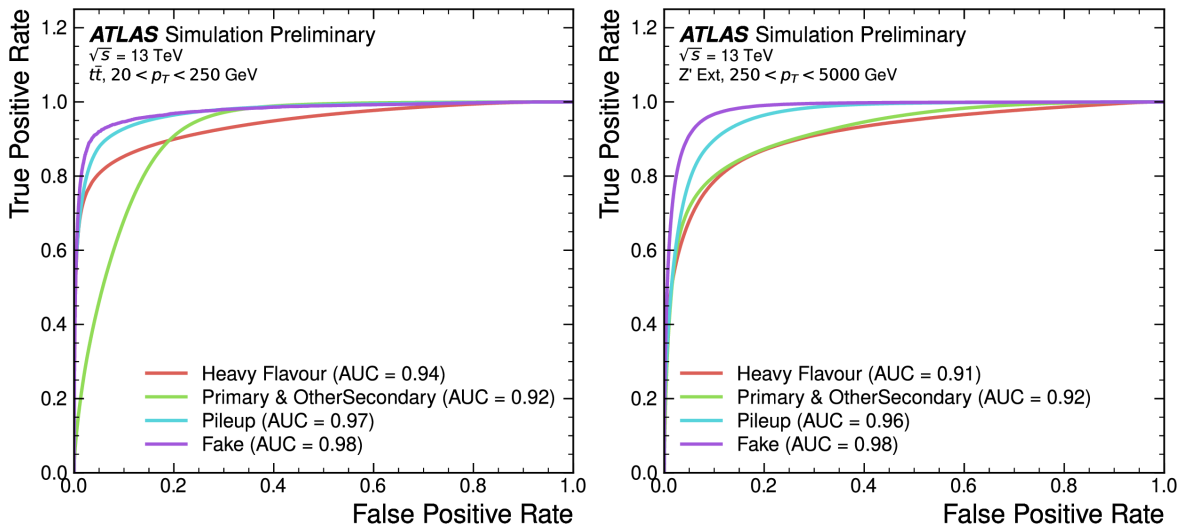


Figure 5.30: ROC curves tracing the false positive rate versus the true positive rate of the truth origin classification on the $t\bar{t}$ (left) and Z' (right) test samples [4]. Heavy Flavour is a weighted combination of the FromB, FromBC, and FromC by their relative abundance.

advanced core unit. 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.

5.3.2 GN2: Transformer Encoder for Flavour Tagging

GN2 is a fine-tuned modification of GN1 built with the same conceptual processing chain but easier to train and simpler to scale in parameters. The main modification with respect to GN1 is the replacement of the computationally complex and expensive GAT layers by a now ubiquitous architecture in machine learning: the transformer [114]. 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.

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. [114] and modified in Ref. [160] is deployed to replace the GAT component of Figure 5.22. 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 [118]. The learning rate scheduler

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 [5].

is based on the one-cycle scheduler of Ref. [161], 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 [162]. The attention computation implemented by the transformer allows similar physics performance to the GAT at a reduced memory footprint and training time [137]. 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:
$$a_{i,j} = \text{softmax}_j\left(e(h_i, h_j)\right).$$

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 5.31. 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 WP. 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 5.32, 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.

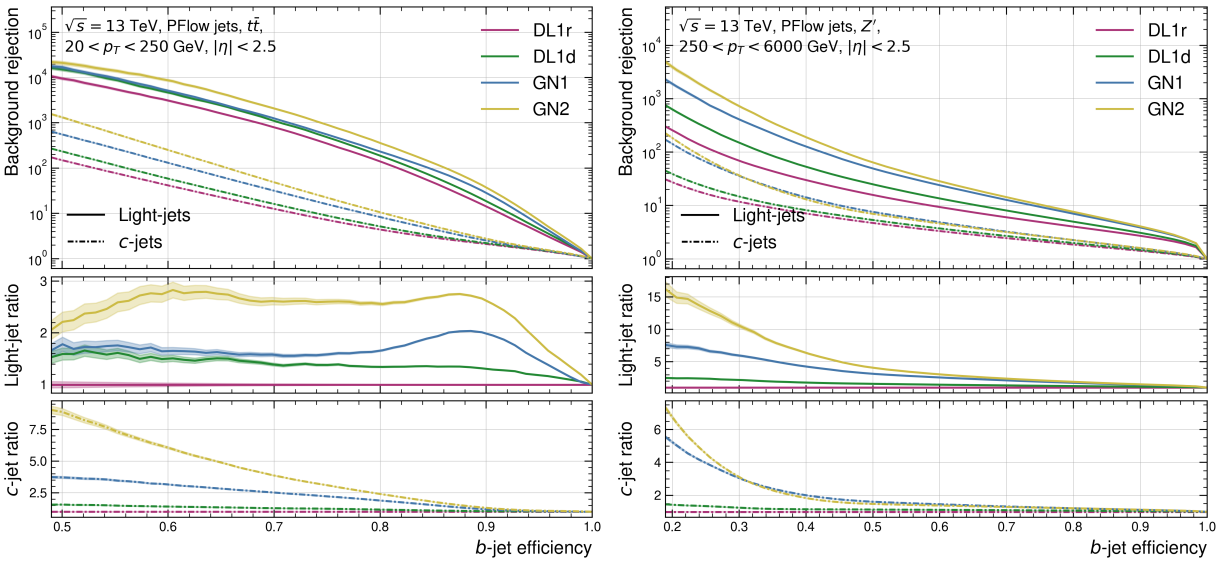


Figure 5.31: The c- and light-rejections as a function of the b-jet tagging efficiency in the $t\bar{t}$ with $20 < p_T < 250$ GeV (left) and Z' with $250 < p_T < 6000$ GeV (right) test samples. Models compared are DL1r in purple, DL1d in green, GN1 in blue, and GN2 in yellow. The bottom plots show the 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.

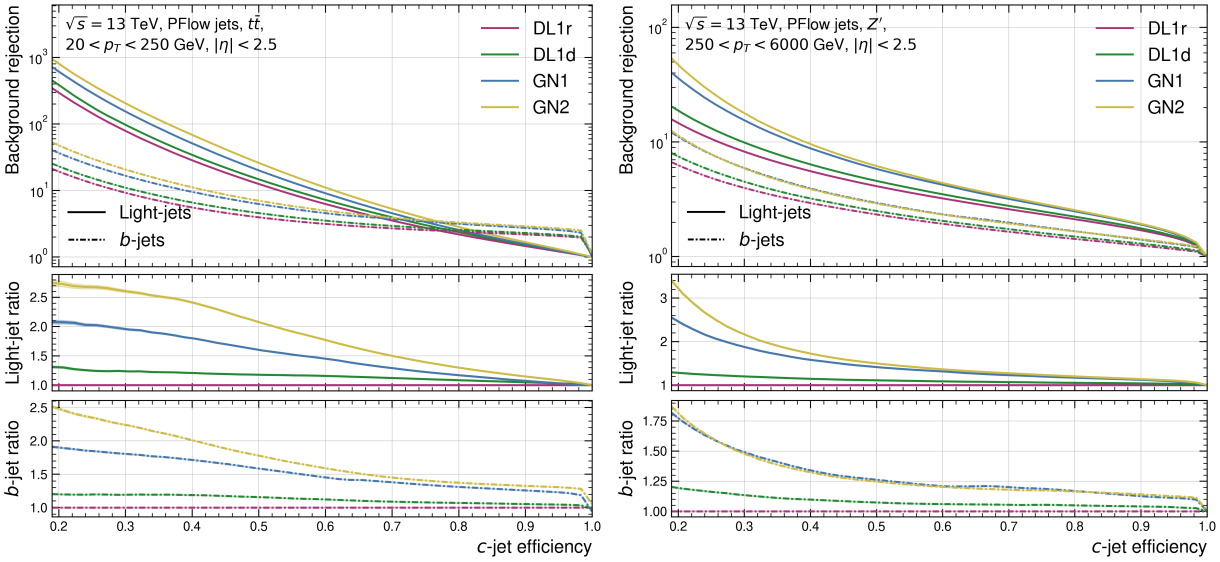


Figure 5.32: The b- and light-rejections as a function of the c-jet tagging efficiency in the $t\bar{t}$ with $20 < p_T < 250$ GeV (left) and Z' with $250 < p_T < 6000$ GeV (right) test samples. Models compared are DL1r in purple, DL1d in green, GN1 in blue, and GN2 in yellow. The bottom plots show the 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.

Fixing the b-tagging performance at the 77% WP for both the $t\bar{t}$ and Z', Figure 5.33 scans the f_c^b flavour fractions for the different models. A clear hierarchy of performance 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 increased 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.

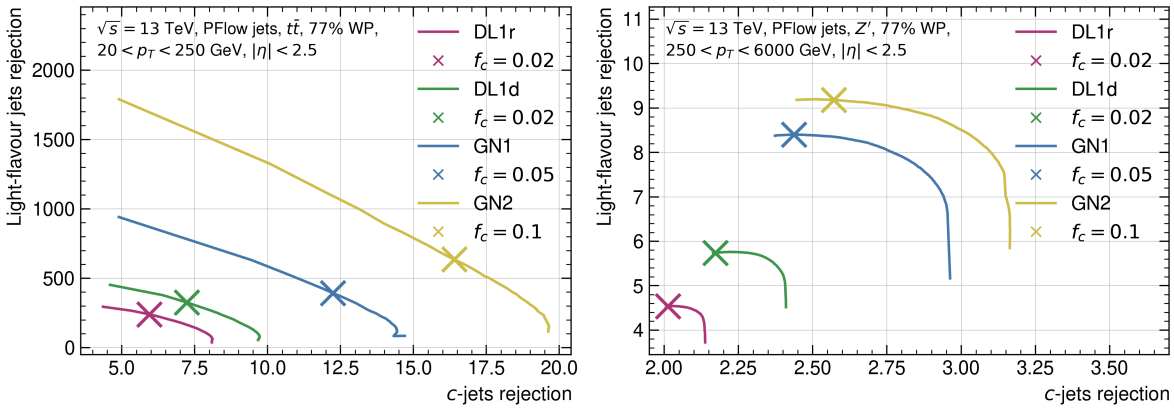


Figure 5.33: The flavour fraction f_c^b scans for b-tagging at a fixed WP of 77% of the different models considered evaluated on the $t\bar{t}$ (left) and Z' (right). The chosen values are marked on the curves, displaying on the x-axis the c-rejection vs the light-rejection on the y axis. Increasing f_c^b shifts the marker rightwards along the curves.

Figure 5.34 displays the flavour fraction f_b^c scans for c-tagging at the 30% WP. 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 comparatively easy identification of b-jets giving them an outlying distribution compared to the overlap of c- and light-jets.

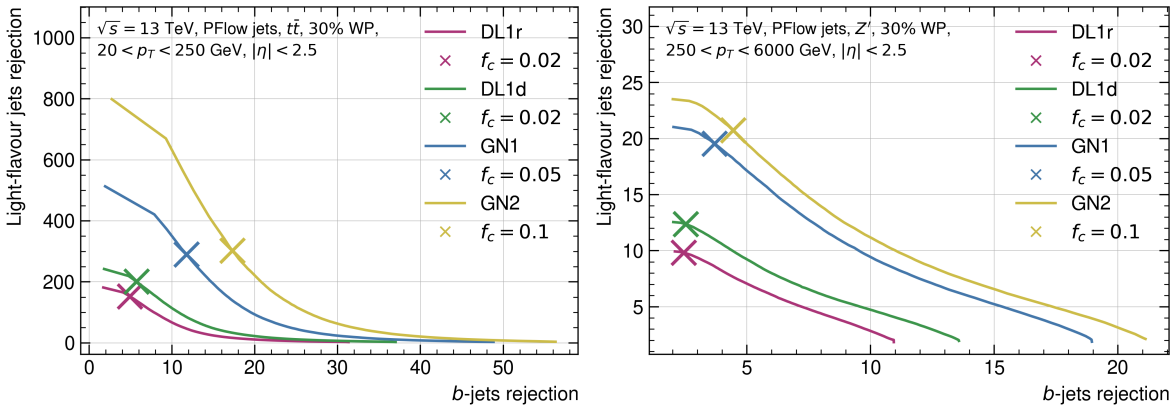


Figure 5.34: The flavour fraction f_b^c scans for c-tagging at a fixed WP of 30% of the different models considered evaluated on the $t\bar{t}$ (left) and Z' (right). The chosen values are marked on the curves, displaying on the x-axis the b-rejection vs the light-rejection on the y axis. Increasing f_b^c shifts the marker rightwards along the curves.

Figure 5.35 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 A.6 in Appendix A.5 displays the same information for c-tagging, leading to the same conclusions.

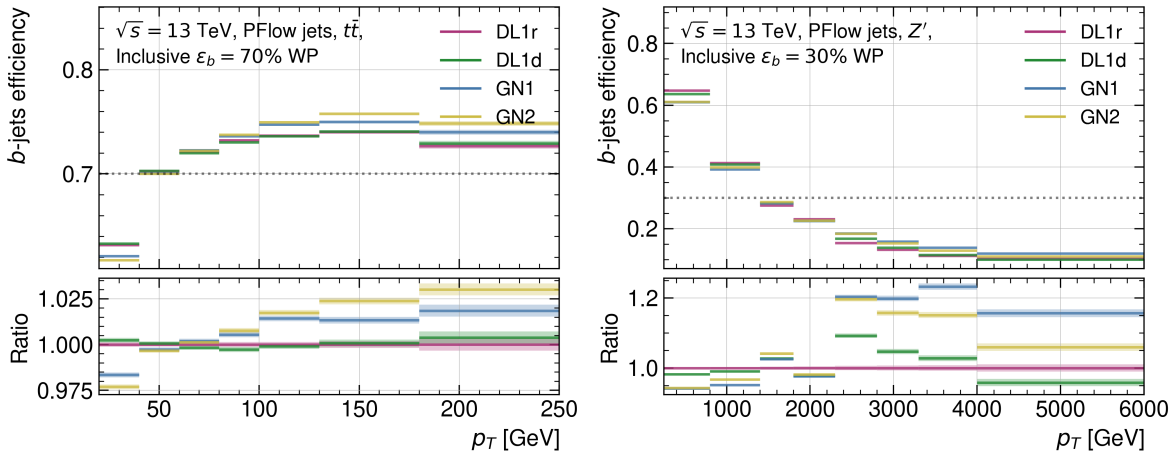


Figure 5.35: Comparing the different models b-tagging efficiency as a function of jet p_T for the inclusive b-tagging 70% WP on the $t\bar{t}$ (left) and 30% WP on Z' (right). The flavour fraction is set at $f_c^b = 0.018$ for DL1r and DL1d, 0.05 for GN1, and 0.1 for GN2.

To avoid biasing the analysis of the results with this per bin performance dependency, Figure 5.36 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 A.7 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 5.37 and 5.38 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 A.8 and A.9 of the appendix. The b- and light-rejection at the 30% c-tagging per bin WP are displayed in Figures 5.39 and 5.40 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 is 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 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 1 million parameters threshold. The latest design of GN2 uses 2.6M parameters, and some tests have raised this number to ∼70M parameters. Expert knowl-

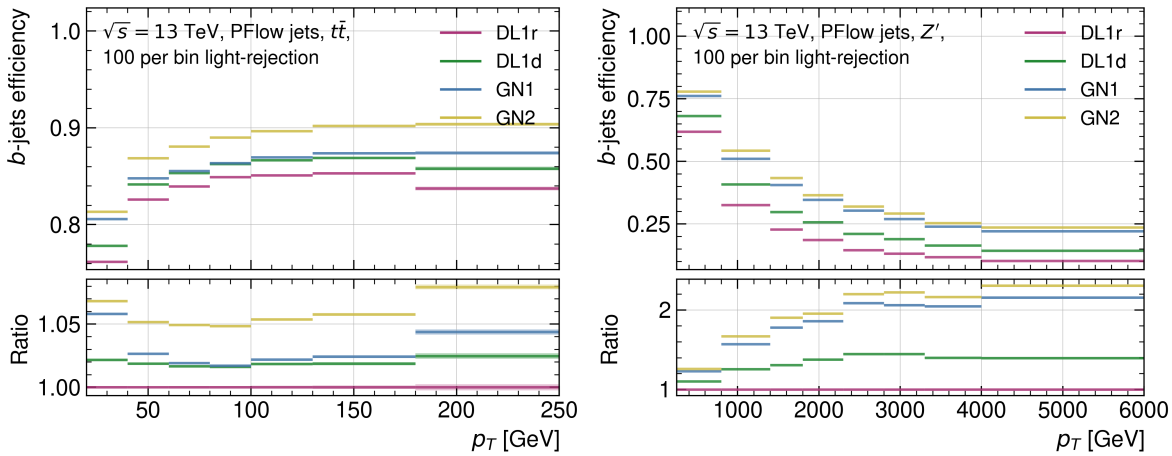


Figure 5.36: Comparing the different models b-tagging efficiency as a function of jet p_T at a fixed 100 light-jet rejection per bin on the $t\bar{t}$ (left) and Z' (right) test samples. The flavour fraction is set at $f_c^b = 0.018$ for DL1r and DL1d, 0.05 for GN1, and 0.1 for GN2.

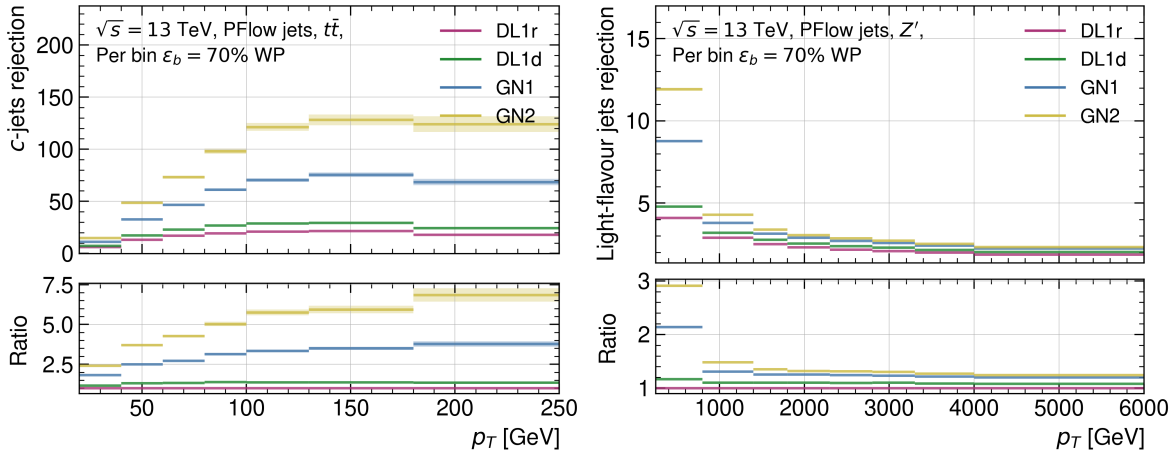


Figure 5.37: Comparing the different models c-rejection as a function of jet p_T for the b-tagging 70% WP per bin on the $t\bar{t}$ (left) and the 30% WP per bin on Z' (right). The flavour fraction is set at $f_c^b = 0.018$ for DL1r and DL1d, 0.05 for GN1, and 0.1 for GN2.

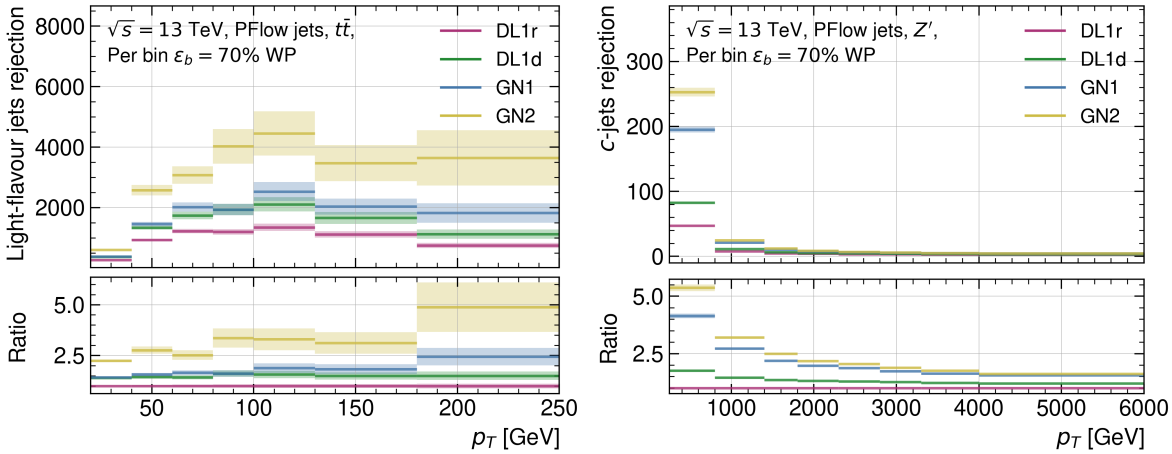


Figure 5.38: Comparing the different models light-rejection as a function of jet p_T for the b-tagging 70% WP per bin on the $t\bar{t}$ (left) and the 30% WP per bin on Z' (right). The flavour fraction is set at $f_c^b = 0.018$ for DL1r and DL1d, 0.05 for GN1, and 0.1 for GN2.

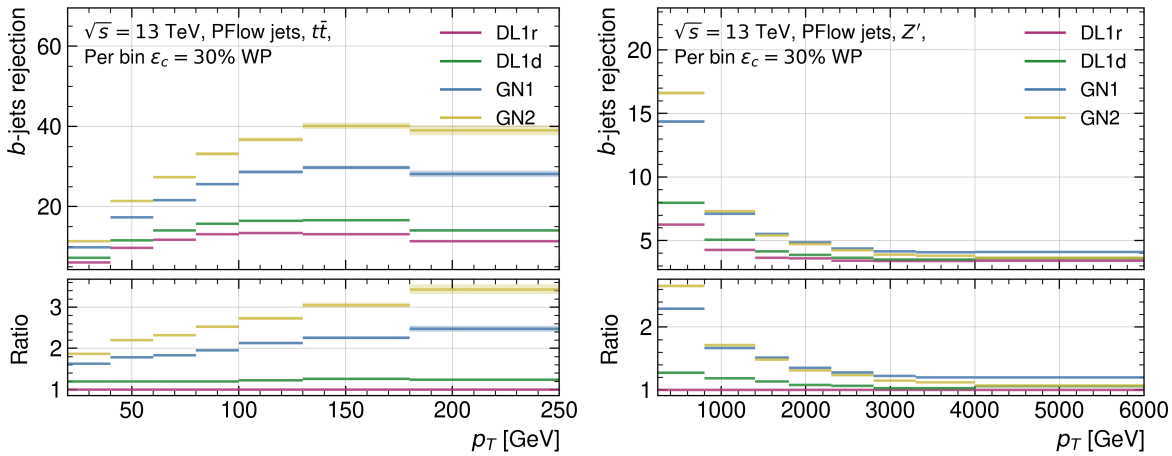


Figure 5.39: Comparing the different models b-rejection as a function of jet p_T for the c-tagging 30% WP per bin on the $t\bar{t}$ (left) and Z' (right). The flavour fraction is set at $f_b^c = 0.2$ for all taggers.

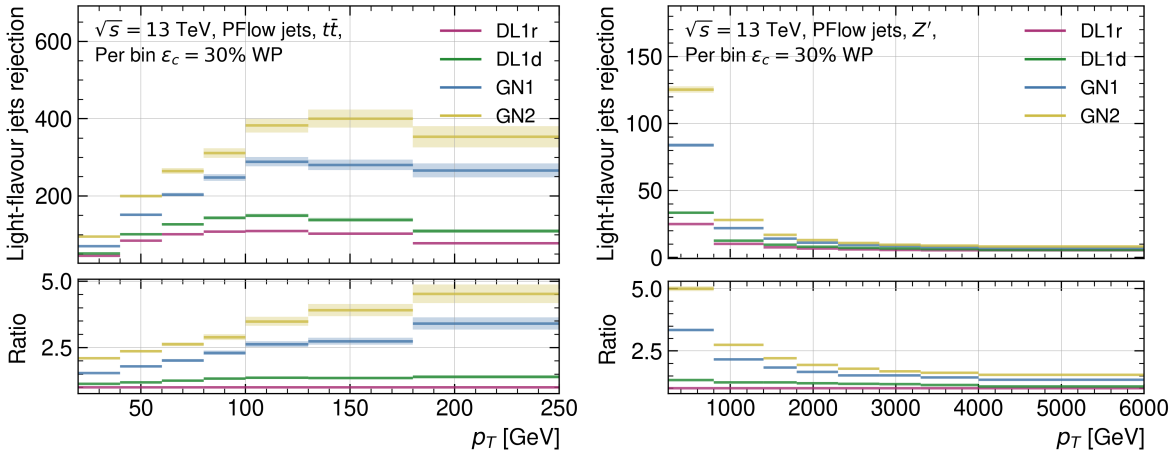


Figure 5.40: Comparing the different models light-rejection as a function of jet p_T for the c-tagging 30% WP per bin on the $t\bar{t}$ (left) and Z' (right). The flavour fraction is set at $f_b^c = 0.2$ for all taggers.

edge 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. Thanks to the flexibility of the SALT software, GN2 has been successfully specialised for boosted object tagging, with the GN2X tagger presented in Appendix A.6 designed to replace X_{bb}.

5.3.3 GN2 Hyperparameter 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.

- 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 [26]. 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 batch size 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 batch size, 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.

Complex network require hardware accelerators such as GPUs to be effectively trained. 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* [163]. KubeFlow 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 to automate and refine the strategy to test and converge on the best hyperparameters [164]. 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 has been 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 currently accessible to LHC 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 synchronised on the different GPUs. A fast connexion between these GPUs is essential, as is having fast read access to the full dataset. Distributing the computation across HPCs 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. Most private and public cloud providers, such as Google Cloud, Amazon Web Service, and Microsoft Azure, are KubeFlow-compatible and host a larger amount of state-of-the-art GPUs. SALT can be effectively deployed on the infrastructure of these cloud providers or 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 [165]. 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. [166] 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 [120]. The default or *standard* parametrisation (SP) follows the so-called LeCun parametrisation [167]. This parametrisation, routinely deployed in ML frameworks such as PyTorch [100], 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	$\eta \times \boldsymbol{d}^{\text{out}}_{\boldsymbol{L}_{\text{inp}}}$
$w^{L_{\text{hid}}}$	$\sim \mathcal{N}\left(0, \frac{1}{d^{\text{in}}_{L_{\text{hid}}}}\right)$	$\sim \mathcal{N}\left(0, \frac{1}{d^{\text{in}}_{L_{\text{hid}}}}\right)$	η	$\eta / \boldsymbol{d}^{\text{in}}_{\boldsymbol{L}_{\text{hid}}}$	η	η
$w^{L_{\text{out}}}$	$\sim \mathcal{N}\left(0, \frac{1}{d^{\text{in}}_{L_{\text{out}}}}\right)$	$\sim \mathcal{N}\left(0, \frac{1}{d^{\text{in}}_{L_{\text{out}}} \times \boldsymbol{d}^{\text{in}}_{\boldsymbol{L}_{\text{out}}}}\right)$	η	$\eta / \boldsymbol{d}^{\text{in}}_{\boldsymbol{L}_{\text{out}}}$	η	$\eta / \boldsymbol{d}^{\text{in}}_{\boldsymbol{L}_{\text{out}}}$
$b^L \; \forall L$	0	0	η	η	η	$\eta \times \boldsymbol{d}^{\text{out}}_{\boldsymbol{L}}$

Table 5.9: Comparing the Standard Parametrisation (SP) to the Maximal Update Parametrisation (μP), as defined in Ref. [165] based on the work of Ref. [166].

This particular derivation of μP, taken from Ref. [165], is equivalent to the original μP derivation introduced in Ref. [166]. 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 $\sqrt{d_k} \to d_k$ to properly scale with width [165]. Figure 5.41 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 absolute 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 $lr = 10^{-2}$).

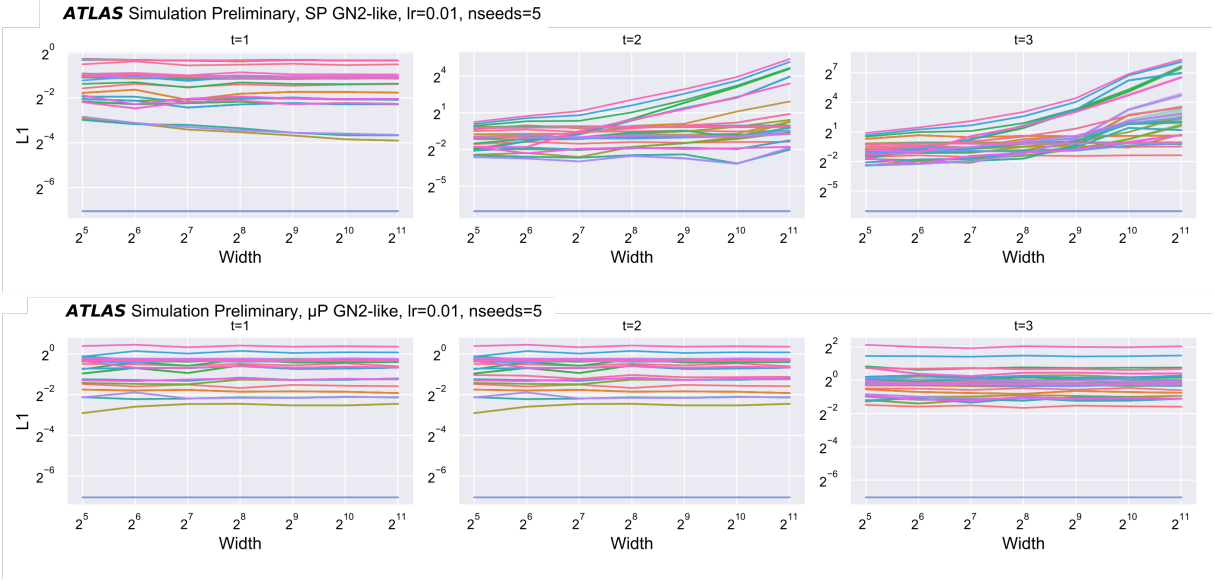


Figure 5.41: The sum of the absolute value of the pre-activation weights for the different layers in the initialiser and transformer parts of a GN2-like model in standard parametrisation (SP - top) and in μP parametrisation (bottom), at three timesteps: initialisation ($t = 1$ - left), after one training step with $lr = 10^{-2}$ ($t = 2$ - centre), and a second training step ($t = 3$) [7]. 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 5.41, 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 [166]. 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) [165]. 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 over-training 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 [165]:

- 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[2], the depth, and the batch size. Only the scaling along width is theoretically proven thanks to μP, while the others are empirically observed to hold [165].

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.

To demonstrate the effect of μP on GN2, a hyperparameter optimisation campaign of the initial and maximal value of the learning rate[3] is performed using the standard and maximal update parameterisation, SP vs μP. Three embedding widths are considered: the nominal 256, defining a GN2 model with 2.3M parameters, a mid-size 128 width with 0.72M parameters, and a small 64 width model with 0.23M parameters. Interestingly, this smaller model with an embedding width 1/4 of the full model only has a 1/10 of the parameters. Furthermore, the small model is 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 jets[4] for 40 epochs with batch size 1024. Parameters not mentioned are kept similar

[2] Number of neurons per layer, number of attention heads in a transformer, ...
[3] The final value, an LR end of 10^{-5}, is kept fixed in all tests due to the limited compute available.
[4] Composed of 60% $t\bar{t}$ and 40% Z'.

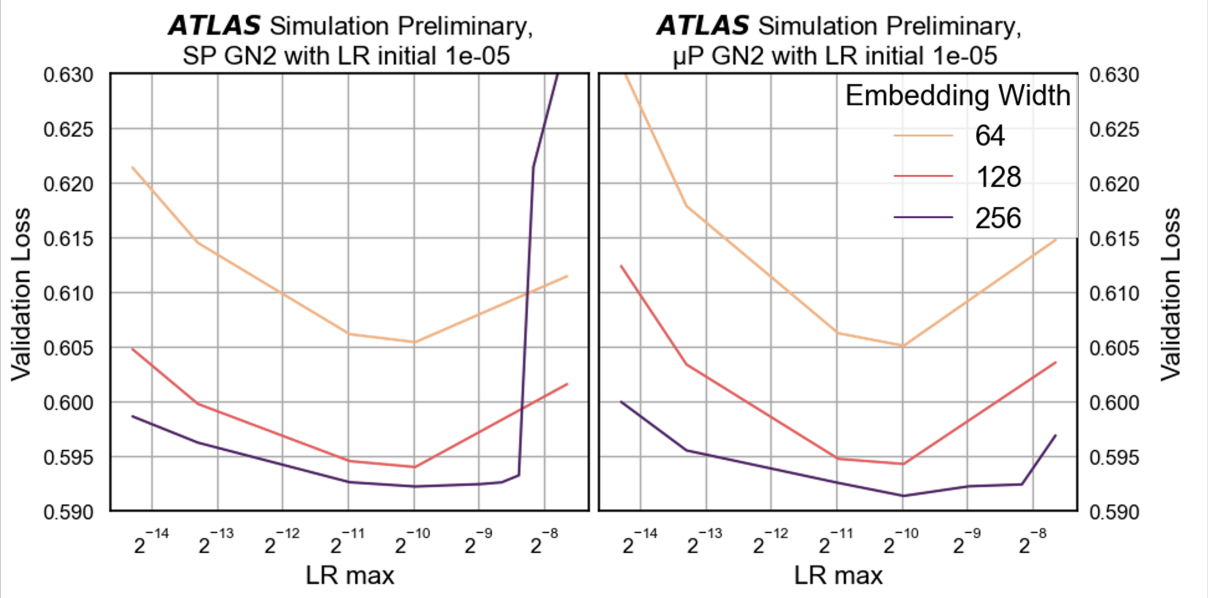


Figure 5.42: Comparison of a maximal learning rate value scan at an initial learning rate value of 10^{-5} for an SP (left) and a μP GN2 models (right) for three different embedding widths: 64 (yellow), 128 (red), and 256 (purple). The y-axis displays the validation loss attained. Taken from [7].

between embedding widths and parametrisations, and the epoch giving the lowest validation loss is chosen from each run. Figure 5.42 outlines the main result of this study, displaying the various LR max considered at the best LR initial found (10^{-5}). Three main observations are drawn from these results:

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 5.43) are shared across widths with μP, while no such behaviour is guaranteed for SP - but is observed in the present case.

The full LR init vs LR max scans can be found in Figure 5.43 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 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

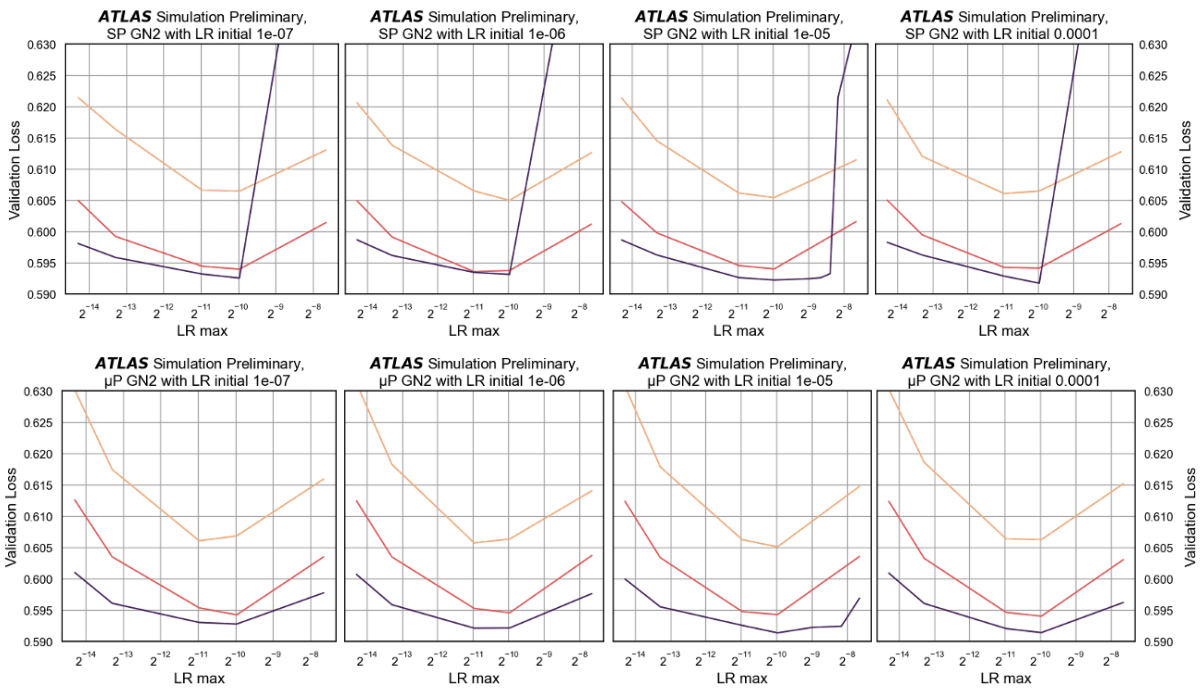


Figure 5.43: Scan of the maximal learning rate (x-axis) versus initial learning rate (individual column) as measured by the validation loss (y-axis) of SP models (top) and the μP model (bottom) with three different embedding widths: 64 (yellow), 128 (red), and 256 (purple). Taken from [7]. The scan at LR initial = 10^{-5} benefitted from more tests to capture the sudden rise in validation loss at larger LR max for SP.

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 = 5×10^{-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 5.44, with the suboptimal one being the worst performing full-width model (LR max = 5×10^{-5}, LR init = 10^{-7}) and the optimal one the best performing one (LR max = 5×10^{-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 WP 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 WP of 30% - which is equivalent to the 30% WP on $t\bar{t}$.

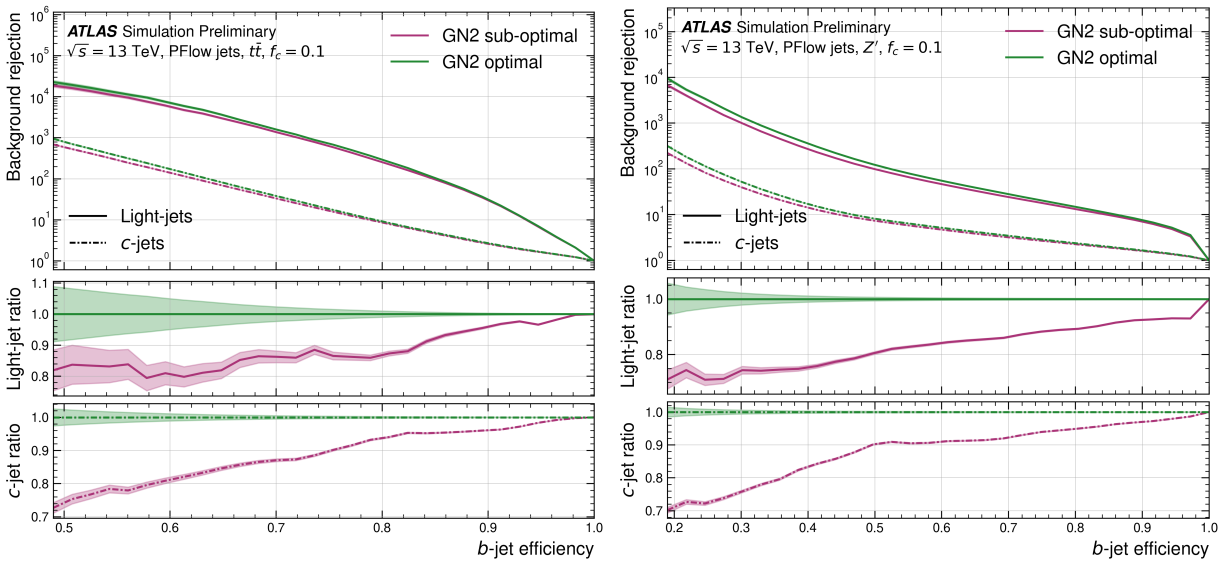


Figure 5.44: The c- and light-rejections as a function of the b-jet tagging efficiency in the $t\bar{t}$ (left) and Z' (right) test samples [7]. Models compared are the optimal μP GN2 (LR max $= 5 \times 10^{-5}$, LR init $= 10^{-5}$) and the suboptimal μP GN2 (LR max 5×10^{-5}, LR init 10^{-7}), all 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 [165]. 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 5.9.

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.

5.4 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 [150] 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, as suggested in Figure 5.45.

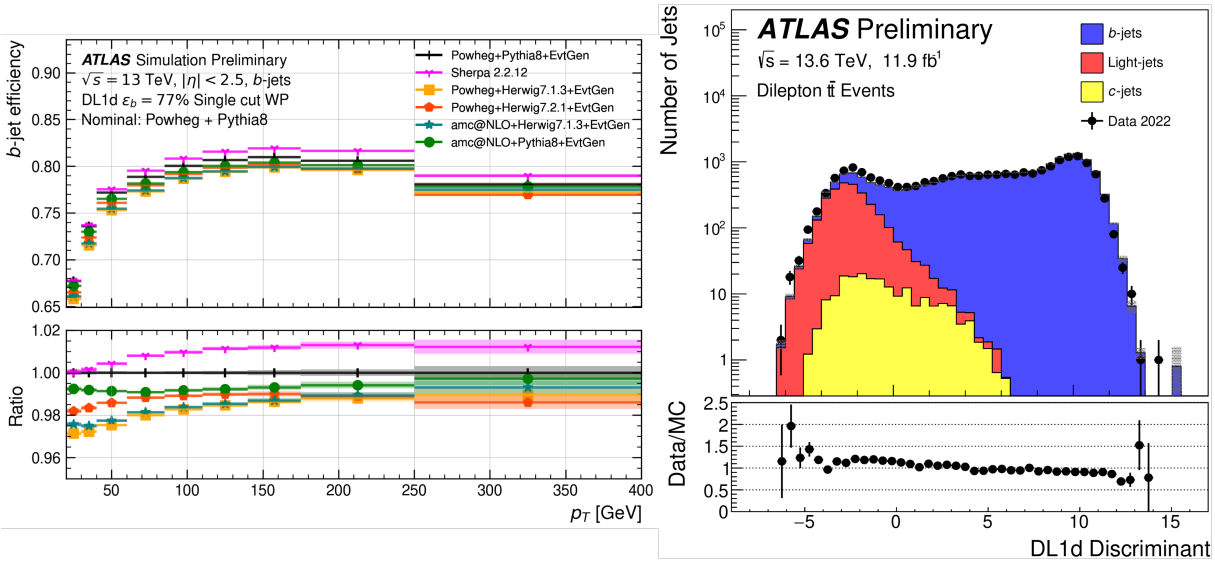


Figure 5.45: Generator dependency of the b-efficiency as a function of jet p_T (left) and Data/MC agreement of the b-discriminant D_b on a dileptonic $t\bar{t}$ Run 3 sample (right) for a calibrated DL1d b-jet tagger [5]. Changing the generator a noticeable impact on the performance of the tagger, and the ratio of alternative generator to the nominal is used to derive MC-MC scale factors. Data-MC scale factors are derived from the Data/MC agreement of the DL1d discriminant on dedicated samples enriched in different jet species, such as the dileptonic $t\bar{t}$ process for b-jets. The different generators are described in Section 6.4.

