School of Physics and Astronomy
The Raymond and Beverly Sackler
Faculty of Exact Sciences
Tel Aviv University

בית הספר לפיזיקה ולאסטרונומיה
הפקולטה למדעים מדויקים
ע"ש ריימונד ובברלי סאקלר
אוניברסיטת תל אביב

Research Thesis for a Degree "Doctor of Philosophy"

LEVERAGING MACHINE LEARNING FOR ENHANCED FPGA-BASED TAU TRIGGERING AND COMBINED $H \to (bb/cc)$ ANALYSIS IN THE ATLAS EXPERIMENT

Submitted to the Senate of Tel Aviv University by

DAVID REIKHER

The work was carried under the supervision of

PROF. EREZ ETZION

ABSTRACT

The exceptionally good agreement between the predictions of the Standard Model of particle physics and the measurements made by the experiments at the Large Hadron Collider at CERN requires collecting more and more data to set increasingly tighter upper limits on observables in the search for physics beyond the Standard Model. Ongoing upgrades aimed at increasing the amount of data delivered to the experiments are pushing the limits of the currently used algorithms that specialize in separating signal and background in order to produce clean samples for physics analyses. Recent advances in machine learning technology allow the development of a new generation of these algorithms that, more often than not, outperform their heuristic-based counterparts in online, real-time triggering scenarios and offline data analysis. In particular, fast inference of trained machine learning models in FPGAs has been recently made achievable by the introduction of new tools.

This thesis explores these next-generation algorithms in the context of online and offline analysis. The online analysis concerns triggering on τ leptons at the Level-1 ATLAS trigger using boosted decision trees, and the offline one explores the utilization of graph neural networks and boosted decision trees in the VHbb(cc) Legacy analysis, which sets the most stringent constraints to date on the coupling of the Higgs boson to b- and c-quarks.

The work presented in this thesis results in the first operational machine learning model used for data taking in the Level-1 trigger of the ATLAS experiment, based on a fully evaluated boosted decision tree, with no special hardware requirements, for the identification of τ leptons, opening the gate for the next generation of trigger algorithms at the ATLAS experiment.

Hofstadter's Law: It always takes longer than you expect, even when you take into account Hofstadter's Law

— Douglas R. Hofstadter [1]

ACKNOWLEDGMENTS

These lines are among the last lines I'm writing before submitting the thesis and I find it hard to believe that I'm actually at the finish line. This was, from a personal perspective, a period of profound change in my life, with the birth of my youngest child Emma and the immigration to and settling in a foreign country, and from a professional perspective, a voyage riddled with days of smooth sailing, days of heavy storms and incidents of running aground which made me question everything, during which the thing that kept me going was the cold awareness of the law of regression towards the mean and, more importantly, the wonderful people that I was surrounded with, whom I would like to thank here.

It seems like life is a long, unplanned random walk, new directions forming by bumping into people we meet along the way. I myself am not a very social person and tend to lock myself in a lab/office and work alone, only reaching out when completely stuck (which, I am aware, is not the best quality). Therefore, I am surprised that, looking back 10 years, I'm so far from my starting point, and in such an unexpected direction. This I attribute to a handful of encounters with key people in my life, of which probably the most influential is my supervisor, *Prof. Erez Etzion*, to whom I wish to express my deepest gratitude, for catching me back when I dropped from my attempt to ascend the high cliffs of theoretical physics during my master's and being my supervisor all the way through, closer to sea level, where I can honestly say I'm much more comfortable. The positive, carefree approach and constant support of the decisions I made with just the right amount of nudges in the right directions gave me the confidence and ability to surmount this formidable task of starting and finishing a Ph.D. in experimental particle physics, while acclimating to a life in another country.

This work contains two large projects - τ identification on FPGAs and the VHbb(cc) Legacy Analysis, none of which would be possible without the support and help of others.