From these studies, two types of SFs are produced. Data-MC SFs are derived by comparing the output of the tagger on a simulated and real dataset using the same selection [126, 168–170]. The efficiencies ϵ^f for each flavour $f \in b, c,$ 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 mistag ging 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. [126]. The SF for c-jet misstagging is calibrated on a $t\bar{t}$ sample decaying to exactly one charged lepton and several jets [170]. Finally, the SF for light-jets is derived in a sample of Z bosons produced in association with jets (Z+jets) [171]. 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.

To probe the effect of different MC simulators, MC-MC SFs are derived between the chosen nominal Monte Carlo simulator used for training and alternative simulators or by changing the tuning [172]. This dependency is measured by applying the same tagger to samples simulated with different generators, mainly Pythia [143], Herwig [173], and Sherpa [174] for variation to the parton shower and hadronisation and MadGraph for variation to the matrix element [175]. The decay chains of b- and c-hadrons in ATLAS are further simulated with the EvtGen package [148]. 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

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

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. [5], 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 WP for both b- and c-tagging compared to the at-the-time main tagger DL1r. Excitingly, a new family of taggers based on a graph attention network for GN1 and a transformer encoder for GN2 is presented, describing the motivation and specificities behind the design. The performance of the different methods are compared, highlighting a promising increase in efficiency for the taggers developed for the Run 3 of the LHC. Efforts to perform to optimise the hyperparameter of GN2 are introduced, addressing the possibilities of a new CERN infrastructure built on KubeFlow as well as the relevance of the maximal update parametrisation to boost the search for the best hyperparameters.

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 was to discover the Brout-Englert-Higgs boson (Higgs - H), a feat achieved by the ATLAS and CMS Experiments in July 2012 [14, 15]. Theorised in 1964 by two independent papers introducing the mechanism of spontaneous symmetry breaking to give mass to the gauge bosons [12, 13], 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. 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 acquire theirs through Yukawa interactions with the Higgs fields [32]. The scale of the interaction for each fermion f is set by an associated Yukawa coupling y_f. These couplings are fundamental parameters of the SM depending on the quark masses and the Higgs field vacuum expectation. This chapter is dedicated to an ATLAS measurement of the y_b and a search of the $H \to c\bar{c}$ decay mode.

6.1 Introduction

The Higgs boson H [12, 13, 176, 177] was discovered in 2012 by the ATlAS and CMS Collaborations using the data of the LHC Run 1 [14, 15]. This triggered a race by both experiments to study the specific properties of the discovered particle, and in particular to observe the different production and decay modes presented in Chapter 2.2. The initial decay channels studied for the discovery were 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 [178, 179]. Additionally, the decay of Higgs bosons to a pair of τ-lepton is now well established and different cross-section measurements have been performed [180, 181]. Importantly, the decay channel of the Higgs boson to a $b\bar{b}$ pair was observed by both ATLAS and CMS [182, 183]. This last decay channel is of particular significance since it has the largest predicted branching ratio of 58% for an SM Higgs with a mass of 125 GeV.

Concerning the second generation of fermions, there is 3 σ evidence of the decay to a $\mu^-\mu^+$ pair by CMS [184] and a 2 σ excess over the background-only hypothesis by ATLAS [185]. Additionally, 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 [129]. This decay mode is the most common Higgs decay mode that has yet to be observed. It is particularly challenging due to the combination of a small predicted 2.9% branching ratio [186], large background rates, and the experimental difficulties in identifying c-jets. It is a fertile ground for new physics beyond the SM due to the smallness of the predicted c-quark Yukawa coupling $y_c \approx 3.99 \times 10^{-3}$ [187] as well as an important test of the validity of the model [16–22]. The Yukawa couplings in the SM are largely added ad-hoc and do not explain the distinct mass hierarchy between the three generations of fermions. This open problem is 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 scrutinises the hierarchy of mass between the b- and c-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 decay branching ratio at the observed Higgs mass, the large multi-jet background in a hadron collider makes this decay mode very challenging. The measurements for both the $b\bar{b}$ and $c\bar{c}$ decay modes are therefore performed in the *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, allowing for efficient triggering and greatly reducing 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 are less sensitive to the Higgs coupling to heavy-flavour quarks. The $VH(H \to b\bar{b})$ and $VH(H \to c\bar{c})$ ATLAS analyses adopt very similar strategies. The main ingredient is to reliably tag the flavour of jets produced in an event to reconstruct the heavy-quark pair produced in the H decay, with the taggers described in Chapter 5.

Using the Run 2 dataset with an integrated luminosity of 139 fb^{-1}, the published $VH(H \to c\bar{c})$ ATLAS analysis obtained an observed (expected) upper limits on the $VH(H \to c\bar{c})$ signal strength of 26 × SM (31 × SM) [130]. The measurement also provided the first constraint on the

Higgs-charm coupling modifier $|\kappa_c| < 8.5$. For comparison, CMS reported an observed (expected) upper limit of 14.4 × SM (7.6 × SM) and a constraint of $|\kappa_c| < 3.4$ on the coupling modifier [131].

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 [188]. 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* [188] and the *boosted* [189] 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 strategy 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 approach: 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 [188]. 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 [189].

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. [130] and the resolved and boosted $VH(H \to b\bar{b})$ combined[1] in Ref. [191]. 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 constraining of important backgrounds such as the V+jets and top-quark processes.

[1] CMS published an analogous combination in Ref. [190].

- 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, the discriminant, and flavour tagging.

This chapter details the current state of the $VH(H \to b\bar{b}/c\bar{c})$ analysis which, at the time of writing, is not yet concluded and still blinded. The stage described corresponds to that attained at the end of the third unblinding approval review. Some modifications to the analysis are expected in the soon-to-be-published final result, in particular to the modelling strategy and the fit frameowrk. The work presented here is largely based on the internal documentation of the experimental team and personal results produced during the duration of the research 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. The regions and boundaries between the different regimes of the analysis are illustrated in Figure 6.1. $VH(H \to b\bar{b})$ and $VH(H \to c\bar{c})$ are separated by the required presence of two b-tagged jets or a c-tagged jets. The p_T^V cut marks the change of the Higgs candidate reconstruction strategy from the resolved to the boosted $VH(H \to b\bar{b})$: two b-tagged small radius (R = 0.4) jets for $p_T^V <$ 400 GeV, otherwise one large radius (R = 1) jets with two b-tagged track-jets associated to the large-R jet.

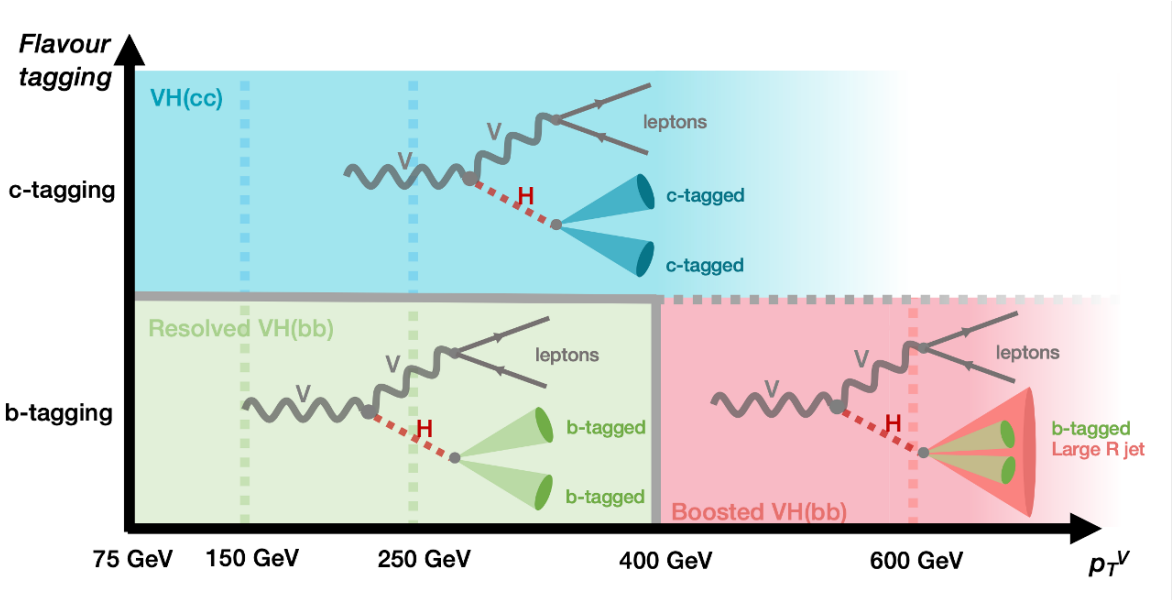


Figure 6.1: The analysis regimes considered in the combined $VH(H \to b\bar{b}/c\bar{c})$ analysis.

For each analysis regime, 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 signals 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 with good statistical significance and offer a test to verify the validity of the strategy adopted. The main backgrounds are the production of a vector boson with additional jets (V+jets) 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). Minor backgrounds are the QCD multi-jet, the single-top process (without an 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. 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-leptons). 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 simulation packages, except for the multi-jet in 1L and the top background in 2L that are estimated from data-driven methods. The multi-jet is only included in the 1-lepton channel, as it is negligible in the other channels.

This chapter is separated into different sections introducing the datasets and MC simulations (Section 6.4), the object and event selection and categorisation (Section 6.5), the analysis discriminants (Section 6.6) the experimental and processes modelling (Section 6.7 and 6.8), the fit framework (Section 6.9.1), and finally the main results (Section 6.9.2).

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 140.1 fb^{-1} [47]. Data events passing some quality requirement are selected, ensuring for example that all subdetectors were correctly operating. The analysis requires extensive and accurate MC 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 [149] using GEANT4 [150]. The nominal samples are produced with 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-section precision. 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 [66]. The EVTGEN 1.6.0 program is used to simulate the properties of b- and c-hadrons decays[2] [148]. 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 [143]. The rest of this section gives more details about the simulation of the different processes. When

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

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.

relevant, alternative samples generated from a different setup to the nominal samples are introduced. These alternative samples are used to assess modelling uncertainties in Section 6.8.1, as summarised in Table 6.14.

6.4.1 Signal Processes

The analysis targets the $VH(H \to b\bar{b})$ and $VH(H \to c\bar{c})$ processes as *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 [140, 141]. The qq-initiated VH samples are simulated with the POWHEG generator with the multiscale improved NLO (MiNLO) procedure [192], with one-loop amplitudes computed with the GoSam automated software [193]. 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 [143]. Both use the AZNLO tune [194] with PDFs based on the NNPDF3.0NLO for matrix elements [142].

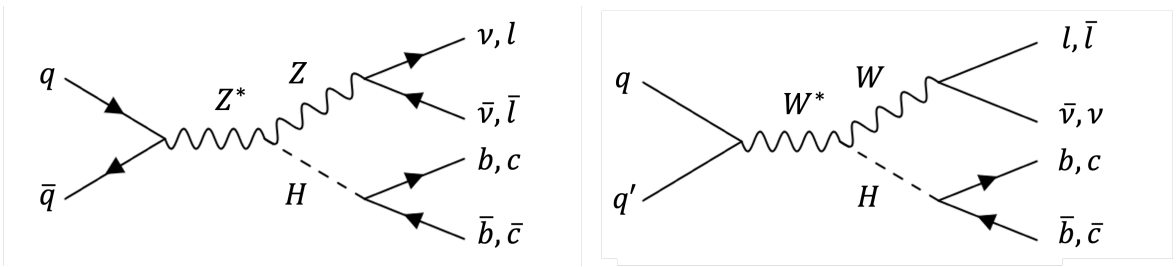


Figure 6.2: Leading order Feynman diagrams for the qq-initiated $VH(H \to b\bar{b}/c\bar{c})$.

The inclusive cross-sections for WH and ZH are calculated at NNLO in QCD [195] and NLO in Electroweak (EW) [142]. The gg-initiaed ZH contribution relies on the LO prediction from

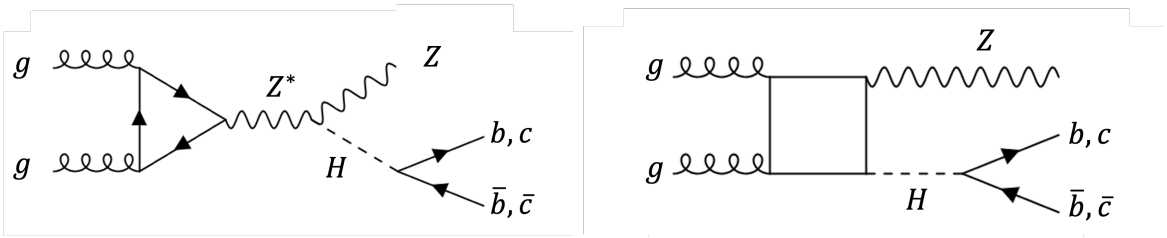


Figure 6.3: Feynman diagrams of the gg-initiated contributions to $ZH(H \to b\bar{b}/c\bar{c})$.

POWHEG instantiated with PYTHIA 8.

Alternative samples are simulated with POWHEG+MiNLO+HERWIG 7.0, with the same simulation stack as the nominal samples but replacing PYTHIA 8 by HERWIG 7.0 [173] for the simulation of the PS, hadronisation, UE, and multiple parton interactions.

6.4.2 Background Processes

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 [196], 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 [142]. 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.

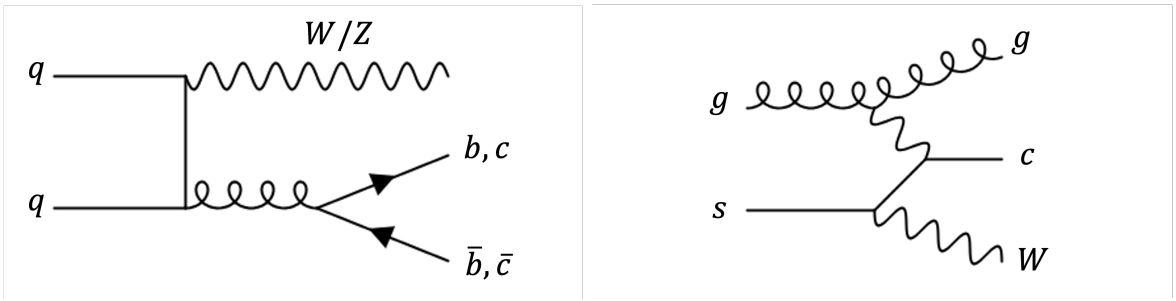


Figure 6.4: Leading order Feynman diagrams of the V+jets process. The left diagram gives jet pairs of the same flavour due to the gluon splitting, while the right one can give mixed flavours.

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 [175]. 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 PS and hadronisation, with the A14 tune and the NNPDF2.3LO PDF set with $\alpha_s = 0.13$.

- SHERPA 2.2.1 [174] samples are used as alternative as they give different p_T^V distributions to SHERPA 2.2.11 [197], an important modification given the observed data-MC disagreements in the p_T^V distributions. These samples are similar to those used in the standalone $VH(H \to c\bar{c})$ [130].

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 [138, 139]. 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 [198].

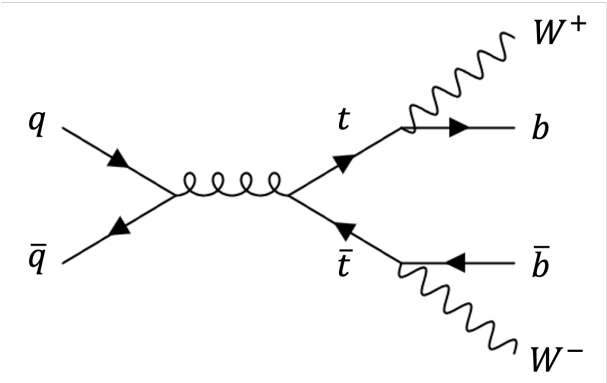


Figure 6.5: Feynman diagrams of the $t\bar{t}$ production and decay.

Alternative samples Several alternatives are simulated for modelling studies:

- Replacing PYTHIA by HERWING 7.0 with H7UE tune [199] while keeping the same nominal POWHEG setup. This sample is used to systematically assess variations to the parton shower, hadronisation, and underlying event modelling.
- Replacing POWHEG by MADGRAPH5_AMC@NLO [175] 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 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:
  - High- and low-variations of Initial State Radiation (ISR), where the μ_R and μ_F scales are doubled and halved.
  - Up- and down-variations of Final State Radiation (FSR), obtained by doubling (halving) the renormalisation scale μ_R.

Single-top Production

The single-top process combines different channels, with the leading Feynman diagrams depicted in Figure 6.6. The dominant contribution is the associated top-production Wt channel, with the $t \to Wb$. The two other contributions are the t- and s-channel, with the former having a larger cross-section than 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 t- and s-channels [200, 201], and with approximate NNLO accuracy from NNLL soft-gluon resummation for the fiducial Wt production cross-section [202, 203].

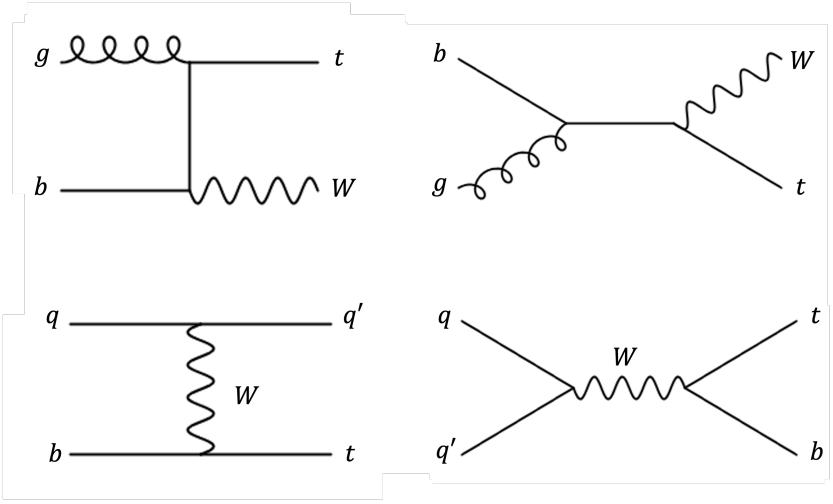


Figure 6.6: Feynman diagrams of the Wt-production (top) and the single top production (bottom) in the t-channel (left) and the s-channel (right).

The Wt production 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 $t\bar{t}$ contributions in the NLO Wt cross-section calculation [204].

Alternative samples are produced for the single-top Wt- and t-channels[3]:

- 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 a different overlap removal procedure is produced with the diagram removal (DR) scheme [204] to systematically model the overlap 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.

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 samples

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

are simulated similarly to the V+jets, using SHERPA 2.2.11 [196]. The gg-initiated processes are simulated with the older SHERPA 2.2.2 version. The cross-sections are computed at NLO precision, with the NNLO PDFs based on NNPDF3.0NNLO [142] for both the matrix element and parton shower.

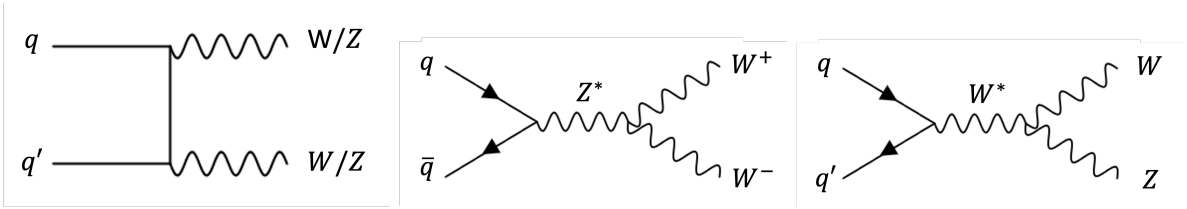


Figure 6.7: Feynman diagrams of the diboson production in the t- (left) and s-channel (centre & right). The t-channel can lead to any combination of W and Z depending on the initial quark-pair.

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

As described in Chapter 3.2, the data collected by ATLAS consists of low-level information measured from various subdetectors. Different processing steps, collectively referred to as *reconstruction*, are applied to unlock high-level physically interpretable objects. This section introduces the specific object recontruction techniques used in the analysis. From this, the 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 the different regions of the analysis regions.

6.5.1 Object Selection

As outlined in Chapter 3.3, the ATLAS software supports several object reconstruction techniques. The reconstruction strategies relevant to the $VH(H \to b\bar{b}/c\bar{c})$ analysis are presented in this section.

Primary Vertex: all events considered in the analysis are required to have at least one primary vertex reconstructed from tracks in the ID [75], as detailed in Section 3.3.2.

Electrons: are reconstructed by matching a deposit in the electromagnetic calorimeter with a track in the ID [77, 205], as described in Section 3.3.3. Electrons are required to have $p_T > 7$ GeV and $|\eta| < 2.47$. They are identified with a *loose* working point of the 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 calorimeters. In 1L, the *tight* likelihood criterion is used with stricter calorimeter isolation requirements to better reject the multi-jet background. Additional requirements on the electron selection depend on the lepton channel, as 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}	$\lvert\Delta z_0 \sin\theta\rvert$	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 Muon Spectrometer (MS) with information from the ID [206], as detailed in Section 3.3.4. 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}	$\lvert\Delta z_0 \sin\theta\rvert$	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 [85], as presented in Section 3.3.6. 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: as described in Section 3.3.7, neutrinos are not detectable in ATLAS and their presence is inferred from momentum imbalance in the transverse plane. E_T^{miss} is calculated as the negative vectorial sum of the transverse momentum of physics objects (electrons, muons, hadronic τ, and jets), with an additional track-based *soft term* from unassigned good-quality tracks [86].

Jets Three types of jets introduced in Section 3.3.5 are used in the analysis, all reconstructed with the anti-k_t algorithm [78]:

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 [207]. *Tight* jet cleaning criteria are applied to suppress non-collision background. Central jets are used in the resolved regime to reconstruct the Higgs candidate with flavour tagging.

2. *Large-R jets*: similar to small-R jets with a larger radius $R = 1.0$, they are required to have $p_T > 250$ GeV and $|\eta| < 2$, and 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 of the boosted $H \to b\bar{b}$ decay [208]. They must 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.

As outlined in Section 3.3.5, jets benefit from extensive corrections and calibrations to improve their reconstructed mass, energy, and axis direction.

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 for both b- and c-tagging in the resolved and boosted regime [136]. The methodology differs slightly between the two regimes of the analysis due to the different flavour tagging needs.

- In the resolved regime, 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 [130]. The DL1r tagger assigns

a b-tagging and a c-tagging discriminant score[4] from Equations 5.1 and 5.2 to selected central jets. To tag a jet, the associated score must be higher than a specific cutoff value defining a Working Point (WP) selection efficiency. Jets can be assigned one of 5 possible labels based on 2 b-tagging and 2 c-tagging WPs, as outlined in Figure 6.8. These WPs are tested in a strict successive order, with first a 60% *tight* b-tagging point (bin 4) followed by a *looser* 70% b-tagging WP (bin 3). A jet passing these selections is labelled B[5]. 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 any WP is not tagged and labelled N (bin 0). The b-tagging WPs correspond to official ATLAS ones for DL1r [136], 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 analysis following the methodology described in Ref. [136].

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 efficiencies for b-, c-, light- and τ-jets in the Pseudo-Continuous Flavour Tagging (PCFT) scheme, measured in a POWHEG+PYTHIA 8 sample of semi-leptonic $t\bar{t}$ events.

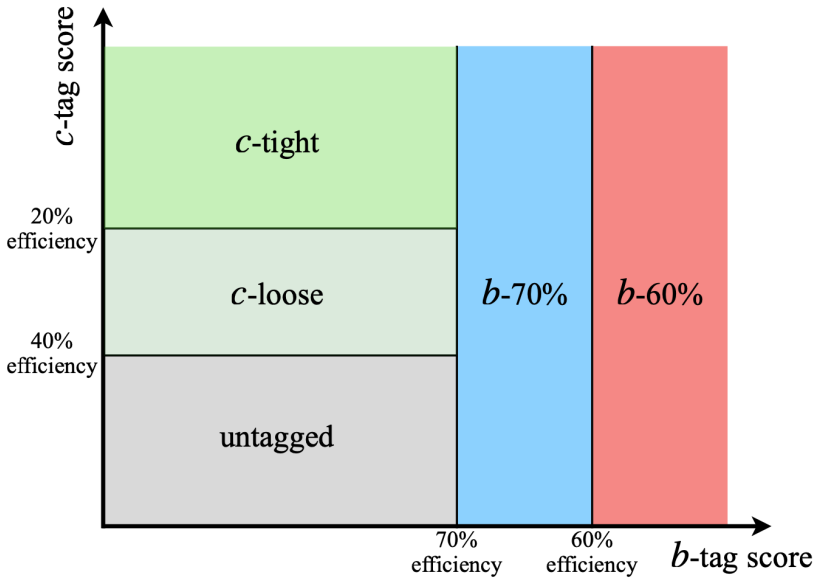


Figure 6.8: The pseudo-continuous flavour tagging scheme defining simultaneously 2 b-tagged, a tight c-tagged, a loose c-tagged, and a non-tagged bins.

- 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 [136]. The track-jets associated with

[4] With flavour fractions set as $f^b{}_c = 0.018$ and $f^c{}_b = 0.3$, respectively.
[5] The difference between these b-tagged is used in the discriminant MVAs of the analysis

the leading large-R jet are given a b-tagging score based on the per-flavour probabilities predicted 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. [136] is used and extended to higher p_T with uncertainty extrapolation due to the large range of p_T probed in the analysis.

The superior single-jet GNN taggers introduced in Chapter 5 or the boosted decay tagger GN2X [6] presented in Appendix A.6 were not available during the analysis, and their calibration is still an ongoing effort. Leveraging the improvements of these GN2-based taggers represents an exciting avenue of progress for future iterations of this study, for both the resolved and boosted regimes.

Object Overlap: is applied to avoid double-counting electrons, muons, small-R and large-R jets, and hadronic τ-leptons passing the object selection.

6.5.2 Event Selection

A subset of all ATLAS recorded events during Run 2 is selected for the analysis based on specific triggers. The trigger selections of $VH(H \to b\bar{b})$ and $VH(H \to c\bar{c})$ are harmonised for the combined analysis, and specified per lepton channel. In 0L, the lowest un-prescaled E_T^{miss} trigger is used 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. The lowest un-prescaled single-electron trigger is deployed for the e-channel. For muons, the E_T^{miss} trigger of 0L is applied for events with $p_T^V > 150$ GeV, while the lowest un-prescaled single-muon trigger is used at 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 increases the signal acceptance by approximately 5%. For leptonic triggers, reconstructed leptons in the event are required to match the triggered leptons.

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 tag hierarchy is introduced, following the ordering: $B > T > L > N$. The pair of candidates is selected from the two central jets having the highest tags, or 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 at most the 3 associated 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, with regime specific selections.

Resolved Higgs candidates: for $VH(H \to b\bar{b})$, the two candidates must be b-tagged (bins 3 or 4) with no additional B- and tight c-tagged jets allowed[6], 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 detailed in the next section, two Control Regions (CRs) are defined by changing this flavour selection: a Top CR, combining at least 1 B-tag with at least 1 T-tag, and the $V + l$ CR 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 candidate. 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 m_{bb} (m_{cc}) must be above 50 GeV before applying energy corrections, to avoid some low-mass V+jets gluon splitting mismodelling.

Boosted Higgs candidates: 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 large-R jet $p_T > 250$ GeV.

The small overlap between the boosted $VH(H \to b\bar{b})$ and resolved $VH(H \to c\bar{c})$ selected events was found to be negligible. 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 6.5 and Table 6.6 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 0L, no VH-loose lepton is allowed and E_T^{miss} should be > 150 GeV (> 250 GeV) in the resolved (boosted) regime, to identify the decay $Z \to \nu\nu$. Additionally, 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 mismodelled region in simulation due to the triggers. In a decay of a $W \to \tau\nu$ followed by a hadronic decay of the τ-lepton 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 in the resolved regiem if at least 1 hadronic τ is reconstructed. The 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* are also applied in all regimes:

- In resolved only, the azimuthal angle between the candidate jets must satisfy $|\Delta\phi(j_1, j_2)| < 140°$.
- The azimuthal angle between E_T^{miss} and the H must satisfy $|\Delta\phi(E_t^{\text{miss}}, H)| > 120°$.

[6] In 2L, additional T-tagged jets are permitted due to the low statistics and the different derivation of the top CR.

- The minimum azimuthal angle between E_T^{miss} and the jets 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 this background negligible in the 0-lepton channel.

Selection specific to the 1-lepton channel

In the 1L channel, the targeted vector boson decay 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[7]. The vector boson is reconstructed from the vectorial sum of the E_T^{miss} and the lepton transverse momentum p_T^lidentified in the event, with $p_T^V > 75$ 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 transverse momentum p_T^V below 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 $Z \to \ell\ell$ bosonic decay, with the Z reconstructed from two VH-loose leptons required to have the same flavour and at least one lepton passes the ZH-signal lepton requirements. In the di-muon channel, the leptons are further required to be of opposite charges[8]. The invariant mass of the di-lepton system is required to be consistent with the Z mass with $81 < m_{\ell\ell} < 101$ GeV in the resolved and $66 < m_{\ell\ell} < 111$ GeV in the boosted regime, to suppress non-resonant lepton-pair producing backgrounds such the $t\bar{t}$ and multi-jet processes. 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.

[7] The first VH-loose lepton corresponds to the WH-signal lepton.
[8] This is not applied to the di-electron channel due to a significantly higher charge misidentification.

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}	
Jets	≤ 4 jets	≤ 3 jets
Additional jets tagging	no T-tag	no B-tag
Top CR tagging	≥ 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	
	μ-channel: single-muon ($p_T^V < 150$ GeV)	
	and 0L E_T^{miss} ($p_T^V > 150$ GeV)	
Jets	≤ 3 jets	
Additional jets tagging	no T-tag	no B-tag
Top CR tagging	≥ 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 6.5: Summary of the event selection in the resolved $VH(H \to b\bar{b}/c\bar{c})$ regime. The resolved 1L and 2L Top CR BT 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).

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 6.6: Summary of the event selection in the boosted $VH(H \to b\bar{b})$ regime.

6.5.3 Event Categorisation

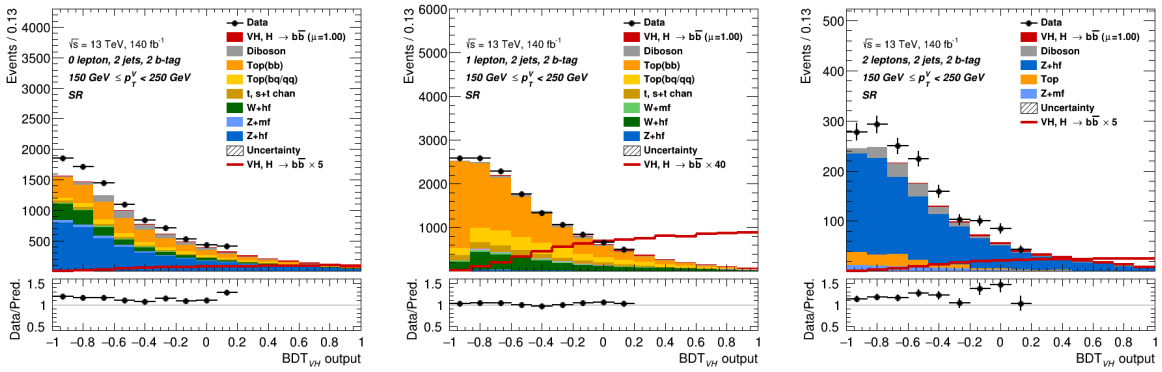
Selected events are finely categorised following a successive decomposition into regions of defined tag, vector boson V transverse momentum p_T^V, and number of jets N_{jet}. The full categorisation gives rise to signal and control regions that enter the statistical analysis defined in the fit framework of Section 6.9. The control regions are defined to help constrain the modelling of specific backgrounds. 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 analysis landscape.

Resolved Regime Categorisation

In the resolved regime, the number of central and 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 to determine the jet multiplicity for the categorisation. This requirement limits the signal migration across STXS bins in $VH(H \to b\bar{b})$, with almost no impact to the $VH(H \to c\bar{c})$ sensitivity. All distributions of the resolved regime regions with processes normalised to their postfit expectations from the fit described in Section 6.9 are presented in Appendix B.5. 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 6.6. The precise definitions of the analysis regions are reviewed in this section.

Resolved $VH(H \to b\bar{b})$ SRs: require 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. 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

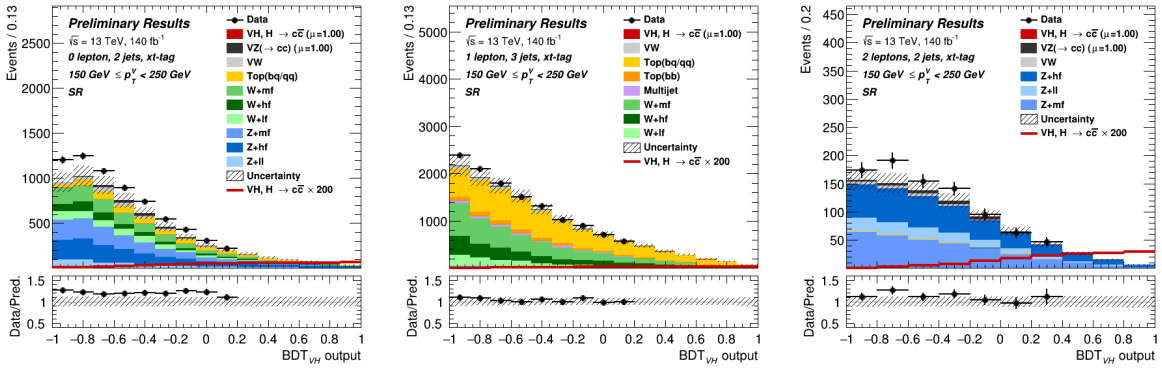
of p_T^V as [75, 150] GeV[9], [150, 250] GeV, and [250, 400] GeV. Some selected $VH(H \to b\bar{b})$ signal regions in the analysis are presented in Figure 6.9.



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

Figure 6.9: The BB-tagged 2-jet 150 GeV $< p_T^V <$ 250 GeV signal regions.

Resolved $VH(H \to c\bar{c})$ SRs: adopt a similar event categorisation to the resolved $VH(H \to b\bar{b})$, with at least one candidate jet being tight c-tagged T. The categorisation of the signal region is then split based on the remaining candidate tag into a 2 c-tags region and a 1 c-tag region. The former requires an extra tight (TT) or loose c-tag (LT)[10], the latter an untagged 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 $\geq$ 250 GeV given the limited impact of the overlap with the boosted $VH(H \to b\bar{b})$. Adding the p_T^V region above 400 GeV was found to increase the total $VH(H \to c\bar{c})$ sensitivity by 10%. The jet multiplicity N_{jet} defines a 2 and a 3 jets categories, with the latter extended to 3 or more jets (3p or $\geq$3) 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 (SRs). The 1 c-tag SRs in the 75 GeV $< p_T^V <$ 150 GeV range are not included in the fit because of their significant background contamination.

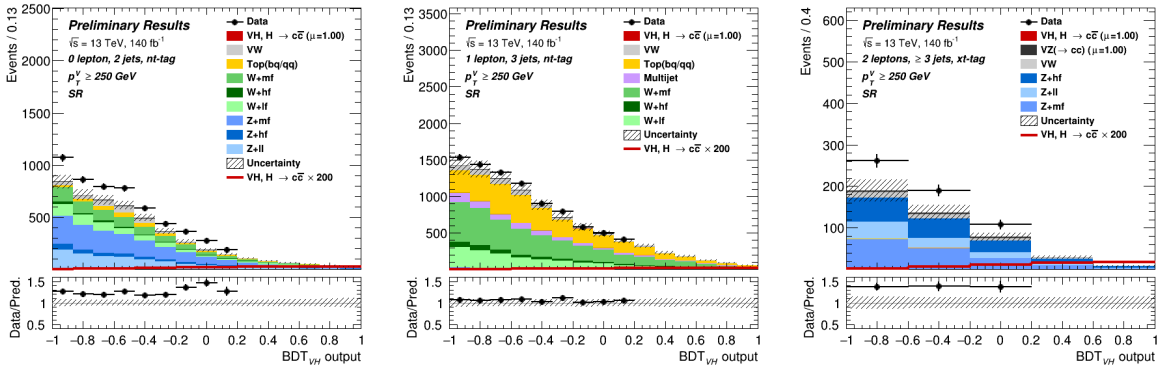


(a) 0-lepton, 2-jet. (b) 1-lepton, 3-jet. (c) 2-lepton, 2-jet.

Figure 6.10: Selection of 2 c-tagged (TT + LT) 150 GeV $< p_T^V <$ 250 GeV signal regions.

High ΔR Control Regions: are designed 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

[9]Not included in 0L due to the trigger threshold on E_T^{miss}.
[10]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.

by a further split from the SRs based on the angular separation $\Delta R(j_1, j_2)$, shortened as ΔR, 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 cuts are defined in Table 6.7 and illustrated in Figure 6.12, with their derivation detailed in Appendix B.1.1. 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 mismodelling effect at high ΔR and to 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, with the TT- and LT-tagged events separated in the CRHighs to respectively constrain the V+hf and V+mfinstead of being merged as in the SRs. In $VH(H \to b\bar{b})$, the CRHighs are used to extract the normalisation of the backgghs 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 6.6. 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 and CRLow. The inequalities are set to enter the control regions, with p_T^V expressed in GeV.

Low ΔR Control Regions: Low ΔR CRs (*CRLow*) are defined in the 1-lepton channel of $VH(H \to b\bar{b})$ to better constrain the W+hf background. 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 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 2L $VH(H \to b\bar{b})$, 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 in Figure 6.15a.

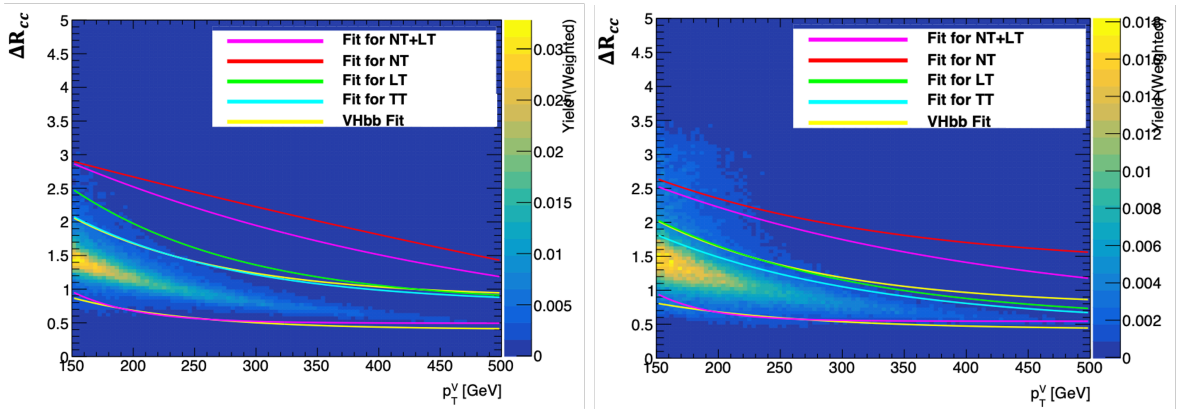
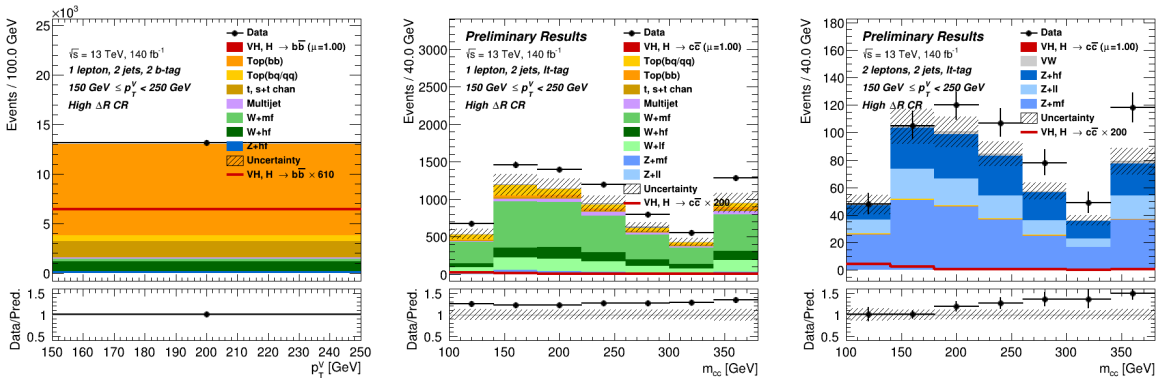


Figure 6.12: The p_T^V-$\Delta R_{c\bar{c}}$ 2D signal yield map of the 1L $VH(H \to c\bar{c})$, for the 2-jet (left) and 3-jet (right) regions. The lines are the results of fitting the high and low $\Delta R_{c\bar{c}}(p_T^V)$ cuts for various signal tags, with the yellow curve showing the $\Delta R_{b\bar{b}}$ cut from $VH(H \to b\bar{b})$that is used in the analysis. The CRHigh is above the top yellow line and the SR below. A Low ΔR CR can be defined by the bottom lines, splitting this region from the SR for the $VH(H \to b\bar{b})$ only.



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

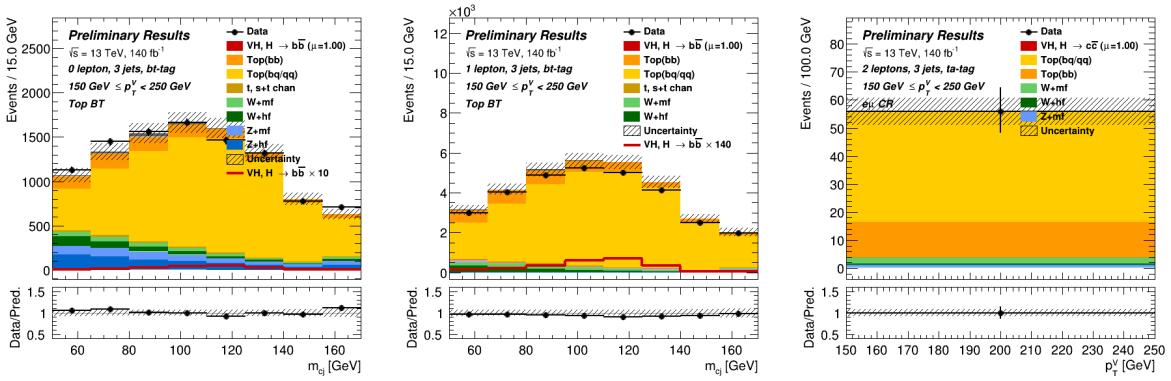
Figure 6.13: Some High ΔR CRs (CRHigh) with 2 jets and 150 GeV $< p_T^V <$ 250 GeV.

Top Control Regions in 0L and 1L: are defined to constrain the Top background Top(bc) and Top(bl) components[11]. The so-called *Top BT CRs* are shared by the resolved $VH(H \to b\bar{b})$ and $VH(H \to c\bar{c})$, with similar p_T^V and jet multiplicity categorisation as the SRs. They are defined for 0L and 1L by requiring events to have at least one B-tag and at least one tight c-tag T, making them orthogonal to the 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 that is a major background in $VH(H \to b\bar{b})$ is controlled from the previously defined CRHighs, 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.

Top Control Regions in 2L: the Top background in 2L is mostly made of di-leptonic $t\bar{t}$ decays, with both subsequent W decaying leptonically. High purity Top CRs are derived for the

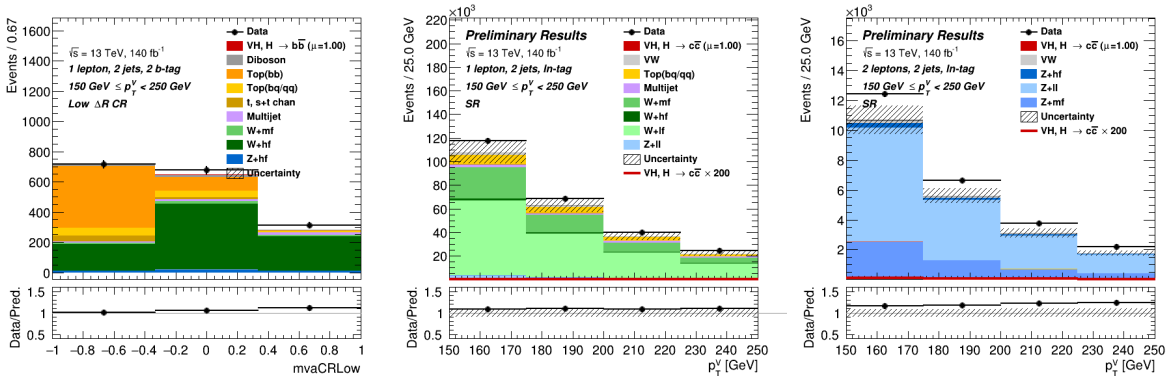
[11] The component in the parenthesis refers to the flavour of the Higgs-candidate jets. As explained later in this chapter, they are floated together in the fit as the Top(bq/qq).

2-lepton channels by requiring leptons of different flavours ($e\mu$ / μe) instead of the same flavour (ee / $\mu\mu$). 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 background in 2L due to the flavour tagging requirements, and the Top $e\mu$ CRs contribute to the fit as single-bin CRs defined per p_T^V and jet multiplicity, with at least one T-tag jet. An example of these CRs is presented in Figure 6.14c.



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

Figure 6.14: Top BT-tagged (left and centre) and $e\mu$ (right) CRs, with 3 jets and 150 GeV $< p_T^V <$ 250 GeV.



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

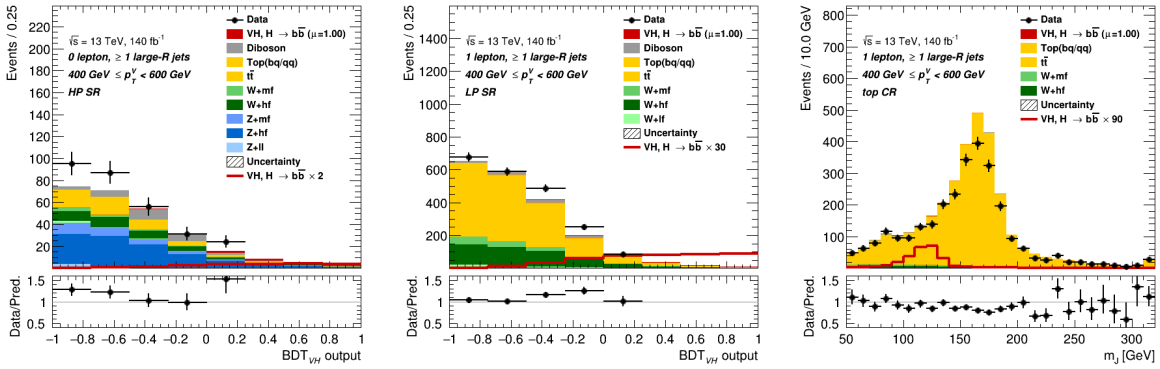
Figure 6.15: A BB-tagged Low ΔR CR (left) and 2 LN-tagged $V+l$ CRs (centre and right), both with 2 jets and 150 GeV $< p_T^V <$ 250 GeV.

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, called $V+l$ *CR*, in the 1L and 2L channels target the W+*lf* and Z+*lf* backgrounds[12]. They are defined by requiring exactly one loose L-tag c-jet without any T- nor B-tagged jet in the event. The selection is otherwise similar to that of the 1 c-tagged signal regions[13], 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.

[12] V+*lf* is a grouping of the V+jets with light-jets, introduced in Section 6.8.3.
[13] Similarly to these SRs, there is no 1L $V+l$ CR for 75 GeV $< p_T^V <$ 150 GeV.

Boosted Regime Categorisation

In the boosted $VH(H \to b\bar{b})$, two p_T^V bins are defined at [400, 600] GeV and $\geq$ 600 GeV to avoid overlap with the resolved $VH(H \to b\bar{b})$. The SRs are defined by requiring exactly 2 of the at most 3 leading track-jets associated with a single leading large-R jet to be b-tagged, with no additional B-tagged track-jet outside the large-R jet to enhance the top background rejection. All boosted regions, with processes normalised to their postfit expectations, are presented in Appendix Section B.5. In the plots, the SRs are further separated into a high- (HP) and low-purity (LP) SRs, respectively when there are 0 or $\geq$ 1 additional small-R jet not associated to the Higgs-candidate large-R jet. These regions are however combined into combined signal regions in the final fit.



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

Figure 6.16: Some boosted BB-tagged in the 400 GeV $< p_T^V <$ 600 GeV signal regions (left and centre) and boosted Top CR (right).

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.

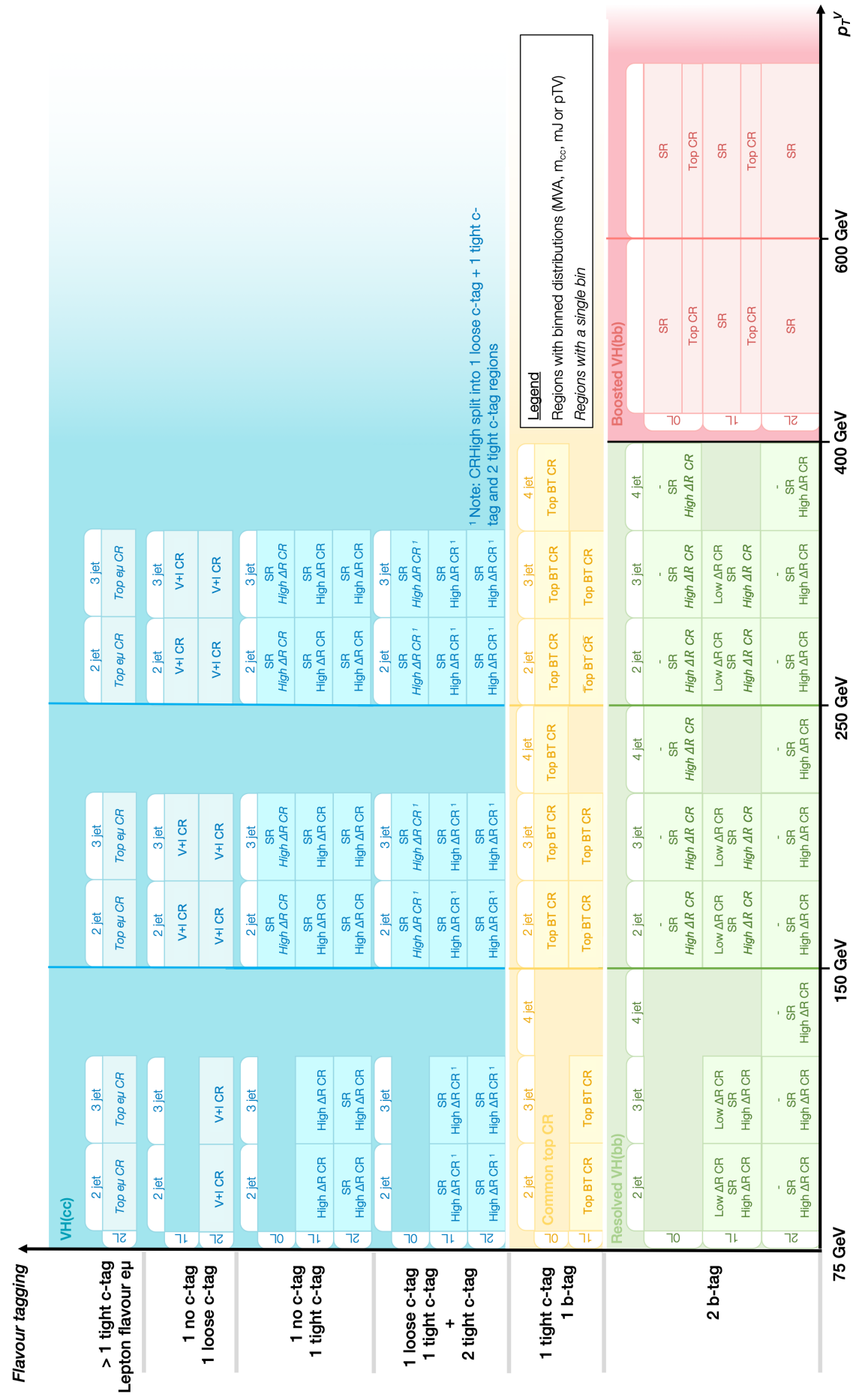


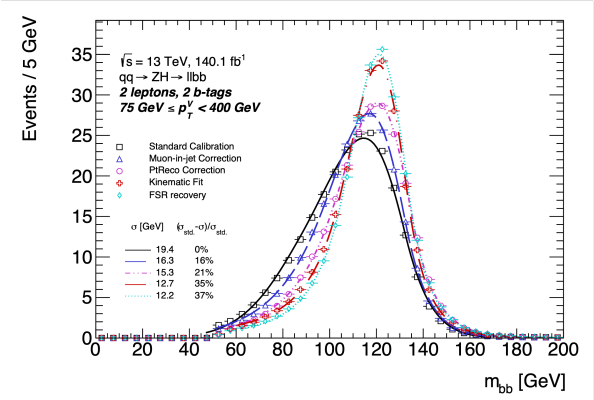
Figure 6.17: The combined $VH(H \to b\bar{b}/c\bar{c})$ analysis regions, showing the Signal Regions (SR), High and Low ΔR control regions (CRHigh and CRLow), the Top BT CR, the Top $e\mu$ CR, the $V + l$ LN-tagged CR, and the boosted Top CR. $VH(H \to c\bar{c})$ is in blue, the shared Top BT CR in the 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.

6.5.4 Tagged-jets Corrections

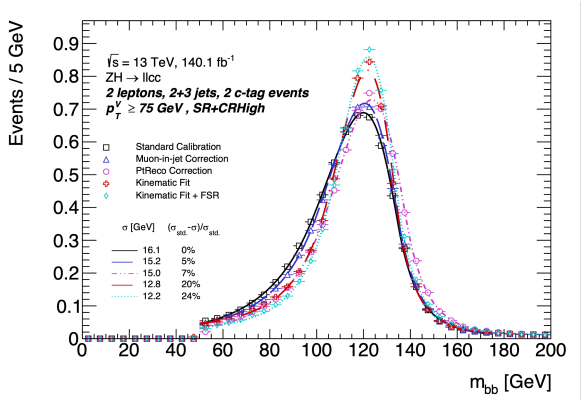
Several corrections to the energy are applied to tagged jets from the previously introduced selection, to improve the energy resolution of the pair of jets selected to form the Higgs candidate. All jets benefit from a standard jet energy calibration introduced in Section 3.3.5. Additional correction for b- and c-jets, summarised in Table 6.8, leverage the unique properties of these heavy-flavour jets. The effects of the different reconstruction techniques are illustrated in Figure 6.18 for some selected 2-lepton resolved and boosted distributions.

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	✓		✓	

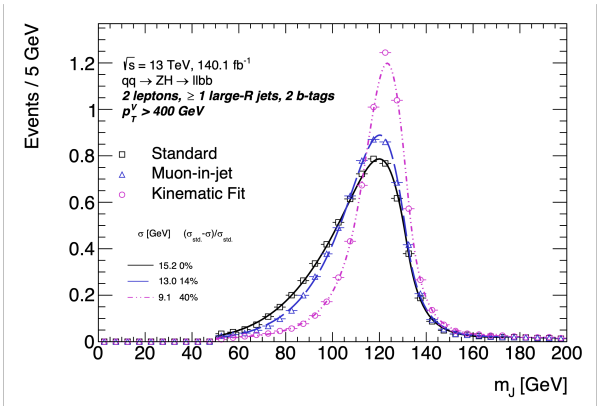
Table 6.8: The different Higgs candidate jet energy correction.



(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.

***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 4-momentum p_T^μ is added to the jet if its angular separation from the jet axis is

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

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.

***P_T-Reco* correction** accounts for missing energy from neutrinos in the semi-leptonic decays and 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. It 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 previous corrections. The fit relies on a likelihood function with terms covering the object resolution, the jet transfer function, a Z mass constraint, and system p_T balance. The boosted 2L channel has a similar kinematic fit based on a Gaussian term. 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. This correction is not applied to 0L or 1L due to the possible increased acceptance of the $t\bar{t}$ background in the sensitive signal regions from the reduced jet multiplicty.

6.6 Discriminant Variables

The analysis leverages a varied set of reconstructed variables in the fit to constrain the different processes and control mismodelling effects. Figure 6.19 displays the chosen variable used for each region in the fit in the resolved regime. The reconstructed Higgs mass offers some separation power of the signals from their major backgrounds, hence some control regions such as the Top

BT CRs and CRHighs are based on the relevant distributions m_{bb}, m_{cc}, or 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 CRHighs 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 mismodelling in the p_T^V distributions of the SHERPA 2.2.11 V+jets samples, as detailed in Section 6.8. To optimise signal and background separation in the statistical analysis, dedicated BDTs, also called MVAs, are trained for the signal regions of the combined analysis with the TMVA ROOT software [209]. 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 this section.

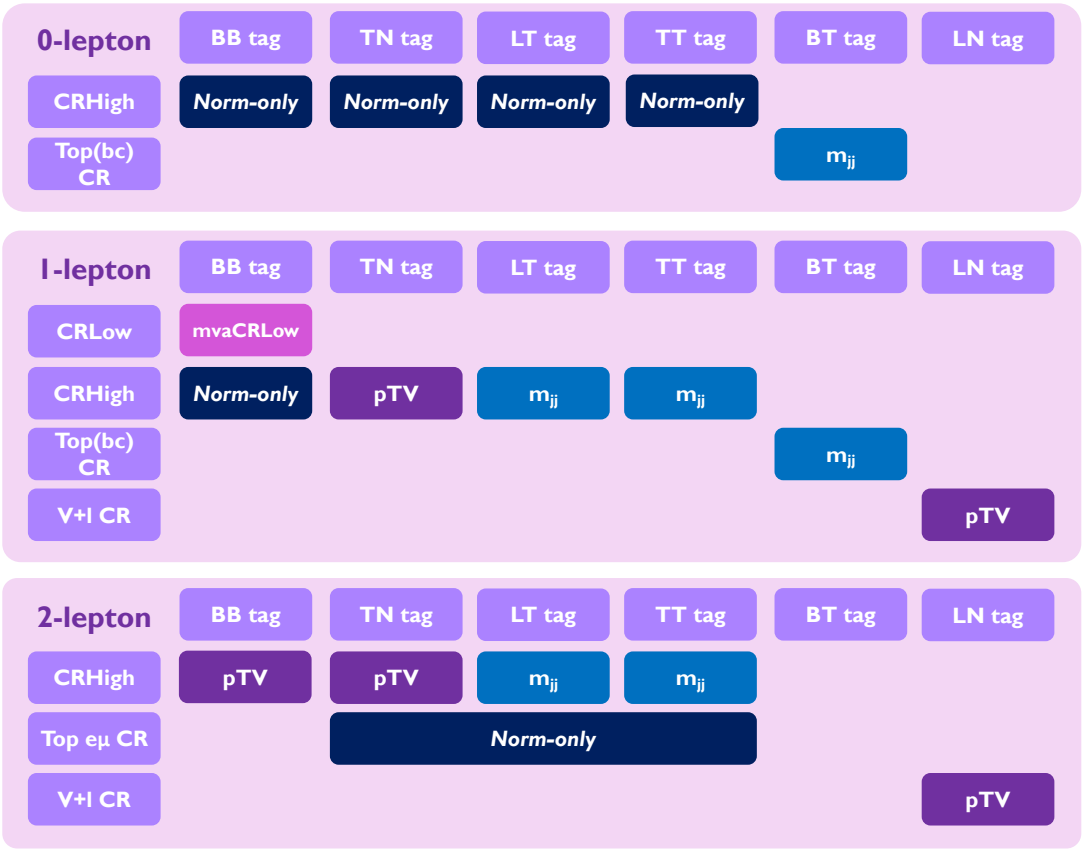


Figure 6.19: Illustration of the discriminant variables used per control regions of the resolved regime in the fit. Norm-only indicates a region to extract a global normalisation and not binned by a variable.

6.6.1 Multivariate Analysis

Three sets of discriminants are trained for the analysis: *MVA* discriminants for the signal region modelling, a specific set *mvaCRLow* for the CRLow distribution in the resolved $VH(H \to b\bar{b})$, and a set of BDTs for the diboson cross-check analysis. For the last one, 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.

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 $VH(H \to b\bar{b})$ [188]. Before training, the object and event selections of Section 6.5 and the jet energy corrections of Section 6.5.4 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. The BDTs are trained to discriminate the respective signal of the different targeted decays [14] from background samples, including V+jets, $t\bar{t}$, single-top, and diboson. BDTs are specifically trained in the following categories, covering the fine analysis categorisation to guarantee sufficient statistics and avoid overtraining:

- **Resolved $VH(H \to b\bar{b}/c\bar{c})$:** BDTs are trained separately for the BB-, 2 c-, and 1 c-tags. Separate trainings are run for each lepton channel and for the following jet multiplicities and p_T^V bins.
  - **0L**: BDTs are trained for the 2-, 3-, and 4-jet categories, in an inclusive $p_T^V \geq 150$ GeV region.
  - **1L**: BDTs are trained for the 2- and 3-jet categories, in $p_T^V \in [75, 150]$ GeV and $p_T^V \geq 150$ GeV.
  - **2L**: BDTs are trained for the 2- and $\geq$3-jet categories in $p_T^V \in [75, 150]$ 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 $VH(H \to b\bar{b}/c\bar{c})$:** a BDT is trained per lepton channel in an inclusive bin of p_T^V.

For training, the full MC samples statistics is leveraged 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 per event weight representing its probability of passing the tagging selection. The result is a weighted distribution possessing the statistical precision of the full MC-samples but distributed per the direct tagging scheme. The weights in this truth tagging procedure are predicted by graph neural networks passed event-level information, as detailed in Appendix B.1.3. The truth tagging procedure is applied separately to BB, TT, TL, and NT events.

The BDTs are trained with specific sets of features defined per lepton channel, as listed in Table 6.9 with prcise variable definitions given in Appendix B.2. Features with long tails are clipped to contain 99% of the centred distributions. Variables undefined for an event are given a default values. The sets of features used are the result of hyperparameter optimisation campaigns, with other variables tested but eventually not included due to their negligible impact on the performance.

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

	$VH(H \to b\bar{b}/c\bar{c})$ 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 B.2.

The architecture of the different BDTs is 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 and large tails in the distributions, the AdaBoost method introduced in Section 4.2.2 is adopted to help stabilise the training [91]. 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.

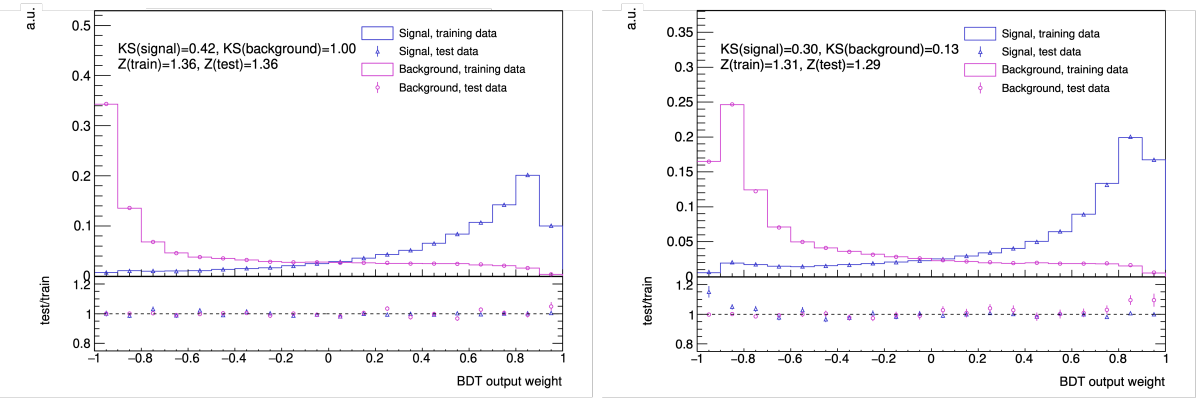
	Resolved $VH(H \to b\bar{b})$			Boosted $VH(H \to b\bar{b})$		
Settings	0L	1L	2L	0L	1L	2L
Boost type	Gradient boost	Gradient boost	Gradient boost	Adaboost	Adaboost	Adaboost
Number of trees	200	600	200	800	800	400
Maximum depth	3	4	4	3	3	3
Learning rate (β)	0.5	0.5	0.5	0.5	0.35	0.3
Number of cuts	100	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.

	$VH(H \to c\bar{c})$		$VZ \to c\bar{c}$
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. Each BDT is therefore 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. Additional overtraining checks are performed on each fold comparing the trained distribution to a test distribution obtained by applying the BDT on the held-out 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 variables.



(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, additional MVAs are trained for the $VH(H \to b\bar{b})$ resolved 1L channel in the Low ΔR CR. 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 events. They are 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 combining the 2- and 3-jet categories. The typical AUC of these discriminants is ~ 0.84, with no overtraining observed.

6.6.2 Output Variable Transformation

The BDTs outputs from the previous section are a finely-binned MVA variables 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 and larger 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 analysis relies on 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 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 starts again from the highest bin not yet recombined.

p_T^V	[75, 150] GeV	[150, 250] GeV	[250, 400] GeV	[400, 600] GeV	$p_T^V > 600$ GeV
$VH(H \to b\bar{b})$	$z_s = 10,\ z_b = 5$		$z_s = 5,\ z_b = 3$		0L/1L: $z_s = 3, z_b = 2$ 2L: $z_s = 2, z_b = 2$
$VH(H \to c\bar{c})$	TT: $z_s = 5, z_b = 3$ Else: $z_s = 10,\ z_b = 5$	0L/1L: TT: $z_s = 5, z_b = 3$; Else: $z_s = 10,\ z_b = 5$ 2L: TT: $z_s = 2,\ z_b = 2$; LT/XT: $z_s = 5,\ z_b = 5$; Else: $z_s = 10,\ z_b = 5$	TT: $z_s = 2, z_b = 2$ LT/XT: $z_s = 5,\ z_b = 3$ Else: $z_s = 10,\ z_b = 5$		

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.

The z_s and z_b parameters are manually tuned for each analysis regime and lepton channel, giving signal regions with a final number of BDT bins varying from 4 to 15, as displayed in the postfit plots of Appendix B.5. An additional protection is added to avoid bins with too small data or MC statistics, 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.

6.7 Experimental Uncertainties

While a lot of effort goes into correctly simulating the collection and reconstruction of information, inaccuracies permeate this procedure and must be accounted for in the statistical analysis of Section 6.9. Several types of experimental uncertainties are considered in the analysis, to cover the systematics effects from to the detector performance, the reconstruction of objects, and the effects of flavour tagging. Table 6.13 summarises the different sources of uncertainty, which are further detailed in this section.

Luminosity & Pile-up The measured Run 2 luminosity for ATLAS is 140.1 $\pm$1.2 fb^{-1} with an uncertainty of 0.83% [47]. The measurement relies on $x-y$ beam separation scans combined with information from dedicated luminosity-sensitive detectors. The PU 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 samples match data at a higher μ than used in their simulation. This rescaling factor is used to reweight the data, matching a simulated-μ of 1.0 to a data-μ of 1.09, a rescaling summarised as 1.0/1.09. A 1σ uncertainty on the average PU 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 from W+jets events, taking into account the statistics of the dataset, assessing systematics effects by deriving scale factors with alternative top and Z+jets samples, and modelling the efficiency dependency on the scalar sum of all final state jets.

Systematic uncertainty name	Description	Regime
	Luminosity and Pile-up	
LUMI_2015_2018	Uncertainty on total integrated luminosity	All
PRW_DATASF	Uncertainty on pile-up modelling	All
	E_T^{miss} and $E_{\text{T,trk}}^{\text{miss}}$	
MET_SoftTrk_ResoPara(Perp)	Soft term longitudinal (transverse) resolution uncertainty	All
MET_SoftTrk_Scale	Soft term scale uncertainty	All
MET_JetTrk_Scale	$E_{\text{T,trk}}^{\text{miss}}$ 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.

Leptons and E_T^{miss} Leptons and E_T^{miss} are calibrated in dedicated analyses, with a reduced set of uncertainties propagated here consisting of:

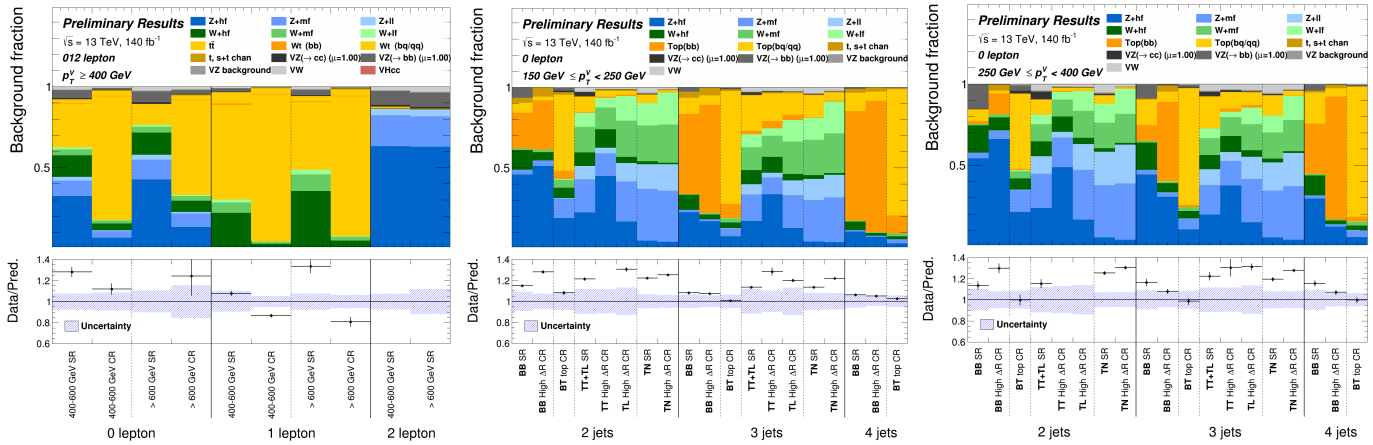
- E_T^{miss}: scale factors account for the direction of the E_T^{miss} and the soft term contribution.
- *Electrons*: uncertainties on the reconstructed values, the identification efficiency, isolation efficiency, and the energy scale and resolution are derived by comparing data and simulations in kinematic distributions of $Z \to e^+e^-$, $W \to e\nu$ and $J/\psi \to e^+e^-$ events [77].
- *Muons*: uncertainties on the reconstruction and identification efficiencies of muons with $p_T >$ 15 GeV and $p_T <$ 15 are included separately, using respectively samples of $Z \to \mu^+\mu^-$ and $J/\psi \to \mu^+\mu^-$ [206]. 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 uncertainties on the reconstruction and RNN-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 [85, 210, 211].

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 [83, 212], as described in Section 3.3.5. The reduced set is derived from a Principal Component Analysis (PCA) to preserve the largest correlations in certain regions of jet kinematics. Large-R jets uncertainties for the energy scale and resolution are similarly estimated from data [213]. 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, and the general flavour tagging 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 [214]. For each jet flavour, there is one uncertainty per p_T bin in the calibration. A τ-jet uncertainty is derived from the c-jet values. PCA 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 extrapolation of the performance to high-p_T jets. GNN truth tagging uncertainties are covered by these flavour tagging uncertainties, so no dedicated uncertainties are considered.

6.8 Signals and Backgrounds Modelling

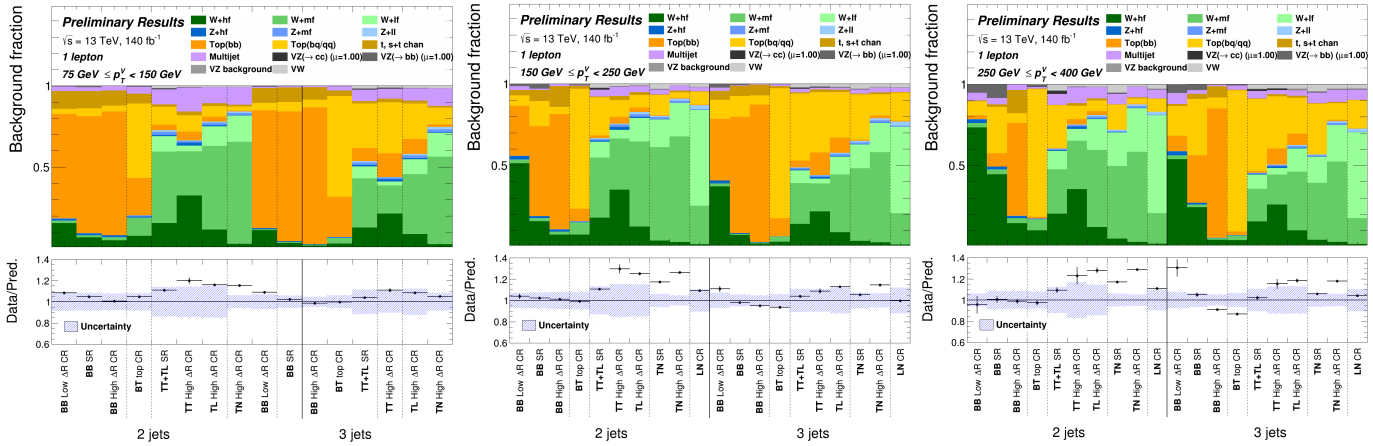
Similarly to the experimental process, the simulations of the signals and backgrounds cannot entirely be accurate and mismodellings are to be expected in the derived samples. These inaccu-



(a) Boosted regime (≥ 400 GeV).

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

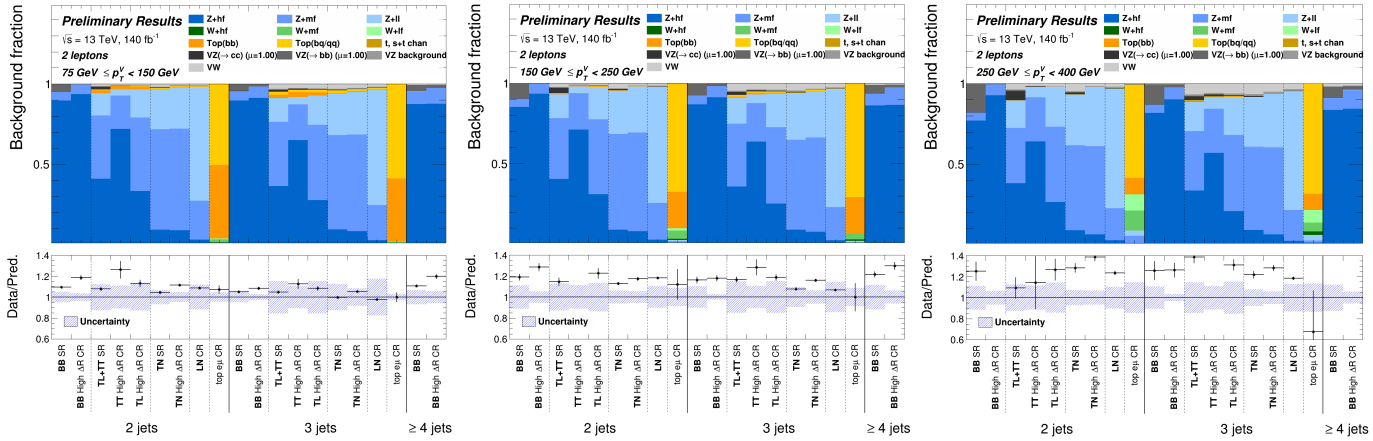
(c) 0L, $p_T^V \in [250, 400]$ GeV.



(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.

racies must be taken into account in the fit framework to avoid introducing bias. The modelling strategy of the signals and backgrounds at the time of writing is discussed in this section. The background composition depends on the lepton channel, the analysis category, and the p_T^V and N_{jet}, as highlighted in Figure 6.21. The V+jets processes are the dominant backgrounds in the signal regions of the 0-lepton and 2-lepton channels, while the top processes contribute more in the 1-lepton channel, and globally at larger jet multiplicities and lower p_T^V. From the flavour tagging requirements, $VH(H \to b\bar{b})$ primarily selects the bb-component of the background while

$VH(H \to c\bar{c})$ has a more diverse flavour composition with the 2 c-tag as an intermediate step between the BB and 1 c-tag. This translates into increased fractions of V+hf in $VH(H \to b\bar{b})$, and V+mf and V+lf in $VH(H \to c\bar{c})$. To summarise the largest backgrounds per lepton channels:

- **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 miss-identified hadronic τ. In $VH(H \to b\bar{b})$, the top background significantly contributes and dominates in 3- and 4-jets. Finally, there is some diboson contribution, primarily for BB-tagged events.
- **1-lepton**: the top process is dominant for $VH(H \to b\bar{b})$, while for $VH(H \to c\bar{c})$ the W+jets leads followed by the top and the multi-jet backgrounds.
- **2-lepton**: Z+jets makes up most of the background, followed by the diboson and some residual top process at low p_T^V for $VH(H \to b\bar{b})$.

The changing background composition in the different analysis regions requires an adequate strategy to constrain their modelling in the fit, as outlined in this section.

6.8.1 General Modelling Strategy

The combined analysis adopts some common strategies to model the signals and backgrounds 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 $VH(H \to b\bar{b})$ and $VH(H \to c\bar{c})$ when possible. The normalisations of the major backgrounds, the V+jets and Top, are free to float in the fit, with Floating Normalisations (FNs) 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 uncertainties, comparisons of the nominal samples to the 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
$t\bar{t}$ and single-top	POWHEG+PYTHIA 8	POWHEG+HERWIG 7, MADGRAPH5+PYTHIA 8	ISR, FSR, DS/DR (for 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 overall uncertainties on the yield of a process, computed in and applied to all regions. These uncertainties are considered from expected yield of a background to derive its normalisation from data, for the diboson and single-top s processes primarily.

Acceptance uncertainties : relative acceptance uncertainties cover possible changes in the distribution of events of a 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*

$$\text{Acceptance Unc}_i = \frac{\text{Yield}[\text{Cat.}^B(\text{Alternative}_i\ \text{MC})]}{\text{Yield}[\text{Cat.}^A(\text{Alternative}_i\ \text{MC})]} \Bigg/ \frac{\text{Yield}[\text{Cat.}^B(\text{Nominal MC})]}{\text{Yield}[\text{Cat.}^A(\text{Nominal MC})]}, \tag{6.2}$$

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:

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

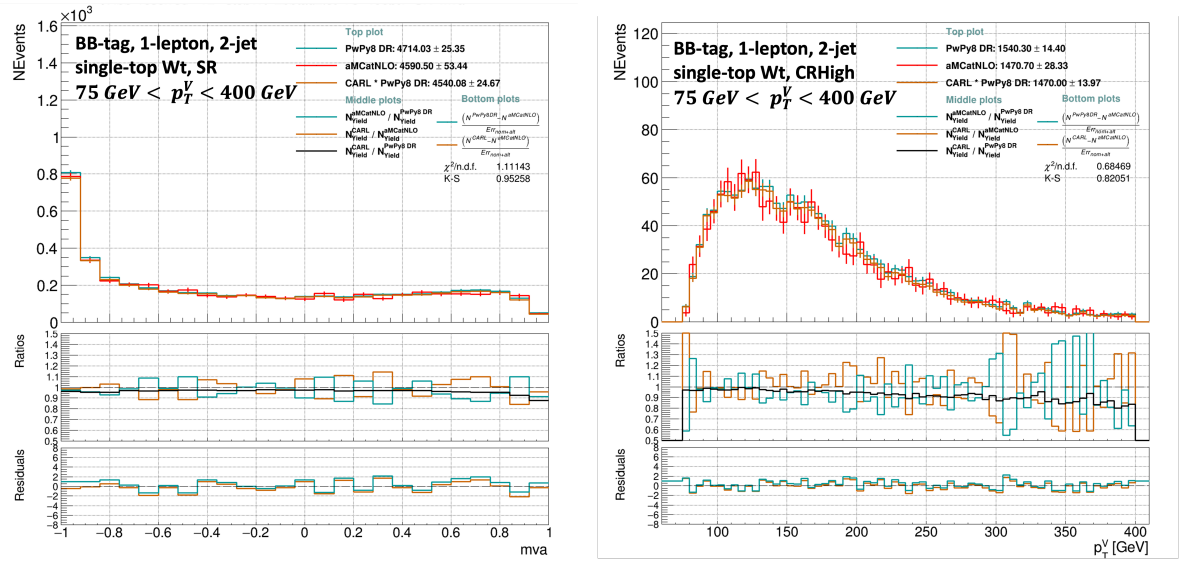
If the extrapolation is across several regions A, B, C ordered by decreasing purity, the acceptance ratio is decomposed into two extrapolations: a first one from $A \to B + C$ followed by an additional $B \to C$ uncertainty. For acceptance uncertainties between distinct analysis regions in the resolved regime, the signal and Top BT control regions are considered jointly due to their similar kinematics. 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 [215]. 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, analogously to truth tagging. The advantage of this technique is that the reweighted nominal distributions benefit from a much larger statistics than the alternative ones, thus smoothing out intra-bin fluctuations and reducing the MC statistics uncertainties. Examples of such derived CARL shape uncertainties modelling the parton shower with MADGRAPH5_AMC@NLO for the single-top Wt process in 1-lepton are presented in Figure 6.22. Additional shape uncertainties are directly derived by comparing samples for 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.

An overview of the signals and backgrounds modelling systematics considered is presented in Figure 6.15 and detailed in Appendix B.4. 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%. If a systematics

Uncertainties	Resolved	Boosted
Signal		
$qqWH$ / $qqZH$ / $ggZH$ normalisations / acceptance	Values from previous analyses [130, 188, 189]	
$H \to bb$ 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 normalisation and acceptance modelling systematic uncertainties. The values given refer to the size of the uncertainty affecting the yield of each background.



(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.

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

6.8.2 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 [216, 217], depicted in Figure 6.23. The bins are defined in successive regions of p_T^V, from truth information in the simulated samples, and the number of additional jets in the event, at 0 or more than 1 additional jet.

The signal samples are finely binned following the STXS prescription, with 5 p_T^V bins for the ZH covering $[75, 150[$ GeV, $[150, 250[$ GeV, $[250, 400[$ GeV, $[400, 600[$ GeV, and $\geq$ 600 GeV. The

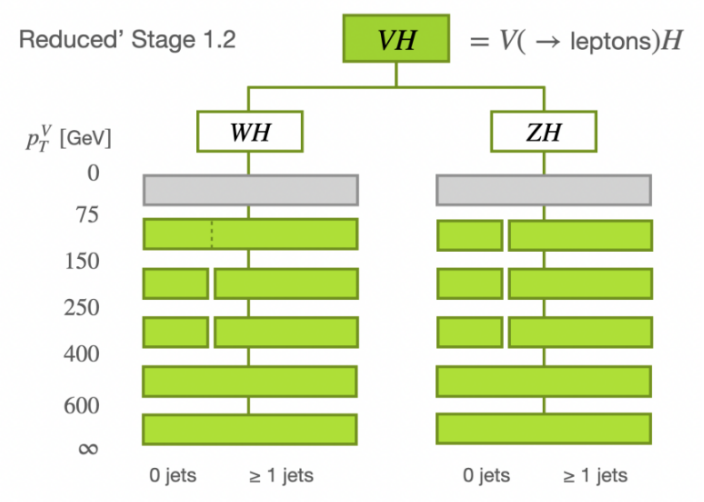


Figure 6.23: The Standard Template Cross-Section scheme in the reduced stage 1.2 [217].

first three bins, corresponding to the resolved regime, are further split with 0 or ≥ 1 additional 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 jet multiplicity split in the $[75, 150[$ GeV bin, giving a total 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 $VH(H \to c\bar{c})$, the templates are later 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. [218]. 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 [219]. 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

considered [33]. 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 adopted, but the full detector response and event reconstruction are simulated with uncertainties covering modifications to the m_J distributions.