On the τ side, I would like to thank Liron Barak, the lead of the Tel Aviv ATLAS τ group, who welcomed me to join the project, gave me the freedom to tinker and experiment and was effectively my "second supervisor", as well as the members of the Tel Aviv τ ML group: Uriel Barron - whose studies on BDTs for hardware τ triggering inspired this work, as well as Boping Chen, Yuval Frid, Luis Pascual Domínguez, Maayan Tamari and Maayan Yaari for the suggestions and discussions.

I would like to thank the incredibly knowledgeable and positive people from Birmingham university, who, if stranded on another planet, would surely be able to build the ATLAS detector from the local minerals - Francesco Gonnella for the guidance and support on the firmware side, Alan Watson for the very informative and detailed e-mails, Juraj Bracinik and Stephen Hillier, for all the suggestions, support and the uplifting positive attitude which kept the wind in my sails and especially Paul Thompson, who guided me from the imaginary world of "works on paper" to the physical one, helped me understand the physical motivation, the inner workings of the trigger system and ATLAS data processing and whose almost every sentence and message carries deep meaning, which I am usually very slow to pick up.

I would like to thank Sioni Summers from CERN for his guidance with the Conifer tool and the following people who contributed many helpful suggestions and provided crucial support - Javier Montejo Berlingen, Ralf Gugel, James Inglis, Rhys Owen, David Sankey and Martin Wessels.

On the analysis side, I would like to thank Jonathan Shlomi, whose idea to use a GNN for signal-background classification inspired me to perform the relevant studies as well as Francesco Armando Di Bello, who guided me during the MVA studies. I would like to thank Milene Calvetti for the support during my first days in the analysis and especially Hannah Arnold and Elisabeth Schopf for the availability to patiently answer any of my silly beginner questions despite the formidable task of leading such a significant physics analysis. I cannot list the entire analysis collaboration, but I am thankful to every one who participated in the writing of the internal documentation from which I quoted plenty when writing the analysis part. I would also like to thank Guy Koren and Hadar Cohen, who, having worked on the previous iteration of the VH(cc) analysis provided a lot of inspiration.

On a more personal note, I would like to thank my parents, Elena and Michael, for instilling in me the unique mixture of theoretical

rigour with high practicality and solemnity with frivolity, in just the right proportions to get me to where I am right now, my kids Emma and Tommy for the change of scenery every time I close my laptop lid at the end of the day and, to save the best for last, my amazing partner and wife Olya who gives me the confidence that no matter what crazy idea either of us has next (excluding travelling to Mars), we will always be on the same boat.

CONTENTS

1

INTRODUCTION

The Large Hadron Collider (LHC) is among the most complex experimental setups ever built and is designed to probe with high precision physical processes at high energies in order to advance our knowledge of fundamental particle physics. The ATLAS experiment is one of the detectors at the LHC. It detects and records vast quantities of proton-proton collision data which is analyzed by multiple teams whose objective is to constrain the parameters of or look for physics beyond the Standard Model (SM). ATLAS has recently concluded Run 2, by the end of which it has collected 139 fb^{-1} of proton-proton collision data for physics analyses at a center-of-mass energy of $\sqrt{s} = 13$ TeV, and, at the time of writing, is in the middle of Run 3 which is expected to collect more data at a higher energy of 13.6 TeV.

The LHC's plan for the future of its operation is to increase as much as possible the amount of interactions per unit time in order to increase the amount of collected data to set more stringent constraints on interesting observables. However, this planned increase in luminosity is a double-edged sword - while delivering more data for analyses and improving the precision, it also raises problems with how to separate the small amount of signal processes from the expected copious amounts of background from processes with a similar experimental signature both *online* - during the process of data taking, and *offline* - when the recorded data is being analysed. In particular, a major problem is the presence of an increasingly higher amount of near-simultaneous collisions, known as *pile-up*.