6.8.3 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. Different components are split based on the flavour composition 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 (FNs) 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 total jet multiplicities N_{jet}[15]. The modelling of Z+jets includes several types of acceptance uncertainties that are applied only in 0L and 2L. In the resolved regime:

- *Channel extrapolation 2L → 0L uncertainties*: for the Z+*hf*, Z+*mf*, and Z+*lf* separately.
- *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.
- *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:

[15] 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 but $VH(H \to c\bar{c})$ is not: to account for 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 B.3 of Appendix B.4. 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 mismodelling of p_T^V in SHERPA 2.2.11.
- QCD scale shape uncertainties by varying μ_R and μ_F.
- EW shape variations, although they are typically small.

Boosted regime The modelling strategy is roughly the same as in the resolved regime, with the uncertainties fully detailed in Appendix Table B.4. The Z+*hf* component is left free-floating in 0L and 2L, while the Z+*mf* and Z+*lf* components both have overall acceptance uncertainties of 35%. The Z+*lf* has no other acceptance uncertainty since it is negligible in the boosted regime. Flavour acceptance uncertainties for Z+*hf* and Z+*mf* are applied in 0L and 2L. They also have channel acceptance uncertainties and SR → Top CR acceptance ratios, both applied in 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 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 $VH(H \to c\bar{c})$. The FNs are decorrelated in p_T^V and jet multiplicities N_{jet}[16]. Acceptance uncertainties, listed in the Appendix Table B.5, are applied in 0L and 1L. They include:

- *Channel extrapolation 1L → 0L uncertainties*: applied in 0L for all components separately.
- *Flavour composition uncertainties*: include a comparison of cc to bb for W+*hf*, of [bc, bl, $c\tau$, $b\tau$] to cl for W+*mf*, and of [$l\tau$, $\tau\tau$] to l for W+*lf*. They are decorrelated in p_T^V and N_{jet}.
- *Region extrapolation uncertainties* are defined differently for the combinations:

[16]The only exception is the 1L W+*lf* in 75 GeV < p_T^V < 150 GeV: it has a 25% normalisation uncertainty.

- W+hf: constrained mostly in the SR and the BB-tagged CRLow[17], 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, applied in SR and CRLow[17].
- W+lf: constrained mostly in the SR and the 1 LN-tagged $V+l$ CR, applied in CRHigh.

- *N_{jet} acceptance*: FNs are left free-floating in N_{jet} for 2- 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.

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

Boosted regime The same modelling strategy as Z+jets is applied, with the uncertainties fully detailed in Appendix Table B.6. The W+hf component is left free-floating in 0L and 1L, 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. The different components 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.

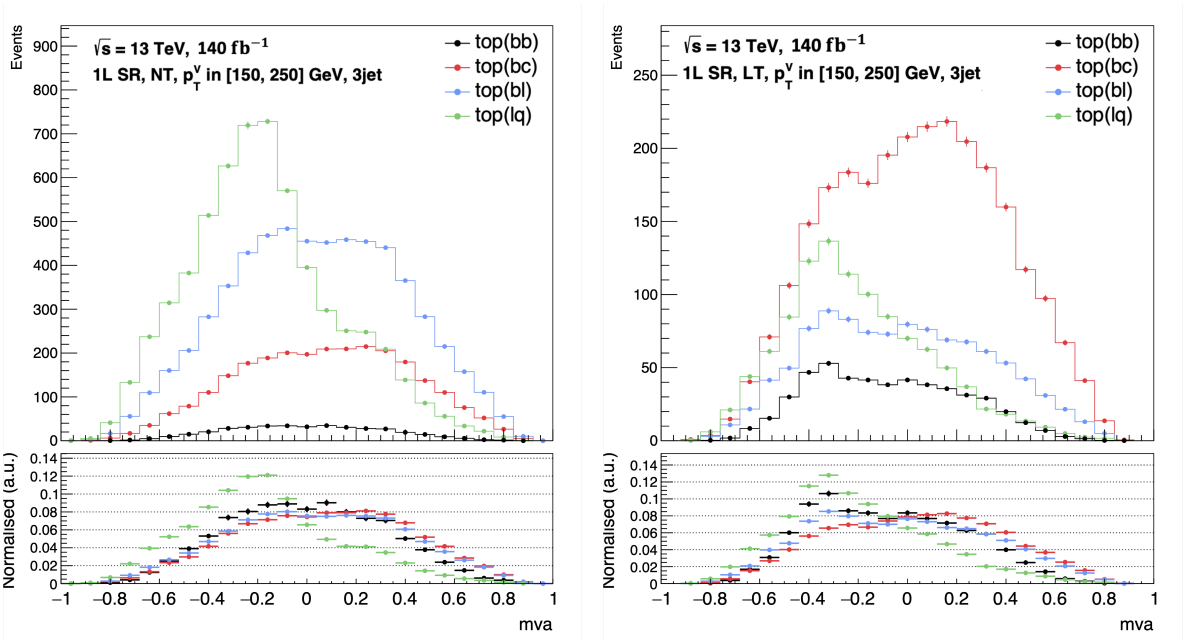
6.8.4 Top Modelling

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[18] 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 due to the large angle between the emitted top-quark that is passed to the b-quarks.
- Top(bq): combining Top(bc) and Top(bl), is mostly in the $VH(H \to c\bar{c})$ phase space and is well selected by the Top BT CR.
- Top(qq): combining Top(cc), Top(cl) and Top(ll), where l is a light-jet (u, d, s, a gluon, or a τ), is mostly in the NT and LT regions of the $VH(H \to c\bar{c})$.

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 because they share the same kinematics, as illustrated in Figures 6.24 in the signal regions of $VH(H \to c\bar{c})$. The Top(bq) background

[17] The CRLow is always only considered in $VH(H \to b\bar{b})$ 1L.
[18] Throughout this chapter, Top will refer to the combination of the $t\bar{t}$ & Wt processes.



(a) (b)

Figure 6.24: The MVA distributions of the top background components (direct tagged) in the $VH(H \to c\bar{c})$ signal regions (NT-tagged on the left, LT-tagged on the right) with 150 GeV $< p_T^V <$ 250 GeV and 3 jets, before rebinning. Top(bb) in black, Top(bc) in red, Top(bl) in blue, and Top(qq) in green. The bottom panels show the normalised distributions.

is particularly important in $VH(H \to c\bar{c})$ as it peaks near the signal mass (having a mass $\sim (m_t + m_W)/2 \approx m_H$) 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 subcomponents shapes modelled by flavour composition uncertainties. 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 in 0L & 1L Resolved 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 primarily in the BB-tagged High ΔR CR and the Top BT CR. These FNs are separated in jet multiplicity N_{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 B.7:

- *Channel extrapolation 1L → 0L uncertainties*: the Top is dominant in 1L, hence the FNs derivation is driven by the 1-lepton channel and applied to the 0L. This uncertainty is split in p_T^V, with 2% in [150, 250] GeV and 8% in [250, 400] GeV.
- *Flavour composition uncertainties*: the Top(bq/qq) includes differently shaped subcomponents. Uncertainties are derived from the alternative samples as double ratios comparing the bl (5%) and qq (10%) to the bc.

- *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 with double ratios from alternative samples.
  - Top(bb): extrapolation uncertainties are derived from the CRHigh and applied in the SR, the Top CR and the CRLow[17]. Additional uncertainties are applied from the SR to the Top CR and CRLow[17]. All uncertainties are split in p_T^V.
  - Top(bq/qq): the uncertainties are derived from the SR + Top CR + CRLow[17], due to their shared kinematic, and applied to the CRHigh. Additional uncertainties are applied from the SR + Top CR to the CRLow[17]. All uncertainties are split in 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 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 from $t\bar{t}$. 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 analysis as a normalisation uncertainty is simultaneously applied to account for the different yields of the DS- and DR-schemes.
- ISR and FSR shape uncertainties are derived by varying the μ_R and μ_F scales. Up- and a down-variations are considered for each of them, with symmetric variations for the ISR while the down-variation of FSR is smaller than its up-variation.

The Single-Top t- & s-channels in 0L & 1L Resolved 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 channel. The importance of single-top t quickly reduces with increasing energy[19]. 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.

[19] Except in the CRHigh region where the ratio stays in the 7%-9% range.

- *Region extrapolations uncertainties*: depend on the p_T^V. For $p_T^V < 150$ GeV, the uncertainty is applied from SR → CRLow+CRHigh, with an additional CRHigh → CRLow uncertainty in 1L. For the higher p_T^V regions, the extrapolations are instead from CRHigh → SR+CRLow[17], with an additional SR → CRLow[17] uncertainty.

- N_{jet} *acceptance*: 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 additional [150, 250] GeV to [250, 400] GeV uncertainties.

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 B.9 of the Appendix details the various single-top uncertainties considered.

Top Backgrounds in 2L Resolved Modelling

Data-driven estimates are used for the 2L channel in the resovled regime only. In $VH(H \to b\bar{b})$, templates are derived 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 to let the Top background left free-floating, with at least one tight c-tagged required.

Top Backgrounds Boosted Modelling

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[20]. The modelling in the boosted regime, detailed in the Appendix Tables B.8 and B.10, covers:

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

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

Boosted shape uncertainties are considered similarly to what is done in the resolved regime.

6.8.5 Diboson Modelling

The diboson production backgrounds consist 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 the W hadronically decaying and the Z leptonically decaying, while in 1L the WW process contributes the most. Again, both contribute similarly to

[20] Studies are, at the time of writing, ongoing to merge these two processes in the boosted regime.

0L. The resolved and boosted acceptance uncertainties are detailed in Table B.12 and Table B.11.

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 (30%) for the gg-initiated $ggVV$. 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 $VZcc$ throughout the modelling. The rest of the WW, WZ, and VZ are classified as background components, denoted as VVbkg. Acceptance uncertainties are summarised in Table 6.14 and detailed in the Appendix Table B.12. For the signal components, the uncertainties are split between the ZZ and WZ and include:

- *Channel extrapolation uncertainties*: two sets are considered due to the differences between the components. One covers the 1L → 0L for WZbb and $WZcc$, and the other the 2L → 0L for $ZZbb$ and $ZZcc$. They are split by N_{jet}.
- *Region extrapolation uncertainties*: from the SR to the CRHigh and CRLow[17], 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.
- N_{jet} *acceptance*: are considered from 2-jet to higher jet-multiplicities. 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.
- p_T^V *extrapolation uncertainties*: the 150 GeV $< p_T^V <$ 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 processes. They are modelled by QCD scale variations.

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

- *Channel extrapolation uncertainties*: 2 sets covering 1L → 0L (for WW and $W_{\text{lep}}Z_{\text{had}}$) and 2L → 0L (for ZZ and $W_{\text{had}}Z_{\text{lep}}$) are included due to difference in purities.
- *Region extrapolation uncertainties*: from the SR to the CRHigh, as to the diboson purity is higher in the SR, separately for the different channels.
- *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.
- p_T^V *extrapolation uncertainties*: all extrapolation go from the 150 GeV $< p_T^V <$ 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 mismodelled p_T^V shape. These uncertainties are applied to all regions.
- 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 The modelling is similar to the resolved regime, with the uncertainties fully detailed in Table B.11 of the Appendix. Small contributions from misidentified W decays as jets or misreconstructed leptons are taken into account. The ZZ and WZ have normalisation uncertainties of 17% and 27% respectively. Acceptance uncertainties are included to cover the lepton channel acceptance, p_T^V acceptance, and STXS uncertainties on the p_T^V and N_{jet} bins, as is done in the resolved regime.

6.8.6 Multi-jet Modelling

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

6.9 Statistical Analysis

After collecting the data and the 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 POIs* with the modelling strategy defined in the previous section. The combined analysis targets several deliverables:

- $VH(H \to b\bar{b})$:
  - 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 6.8.2 and Figure 6.23: 15 POIs, 8 for ZH and 7 for WH.
  - Constraints on the y_b Yukawa bottom coupling modifier κ_b.
- $VH(H \to c\bar{c})$:

  - Inclusive signal strength $\mu_{VH_{cc}}$ upper limits at the 95% Confidence Level (CL) : 1 POI.
  - Constraints on the y_c Yukawa 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_b.

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

6.9.1 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 a binned-likelihood function in all 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 representing, respectively, the number of events per bin $\mathcal{L}_{\text{Events}}$, the impact of systematics $\mathcal{L}_{\text{Systematics}}$, and the impact of the limited statistics of the simulated sample $\mathcal{L}_{\text{MC-stats}}$. They are combined into the likelihood function

$$\mathcal{L} = \mathcal{L}_{\text{Events}} \times \mathcal{L}_{\text{Systematics}} \times \mathcal{L}_{\text{MC-stats}}. \tag{6.3}$$

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 Parameters (NPs) $\boldsymbol{\theta}$, accounting for possible perturbations in each bin to the expected signal and background yields $\{s_i, b_i\} \to \{s_i(\boldsymbol{\theta}), b_i(\boldsymbol{\theta})\}$. The NPs are statistically modelled as standard Gaussian $\mathcal{N}(0,1)$ penalties of 0 mean 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 6.7 and 6.8. 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 defined as a *pull*

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

where the prefit values are $\theta_0 = 0$ and $\sigma_{\theta_0} = 1$. The *constraint* indicates the change in certainty on the NP after the fit, estimated by the variance $\hat{\sigma_\theta}$ measured from the inverse Hessian matrix at the maximal likelihood point $\hat{\theta}$. 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 of the likelihood covers the uncertainties linked to the limited statistics of the Monte Carlo 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

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

where $\beta_i = 1/\sigma^2_{\text{rel}}$ introduces the relative statistical uncertainty σ_{rel} on the expected yield b_i of the sum of backgrounds in bin i.

The full likelihood function of Equation 6.3 is therefore defined as

$$\mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\theta}, \boldsymbol{\gamma}) = \prod_{i \in \text{bins}} \mathcal{P}(N_i \,|\, \boldsymbol{\mu} \boldsymbol{s}_i(\boldsymbol{\theta}) + b_i(\boldsymbol{\theta}, \boldsymbol{\gamma})) \times \prod_{\theta \in \boldsymbol{\theta}} \mathcal{N}(\theta \,|\, 0, 1) \times \prod_{i \in \text{bins}} \mathcal{N}(\beta_i \,|\, \gamma_i\beta_i, \sqrt{\gamma_i\beta_i}). \quad (6.4)$$

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 $\{\hat{\hat{\boldsymbol{\theta}}}, \hat{\hat{\boldsymbol{\gamma}}}\}$. A profile likelihood ratio $\lambda(\boldsymbol{\mu})$ is defined from these 2 sets to test a 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 bounded in the [0, 1] range, with higher values implying a good agreement between the data and the hypothesised $\boldsymbol{\mu}$ while lower values are signs of disagreements. This pattern permits the construction of a likelihood ratio test statistics $t_{\boldsymbol{\mu}}$, defined as [220]

$$t_{\boldsymbol{\mu}} = \begin{cases} -2 \ln \lambda(\boldsymbol{\mu}) & \hat{\boldsymbol{\mu}} \geq \boldsymbol{\mu} \\ 0 & \hat{\boldsymbol{\mu}} < \boldsymbol{\mu} \end{cases}, \quad (6.6)$$

as the signal can only have a positive contribution to the yield in the present case. 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 test, also called the null hypothesis, the p-value quantifies

the compatibility of the observed data with the background-only hypothesis ($\mu = 0$)

$$p_{\boldsymbol{\mu}} = \int_{t_{\boldsymbol{\mu},\text{obs}}}^{\infty} f(t_{\boldsymbol{\mu}} \,|\, \mathbf{0}) dt_{\boldsymbol{\mu}}, \tag{6.7}$$

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 probability density function of the test statiscs $t_{\boldsymbol{\mu}}$ assuming $\boldsymbol{\mu} = \mathbf{0}$. The p-value is the probability of finding data that is at least equally 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

$$Z = \Phi^{-1}(1 - p_{\boldsymbol{\mu}}), \tag{6.8}$$

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 [220, 221], 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 [220]. Two fits are considered: a *prefit* Asimov where the nuisance parameters are constrained to their initial values[21], and a *postfit* Asimov where the NPs take their best-fit values from the fit to the real dataset. The postfit Asimov can be used to define expected results, to 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.

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

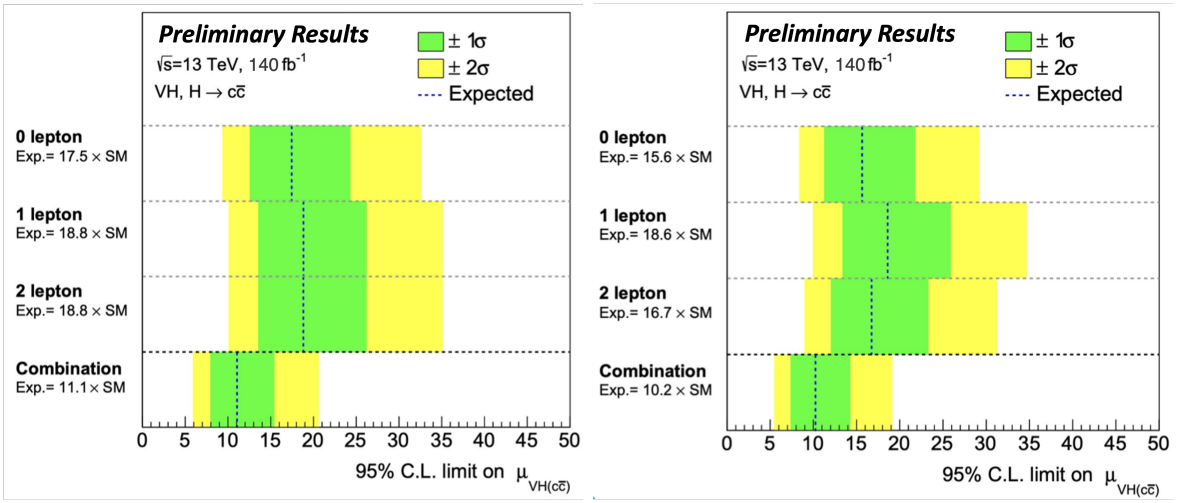
There are 15 POIs for the $VH(H \to b\bar{b})$ side and 1 POI for $VH(H \to c\bar{c})$. The binning used and regions included as well as the variables defining the underlying distributions entering the fits are detailed in Sections 6.5 and 6.6. A dense summary of the full categorisation is presented 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 CRHighs), 48 in the resolved $VH(H \to b\bar{b})$ (21 SRs, 6 CRLows, 21 CRHighs), 12 BT-tagged Top CRs shared in

[21] 0 for all NPs but the FNs, which are set at 1.

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 mismodelling and avoid biasing the fit, as described in Sections 6.7 and 6.8. 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 bins of the signal regions most sensitive to the signal hidden. For m_{bb} or m_{cc} distributions, the Higgs 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 hidden. For conditional fits, where the signal strength are fixed at 1, 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. The following results are therefore partial and temporary, but already highlight an appreciable increase in sensitivty.

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

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) Postfit Asimov. (b) 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 postfit (left) and prefit (right) on the Asimov dataset.

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 ATLAS published $VH(H \to c\bar{c})$ result [130], to 11.1 × SM (10.2 × SM) in the postfit (prefit) Asimov fit of the combined analysis, a factor 2.8 improvements in sensitivity. Gains are expected to be made in all lepton channels, which now have 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 1-lepton from 60 × SM → 18.8 × SM, and the 2-lepton from 51 × SM → 18.8 × SM [130]. 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 latest ATLAS published expected significance of 6.3 σ [191]. 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 now distinctively the 1-lepton channel.

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 ZH 150 GeV $< p_T^V <$ 250 GeV without additional jet bin. The impact of systematics and statistical uncertainties on the signal strengths of the different bins is shown in Figure 6.26. The measurement is particularly statistically limited.

VH	Truth p_T^V	0 additional N_{jet}	≥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	≥ 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	≥ 600 GeV	1.07 σ	

Table 6.16: The expected prefit significances 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 $VZ(\to b\bar{b})$ part, the postfit expected significance reaches a large value of 15.1 σ when combining lepton channels. The 0-lepton, 1-lepton, and 2-lepton channels respectively reach postfit sensitivities of 11.2 σ, 6.2 σ, and 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 published 2.2 σ expected result [130]. The

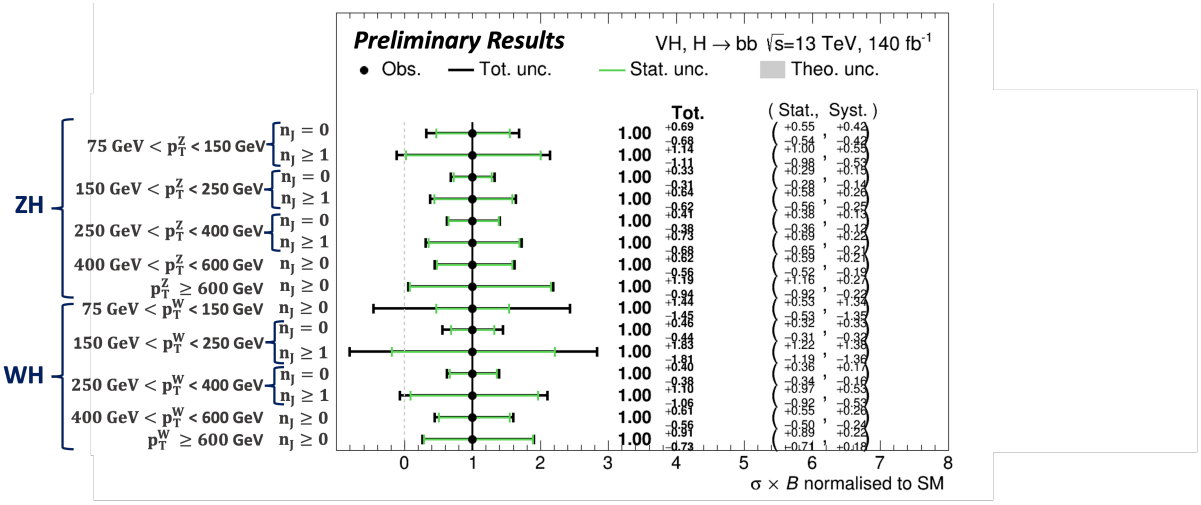


Figure 6.26: The constraints on the prefit STXS signal strengths.

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.

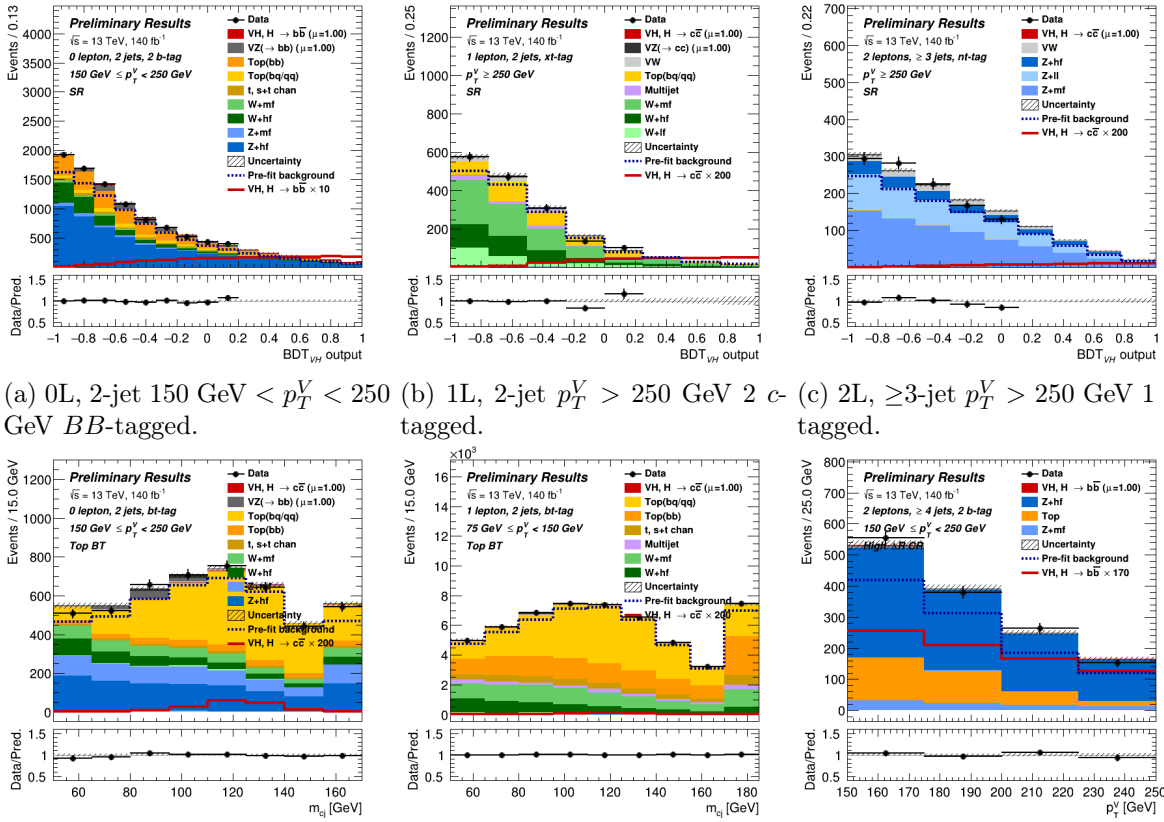
Additional Fit Results

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 that the Monte Carlo samples correctly model the data after the fit, some postfit plots are presented in Figure 6.27 for selected signal and control regions. All the postfit analysis distributions are listed in Appendix B.5. Interestingly, good agreement between the data and postfit MC samples is also observed in validation regions not directly constrained in the fit. Figure 6.28 displays postfit distributions for a BL-tagged region analoguous to the included BT-tagged Top CR, an LL-tagged region similar to the c-tagged signal region, and the p_T^V spectrum of the inclusive 2L BB-tagged 2-jet signal region. The good agreement observed between data and MC-samples in all regions is evidence of a sufficient fit constraining.

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 of the nominal maximal likelihood fit as $\hat{\mu}$ with uncertainty $\hat{\sigma_\mu}$, and of a re-run fit with a group of NPs fixed at $\hat{\mu}'$ with uncertainty $\sigma_{\hat{\mu}'}$. The impact of the fixed group of NPs is defined as the change in uncertainty measured by

$$\text{Impact} = \sqrt{\hat{\sigma_\mu}^2 - \hat{\sigma_\mu'}^2}. \tag{6.10}$$

To evaluate 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



(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.

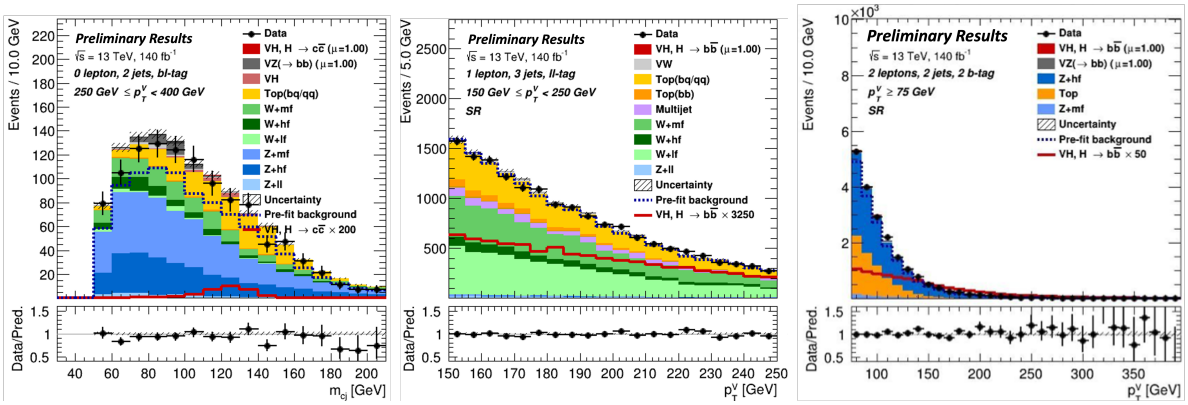
(c) 2L, ≥3-jet $p_T^V >$ 250 GeV 1 c-tagged.

(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 postfit signal regions (top row) and control regions (bottom row), for the 0L (left), 1L (centre), and 2L (right).



(a) 0L BL-tagged, TopCR-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.

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.

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.

For $VH(H \to c\bar{c})$, similar observations are made with several nuances. On the modelling side, the signal modelling is less paramount, with the top and multi-jet processes contributing more significantly. Additionally, the Z+jets uncertainties are now clearly leading, 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-jets are now dominant, with the jet reconstruction 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 $\hat{\sigma_\theta}$ are measured: one with the NP fixed at $\hat{\theta} + \hat{\sigma_\theta}$ and one with $\hat{\theta} - \hat{\sigma_\theta}$. 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 (at 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 those systematics highly ranked. Shape uncertainties associated with the diboson process and Higgs modelling uncertainties as well as the Wt DS-DR shape uncertainty and b-jet tagging uncertainties 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 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, the 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 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 the final unblinded conditional fit to data.

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 their 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 normalisations, with some common trends per process highlighted. Concerning the Top backgrounds, in 0L and 1L it seems mostly overestimated in the MC simulations, 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

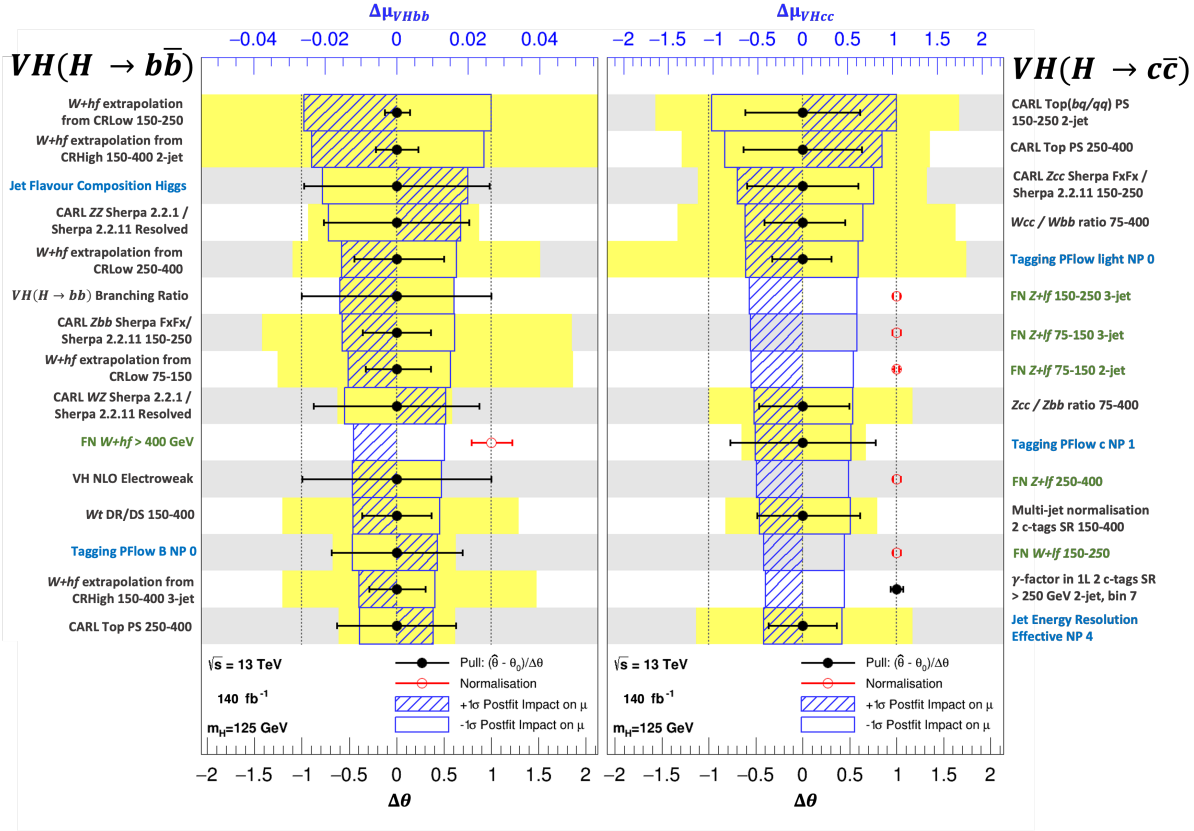


Figure 6.29: The 15 most highly ranked Asimov postfit nuisance parameters for the $VH(H \to b\bar{b})$ (left) and $VH(H \to c\bar{c})$ (right) signal strengths. Modelling NPs are written in black, experimental NPs in blue, and floating normalisation (and γ-factor) in green, with values indicated by the 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 NP. The effect of changing the NP by +1 σ (-1 σ) induces the change in signal strength $\Delta\mu$ shown by the hashed (empty) blue rectangle, defined with respect to the top axis.

multiplicities. The Top(bq/qq) and Top(bb) have generally similar FN values. Concerning the W+jets, the W+hf is well modelled in 2-jet across p_T^V but less so in the ≥3-jet category, where the underestimation of the simulations grows with p_T^V. The boosted W+hf normalisation in the ≥ 400 GeV range is significantly distant higher than unity. The same observations hold for W+lf, which is well-modelled in 2-jet but gets higher FNs in 3-jet. The W+mf component requires similar large corrections from the fit, with FN values ∼ 1.3 across 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. The Z+jets yield is globally corrected upwards, 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" are only applied to the 3-jet category. There is therefore some overlap, with the latter set of FNs used to correct downwards the large normalisation of the ≥ 3-jet. Similarly to the W+hf, the boosted Z+hf FN values are significantly pulled away from unity.

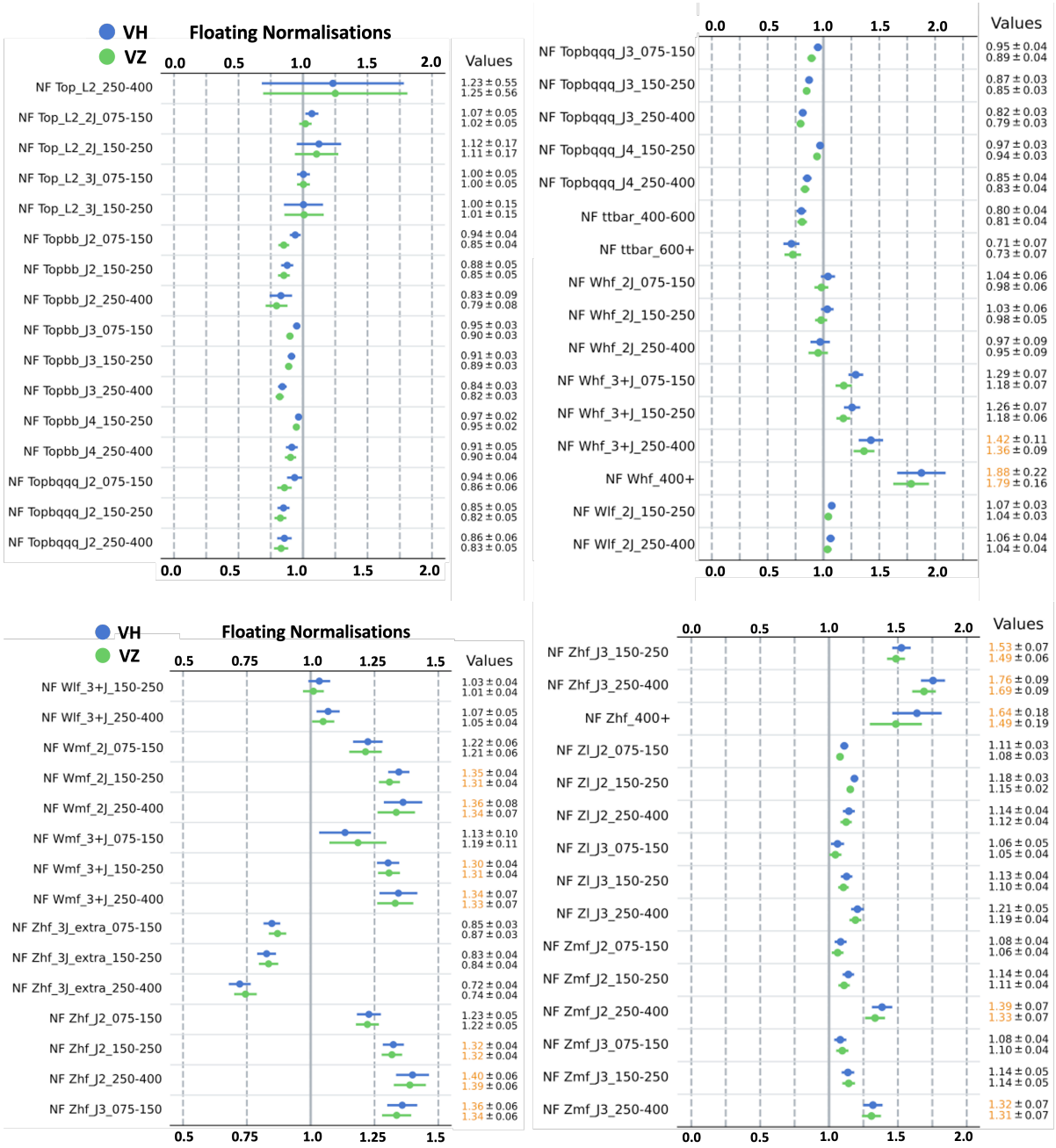


Figure 6.30: The floating normalisations of the major background in the combined analysis targeting the $VH(H \to b\bar{b}/c\bar{c})$ in blue, versus the cross-check analysis $VZ(\to b\bar{b}/c\bar{c})$ in green.

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 FNs of the same process, from different p_T^V and N_{jet} categories. Some striking exceptions are visible: the boosted $t\bar{t}$ displays some small uncorrelations 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 to a lesser extent) 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 Top BT CR. The V+hf normalisations are slightly anti-correlated to the V+lf, and the V+lf are strongly correlated between the W and Z.

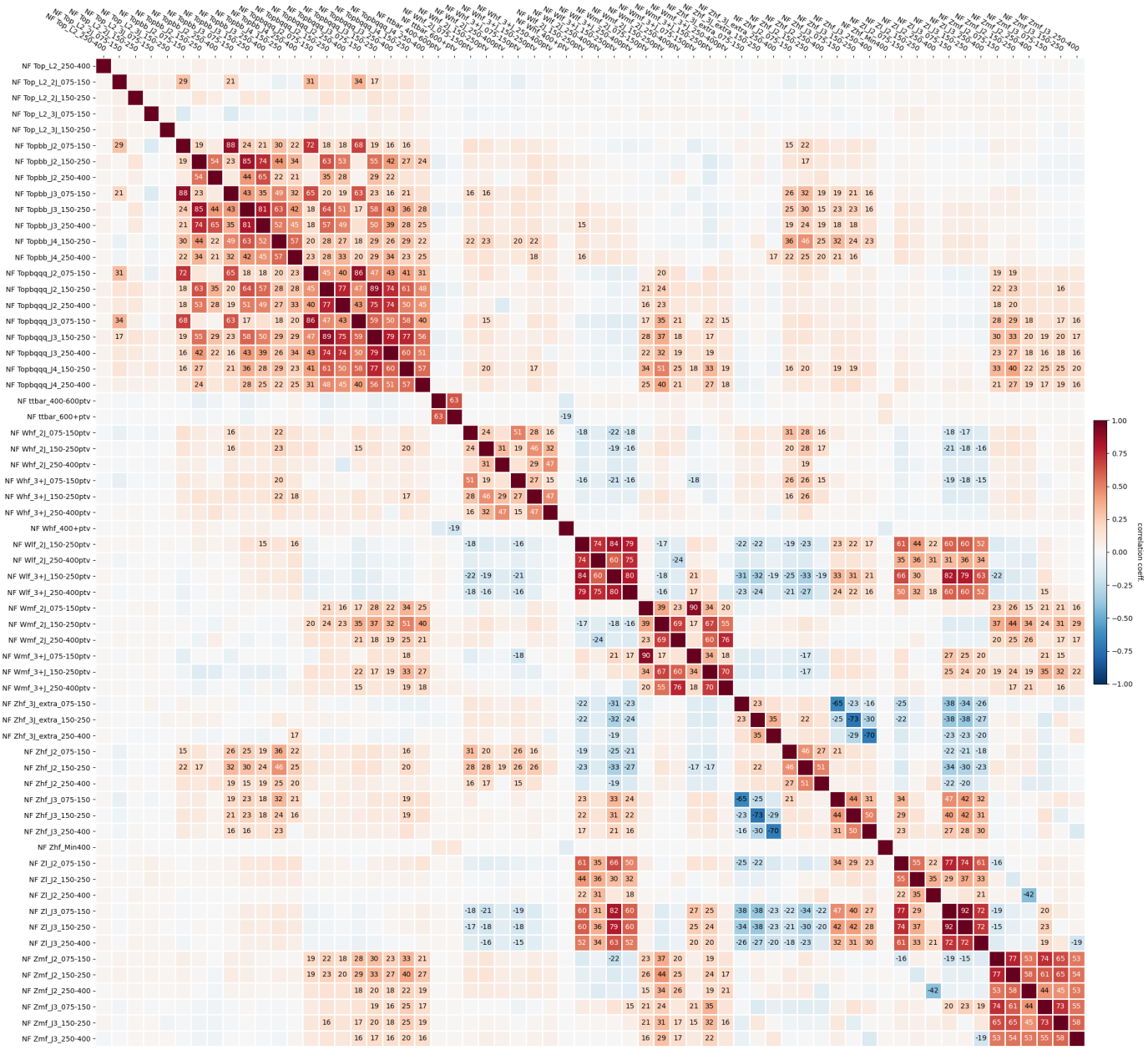


Figure 6.31: The correlations between the floating normalisations of the major background in the combined $VH(H \to b\bar{b}/c\bar{c})$ analysis.

6.10 Conclusion

This chapter introduces the combined $VH(H \to b\bar{b}/c\bar{c})$ ATLAS analysis using the 140 fb^{-1} of data collected during Run 2, from 2015 to 2018. The state after its third unblinding approval meeting is presented, as the analysis is not yet concluded. Events are separated based on their p_T^V into a resolved or boosted regime, as $VH(H \to b\bar{b})$ or $VH(H \to c\bar{c})$ depending on the two highest tags of their jets or sub-jets. Flavour tagging is performed with the ML-based DL1r tagger, with a hierarchy of tags ranked as b-tagged > tight c-tagged > loose c-tagged > untagged. This is followed by a split into leptonic channels based on the number of charged lepton ℓ (e, μ) in the final state, to separate the $Z(\to \nu\nu)H$, $W(\to \ell\nu)H$, and $Z(\to \ell^+\ell^-)H$, with $H \to b\bar{b}$ or $H \to c\bar{c}$.

To boost the sensitivity, a fine categorisation further splits the analysis 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, the latter grouping the production of $t\bar{t}$ pair and the single-top Wt. Backgrounds are constrained from data in dedicated control regions, defined respectively by a cut on the angular separation of the Higgs-candidate jets and by an alternative event-tagging selection. The cross-check analysis targeting the $VZ(\to b\bar{b}/c\bar{c})$ is performed to validate the adopted strategy.

The analysis promises to significnatly increase the sensitivity of the ATLAS search for the $H \to c\bar{c}$ process as well as delivering the finest measurements to date of the differential cross-section of the $H \to b\bar{b}$. MVA discriminants are introduced throughout the different regions to improve the sensitivity to the sought signals. The adoption of upgraded flavour tagger and the pseudo-continuous joint-tagging approach paved the way for this coherent joint measurements of the VH to heavy-flavour quarks decay. New MC samples with more statistics contribute to reducing the importance of uncertainties plaguing the final fit performance. Some final studies on the modelling strategy and the fit framework are still underway at the time of writing. Some adjustments are required to stabilise the complex fit introduced in this thesis and understand the constrains on the different processes before unblinding the analysis.

This study serves as the join combined legacy $VH(H \to b\bar{b}/c\bar{c})$ analyses of ATLAS on the full Run 2 dataset. 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). Additional improvements to the experimental tools and the analysis strategy are required to reach this threshold. The former will primarily rely on the improved flavour tagging abilities presented in Chapter 5: from the single tagger GN2 to the boosted $X \to b\bar{b}/c\bar{c}$ decay tagger *GN2X* [6]. The adoption of transformer-based neural networks is promising a significant increase in tagging performance. These will be available for the next iteration of the VH analysis, and is expected to reverberate into an improved signal acceptance and a better background rejection. The larger volume of data to be collected in ongoing Run 3 and following Run 4 of the LHC as well as future data-taking campaigns will significantly imrpove the prospects of this severely statistically-limited analysis.

CHAPTER 7

Conclusion and Outlook

This thesis aims to follow a logical order, starting from the theory underpinning the modern edifice of particle physics in Chapter 2. The SM has been extensively validated by many experiments across the world, in particular by the ATLAS Collaboration using data collected from proton-proton collisions in the LHC as presented in Chapter 3. Many properties of this particle discovered 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 anc challenge the SM in evermore complex measurements, searching for any possible discrepancy between observations and theoretical predictions. This mission requires state-of-the-art detectors and reconstruction software. At its core, a particle physics analysis is statistical data analysis that is well suited to modern machine learning and artificial intelligence, as reviewed in Chapter 4. The recent progress in this field provides an exciting avenue of development for ATLAS, helping the Collaboration propel the performance of its software to new heights by designing effective network-based models for specific purposes.

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 taggers of the ATLAS Collaboration, relies on a single multimodal network exploiting a Transformer Encoder at its core, with multiple tasks targeted to distil expert knowledge in the network. The state-of-the-art performance it delivers 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 joined into the combined $VH(H \to b\bar{b}/c\bar{c})$ 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 however, there are hints

of great progress 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 of 2.8 to 11.1 $\times$ SM expectations, compared to the last published ATLAS result [130]. Similarly, great progress is made in the precision 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 [191]. For the first time, both production modes are expected to be observed at more than 5 σ in the $H \to b\bar{b}$ decay mode, with respective significances 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 across the experiment. Collecting large datasets is crucial to analyses searching for rare signatures and performing precision measurements. The $VH(H \to b\bar{b}/c\bar{c})$ analysis presented here suffers from large statistical uncertainties that will be improved with the addition of data. Collecting more data at higher energies require the detector to be operated i at more challenging conditions: more pile-up is the price of a higher instantaneous luminosity. The subdetectors must be upgraded to deal with this increased activity, with in particular finer-grain measurements expected to help performance. In this regard, the development of the next inner detector system called ITk is a promising avenue [222].

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 future challenging conditions. In this respect, flavour tagging benefits greatly from adopting advanced new NN architecture such as the Transformer, but also from the multimodal input and multitask paradigms to nimbly introduce expert knowledge. Future avenues of progress primarily relies on pursuing this path further, adding additional low-level input information and defining additional tasks to help the main classification objective. Performance is highly correlated with the number of parameters, and reliably training larger networks requires careful design, well-thought training procedures, large datasets, and optimised hyperparameters. Across science and industry, advanced machine learning plays a crucial role in the effort to modernise software capabilities. This is particularly the case in HEP, where the large databases measured from collisions or simulated are effectively exploited to create reliable and precise models. Such networks are trained for all the uses of the field: from generative AI to effectively produced simulated samples, to fast network deployed on FPGAs- or GPU-based triggers, DL to reconstruct physics objects from the rich noisy set of low-level data, and finally ML deploying in analyses to improve signal discrimination from the backgrounds and help constrain the modelling of the different processes.

Bibliography

[1] Daniel Dominguez. "An artistic depiction of the Brout-Englert-Higgs field". In: (2022). URL: https://cds.cern.ch/record/2815837.

[2] *Umami Framework*. https://gitlab.cern.ch/atlas-flavor-tagging-tools/algorithms/umami. Accessed: 2023-04-21.

[3] *Salt Framework*. https://gitlab.cern.ch/atlas-flavor-tagging-tools/algorithms/salt. Accessed: 2024-02-20.

[4] *Graph Neural Network Jet Flavour Tagging with the ATLAS Detector*. Tech. rep. Geneva: CERN, 2022. URL: %5Curl%7Bhttps://cds.cern.ch/record/2811135%7D.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] Matthew D. Schwartz. *Quantum Field Theory and the Standard Model*. Cambridge University Press, 2013.

[11] 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.

[12] 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.

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

[14] 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].

[15] 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].

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

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

[18] 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].

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

[20] 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.

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

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

[23] 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.

[24] *Wikipedia: The Standard Model (of Physics)*. https://simple.wikipedia.org/wiki/Standard_Model. Accessed: 2024-03-15.

[25] C. N. Yang and R. L. Mills. "Conservation of Isotopic Spin and Isotopic Gauge Invariance". In: *Phys. Rev.* 96 (1 1954), pp. 191–195. DOI: 10.1103/PhysRev.96.191. URL: https://link.aps.org/doi/10.1103/PhysRev.96.191.

[26] Particle Data Group. "Review of Particle Physics". In: *PTEP* 2022 (2022), p. 083C01. DOI: 10.1093/ptep/ptac097.

[27] 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.

[28] Steven Weinberg. "A Model of Leptons". In: *Phys. Rev. Lett.* 19 (21 1967), pp. 1264–1266. DOI: 10.1103/PhysRevLett.19.1264. URL: https://link.aps.org/doi/10.1103/PhysRevLett.19.1264.

[29] Abdus Salam. "Weak and Electromagnetic Interactions". In: *Conf. Proc. C* 680519 (1968), pp. 367–377. DOI: 10.1142/9789812795915_0034.

[30] Steven Weinberg. "General Theory of Broken Local Symmetries". In: *Phys. Rev. D* 7 (4 1973), pp. 1068–1082. DOI: 10.1103/PhysRevD.7.1068. URL: https://link.aps.org/doi/10.1103/PhysRevD.7.1068.

[31] John C. Collins, Davison E. Soper, and George Sterman. *Factorization of Hard Processes in QCD*. 2004. arXiv: hep-ph/0409313 [hep-ph].

[32] 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.

[33] 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].

[34] 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].

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] J. Vollaire et al. *Linac4 design report*. Vol. 6/2020. CERN Yellow Reports: Monographs. Geneva: CERN, Sept. 2020. ISBN: 978-92-9083-579-0, 978-92-9083-580-6. DOI: 10.23731/CYRM-2020-006.

[41] K. H. Reich. "The CERN Proton Synchrotron Booster". In: *IEEE Trans. Nucl. Sci.* 16 (1969), pp. 959–961. DOI: 10.1109/TNS.1969.4325414.

[42] J. B. ADAMS. "The Cern Proton Synchrotron". In: *Nature* 185.4713 (1960), pp. 568–572. DOI: 10.1038/185568a0. URL: https://doi.org/10.1038/185568a0.

[43] "The Super Proton Synchrotron". In: (2012). URL: https://cds.cern.ch/record/1997188.

[44] "LHC Machine". In: *JINST* 3 (2008). Ed. by Lyndon Evans and Philip Bryant, S08001. DOI: 10.1088/1748-0221/3/08/S08001.

[45] *Accelerator Complex of CERN*. https://home.web.cern.ch/science/accelerators. Accessed: 2024-03-21.

[46] 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.

[47] 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].

[48] 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.

[49] 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.

[50] *Public ATLAS Luminosity Results for Run-2 of the LHC*. https://twiki.cern.ch/twiki/bin/view/AtlasPublic/LuminosityPublicResultsRun2. Accessed: 2024-03-22.

[51] *ATLAS Schematics*. https://atlas.cern/Resources/Schematics. Accessed: 2024-03-22.

[52] 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].

[53] ATLAS Collaboration. *ATLAS inner detector: Technical Design Report, 1*. Technical design report. ATLAS. Geneva: CERN, 1997. URL: https://cds.cern.ch/record/331063.

[54] 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].

[55] M Capeans et al. *ATLAS Insertable B-Layer Technical Design Report*. Tech. rep. 2010. URL: https://cds.cern.ch/record/1291633.

[56] 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.

[57] ATLAS Collaboration. *IBL Efficiency and Single Point Resolution in Collision Events*. Tech. rep. Geneva: CERN, 2016. URL: https://cds.cern.ch/record/2203893.

[58] 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.

[59] ATLAS Collaboration. *Technical Design Report for the ATLAS Inner Tracker Strip Detector*. Tech. rep. Geneva: CERN, 2017. URL: https://cds.cern.ch/record/2257755.

[60] 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.

[61] 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.

[62] 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.

[63] Francesca Cavallari. "Performance of calorimeters at the LHC". In: *Journal of Physics: Conference Series* 293.1 (2011), p. 012001. DOI: 10.1088/1742-6596/293/1/012001. URL: https://dx.doi.org/10.1088/1742-6596/293/1/012001.

[64] 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.

[65] 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.

[66] ATLAS Collaboration. *The ATLAS Collaboration Software and Firmware*. Tech. rep. Geneva: CERN, 2021. URL: %5Curl%7Bhttps://cds.cern.ch/record/2767187%7D.

[67] ATLAS Collaboration. *ATLAS Computing Acknowledgements*. Tech. rep. Geneva: CERN, 2020. URL: https://cds.cern.ch/record/2717821.

[68] Joao Pequenao and Paul Schaffner. "How ATLAS detects particles: diagram of particle paths in the detector". 2013. URL: https://cds.cern.ch/record/1505342.

[69] ATLAS Collaboration. "Operation of the ATLAS trigger system in Run 2". In: *Journal of Instrumentation* 15.10 (2020), P10004. DOI: 10.1088/1748-0221/15/10/P10004. URL: https://dx.doi.org/10.1088/1748-0221/15/10/P10004.

[70] ATLAS Collaboration. *The Optimization of ATLAS Track Reconstruction in Dense Environments*. Tech. rep. Geneva: CERN, 2015. URL: https://cds.cern.ch/record/2002609.

[71] 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.

[72] 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.

[73] 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].

[74] 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.

[75] 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.

[76] W Lampl et al. *Calorimeter Clustering Algorithms: Description and Performance*. Tech. rep. Geneva: CERN, 2008. URL: https://cds.cern.ch/record/1099735.

[77] 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.

[78] 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].

[79] 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.

[80] 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 2017), p. 072002. DOI: 10.1103/PhysRevD.96.072002. URL: https://link.aps.org/doi/10.1103/PhysRevD.96.072002.

[81] 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.

[82] 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.

[83] 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.

[84] 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.

[85] 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.

[86] 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.

[87] Kevin P. Murphy. *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012. ISBN: 0262018020.

[88] 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.

[89] 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.

[90] 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.

[91] 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.

[92] 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.

[93] 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.

[94] Benedek Rozemberczki et al. "The Shapley Value in Machine Learning". In: *ArXiv* abs/2202.05594 (2022). URL: https://api.semanticscholar.org/CorpusID:246822765.

[95] 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.

[96] 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.

[97] 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.

[98] 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.

[99] Abien Fred Agarap. *Deep Learning using Rectified Linear Units (ReLU)*. 2019. arXiv: 1803.08375 [cs.NE].

[100] 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.

[101] 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.

[102] 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.

[103] 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.

[104] 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.

[105] 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.

[106] 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.

[107] 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.

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

[109] 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.

[110] Petar Veličković et al. “Graph Attention Networks”. In: *International Conference on Learning Representations* (2018). URL: https://openreview.net/forum?id=rJXMpikCZ.

[111] 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.

[112] 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.

[113] 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.

[114] 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.

[115] 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.

[116] Alec Radford et al. “Improving language understanding by generative pre-training”. In: (2018).

[117] 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.

[118] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. *Layer Normalization*. cite arxiv:1607.06450. 2016. URL: http://arxiv.org/abs/1607.06450.

[119] 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.

[120] 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.

[121] 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.

[122] Albert Q. Jiang et al. *Mistral 7B*. 2023. arXiv: 2310.06825 [cs.CL].

[123] Bin Xiao et al. *Florence-2: Advancing a Unified Representation for a Variety of Vision Tasks*. 2023. arXiv: 2311.06242 [cs.CV].

[124] 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.

[125] 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.

[126] 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.

[127] *Representation of a b-jet*. https://tikz.net/jet_btag/. Accessed: 2024-03-20.

[128] *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.

[129] 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.

[130] 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.

[131] 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.

[132] 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.

[133] ATLAS Collaboration. "Optimisation and performance studies of the ATLAS b-tagging algorithms for the 2017-18 LHC run". In: (July 2017).

[134] 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.

[135] 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.

[136] 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.

[137] Arnaud Duperrin. *Flavour tagging with graph neural networks with the ATLAS detector*. 2023. arXiv: 2306.04415 [hep-ex].

[138] 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.

[139] 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.

[140] 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.

[141] 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.

[142] 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.

[143] 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.

[144] ATLAS Collaboration. *ATLAS Pythia 8 tunes to 7 TeV data*. Tech. rep. Geneva: CERN, 2014. URL: %5Curl%7Bhttps://cds.cern.ch/record/1966419%7D.

[145] 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.

[146] 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.

[147] 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.

[148] 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.

[149] 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.

[150] 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.

[151] 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.

[152] "Boosted Object Tagging with Variable-R Jets in the ATLAS Detector". In: (July 2016). URL: %5Curl%7Bhttps://cds.cern.ch/record/2199360%7D.

[153] 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.

[154] ATLAS Collaboration. *Measurement of the $t\bar{t}$ cross-section and $t\bar{t}/Z$ cross-section ratio using LHC Run 3 pp collision data at a centre-of-mass energy of $\sqrt{s}$ = 13.6 TeV*. Tech. rep. Geneva: CERN, 2022. URL: %7Bhttps://cds.cern.ch/record/2842916%7D.

[155] 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.

[156] 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.

[157] Dasol Hwang et al. *Self-supervised Auxiliary Learning with Meta-paths for Heterogeneous Graphs*. 2021. arXiv: 2007.08294 [cs.LG].

[158] Hadar Serviansky et al. *Set2Graph: Learning Graphs From Sets*. 2020. arXiv: 2002.08772 [cs.LG].

[159] Rachel E. C. Smith et al. *Differentiable Vertex Fitting for Jet Flavour Tagging*. 2023. arXiv: 2310.12804 [hep-ex].

[160] Sam Shleifer, Jason Weston, and Myle Ott. *NormFormer: Improved Transformer Pretraining with Extra Normalization*. 2021. arXiv: 2110.09456 [cs.CL].

[161] 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].

[162] Leslie N. Smith and Nicholay Topin. *Super-Convergence: Very Fast Training of Neural Networks Using Large Learning Rates*. 2018. arXiv: 1708.07120 [cs.LG].

[163] *ml.cern.ch*. https://ml.docs.cern.ch. Accessed: 2024-02-16.

[164] Johnu George et al. *A Scalable and Cloud-Native Hyperparameter Tuning System*. 2020. arXiv: 2006.02085 [cs.DC].

[165] 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.

[166] 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.

[167] 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.

[168] 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.

[169] 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.

[170] 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.

[171] 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].

[172] 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.

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

[174] 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.

[175] 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.

[176] 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.

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

[178] 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].

[179] 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].

[180] 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.

[181] 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].

[182] 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].

[183] 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].

[184] 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].

[185] 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].

[186] 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.

[187] 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.

[188] 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].

[189] 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].

[190] 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.

[191] 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.

[192] 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.

[193] 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.

[194] 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.

[195] 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.

[196] 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.

[197] 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.

[198] 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.

[199] 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.

[200] 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.

[201] 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.

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

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

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

[205] 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.

[206] "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.

[207] 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.

[208] 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.

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

[210] 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.

[211] 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.

[212] 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.

[213] 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].

[214] 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.

[215] 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.

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

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

[218] *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.

[219] 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].

[220] 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.

[221] 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.

[222] D. Bortoletto. "ATLAS ITk tracking and readout performance". In: *Nucl. Instrum. Meth. A* 1048 (2023), p. 167912. DOI: 10.1016/j.nima.2022.167912.

[223] 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.

[224] 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].

[225] 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.

[226] 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].

[227] 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.

[228] 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.

[229] ATLAS Collaboration. *Improving jet substructure performance in ATLAS using TrackCaloClusters*. Tech. rep. Geneva: CERN, 2017. URL: %5Curl%7Bhttps://cds.cern.ch/record/2275636%7D.

[230] 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.

[231] 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.

[232] ATLAS Collaboration. *Flavor Tagging Efficiency Parametrisations with Graph Neural Networks*. Tech. rep. Geneva: CERN, 2022. URL: https://cds.cern.ch/record/2825433.

Appendices

APPENDIX A

Flavour Tagging

This Appendix lists some additional results in support of Chapter 5.

A.1 Understanding DIPS

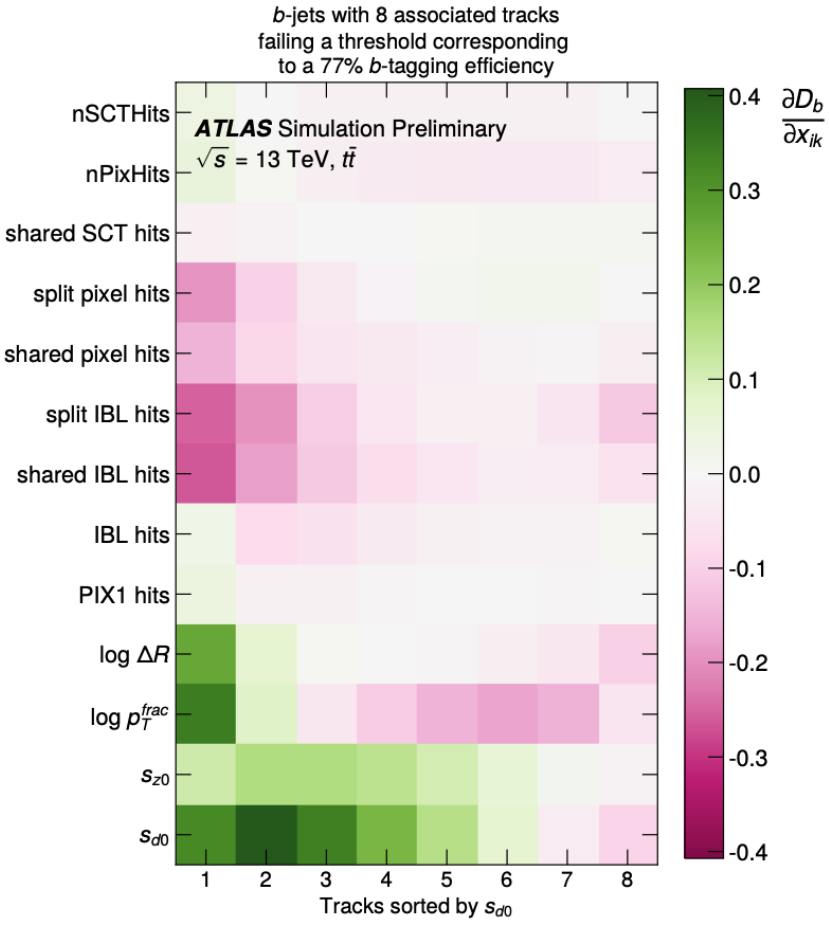


Figure A.1: Saliency map for b-tagging with 8 tracks sorted by $|S_{d_0}|$ and indexed by i, showing the gradient of the discriminant D_b with respect to the k track features x_{ik} [135].

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 A.1 presents the result of applying the *saliency maps* technique [223]. 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:

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

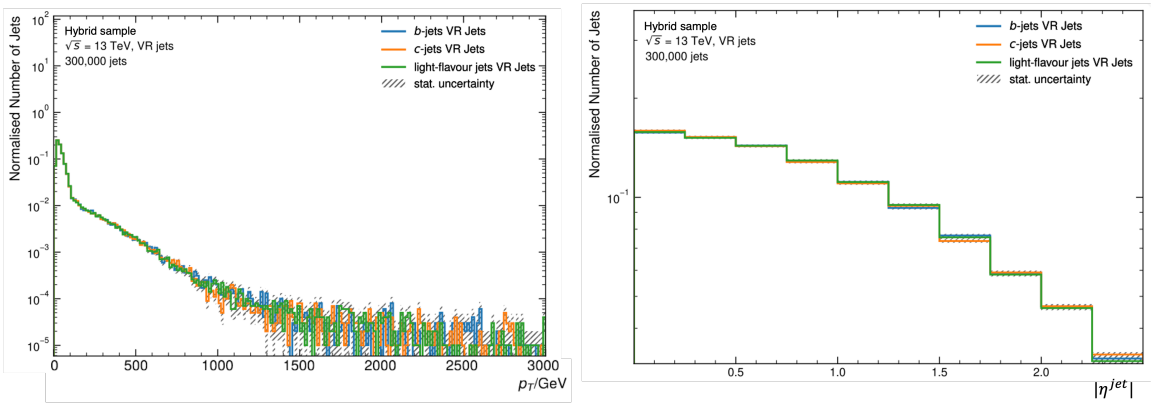
where i indexes the 8 tracks, j indexes the jet in the sample of size N, x_{ik} is the k^{th} feature of the i^{th} track [135]. This process effectively probes the linear sensitivity of the discriminant on the track features. Using the saliency map, one can infer what features to modify to correct the 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.

A.2 DIPS with Variable Radius Jets

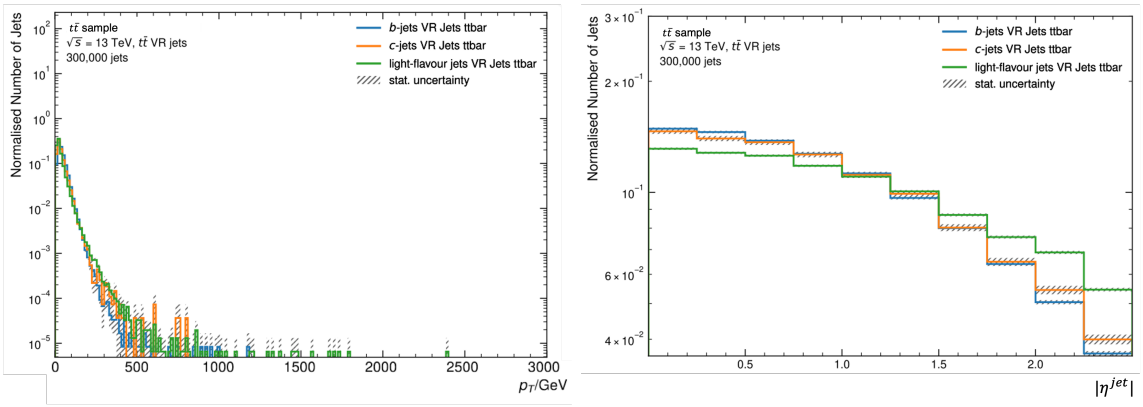
This section of the Appendix displays more information on the variable radius (VR) jet training of DIPS. The samples distributions are shown in Figure A.2.

A.3 DL1d with Variable Radius Jets

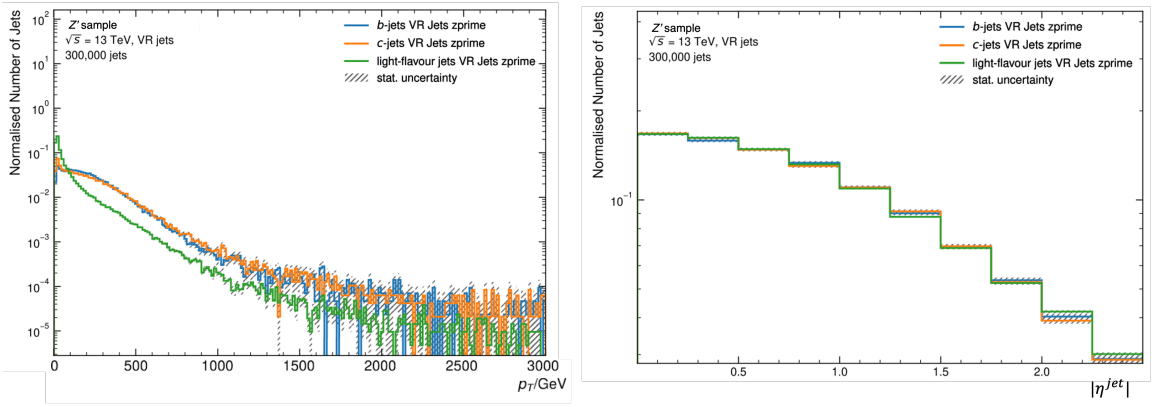
This section presents some plots on the VR-training of DL1d. Figure A.3 displays some flavour fractions scans for the b-tagging and c-tagging.



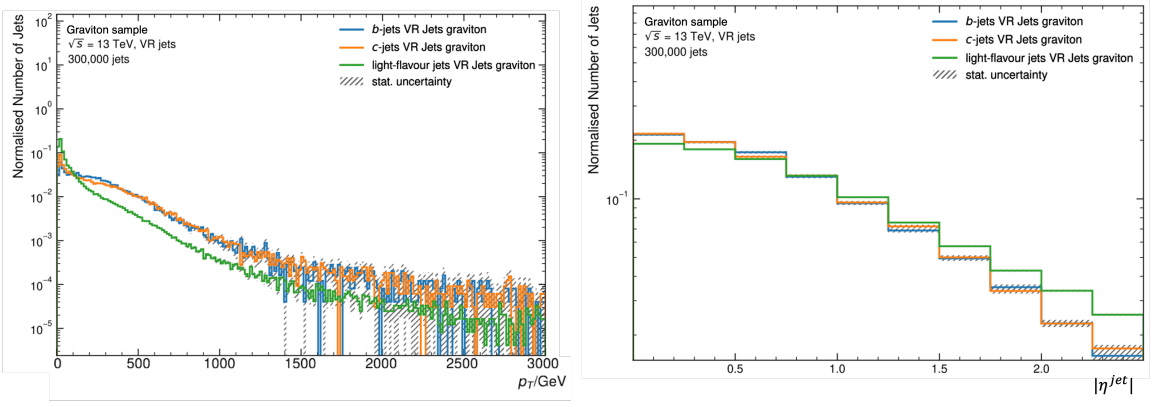
(a) Hybrid sample.



(b) $t\bar{t}$ sample.

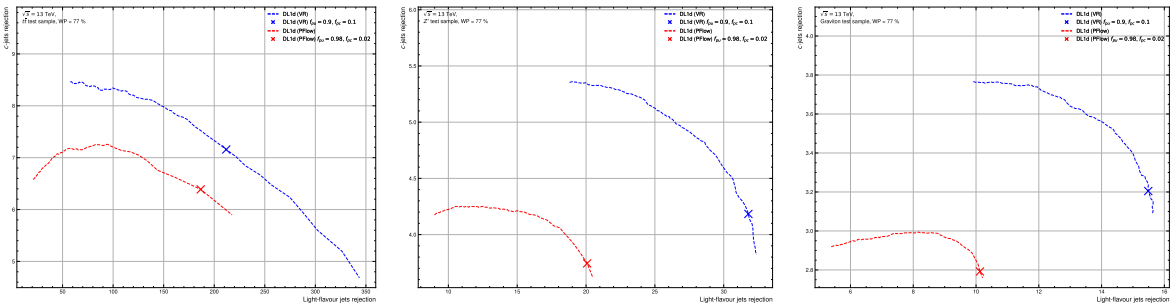


(c) Z' sample.

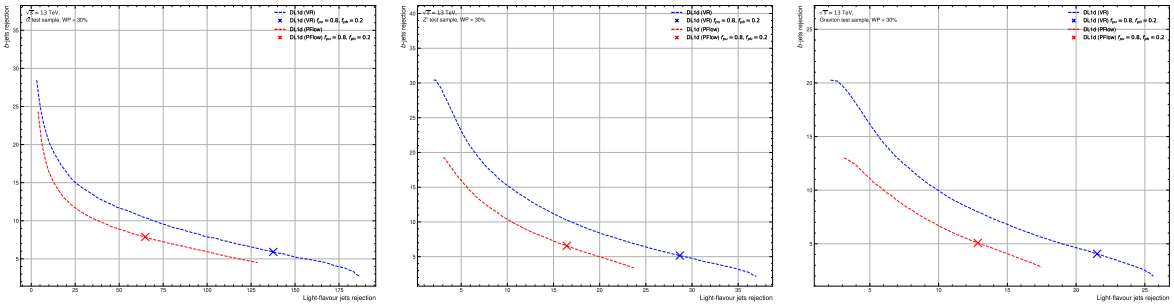


(d) Graviton sample.

Figure A.2: Distributions for the VR-jet training of jets p_T (left) and $|\eta|$ (right).



(a) Flavour fraction f_c^b for b-tagging scans.



(b) Flavour fraction f_b^c for c-tagging scans.

Figure A.3: The flavour fractions scans of the VR- and PFlow-trained DL1d model in blue and red respectively: left is $t\bar{t}$, centre Z', and right the graviton test samples. 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.

A.4 GN2 public plots

A comparison of the global performance of this GN2 model to the DL1d and GN1 models is displayed in the b- and c-tagging ROC curves of Figures 5.31 and 5.32. These results are taken from Ref [5], for which the DL1d model was retrained on the same dataset as GN2, and the DL1r and GN1 models are taken from Chapter 5.3.1. GN2 delivers yet another significant boost to performance, drasticaly surpassing the GN1 rejections at all efficiencies considered. The largest improvement is again 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 $\times 1.75$ ($\times 1.2$) 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 $\times 3.4$ ($\times 4$) on Z'. Concerning c-tagging, a similar large performance gained is obtained by the new GNN family over DL1d, although the change on the $t\bar{t}$ is more impressive for the b-jet ratio than for light-jet. This indicates a non-optimal choice for the flavour fraction f_b^c, which was set at 0.2 for all models.

A.5 GN2 supporting plots

This section presents more plots in support of Chapter 5.3.2. Figure A.6 presents the c-tagging efficiency per bin for an overall c-tagging working point of 30% per region displayed. Figure A.7 presents the c-tagging efficiency per bin for a per bin light-rejection of 50 for $t\bar{t}$ and 10 for Z'. The GN2 performance dominates across the board, except for the highest energy bin of the Z'. Figures A.7 presents the c- and light-rejection at an inclusive 70% b-tagging WP. The equivalent information for c-tagging at a c-tagging WP of 30% is displayed in Figures A.10 and A.11 for b- and light-rejection.

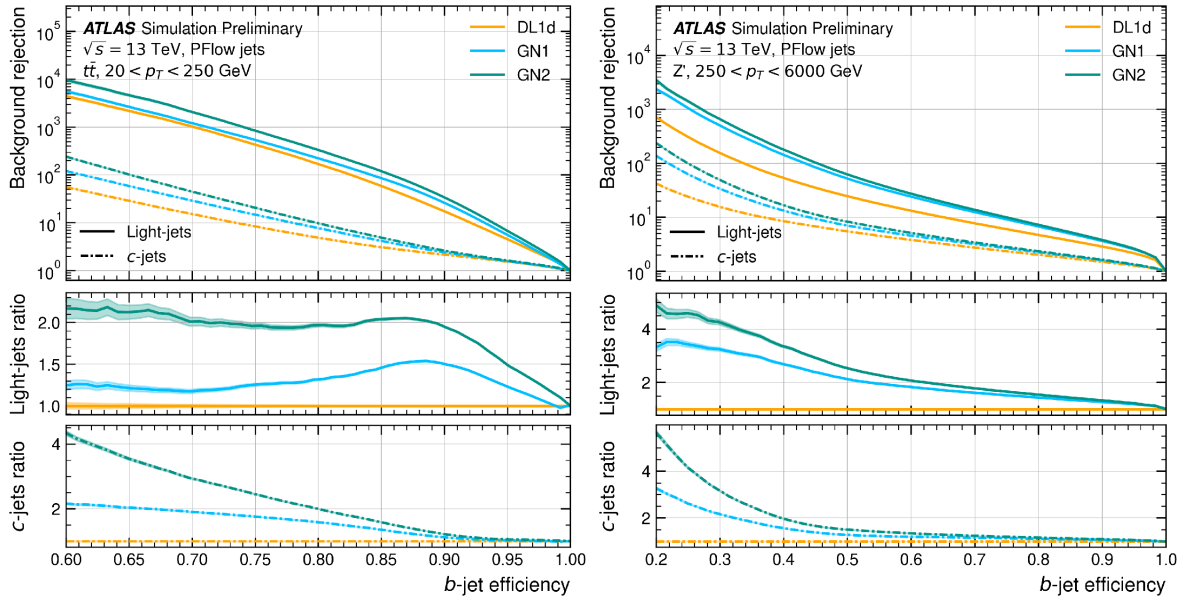


Figure A.4: The c- and light-rejections as a function of the b-jet tagging efficiency in the $t\bar{t}$ with $20 < p_T < 250$ GeV (left) and Z' with $250 < p_T < 6000$ GeV (right) test samples, from [5]. Models compared are DL1d in orange, GN1 in turquoise, and GN2 in blue. The bottom plots show the ratio with respect to the DL1d performance. Flavour fractions are set at $f_c^b = 0.018$ for DL1d, 0.05 for GN1, and 0.1 for GN2. Shaded regions represent the binominal error band.

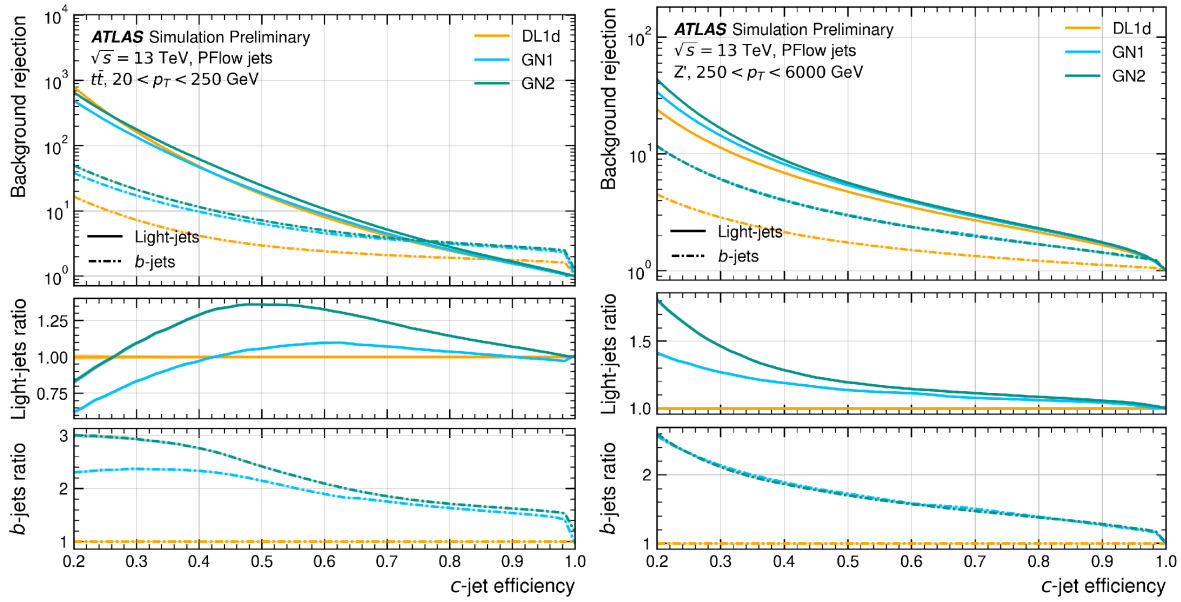


Figure A.5: The b- and light-rejections as a function of the c-jet tagging efficiency in the $t\bar{t}$ with $20 < p_T < 250$ GeV (left) and Z' with $250 < p_T < 6000$ GeV (right) test samples, from [5]. Models compared are DL1d in orange, GN1 in turquoise, and GN2 in blue. The bottom plots show the ratio with respect to the DL1d performance. Flavour fractions are set at $f_b^c = 0.2$ for all taggers. Shaded regions represent the binominal error band.

A.6 GN2X: GN2 Variant for Boosted Higgs Decays to Heavy Flavours

This section presents an interesting application of the GN2 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

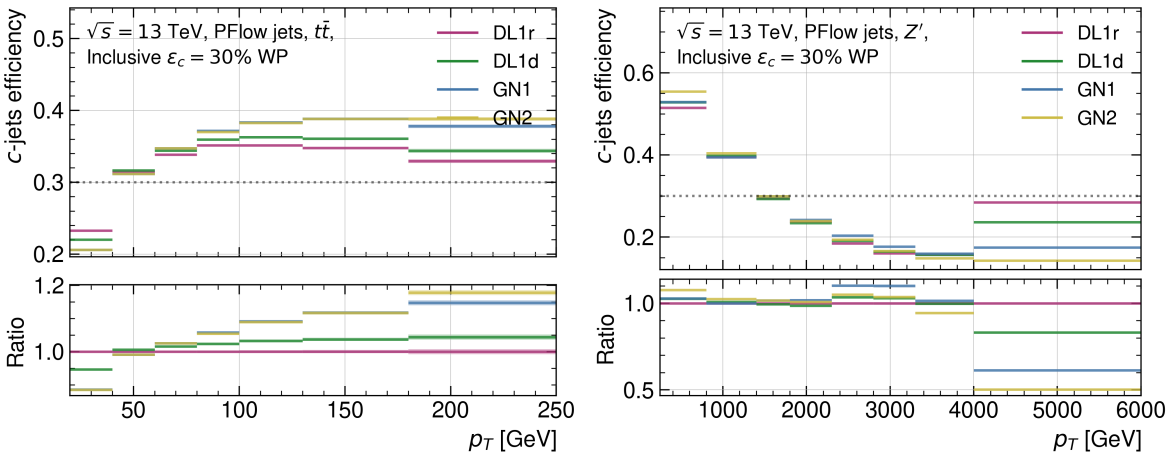


Figure A.6: Comparing the different models c-tagging efficiency as a function of jet p_T for the inclusive c-tagging 30% working point on the $t\bar{t}$ (left) and Z' (right). The flavour fraction is set at $f_b^c = 0.2$ for all taggers.

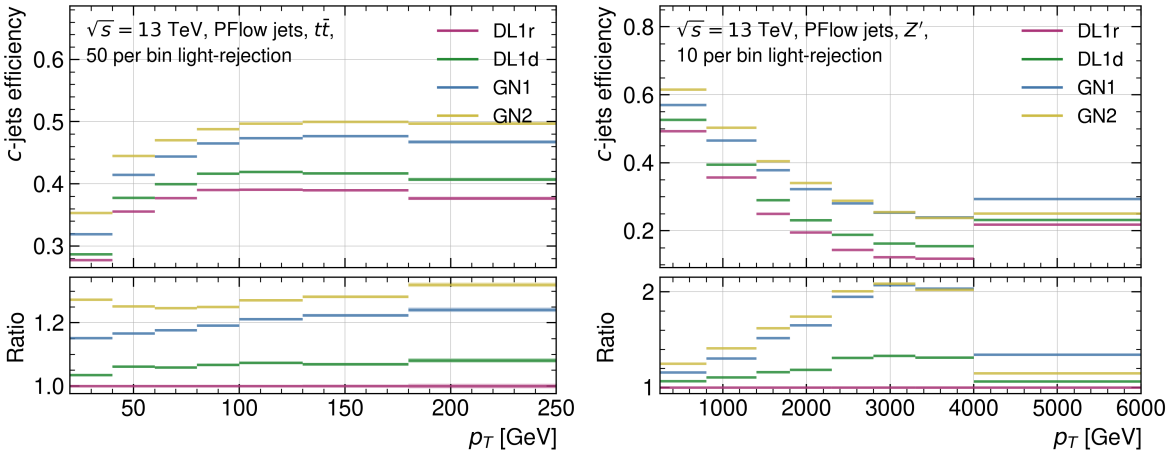


Figure A.7: Comparing the different models c-tagging efficiency as a function of jet p_T at a fixed light-jet rejection per bin of 50 for the $t\bar{t}$ (left) and 10 for the Z' (right) test samples. The flavour fraction is set at $f_b^c = 0.2$ for all taggers.