The LHC is designed to produce at least one proton-proton collision every 25 ns, which requires high resolution data acquisition at a rate of 40 MHz. In the ATLAS experiment this results in a torrent of data far beyond the processing and storage capacity of current technologies at reasonable costs. Before being analysed offline, these raw data pass through a set of online and subsequent offline filters forming a sieve which allows only the most interesting parts of the most interesting collision events for a physics analysis through.

In ATLAS, the very first, real-time filter through which all collision events must pass is the ATLAS trigger, which reduces the rate of captured events from 40 MHz to around 1 kHz. In Runs 1-3 of the

LHC, this trigger is further sub-divided into the hardware-based Level-1 (L1) trigger, which reduces the rate from 40 MHz to 100 kHz, and the software-based High-Level Trigger (HLT), which further reduces the rate down to the final 1 kHz. The L1 trigger is comprised of several components specializing in the fast identification of one or more physics objects such as e, μ, τ, jets and missing energy, with the HLT running more elaborate higher-latency algorithms to make a more precise identification.

After being selected and stored, as part of an offline physics analysis, the events are analysed in order to compute observables on a subset of the collected data that contains a *signal*, which is a physics process of interest, in order to perform a statistical fit of functions that model these observables and that are parameterized by the physics parameters of interest, related to this process, that we want to constrain. Usually, the bulk of the effort in such analyses goes into designing a selection algorithm that produces a sub-sample of the experimental data which has high purity in terms of the desired signal while at the same time having a high signal selection efficiency. Nevertheless, even after careful selection, the data will be a mixture of signal and background, so the chosen fit observables must be maximally sensitive to the signal.

In both the online and offline cases there is increasing interest in trainable machine learning models, which more often than not show superior performance over simple heuristics-based algorithms in separating signal from background. These algorithms become more and more crucial as the amount of bandwidth and data in both cases increases and present a solution to the signal-background separation problem in noisy environments.

One of the main goals of the ATLAS experiment is to make precision measurements of SM parameters and set limits on them, which may provide clues to physics beyond the SM. In 2012 a particle consistent with the theoretically predicted SM Higgs boson was discovered by the ATLAS and CMS experiments [2, 3]. Since then, the analysis of data collected at center-of-mass energies of 7, 8 and 13 TeV during Runs 1 and 2 has led to the observation and measurement of the four main Higgs production modes - gluon-gluon fusion, vector boson fusion and associated production with a weak gauge boson or a pair of top quarks, as well as several decay channels predicted by the SM [4]. Among these a 6.4σ observation of $H \to \tau\tau$ in agreement with the SM [5] and an upper limit of 26 times the predicted SM cross section times branching fraction for a Higgs produced in association with a vector boson and

decaying to a pair of charm quarks [6] were recently obtained. Both of these decay modes are sensitive to contributions from, and are therefore promising probes of physics beyond the standard model and therefore it is crucial to improve their sensitivity to their respective signals. The analyses that target these processes can benefit from machine-learning based online and offline improvements.

The work presented here documents the contribution of the author to offline ATLAS physics analysis and online data acquisition. The former is done in the context of the VHbb(cc) Legacy analysis, aiming to combine several previous analyses of the Higgs boson decaying to b or c quarks in association with a vector boson in order to improve the signal sensitivity. The latter is done in the context of τ lepton identification in the ATLAS L1 trigger and results in the world-first full evaluation of a machine learning model in a high-energy physics experiment L1 trigger algorithm, without requiring special hardware. The algorithm has become the baseline τ identification algorithm, operating during Run 3 and opening the gate to the usage of machine learning algorithms in the L1 trigger in ATLAS.

The work presented here is divided into four parts. Part 1 lays out the required theoretical background, Part 2 describes the LHC and the ATLAS experiment, Part 3 details the development, implementation and tuning of the L1 τ identification algorithm and Part 4 describes the VHbb(cc) Legacy analysis and the relevant contributions by the author.