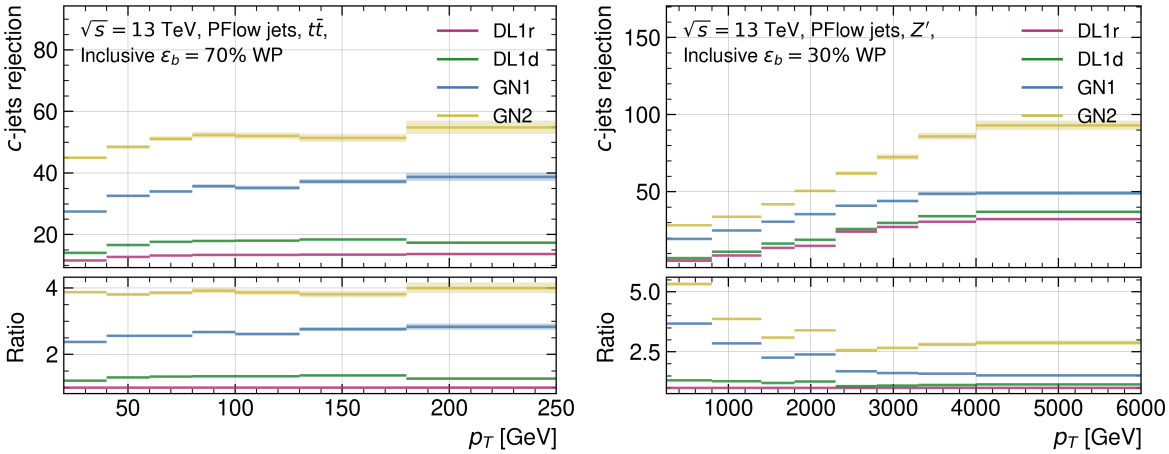


Figure A.8: Comparing the different models c-rejection as a function of jet p_T for the b-tagging inclusive 70% working point on the $t\bar{t}$ (left) and 30% working point on Z' (right). The flavour fraction is set at $f_c^b = 0.018$ for DL1r and DL1d, 0.05 for GN1, and 0.1 for GN2.

Higgs particles to a $c\bar{c}$ pair [224], for the precise measurement of the Higgs boson p_T spectrum

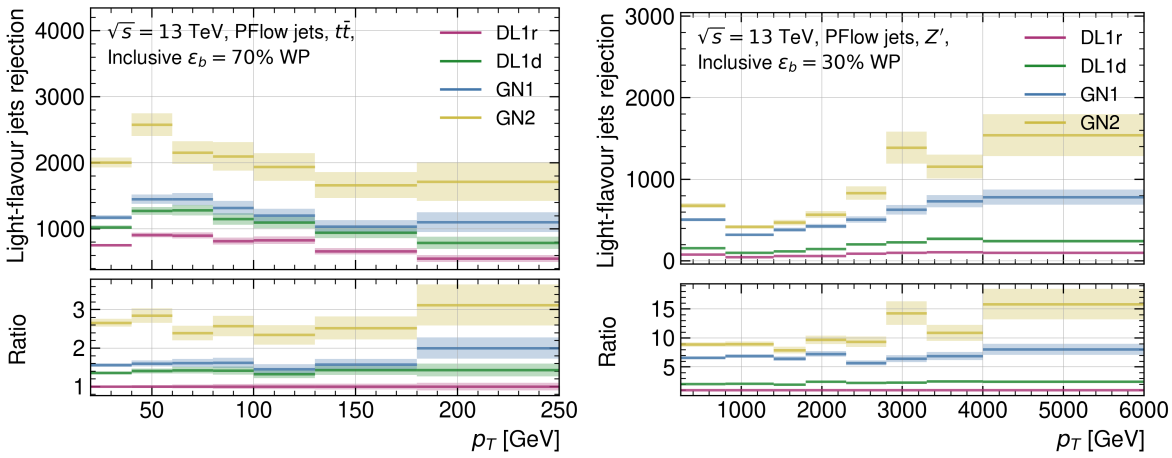


Figure A.9: Comparing the different models light-rejection as a function of jet p_T for the b-tagging inclusive 70% working point on the $t\bar{t}$ (left) and 30% working point on Z' (right). The flavour fraction is set at $f_c^b = 0.018$ for DL1r and DL1d, 0.05 for GN1, and 0.1 for GN2.

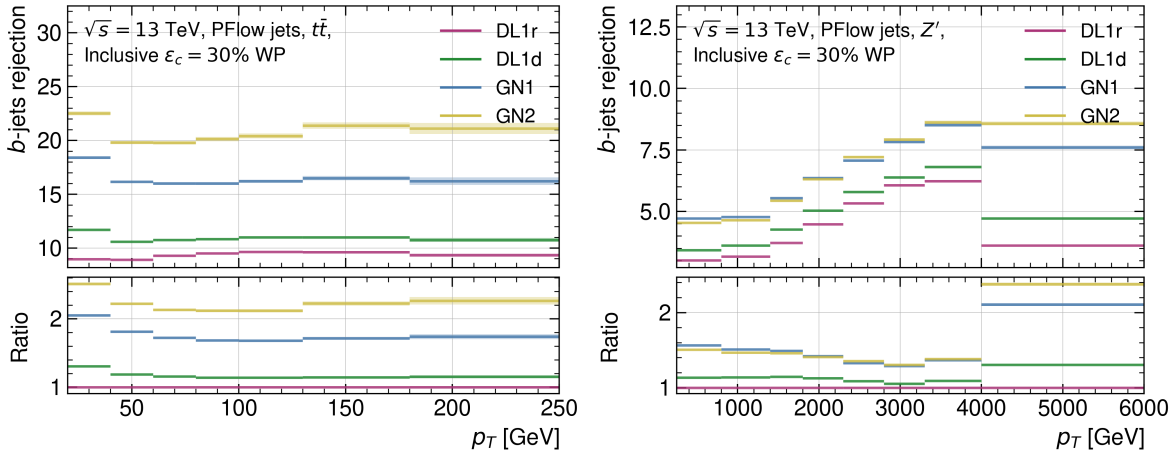


Figure A.10: Comparing the different models b-rejection as a function of jet p_T for the c-tagging inclusive 30% working point on the $t\bar{t}$ (left) and Z' (right). The flavour fraction is set at $f_b^c = 0.2$ for all taggers.

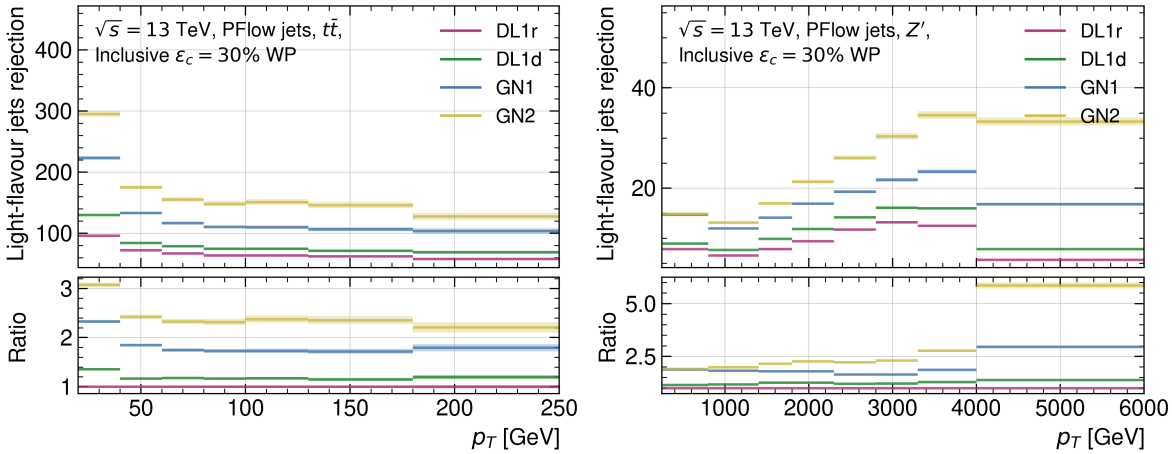


Figure A.11: Comparing the different models light-rejection as a function of jet p_T for the c-tagging inclusive 30% working point on the $t\bar{t}$ (left) and Z' (right). The flavour fraction is set at $f_b^c = 0.2$ for all taggers.

[225], and for beyond the SM measurements [226]. To perform this task, a new algorithm labelled

GN2X is introduced based on the design of GN2 [6]. 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 [227]. UFO combines PFlow [228] and Track-Calo clusters objects [229], thereby including neutral and charged components in the reconstruction. UFO large-R jets are reconstructed with the anti-k_T algorithm with a radius $R = 1.0$ [78].

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. [6]. 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 [230, 231]. 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.

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:

$$D_{H_{b\bar{b}}} = \log \frac{p_{H_{b\bar{b}}}}{f_{H_{c\bar{c}}} \cdot p_{H_{c\bar{c}}} + f_{\text{top}} \cdot p_{\text{top}} + (1 - f_{H_{c\bar{c}}} - f_{\text{top}}) \cdot p_{\text{QCD}}}, \tag{A.2}$$

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:

$$D_{H_{c\bar{c}}} = \log \frac{p_{H_{c\bar{c}}}}{f_{H_{b\bar{b}}} \cdot p_{H_{b\bar{b}}} + f_{\text{top}} \cdot p_{\text{top}} + (1 - f_{H_{b\bar{b}}} - f_{\text{top}}) \cdot p_{\text{QCD}}}, \tag{A.3}$$

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 A.12. 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,

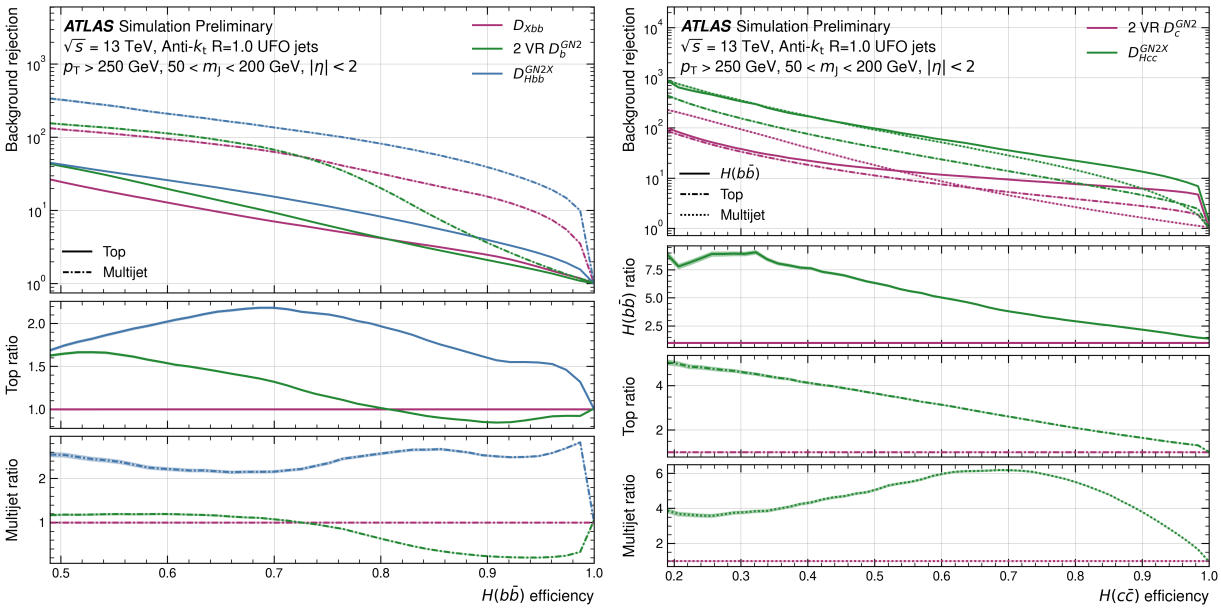


Figure A.12: The ROC curves for $H(b\bar{b})$ (left) and $H(c\bar{c})$ tagging (right) on an SM simulated test samples [6]. The respective tagging efficiency is displayed versus the top and multi-jet rejections, for jets with a $p_T > 250$ GeV and a mass $50 < m_J < 200$ GeV. Models compared are the 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.

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) [6]. 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.

APPENDIX B

COMBINED $VH(H \to b\bar{b}/c\bar{c})$ ANALYSIS APPENDIX

This Appendix lists some additional results in support of Chapter 6.

B.1 Analysis Categorisation

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

B.1.1 The ΔR Cut Between Higgs Candidate Jets

The angular separation between the two candidate jets $\Delta R(j_1, j_2)$, as defined in Equation 3.5, can be used to define a control region enriched in V+jets and $t\bar{t}$ backgrounds since these two processes give candidate jets with a flat angular spectrum while the signal peaks at low values of ΔR. A *high* ΔR control region (High ΔR CR) is defined using parametrised cuts on ΔR between the Higgs candidate jets as a function of p_T^V. An additional *low* ΔR control region (Low ΔR CR) for the 1L channel in the resolved $VH(H \to b\bar{b})$ is also introduced (for $VH(H \to c\bar{c})$, it is merged with the signal region). The philosophy behind the parametrisation of this function is to adapt the cut on the expected angular separation between the two Higgs candidate jets as a function of how boosted they are, as described by the p_T^V variable. For signal events, we expect the H and V to be approximately back-to-back hence p_T^V is a good proxy for p_T^H while benefiting from better experimental resolution, as it is reconstructed from leptons p_T and/or E_T, depending on the channel. From physical principals, boosted candidate jets are indeed expected to have a lower angular separation. The cuts are defined by fitting a template function $c_1 \times e^{c_2 + c_3 \times p_T^V}$ to the $VH(H \to b\bar{b})$ selected events, so that:

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

The results of these fits for the 1L channel are displayed in Figure 6.12, showing the signal yield in a 2-dimensional histogram (p_T^V vs $\Delta R_{c\bar{c}}$) for different tags applied. Cuts derived on the $VH(H \to c\bar{c})$ selected events showed good agreement with the $VH(H \to b\bar{b})$ derived cuts.

The $VH(H \to b\bar{b})$ cuts is chosen so that the kinematic selection of the two analyses is harmonised.

B.1.2 Resolved Top Control Region in 0L and 1L

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

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

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

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

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

- top(bb): in this case, the two b-jets produced during the $t\bar{t}$ decay are selected. This is a small component in the signal regions of the $VH(H \to c\bar{c})$ analysis, due to the 70% efficiency WP for b-tagging and the low mistag rate for b-jets in c-tagging. Naturally, in $VH(H \to b\bar{b})$ it is the leading contribution. Due to the origin of the candidate jets, a large $\Delta R_{b\bar{b}}$ is expected between the two b-jets so this component is most effectively constrained by the High ΔR CR.

- top(bc): where the b is from a t decay and the c from a subsequent W hadronic decay (or from ISR/FSR though this is less likely). Given the definition of the Top CR, this is the dominating component in that region and the most important to constrain in the signal regions of the $VH(H \to b\bar{b}/c\bar{c})$ analyses due to its signal-like kinematics.

- top(bl): where l stands for anything not b nor c (light jets predominantly but also some mistagged hadronic τ). This component is similar to the top(bc) as it also consists of a b + a jet from the W and can end up in the SRs and Top CRs due to mistags.

- top(lq): where l is as above and q can be any sort of jet except a b. This is a small component that mostly accumulates in the background-like part of the BDT score distribution. It is not constrained in the high ΔR regions nor the Top CRs.

The signal region distributions in the 1L channel in the p_T^V range $[150, 250]$ GeV are displayed in Section B.5 of the Appendix. While the top is not the dominant background, except in the tighter tagged TT 3-jet region, its relative contribution to the background composition increases at signal-like values of the discriminant.

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

- 2-jet: BL: BL; BT: BT

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

In the *AllSignal* strategy, the Higgs candidates in the Top CR are always the highest p_T b- and c-tagged jets. This selection was observed to make the top control region distributions more closely match the distributions in the signal regions. For the fit, only the BT region is used, as it provides sufficient control on the important top background components. The BL region can however be studied to assess the data-MC agreement after correcting the yields of the major backgrounds from the fit, as is shown in Figure 6.28a.

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

B.1.3 Truth Tagging

The tagging method described in Section 6.5, referred to as *direct tagging*, is a cut-based method where a jet passes or fails a threshold cut, as defined by dedicated working points in the Pseudo-Continuous Flavour Tagging (PCFT) or PCBT schemes. These WP have a large rejection for b-tagging due to the good performance of the method. For c-jets, the tagging efficiency is low and many c-jets end up rejected by the selection. This problem is compounded by the event selection criteria, requiring two b-tags or at least one tight c-tag to enter the analysis' regions. Only a part of the events in the simulated samples satisfy these requirements, and most are discarded from the analysis. Having sufficient MC statistics in all regions is essential to effectively model the backgrounds and reduce the MC statistical uncertainty. An alternative approach to direct tagging used in the analysis to retain the large MC statistics is *truth tagging*. Rather than applying a pass-fail decision, truth tagging reweighs events by their probability of being selected at a specific working point, based on truth information only available in the simulated samples. The tagging scale factors are applied in the analysis after truth tagging.

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

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

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

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

the expression for w can be factorised as [232]:

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

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

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

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

Technically, truth tagging was deployed with map-based 2D histograms $p_T - \eta$ parametrising the tagging efficiency of the jets in the latest standalone $VH(H \to b\bar{b})$ and $VH(H \to c\bar{c})$ analysis [130, 188]. Such histograms are called *efficiency map*, leading to the implementation being referred to *map-based truth tagging*. These maps were derived individually for each b-, c-, light-, and τ-jets flavour and each working point. A further possibility is to combine direct tagging with truth tagging into the so-called *hybrid tagging* strategy, in which a portion of the events are direct tagged and the rest is truth tagged. This last approach limits the mismodelling incurred by truth tagging and remove the need to correct for non-closure effects.

A new approach considered for the combined analysis relies on a GNN to perform the so-called *GNN truth tagging* [232]. This removes the statistical dispersion limitation of high-dimension efficiency maps. Interestingly, it also becomes possible to include more variables to parametrise the efficiency, leading to better agreement with the direct tagging distribution comparing to map-

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

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

based truth tagging. The network builds a fully-connected graph with several layers message-passing updates [108], where each node represents a jet in the event[1]. The features per node are the jet-level and event-level variables listed in Table B.1, with the angular separation between the jets set as edge between the nodes. Finally, a fully-connected NN receives the last update graph and outputs all track-jets or jets flavour-tagging efficiencies.

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

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

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

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

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

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

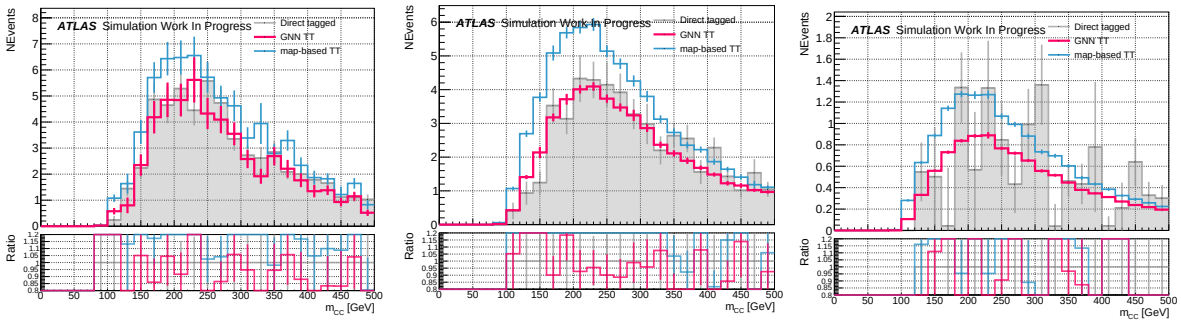
process where hybrid tagging is deployed, with both b-jets being direct tagged at the 70% WP.

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

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

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

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



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

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

B.2 MVA Variables

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

Input variables for the resolved regime:

- p_T^V: transverse energy of the vector boson. In 0-lepton channel it is equivalent to the missing transverse energy (E_T^{miss}); in 1-lepton channel it is the vector sum of E_T^{miss} and the lepton p_T; in 2-lepton channel, it is the vector sum of the 2 charged lepton p_T.
- $p_{\text{T}}^{j_1}$ and $p_{\text{T}}^{j_2}$: transverse momenta of the Higgs candidate jets, j_1 has higher p_T.
- $m_{j_1 j_2}$ or m_J: invariant mass of the reconstructed H system, depending on the analysis regime.
- $\Delta R(j_1, j_2)$: angular distance between the two Higgs-candidate jets, defined as

  $$\Delta R(i,j) = \sqrt{(\Delta\phi(i,j))^2 + (\Delta\eta(i,j))^2}$$

  with $\Delta\phi(i,j) = \phi_i - \phi_j$ the azimuthal and $\Delta\eta(i,j) = \eta_i - \eta_j$ the pseudorapidity distances.
- $m_{j_1 j_2 j_3}$: invariant mass of two Higgs-candidate jets and the remaining jet with highest p_T. When there are only 2 jets in an event, $m_{j_1 j_2 j_3} = m_{j_1 j_2}$.
- $\Delta\phi(V, H)$: azimuthal distance between the reconstructed vector boson V and Higgs boson candidates H.
- $\text{bin}_{\text{DL1r}(j1)}$, $\text{bin}_{\text{DL1r}(j2)}$: variable showing the tagged-bin the jet or track-jet j_1 belongs to (5 possible bins, as defined in Section 6.5) - the untagged N, the loose (70% WP) and the tight (60% WP) b-tagged, and the loose and the tight c-tagged bins. In the MVA, the value of the two Higgs-candidate jets or track-jets are used.
- $\sum_{i \neq 1,2} p_T^{j_i}$: p_T sum of non H candidate jets that have $p_T > 20$ GeV.
- **0-lepton channel variables**:
  - $|\Delta\eta(j_1, j_2)|$: absolute value of the pseudorapidity distances between the two Higgs-candidate jets or track-jets.
  - $\min\{\Delta R(j_i, j)\}_{i=1,2}$: the distance in R between the closest b- or c-tagged Higgs candidate jet and an additional jet with $p_T^V > 20$ GeV.
  - m_{eff}: the scalar sum of the $p_T$of all small-R jets and E_T^{miss} in the event.
- **1-lepton channel variables**:
  - m_T^W: transverse mass of the W boson candidate reconstructed from the lepton and E_T^{miss}, as presented in the 1L-specific selection of Section 6.5.2.
  - E_T^{miss}: missing transverse energy.
  - $\Delta y(V, H)$: rapidity difference between the V and H.
  - $\min\left[\Delta\phi(l, j_i)\right]_{i=1,2}$: distance in ϕ between the lepton and the closest b-tagged (c-tagged) H candidate jet.
  - m_{top}: reconstructed mass of the leptonically decaying top quark. The longitudinal momentum of the neutrino (p_z^ν) is first reconstructed the mass of the W boson, and selected to minimise the reconstructed m_{top} with the 2 Higgs candidates.
- **2-lepton channel variables**:

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

Input variables for the boosted regime:

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

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

  where θ is the angle between the indexed jets, j_3 is the leading additional jet, and j_2 are the H candidate jets.
- $\text{bin}_{\text{DL1r}(j1,\text{trk})}$, $\text{bin}_{\text{DL1r}(j2,\text{trk})}$: corresponds to the tagged-bin the track-jet belongs to (4 possible bins): the 85%, the 77%, the 70% and 60% b-tagging efficiency bins.
- **0-lepton channel specific variables**
  - E_T^{miss}: missing transverse energy, same as p_T^V.
- **1-lepton channel specific variables**
  - $\Delta y(\boldsymbol{V}, \boldsymbol{H})$: same as in the resolved regime.
  - p_T^l: transverse momentum of the lepton.
  - $(p_T^l - E_T^{\text{miss}})/p_T^W$: proxy for the p_T imbalance of the charged lepton and the neutrino of the W-boson.
- **2-lepton channel specific variables**
  - $\Delta y(V, H)$: same as in the resolved regime.
  - $\cos\theta(l^-, \mathbf{Z})$: same as in the resolved regime.

B.3 Top Modelling Uncertainties in the Fit

There are many processes of relevance in a complex analysis such as the $VH(H \to b\bar{b}/c\bar{c})$. These must individually be modelled, with studies of the pulls of the different systematics required to verify the fit correctly accounts for the background's contribution to the analysis. The risk with such a complex fit structure with large numbers of NPs necessary to model a large variety of effects is to give the fit too much freedom and, in a sense, overfit to the data distributions. To highlight the process, some top-related pulls are shown in Figure B.2, with pulls displayed for both top acceptance uncertainties and CARL shape systematics. Again, there is good agreement for most pulls between the VH and VZ analyses. In the acceptance systematics part, a very large significant pull is observed for the so-called "*MetTrigTop*", an E_T^{miss} trigger related experimental uncertainty derived from the top-process, as described in 6.7.

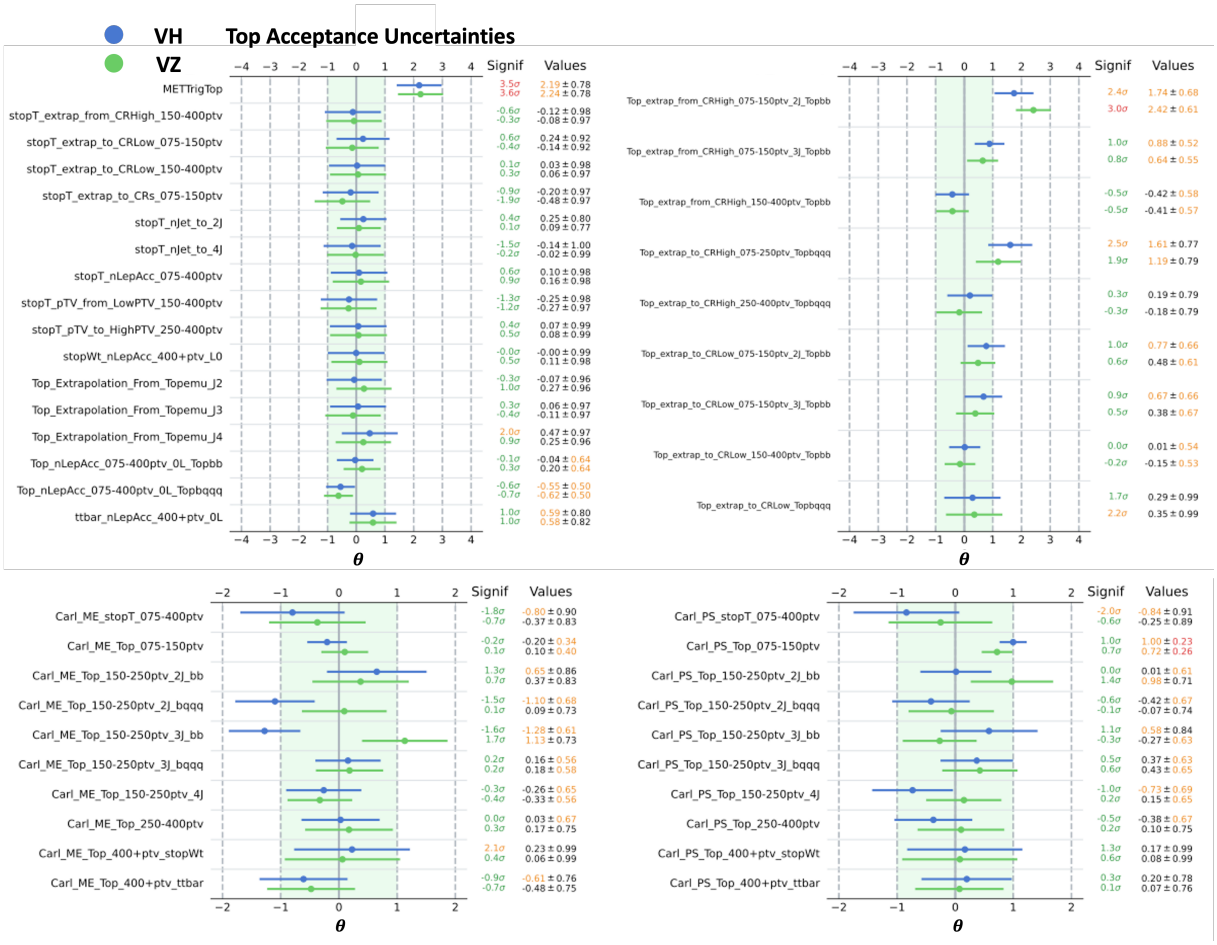


Figure B.2: Some Top nuisance parameters related to acceptance uncertainties (top) and CARL shapes (bottom) in the combined analysis targeting the $VH(H \to b\bar{b}/c\bar{c})$ in blue, versus the cross-check analysis $VZ(\to b\bar{b}/c\bar{c})$ in green.

Other uncertainties presented cover the region extrapolation for the single-top t (left) and combined Top process (right), as well as N_{jet} and p_T^V (for single-top t), lepton channel extrapolations, and the extrapolation from the Top e/μ CR. Most of the NPs are not significantly pulled, with little constraining. This indicates that the fit is not very sensitive nor requires the effect they implement. One exception is the Top(bb) extrapolation from the CRHigh in 75 GeV $< p_T^V <$ 150 GeV with 2-jet: the NP is largely pulled, and even more so in the VZ cross-check analysis. This feature can be understood from the large presence of V+jets and Top in the 0L and 1L CRHigh, leading to an interplay between the two processes when shifting the focus towards the signal part of V+jets. This interplay is visible in the correlation of the boosted $t\bar{t}$ and W+hf in Figure 6.31.

B.4 Signal and Background Modelling

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

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

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

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

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

Acceptance Ratio Name	**Applied**	**Value**
W+hf normalistion	W+hf	floating
W+mf normalistion	W+mf	floating
W+lf normalistion	W+lf	floating
W+lf normalistion	W+lf1L p_T^V 150-250	25%
Wcc/Wbb ratio	Wcc, 1L p_T^V 75-150	20%
Wcc/Wbb ratio	Wcc, 1L p_T^V >150, 2-jet	4%
Wcc/Wbb ratio	Wcc, 1L p_T^V >150, 3-jet	15%
Wcc/Wbb ratio	Wcc, $VH(H \to b\bar{b})$, 0L, 2-jet	4%
Wcc/Wbb ratio	Wcc, $VH(H \to b\bar{b})$, 0L, 3-jet	10%
Wcc/Wbb ratio	Wcc, $VH(H \to b\bar{b})$, 0L, 4-jet	10%
Wcc/Wbb ratio	Wcc, $VH(H \to c\bar{c})$, 0L	25%
Wbc/Wcl ratio	Wbc, p_T^V 75-150	24%
Wbc/Wcl ratio	Wbc, p_T^V 150-250, 2-jet	24%
Wbc/Wcl ratio	Wbc, p_T^V 150-250, 3-jet	14%
Wbc/Wcl ratio	Wbc, p_T^V >250	14%
Wbl/Wcl ratio	Wbl, p_T^V 75-150	29%
Wbc/Wcl ratio	Wbc, p_T^V 150-250, 2-jet	29%
Wbc/Wcl ratio	Wbc, p_T^V 150-250, 3-jet	22%
Wbc/Wcl ratio	Wbc, p_T^V >250, 2-jet	19%
Wbc/Wcl ratio	Wbc, p_T^V >250, 3-jet	12%
Wbc/Wcl ratio	Wbc, 0L, 4-jet	8%
$Wq\tau/Wcl$ ratio	$Wc\tau$, $Wb\tau$	20%
$Wl\tau/Wcl$ ratio	$Wl\tau$, $W\tau\tau$	9%
W+hf CRHigh / SR+CRLow ratio	W+hf, 1L, CRHigh, p_T^V 75-150, 2-jet	3%
W+hf CRHigh / SR+CRLow ratio	W+hf, 1L, CRHigh, p_T^V 75-150, 3-jet	7%
W+hf CRHigh / SR+CRLow ratio	W+hf, 1L, CRHigh, p_T^V 150-250, 2-jet	30%
W+hf CRHigh / SR+CRLow ratio	W+hf, 1L, CRHigh, p_T^V 150-250, 3-jet	10%
W+hf CRHigh / SR+CRLow ratio	W+hf, 1L, CRHigh, p_T^V >250, 2-jet	50%
W+hf CRHigh / SR+CRLow ratio	W+hf, 1L, CRHigh, p_T^V >250, 3-jet	20%
W+hf CRHigh / SR+CRLow ratio	W+hf, 0L, CRHigh, 2-jet	30%
W+hf CRHigh / SR+CRLow ratio	W+hf, 0L, CRHigh, 3-jet	20%
W+hf CRHigh / SR+CRLow ratio	W+hf, 0L, CRHigh, p_T^V 150-250, 4-jet	10%
W+hf CRHigh / SR+CRLow ratio	W+hf, 0L, CRHigh, p_T^V >250, 4-jet	15%
W+hf SR / CRLow ratio	W+hf, 1L, SR, topCR, p_T^V 75-150, 2-jet	33%
W+hf SR / CRLow ratio	W+hf, 1L, SR, topCR, p_T^V 75-150, 3-jet	3%
W+hf SR / CRLow ratio	W+hf, 1L, SR, topCR, p_T^V 150-250, 2-jet	65%
W+hf SR / CRLow ratio	W+hf, 1L, SR, topCR, p_T^V 150-250, 3-jet	7%
W+hf SR / CRLow ratio	W+hf, 1L, SR, topCR, p_T^V >250, 2-jet	20%
W+hf SR / CRLow ratio	W+hf, 1L, SR, topCR, p_T^V >250, 3-jet	13%
W+mf CRHigh / SR ratio	W+mf, 1L, CRHigh, CRLow, p_T^V 75-150, 2-jet	2%
W+mf CRHigh / SR ratio	W+mf, 1L, CRHigh, CRLow, p_T^V 75-150, 3-jet	5%
W+mf SR / CRHigh ratio	W+mf, 01L, SR, topCR, CRLow p_T^V 150-250	7%
W+mf SR / CRHigh ratio	W+mf, 01L, SR, topCR, CRLow p_T^V >250	16%
W+lf CRHigh / SR ratio	W+lf, 1L, CRHigh, p_T^V 75-150, 2-jet	5%
W+lf CRHigh / SR ratio	W+lf, 1L, CRHigh, p_T^V 75-150, 3-jet	10%
W+lf CRHigh / SR ratio	W+lf, 01L, CRHigh, p_T^V 150-250, 2-jet	5%
W+lf CRHigh / SR ratio	W+lf, 01L, CRHigh, p_T^V 150-250, 3-jet	10%
W+lf CRHigh / SR ratio	W+lf, 01L, CRHigh, p_T^V >250, 2-jet, 3-jet	17%
W+hf 4-jet / 3-jet ratio	W+hf, 0L, p_T^V 150-250, 4-jet	12%
W+hf 4-jet / 3-jet ratio	W+hf, 0L, p_T^V >250, 4-jet	20%
W+hf 0L / 1L ratio	W+hf, 0L, p_T^V 150-250, 2-jet	30%
W+hf 0L / 1L ratio	W+hf, 0L, p_T^V 150-250, 3(+)-jet	20%
W+hf 0L / 1L ratio	W+hf, 0L, p_T^V >250, 2-jet	20%
W+hf 0L / 1L ratio	W+hf, 0L, p_T^V >250, 3(+)-jet	13%
W+mf 0L / 1L ratio	W+mf, 0L, p_T^V 150-250, 2-jet	3%
W+mf 0L / 1L ratio	W+mf, 0L, p_T^V 150-250, 3-jet	8%
W+mf 0L / 1L ratio	W+mf, 0L, p_T^V >250	10%
W+lf 0L / 1L ratio	W+lf, 0L	4%

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

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

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

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

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

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

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

Acceptance Ratio Name	Applied	Value
stop-t normalisation	01L, all regions	17 %
stop-t CRLow+CRHigh / SR ratio	1L, 75-150 GeV, CRHigh and CRLow	3 %
stop-t CRLow / CRHigh ratio	1L, 75-150 GeV, CRLow	6 %
stop-t CRLow+SR / CRHigh ratio	01L, SR, TopCR, CRLow, decorr. 150-250 and 250-400 GeV	6 %
stop-t CRLow / SRratio	01L, CRLow, decorr. 150-250 and 250-400 GeV	17 %
stop-t 2-jet / 3-jet ratio	01L, 2-jet region	15 %
stop-t 4-jet / 2+3-jet ratio	0L, 4-jet region	15 %
stop-t p_T^V150-400 / 75-150 ratio	01L, decorr. 150-250 and 250-400 GeV	7 %
stop-t p_T^V250-400 / 150-250 ratio	01L, 250-400 GeV	15 %
stop-t 0L / 1L	0L	6 %
CARL ME stop-t shape	01L	–
CARL PS stop-t shape	01L	–
ISR stop-t shape	01L	–
FSR stop-t shape	01L	–

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

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

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

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

Table B.11: Diboson uncertainties in the boosted regime.

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

Table B.12: Diboson uncertainties in the resolved regime.

B.5 Analysis Postfit Regions

B.5.1 Resolved Postfit Regions

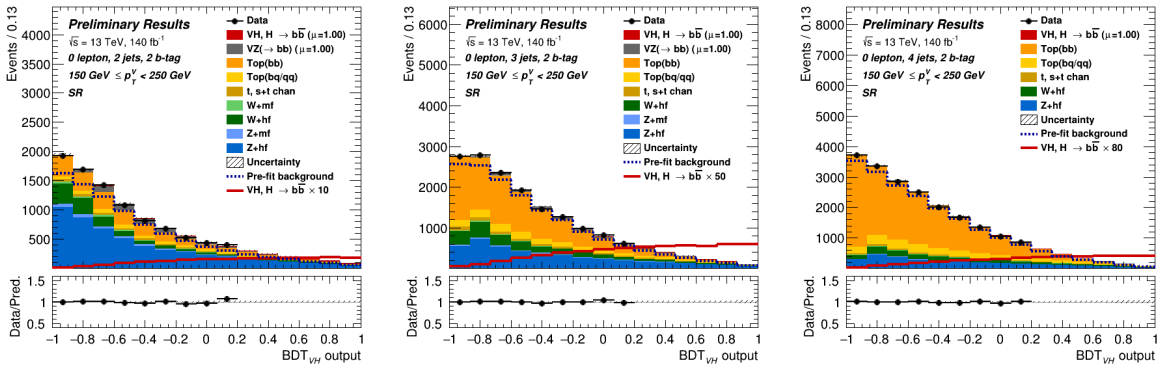
All regions in the resolved regime of the combined $VH(H \to b\bar{b}/c\bar{c})$ analysis after the conditional fit to data of Section 6.9 are presented here, organised by increasing number of charged lepton channel (0L, 1L, 2L). The distributions indicate the pre-fit expactations of the sum of processes in dashed blue lines and highlight multiples of either the $VH(H \to b\bar{b})$ or $VH(H \to c\bar{c})$ signal distributions in red lines. Figures B.3, B.5, and B.8 are the BB-tagged signal regions. The 2 c-tagged SRs are displayed in Figures B.10, B.15, and B.21. The 1 c-tagged SRs are displayed in Figures B.11, B.16, and B.22.

The control regions are presented in:

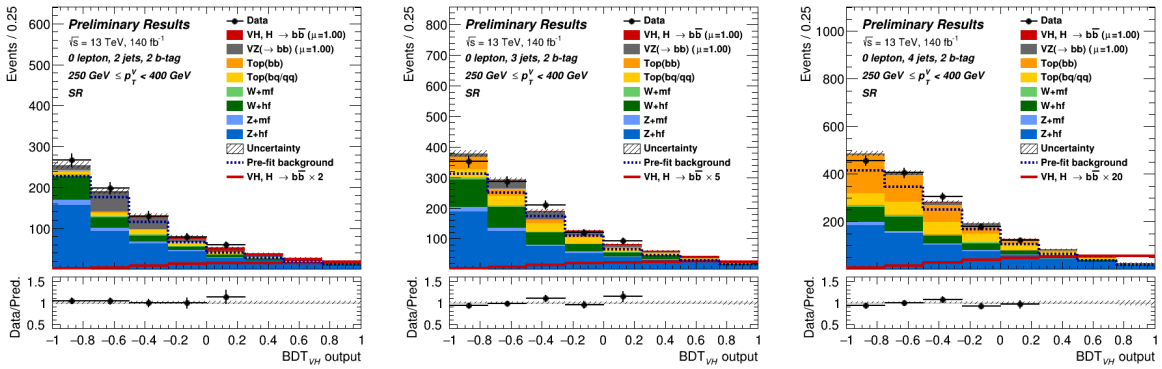
- The BB-tagged High ΔR CRs in Figures B.4, B.6, and B.9.
- The c-tagged (TN, TL, and TT) High ΔR CRs in Figures B.12, B.13, B.17, B.18, B.23, and B.24.
- The 1L BB-tagged Low ΔR CRs in Figure B.7.
- The 1L and 2L $V + l$ CRs (LN-tagged) in Figures B.20 and B.25.
- The 0L and 1L top CRs BT-tagged in Figures B.14 and B.19.
- The 2L top $e\mu$ CRs with ≥ 1 T-tag in Figure B.26.

B.5.2 Boosted Postfit Regions

This section presents the boosted regime regions after the conditional fit, with Figure B.27 presenting the 0L regions, Figure B.28 the 1L regions, and Figure B.29 the 2L regions.

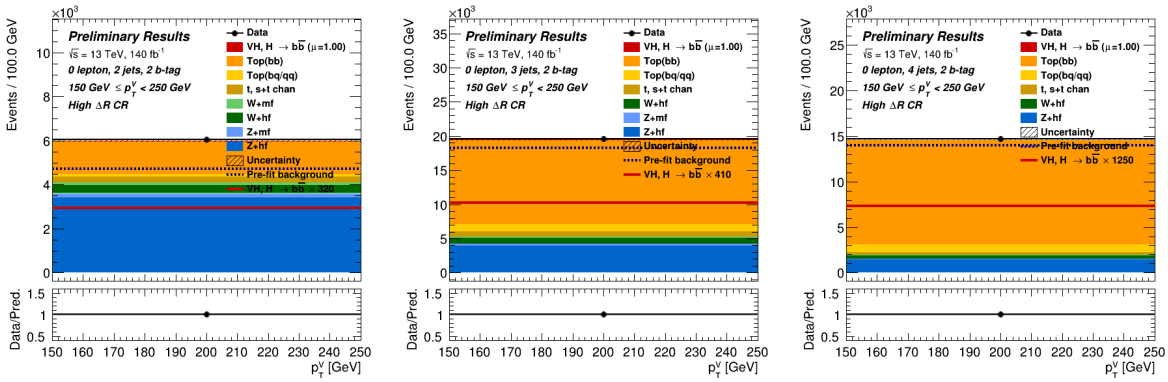


(a) $150 < p_T^V < 250$ GeV.

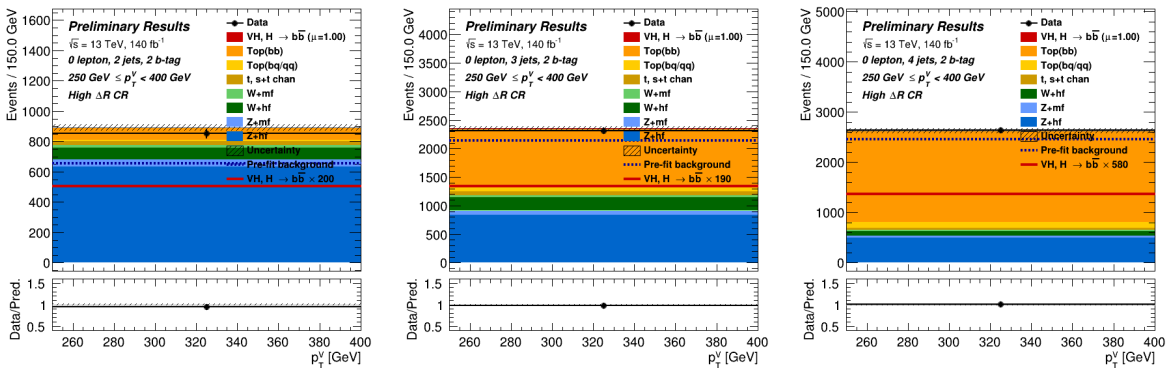


(b) $250 < p_T^V < 400$ GeV.

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

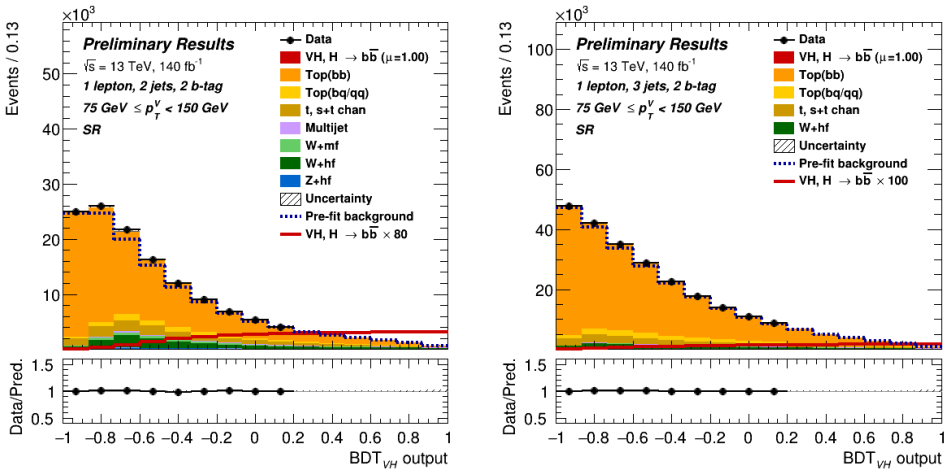


(a) $150 < p_T^V < 250$ GeV.

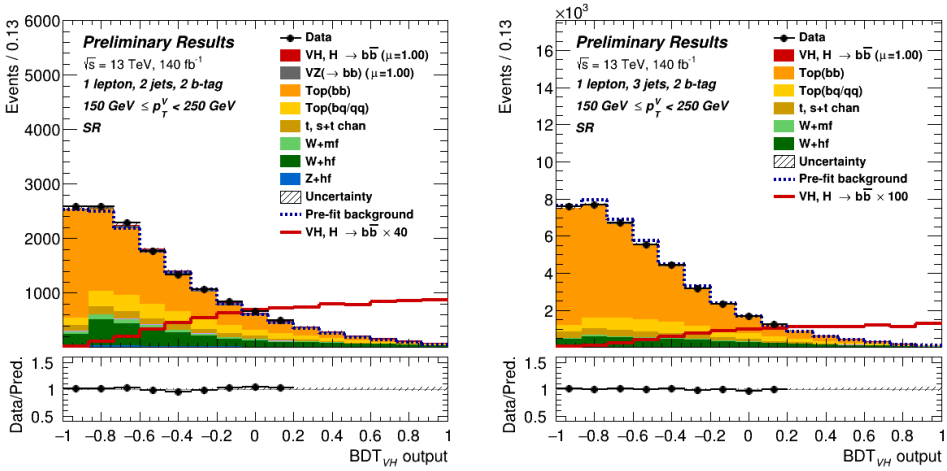


(b) $250 < p_T^V < 400$ GeV.

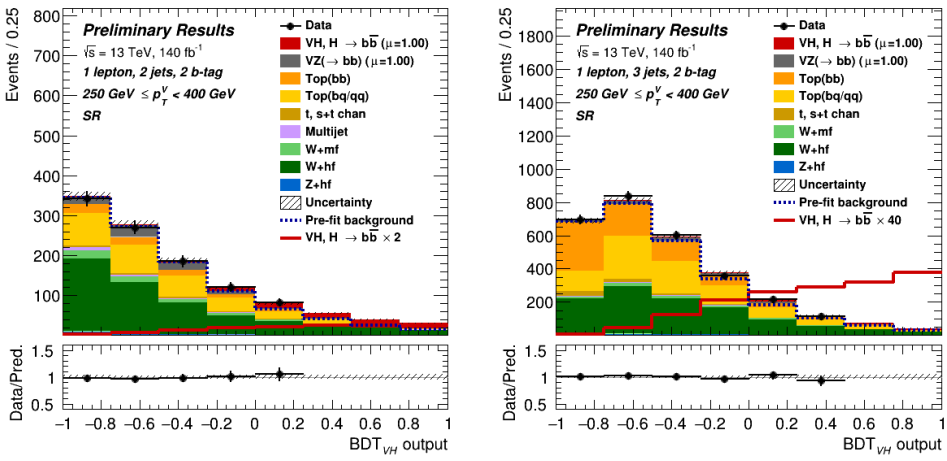
Figure B.4: The 0L High ΔR CR in the BB-tagged 2-jet (left), 3-jet (centre), and 4-jet (right).



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

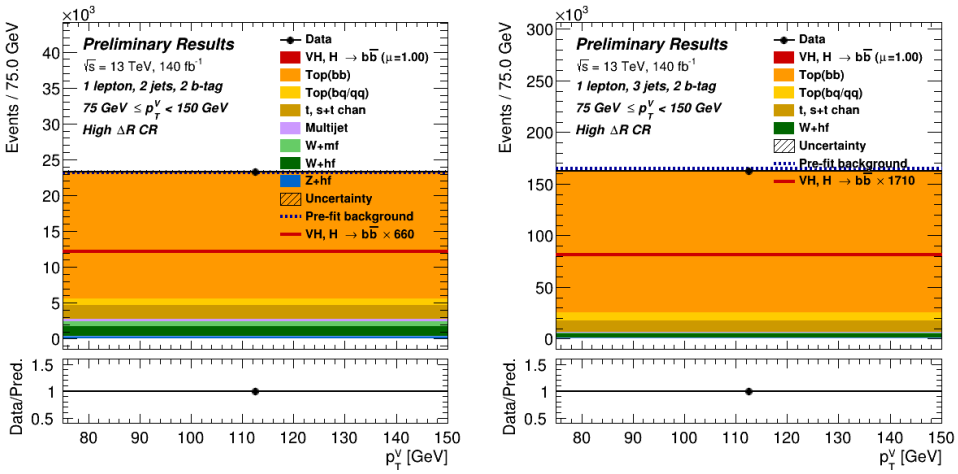


(b) $150 < p_T^V < 250$ GeV.

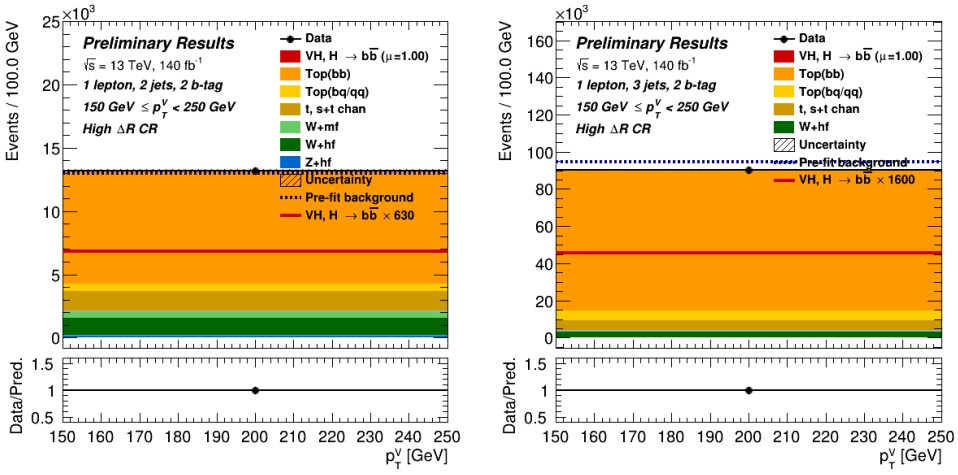


(c) $250 < p_T^V < 400$ GeV.

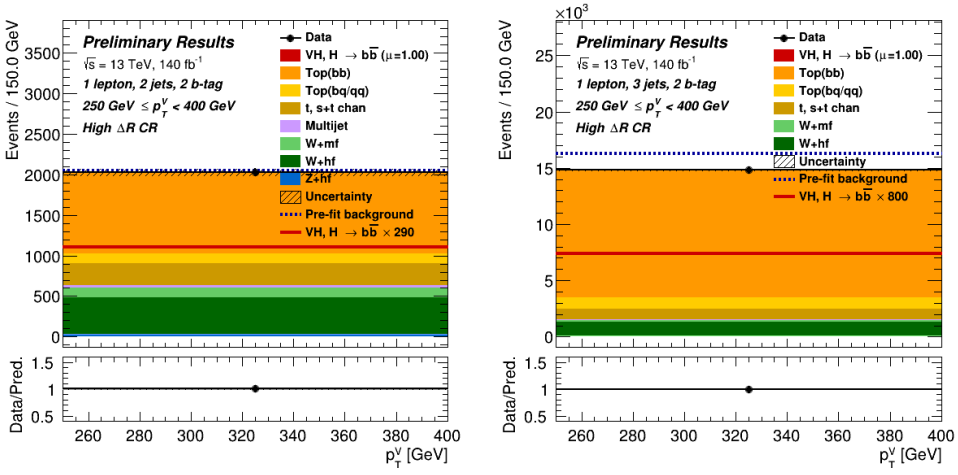
Figure B.5: The 1L signal regions in the BB-tagged 2-jet (left) and 3-jet (right).



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

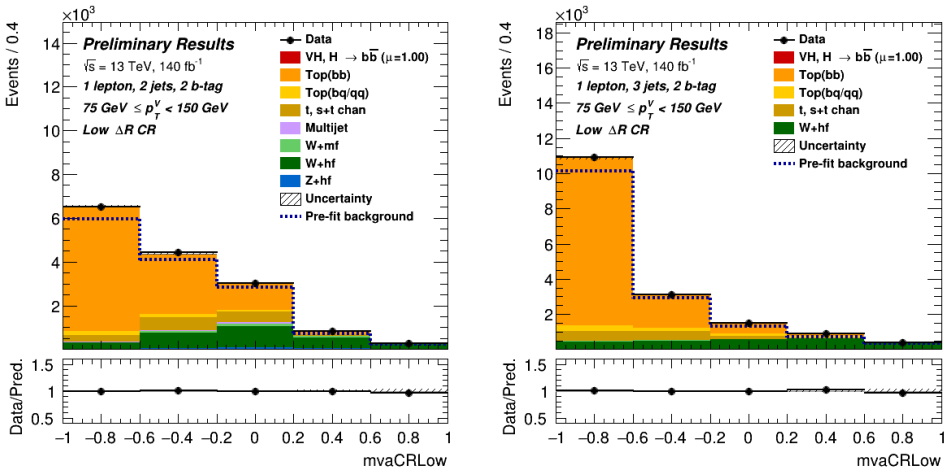


(b) $150 < p_T^V < 250$ GeV.

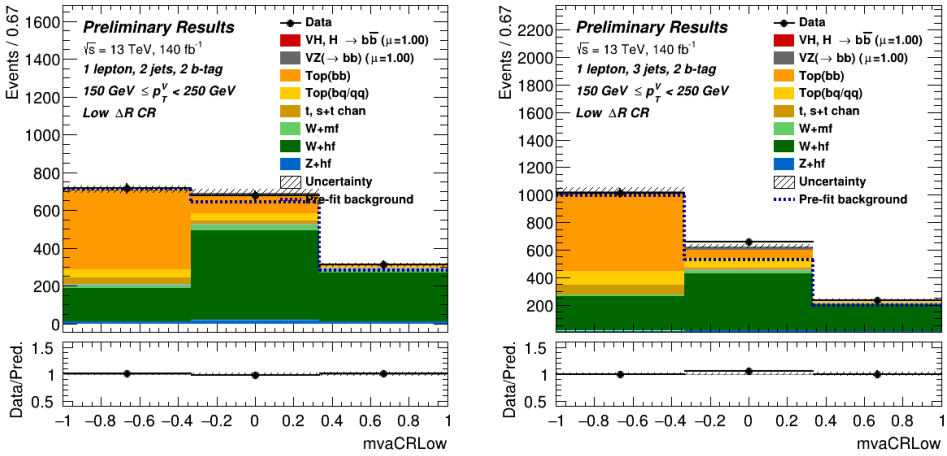


(c) $250 < p_T^V < 400$ GeV.

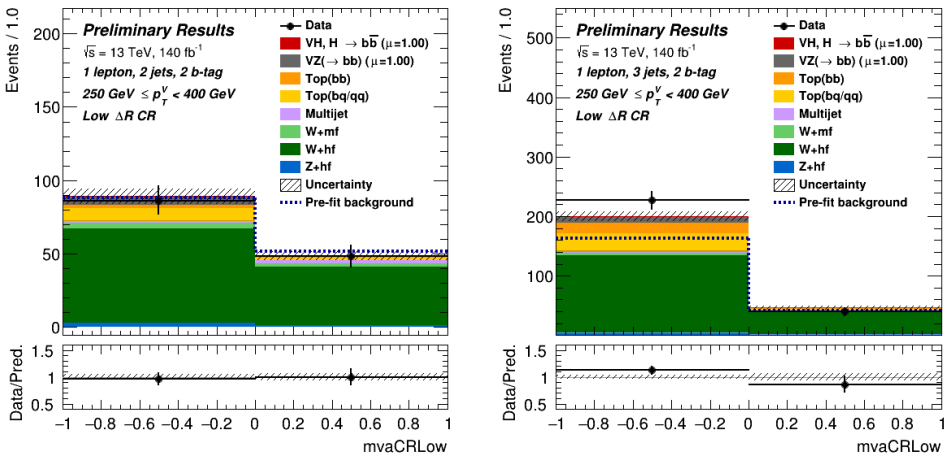
Figure B.6: The 1L High ΔR CR in the BB-tagged 2-jet (left) and 3-jet (right).



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

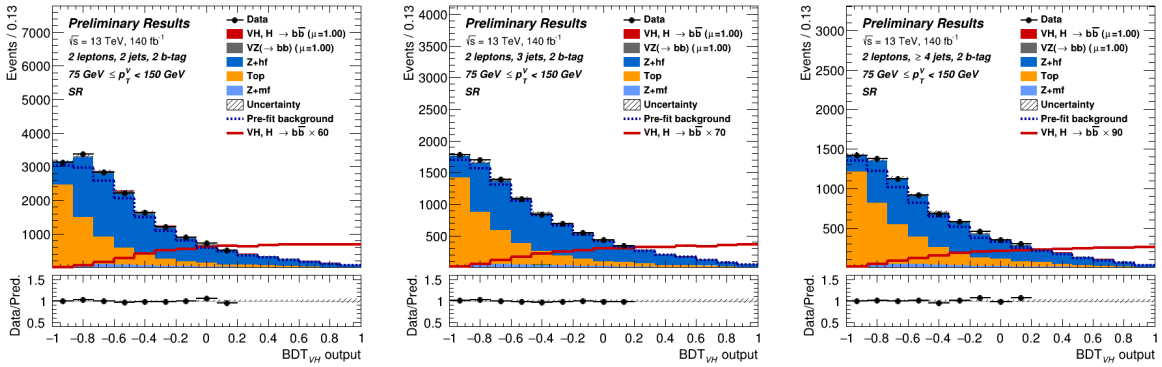


(b) $150 < p_T^V < 250$ GeV.

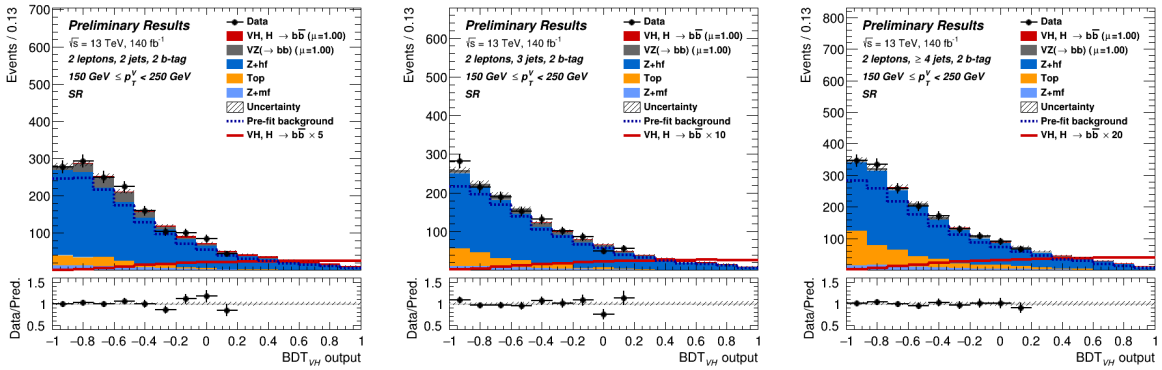


(c) $250 < p_T^V < 400$ GeV.

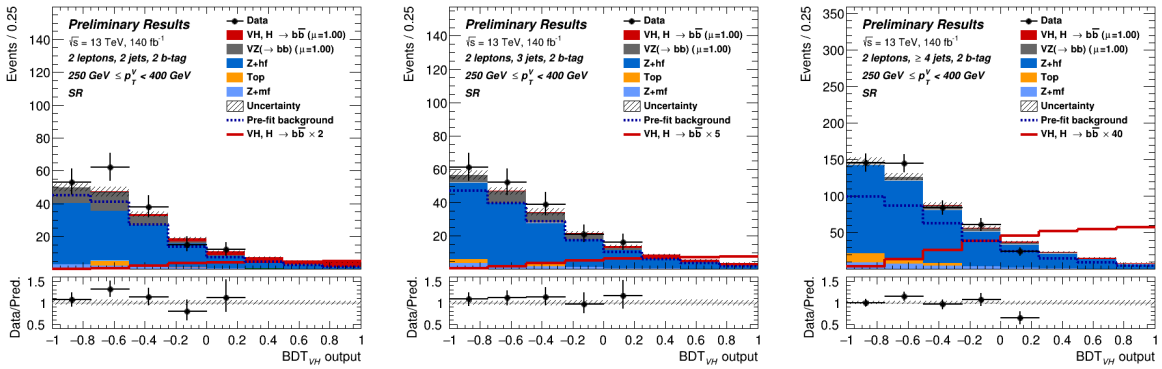
Figure B.7: The 1L Low ΔR CR in the BB-tagged.



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

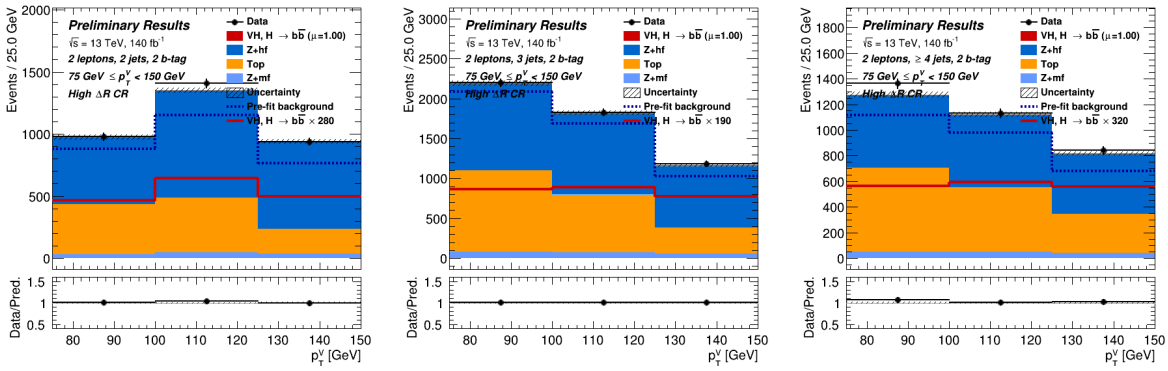


(b) $150 < p_T^V < 250$ GeV.

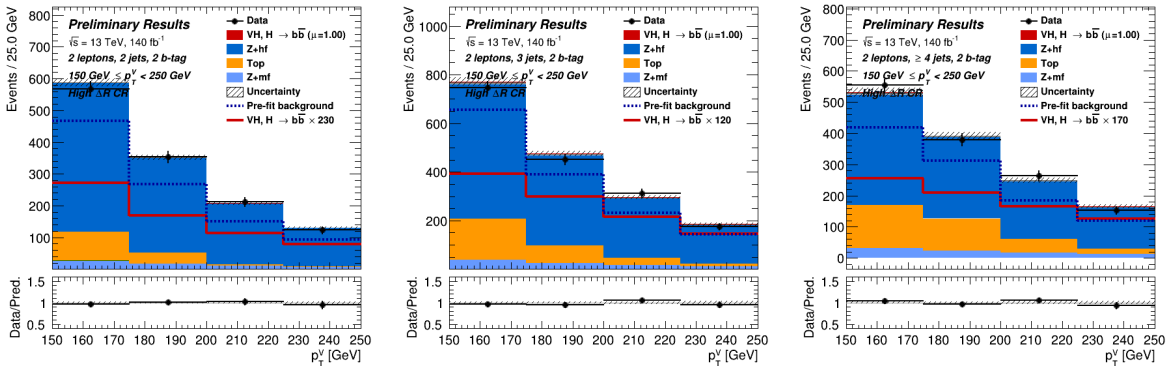


(c) $250 < p_T^V < 400$ GeV.

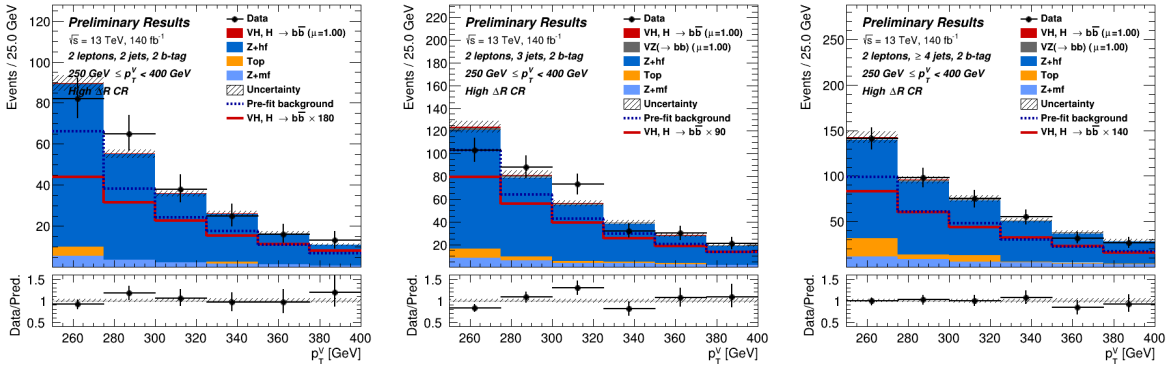
Figure B.8: The 2L signal regions in the BB-tagged 2-jet (left), 3-jet (centre), and ≥4-jet (right).



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

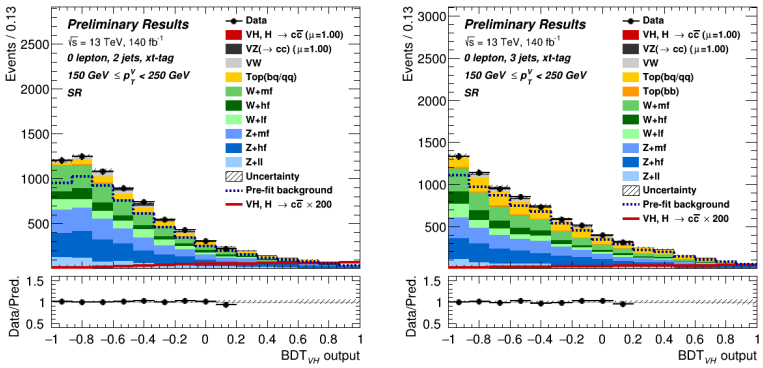


(b) $150 < p_T^V < 250$ GeV.

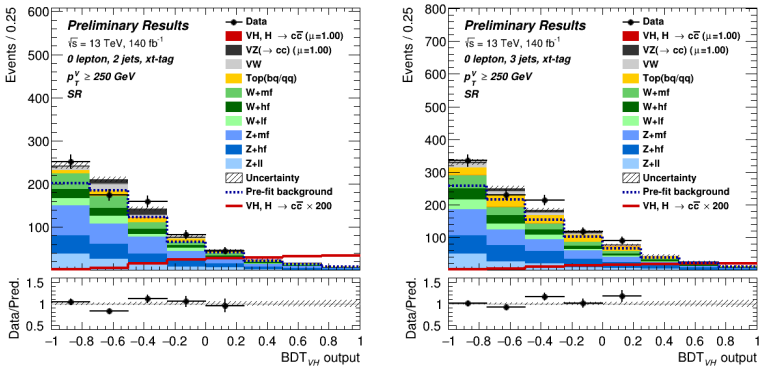


(c) $250 < p_T^V < 400$ GeV.

Figure B.9: The 2L High ΔR CR in the BB-tagged 2-jet (left), 3-jet (centre), and ≥4-jet (right).

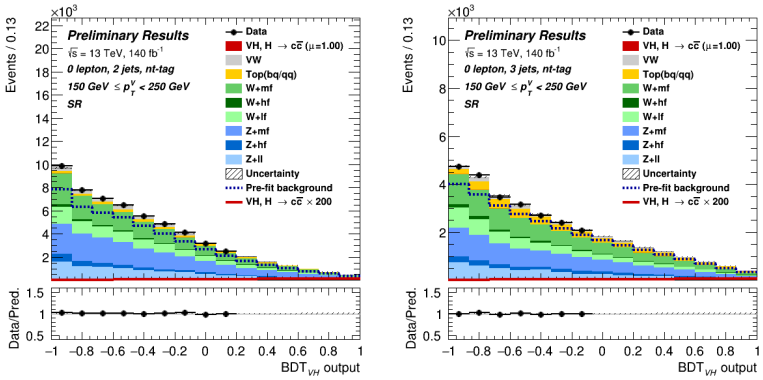


(a) $150 < p_T^V < 250$ GeV.

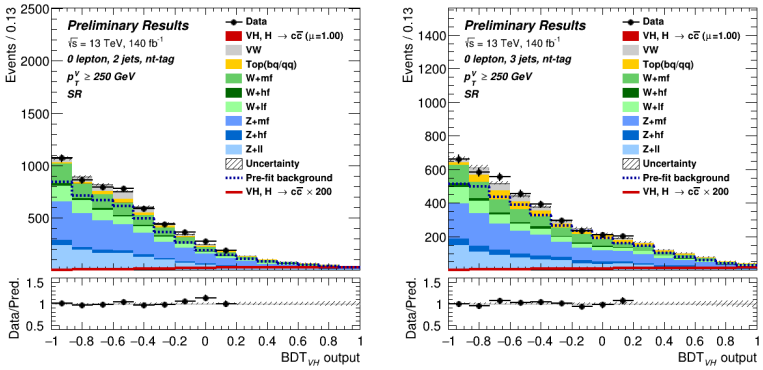


(b) 250 GeV $< p_T^V$.

Figure B.10: The 0L signal regions in the 2 c-tagged 2-jet (left) and 3-jet (right).

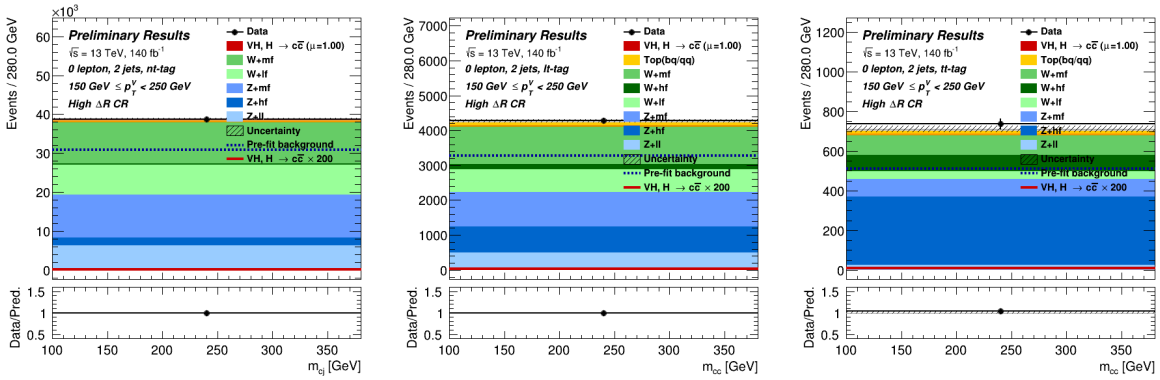


(a) $150 < p_T^V < 250$ GeV.

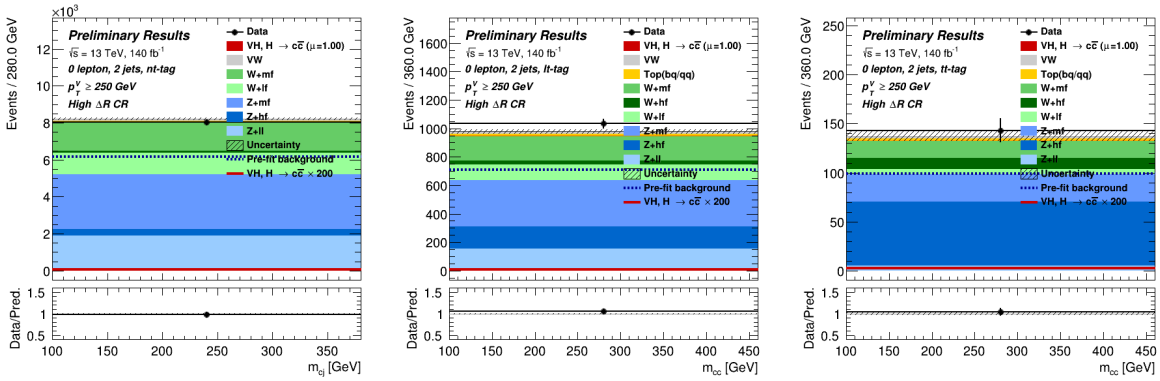


(b) 250 GeV $< p_T^V$.

Figure B.11: The 0L signal regions in the 1 c-tagged 2-jet (left) and 3-jet (right).

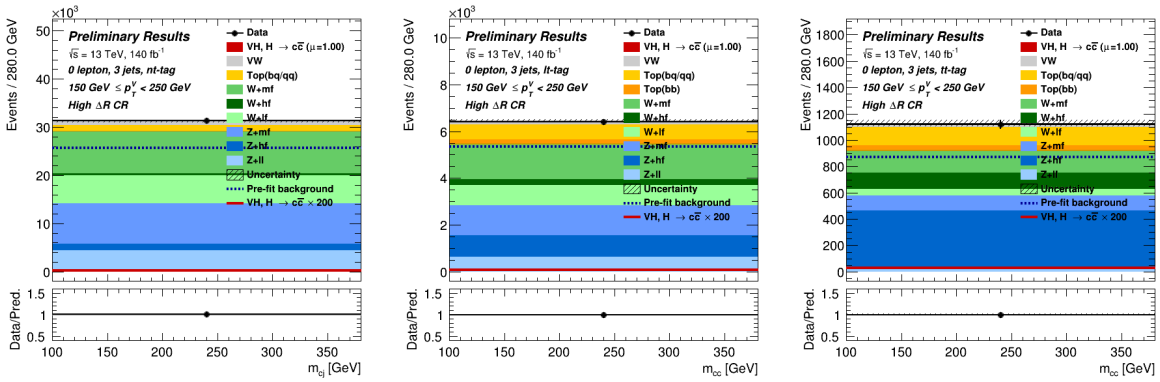


(a) $150 < p_T^V < 250$ GeV.

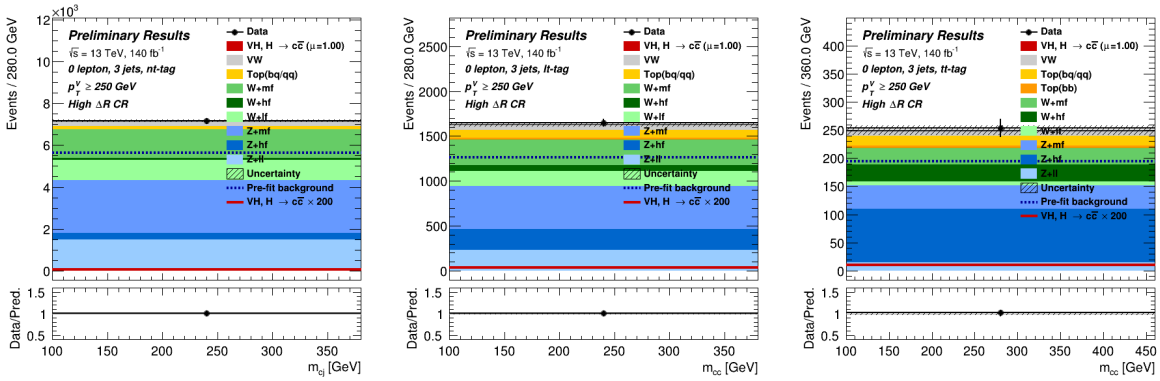


(b) 250 GeV $< p_T^V$.

Figure B.12: The 0L 2-jet High ΔR CR in the TN- (left), LT -(centre), and TT-tagged (right).

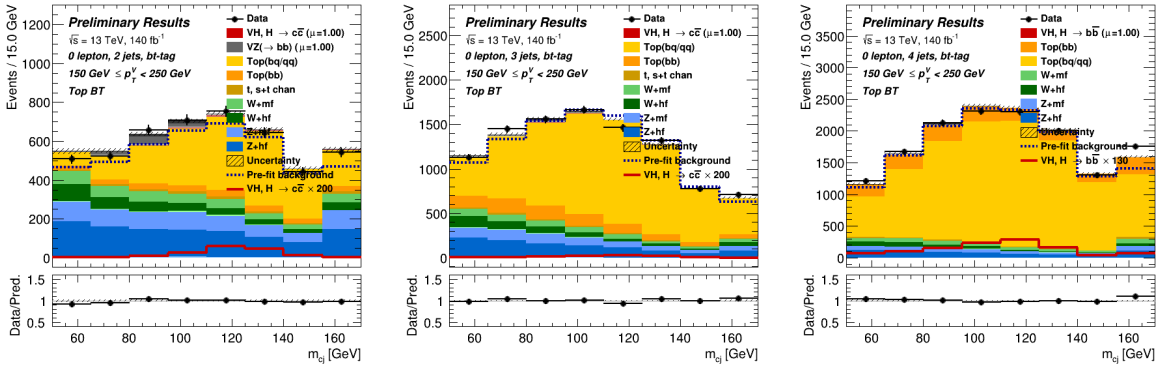


(a) 150 GeV $< p_T^V < 250$ GeV.

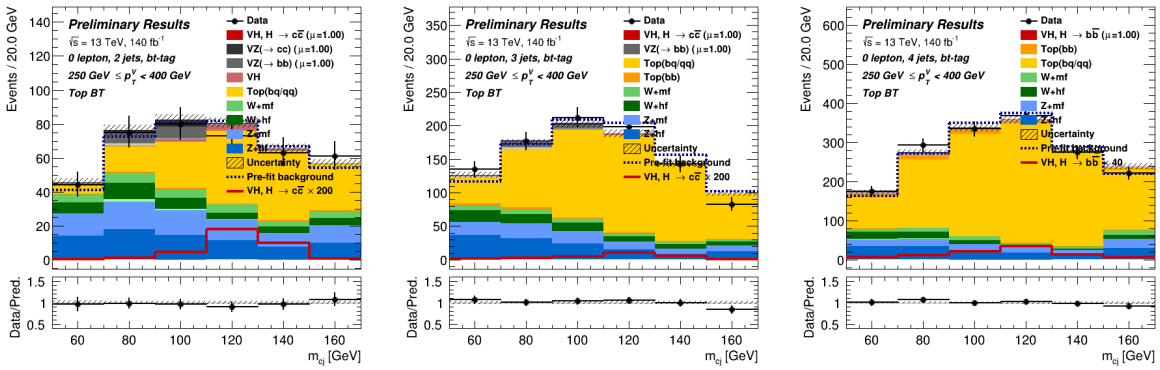


(b) 250 GeV $< p_T^V$.

Figure B.13: The 0L 3-jet High ΔR CR in the TN- (left), LT- (centre), and TT-tagged (right).

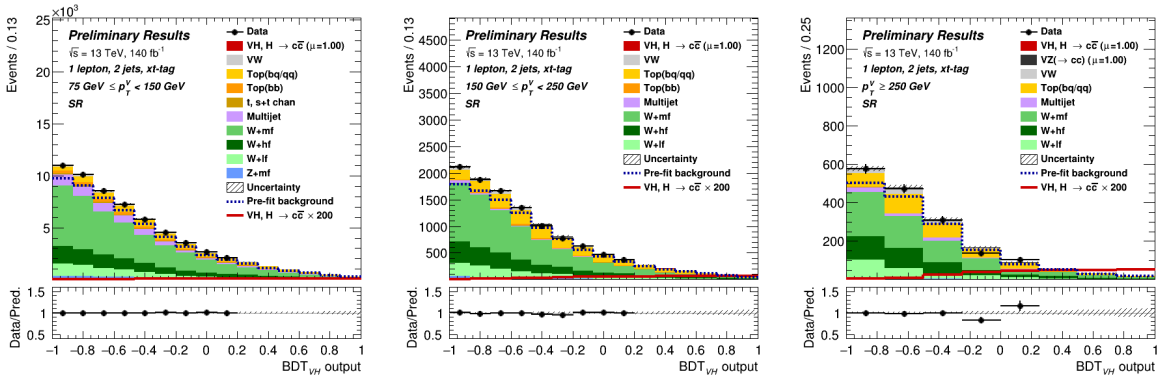


(a) $150 < p_T^V < 250$ GeV.

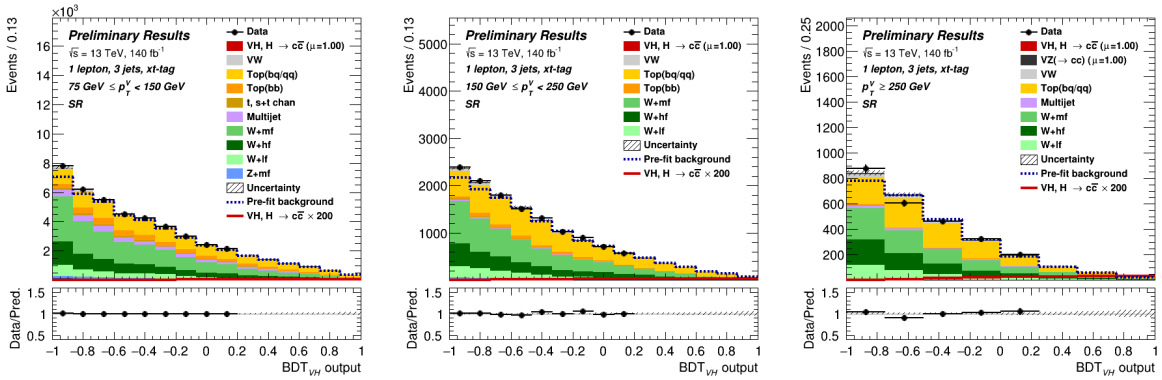


(b) 250 GeV $< p_T^V <$ 400 GeV.

Figure B.14: The 0L Top CR in the BT-tagged 2-jet (left), 3-jet (centre), and 4-jet (right).

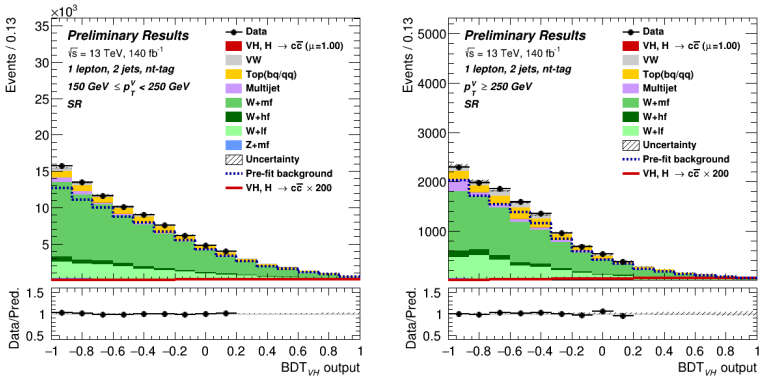


(a) 2-jet, [75, 150] GeV (left), [150, 250] GeV (centre), and 250 GeV $\leq$ (right) p_T^V regions.

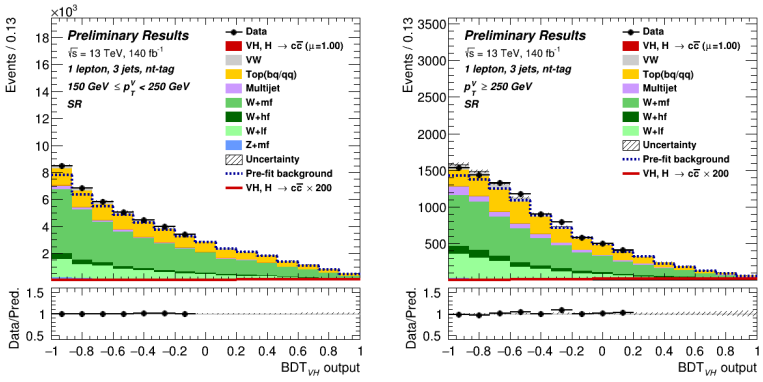


(b) 3-jet, [75, 150] GeV (left), [150, 250] GeV (centre), and 250 GeV $\leq$ (right) p_T^V regions.