1.1 PERSONAL CONTRIBUTIONS

For τ identification at the L1 trigger, almost all aspects of the algorithm's design, implementation both in hardware and in software and subsequent tuning and optimization for physics performance and resource efficiency (aside from professional guidance) were done by the author. That includes implementation of novel proprietary tools in order to tackle the non-trivial task of introducing a machine-learning algorithm into a previously all-heuristic environment.

For the VHbb(cc) Legacy analysis, the author was a member of the analysis team and participated in several of its aspects. One is the contribution to the software frameworks used for event selection and multi-variate analysis. In particular, the author developed a system to facilitate and speed up the running of jobs on the CERN computing grid that has been used by the entire analysis team. Another aspect is

a contribution to training and evaluating the trained boosted decision tree models producing the discriminant used for the final fit in $VH(\rightarrow b\bar{b})$ 2-lepton and $VH(\rightarrow c\bar{c})$ 1-lepton channels. A final contribution is a preliminary assessment of the performance of a GNN-based model for the production of the fit discriminant. A significant part of the materials and plots presented in Part 4 is based on the internal note of the analysis, collaboratively written by the analysis team members, and is provided to give context to the author's contributions.

One contribution not mentioned in detail in this work, as it is not directly physics-related, but which had, nevertheless, required significant effort, is the development and maintenance of a system responsible for collecting and storing measurements and construction details of the sTGC wedges of the ATLAS New Small Wheel during its construction as part of the Phase-I upgrade of the ATLAS detector in 2018. This was done in a context of a qualification task required by all ATLAS members who wish to gain authorship rights. The work lasted one full year and required the author's full attention. The system was deployed around the globe in all laboratories participating in the construction of the wedges in Canada, Chile, China, Israel, Russia and in CERN and was exclusively used for the documentation of the construction.

Part I

THEORETICAL BACKROUND

2

THE STANDARD MODEL

The Standard Model (SM) of particle physics [7–10] is built upon the framework of quantum field theory. The model's assumption is that several fields, each with at least one degree of freedom are superimposed at each space-time point. A Lagrangian, constructed from those fields as the most general one obeying a set of invariance requirements under transformations (representing symmetries of nature) acting on those fields governs their dynamics.

An especially important set of transformations under which the Lagrangian is invariant are global and local gauge transformations. While the requirement of invariance under the former leads to the conservation of certain combinations of those degrees of freedom, known as *charge*, due to Noether's theorem, imposing invariance under the latter leads to the emergence of *gauge fields* that result in forces acting between particles which carry this charge. The force carrier particles are excitations of those fields.