Figure B.15: The 1L signal regions in the 2 c-tagged regionss.

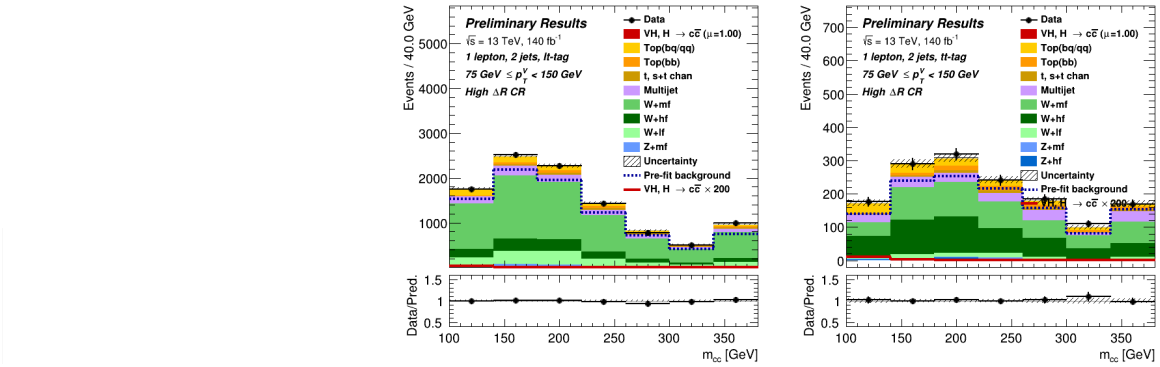


(a) 2-jet, [75, 150] GeV (left), [150, 250] GeV (centre), and 250 GeV $\leq$ (right) p_T^V regions.

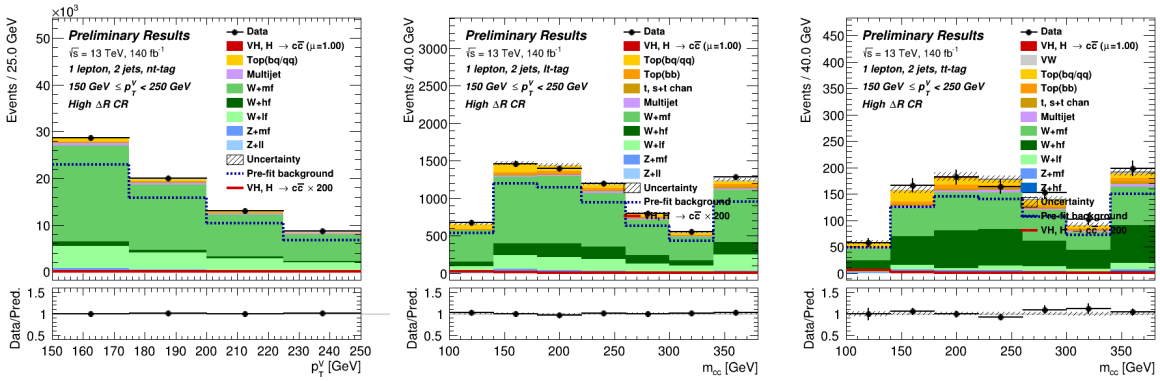


(b) 3-jet, [75, 150] GeV (left), [150, 250] GeV (centre), and 250 GeV $\leq$ (right) p_T^V regions.

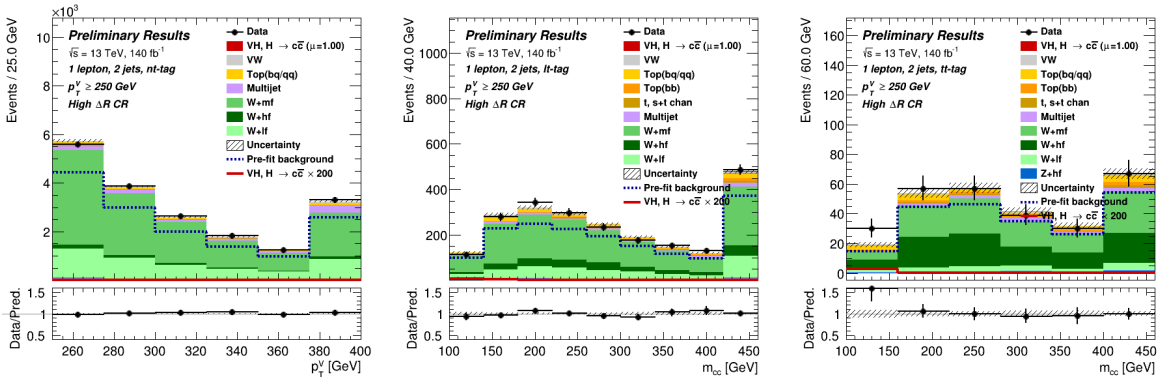
Figure B.16: The 1L signal regions in the 1 c-taggedged regions.



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

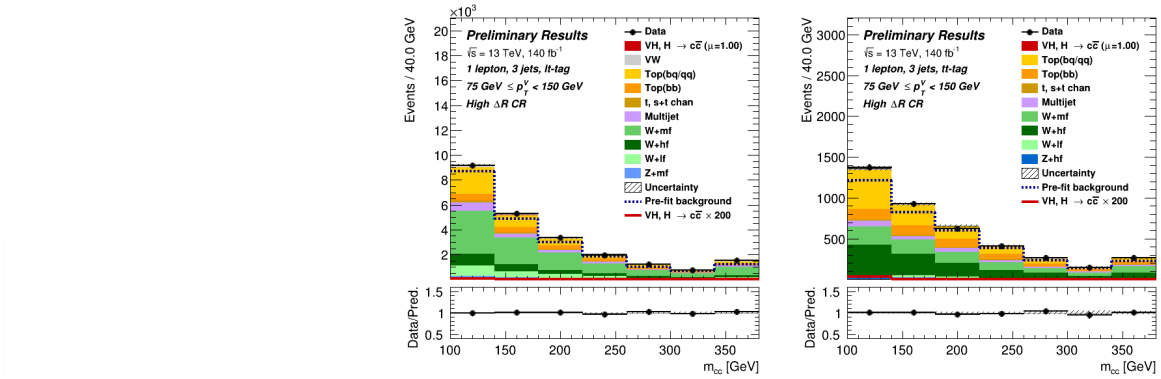


(b) $150 < p_T^V < 250$ GeV.

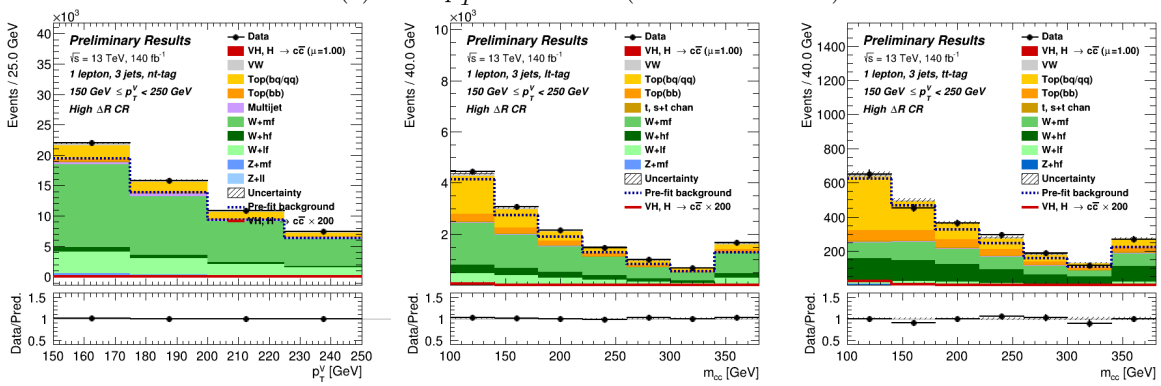


(c) 250 GeV $< p_T^V$.

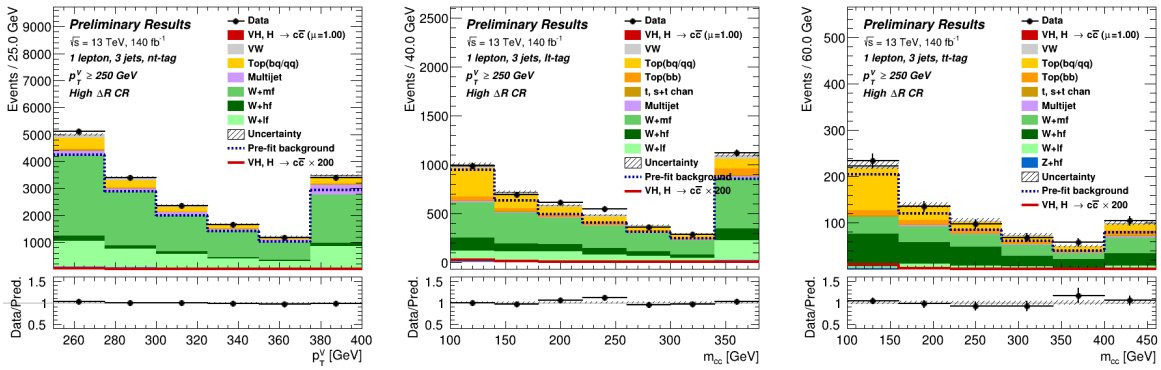
Figure B.17: The 1L High ΔR CR in the 2-jet TN- (left), LT- (centre), and TT-tagged (right) regions.



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

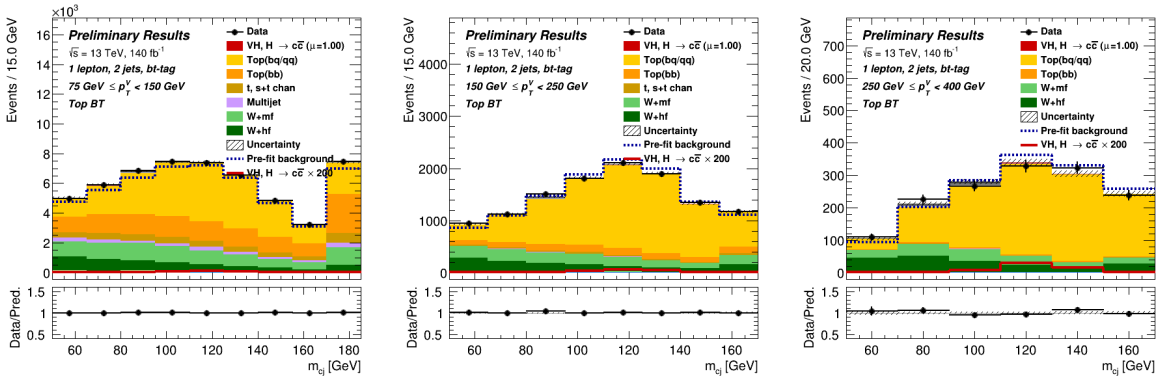


(b) $150 < p_T^V < 250$ GeV.

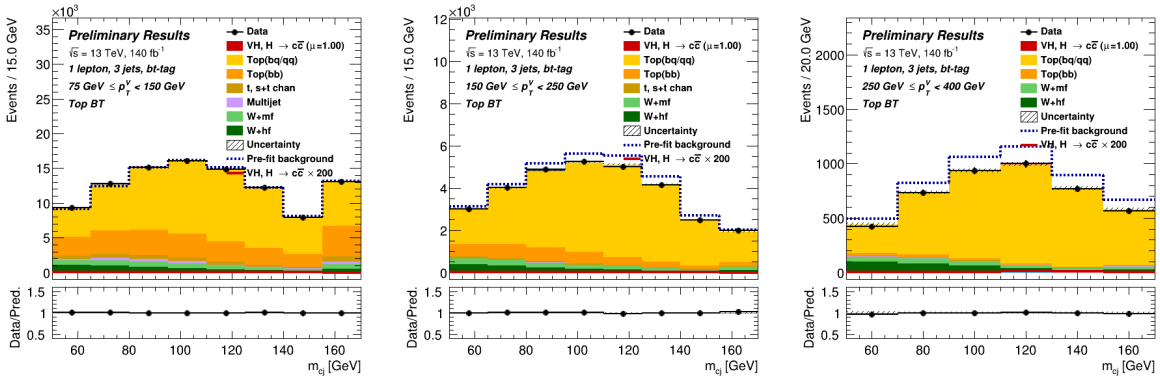


(c) 250 GeV $< p_T^V$.

Figure B.18: The 1L High ΔR CR in the 3-jet TN- (left), LT- (centre), and TT-tagged (right) regions.

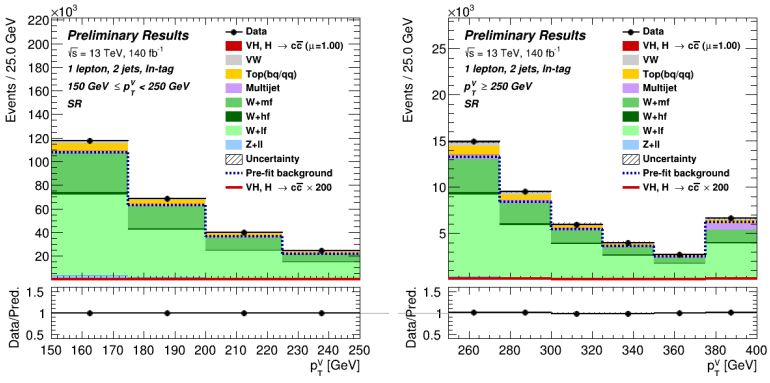


(a) 2-jet, [75, 150] GeV (left), [150, 250] GeV (centre), and 250 GeV ≤ (right) p_T^V regions.

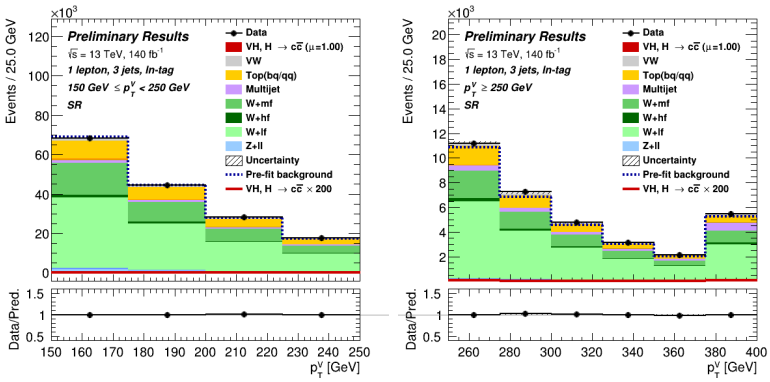


(b) 3-jet, [75, 150] GeV (left), [150, 250] GeV (centre), and 250 GeV ≤ (right) p_T^V regions.

Figure B.19: The 1L Top CR BT-tagged regions.

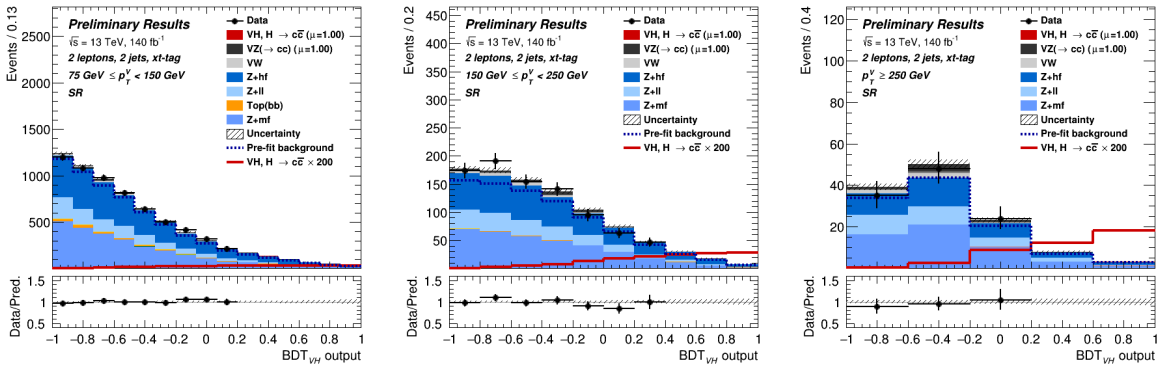


(a) 2-jet.

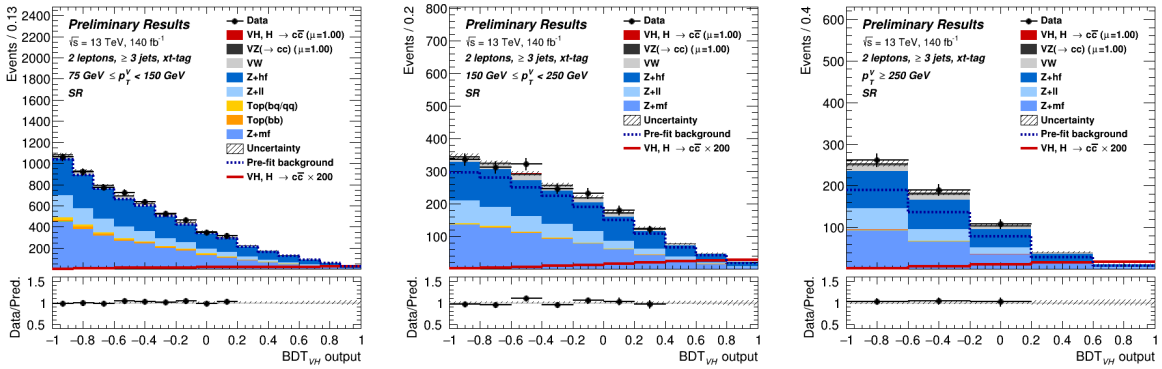


(b) 3-jet.

Figure B.20: The 1L $V + l$ CR in the LN-tagged, [150, 250] GeV (left) and 250 GeV ≤ (right) p_T^V regions.

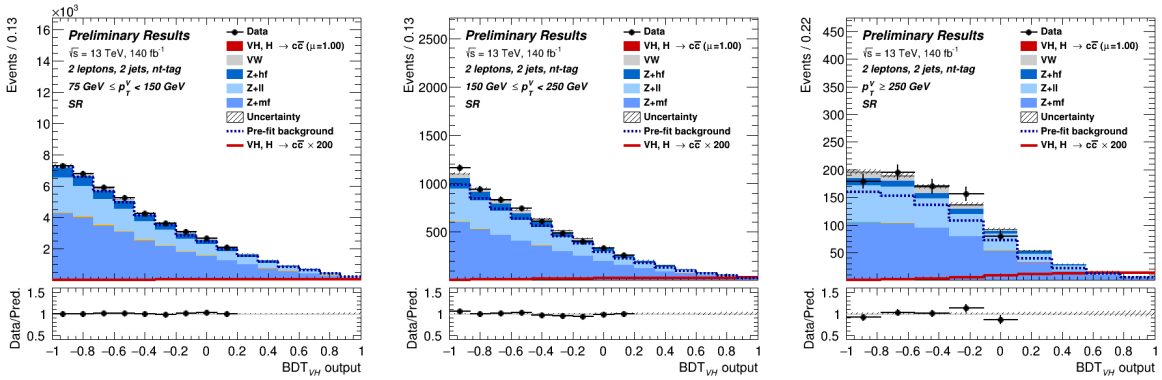


(a) 2-jet, [75, 150] GeV (left), [150, 250] GeV (centre), and 250 GeV ≤ (right) p_T^V regions.

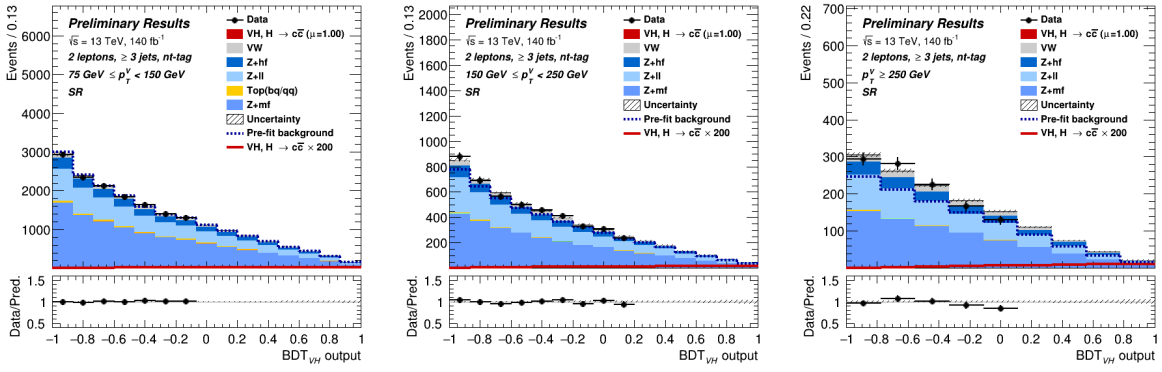


(b) ≥3-jet, [75, 150] GeV (left), [150, 250] GeV (centre), and 250 GeV ≤ (right) p_T^V regions.

Figure B.21: The 2L signal regions in the 2 c-tagged regionss.

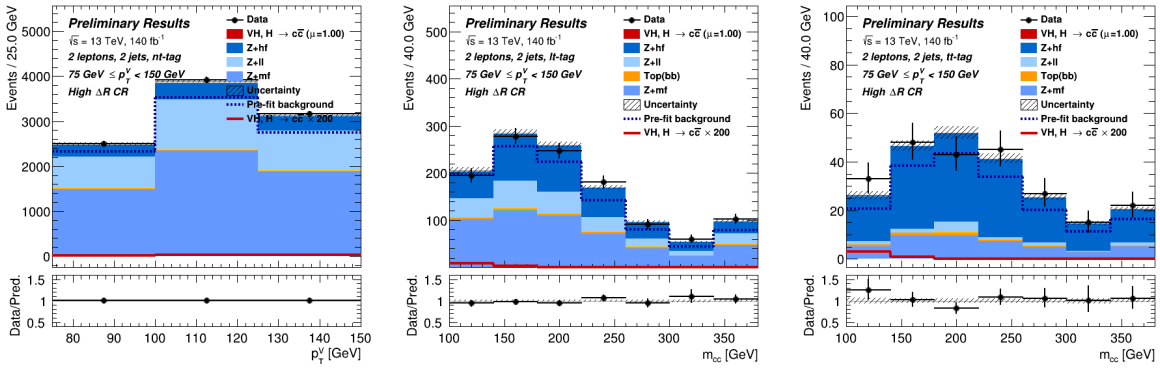


(a) 2-jet, [75, 150] GeV (left), [150, 250] GeV (centre), and 250 GeV ≤ (right) p_T^V regions.

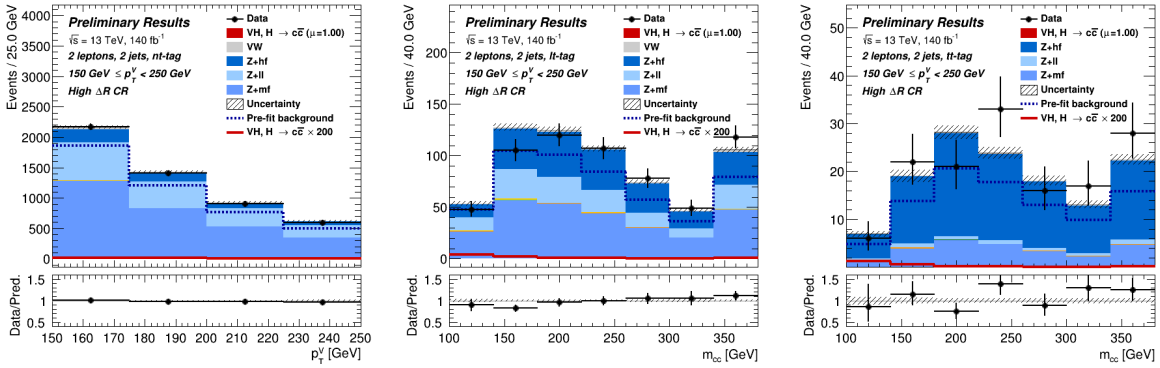


(b) ≥3-jet, [75, 150] GeV (left), [150, 250] GeV (centre), and 250 GeV ≤ (right) p_T^V regions.

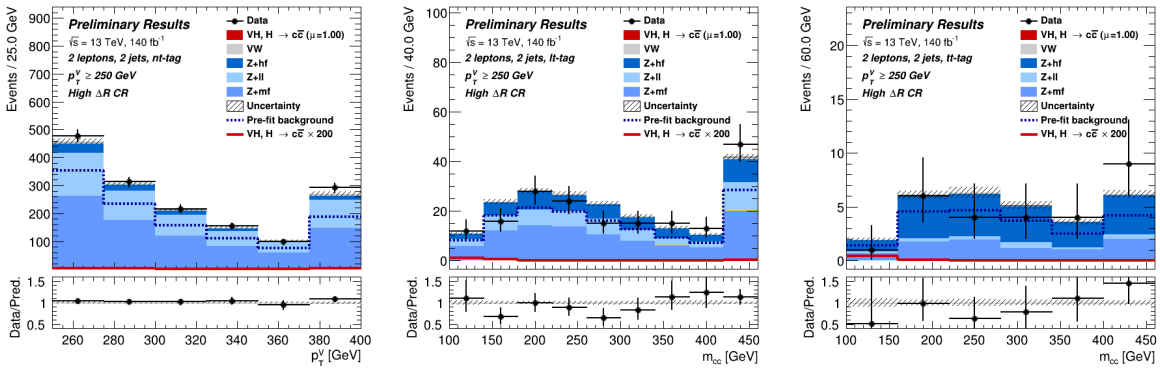
Figure B.22: The 2L signal regions in the 1 c-taggedged regions.



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

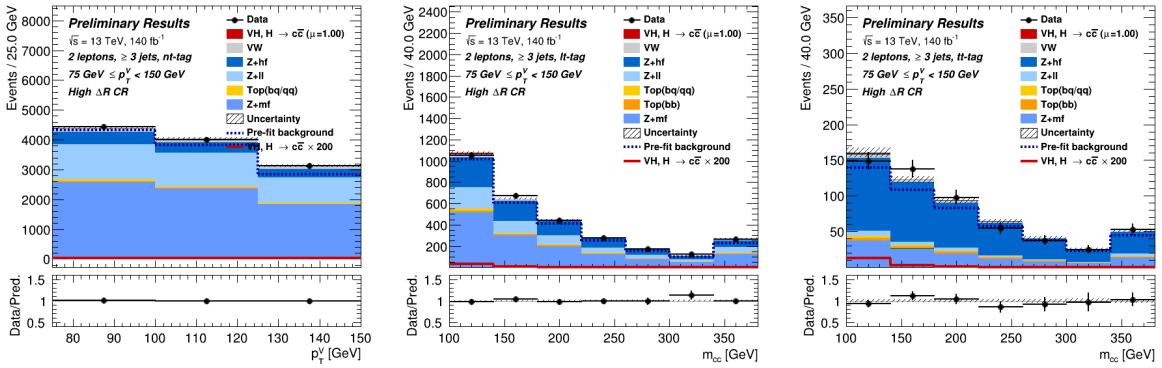


(b) $150 < p_T^V < 250$ GeV.

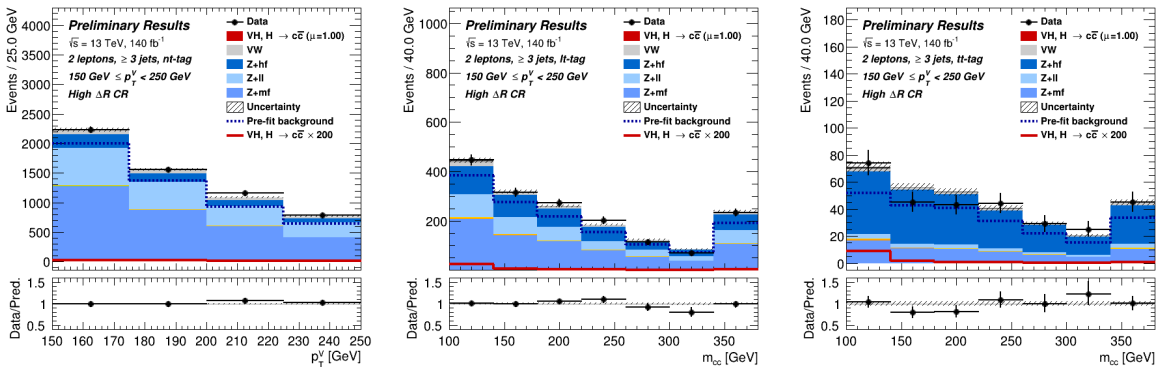


(c) 250 GeV $< p_T^V$.

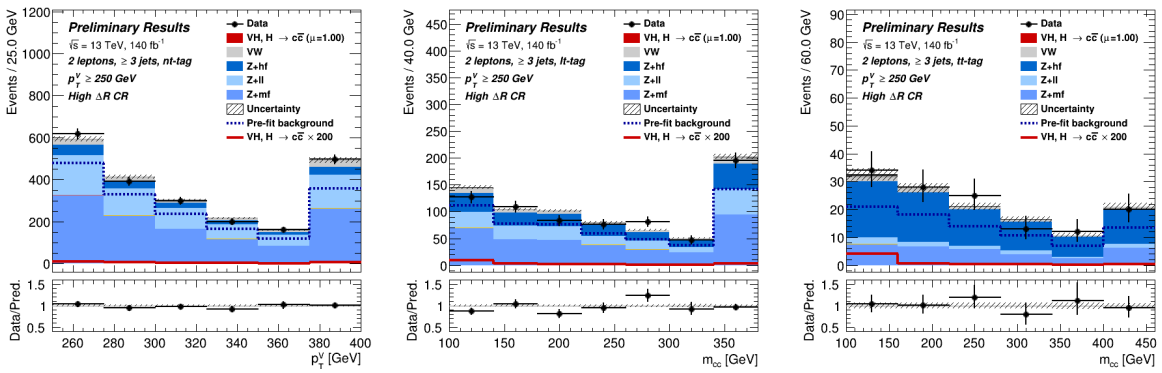
Figure B.23: The 2L High ΔR CR in the 2-jet TN- (left), LT- (centre), and TT-tagged (right) regions.



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

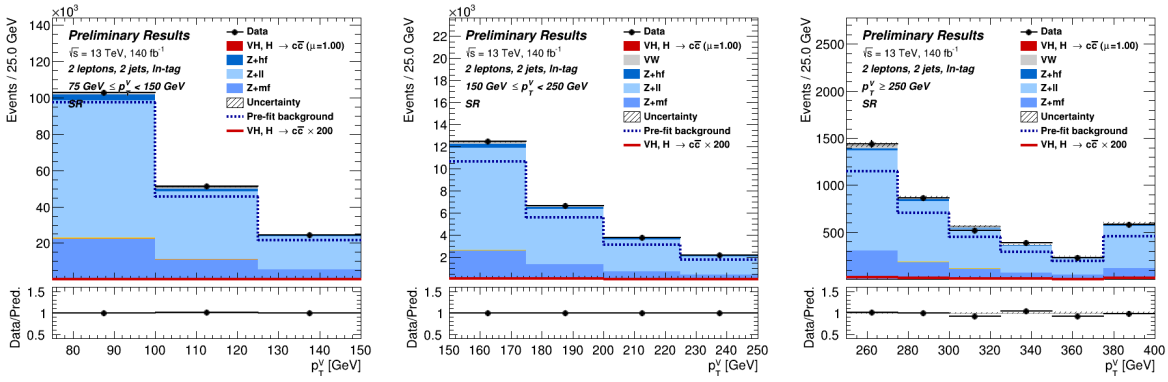


(b) $150 < p_T^V < 250$ GeV.

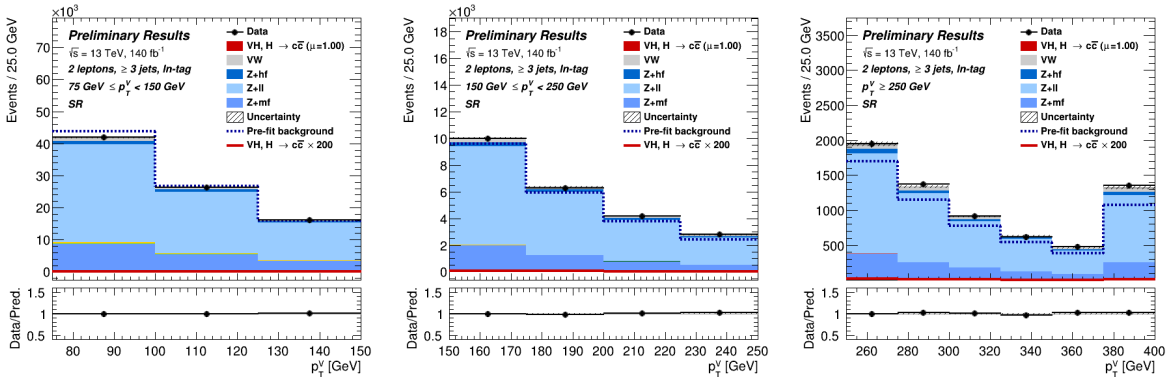


(c) 250 GeV $< p_T^V$.

Figure B.24: The 2L High ΔR CR in the 3-jet TN- (left), LT- (centre), and TT-tagged (right) regions.

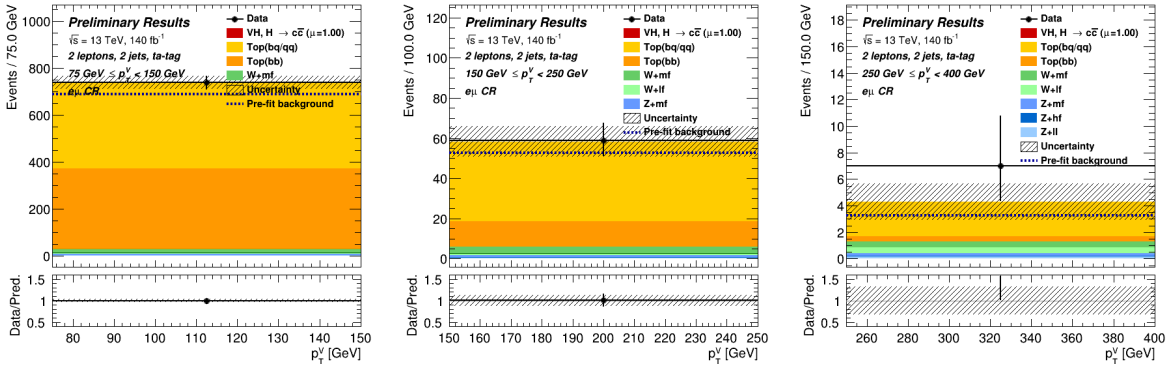


(a) 2-jet, [75, 150] GeV (left), [150, 250] GeV (centre), and 250 GeV $\leq$ (right) p_T^V regions.

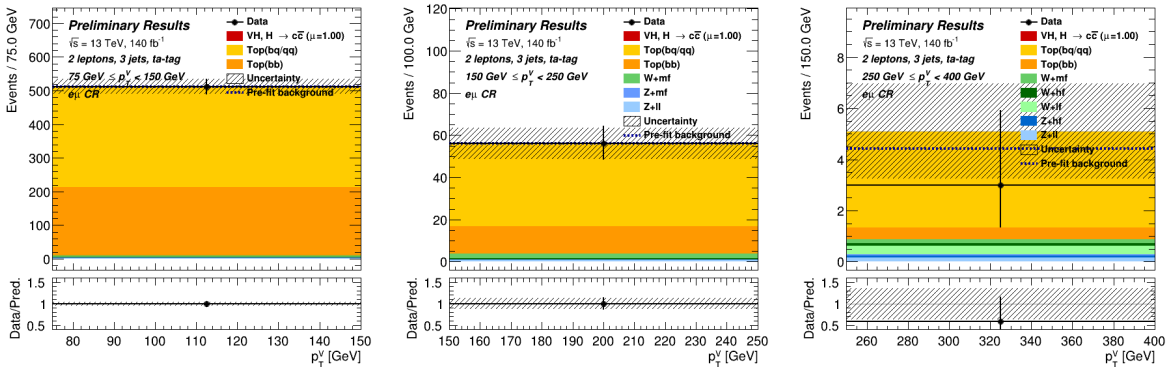


(b) $\geq$3-jet, [75, 150] GeV (left), [150, 250] GeV (centre), and 250 GeV $\leq$ (right) p_T^V regions.

Figure B.25: The 2L $V + l$ CR in the LN-tagged regions.

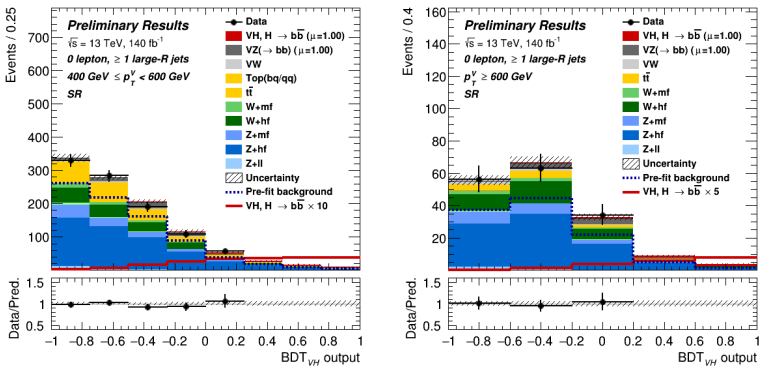


(a) 2-jet, [75, 150] GeV (left), [150, 250] GeV (centre), and 250 GeV $\leq$ (right) p_T^V regions.

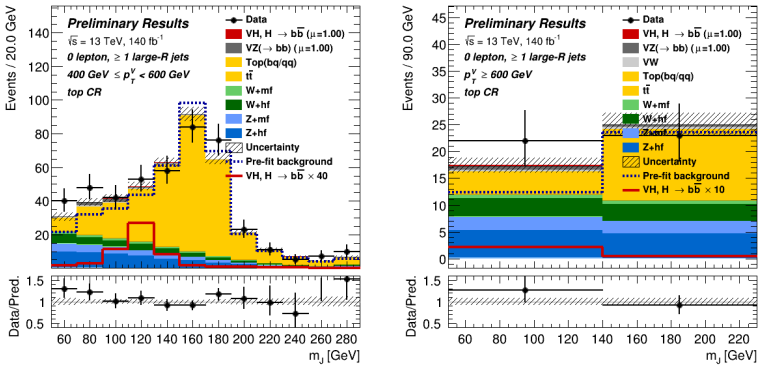


(b) $\geq$3-jet, [75, 150] GeV (left), [150, 250] GeV (centre), and 250 GeV $\leq$ (right) p_T^V regions.

Figure B.26: The 2L Top $e\mu$ CR with $\geq$1 T-tagged regions.

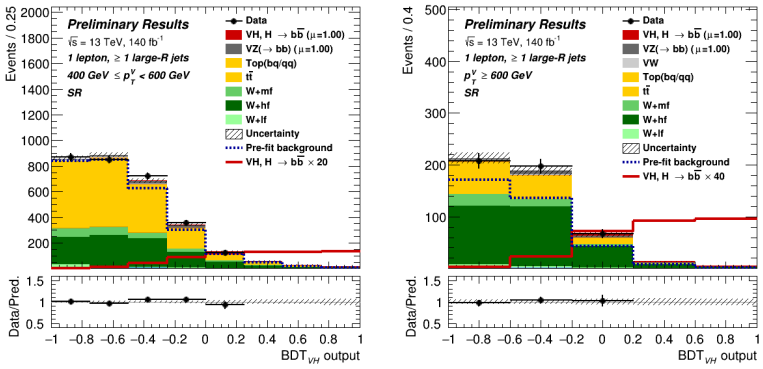


(a) The $p_T^V \in [400, 600]$ GeV (left - combines high- and low-purity combined) and the $p_T^V \geq 600$ GeV (right) signal regions.

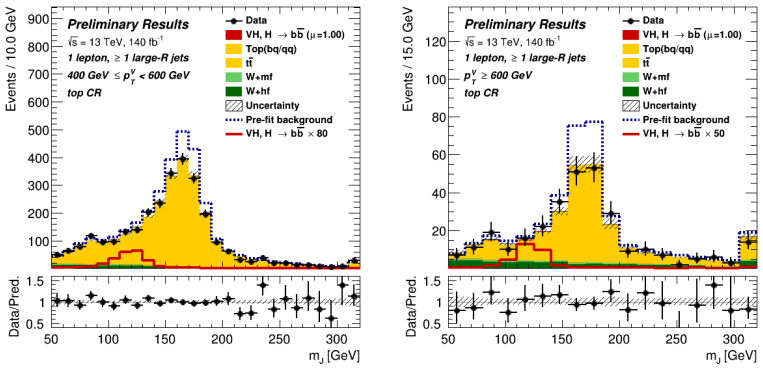


(b) The $p_T^V \in [400, 600]$ GeV (left) and $p_T^V \geq 600$ GeV (right) boosted Top CR.

Figure B.27: The boosted BB-tagged 0L regions.



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



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

Figure B.28: The boosted BB-tagged 1L regions.

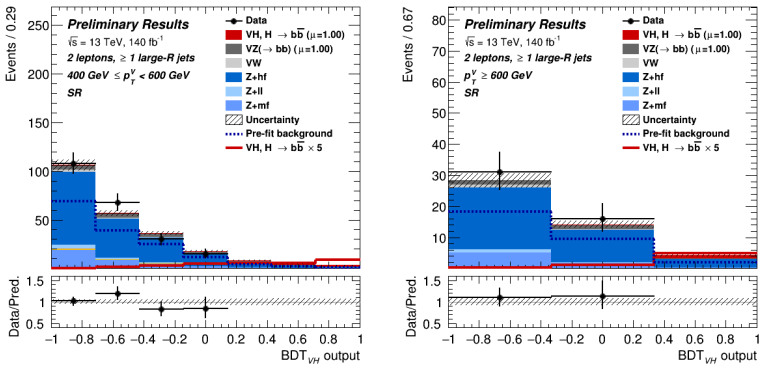


Figure B.29: The boosted BB-tagged 2L signal regions, $p_T^V \in [400, 600]$ (left) and $p_T^V \geq 600$ GeV (right).