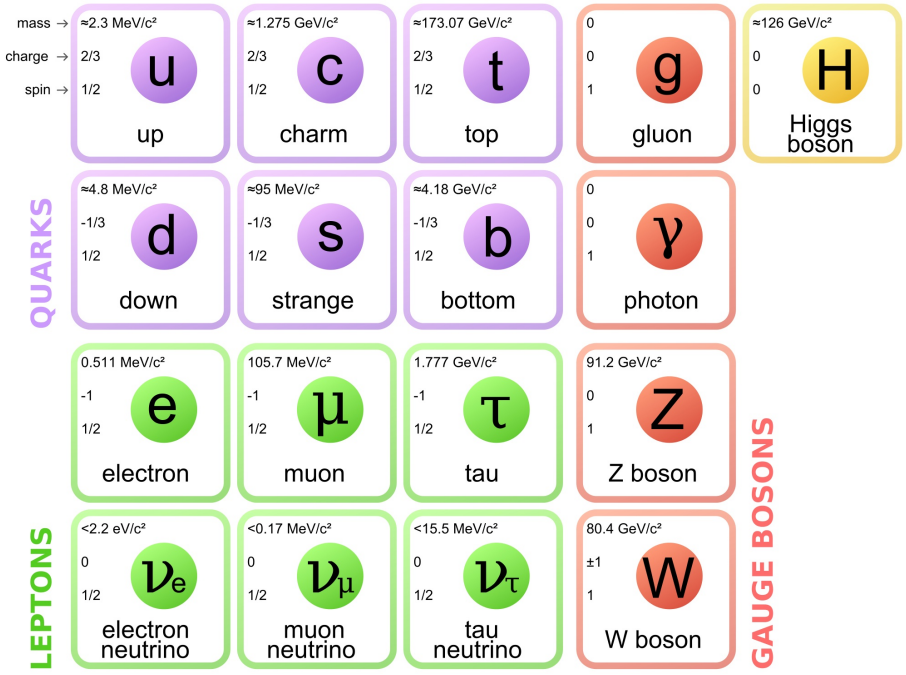


Figure 2.1: Elementary particles of the SM and their properties.

The elementary particles of the Standard Model comprise three generations of leptons where each generation is a pair of an electrically charged massive particle of charge $-e$ with a corresponding massless and electrically neutral neutrino and three generations of quarks where each generation is a pair of an up-type quark with an electrical charge $2e/3$ and a down-type quark with a charge $-e/3$. The quarks also carry a strong color charge and come in three "colors" - red, green and blue.

There are four vector gauge boson charge carrier particles - the neutral massless photon which carries the electromagnetic force, the massive electrically neutral Z and charged W bosons which carry the weak force and the electrically neutral massless gluons which carry a color charge and mediate the strong force.

The massive leptons and quarks interact with photons, W and Z bosons and therefore are influenced by the weak and electromagnetic force. Neutrinos interact only with the W and Z bosons and feel only the weak force.

The quarks, while interacting with photons and the W and Z bosons, interact also with gluons and therefore in addition feel the strong force. The strong interaction confines quarks to bound states of two, termed *mesons* or of three quarks, termed *baryons*, though bound states of more quarks have recently been observed [11]. The proton and the neutron are baryons which, along with electrons are the building blocks of all visible matter.

An additional elementary particle that will be discussed in more detail in the following section is the electrically neutral massive scalar Higgs boson discovered at the LHC in 2012, after having been predicted from theory. It is a necessary component of the SM to allow electroweak (EW) symmetry breaking which unifies the weak and electromagnetic forces. The SM particles and their details are summarized in Figure 2.1.

2.1 THE ELECTROWEAK SECTOR AND THE HIGGS MECHANISM

While in the electromagnetic and strong interactions the emerging gauge fields have no mass terms in the Lagrangian, as they describe the massless gluons and photon, respectively, a problem arises in the weak interaction. Unlike with the strong interaction, where its short range is a consequence of color confinement, the short range of the weak force originates from its massive force carriers, the $W^{\pm}$ and Z bosons

and thus their mass terms must somehow appear in the Lagrangian as self-interaction terms of the form $A^\mu A_\mu$. However, such terms are not invariant under the corresponding local gauge transformation, as such a transformation shifts the gauge fields. Therefore the local gauge symmetry must be broken in some way, while keeping the Lagrangian invariant.

In the 1960s Weinberg, Salam and Glashow proposed the EW theory, in which the electromagnetic and weak force are unified above a certain energy scale. Since the weak interaction violates parity, fermions with left- and right-handed chirality should couple differently to the gauge bosons of the theory. This is achieved by grouping the left-handed fermions into a $SU(2)$ doublet and the right-handed ones into a $SU(2)$ singlet. The group $SU(2)$ has 3 generators, while four are required (one for each gauge boson $W^\pm, Z^0, \gamma$), so the final symmetry group is called $SU(2)_L \times U(1)_Y$. The corresponding charges are the three components of the *weak isospin* $T^{1,2,3}, T^i \equiv \frac{\sigma^i}{2}$ and the *hypercharge* Y. The four gauge fields do not have explicit mass terms in the Lagrangian and masses for three combinations of them are acquired at lower energies via the Higgs mechanism in a process called *spontaneous symmetry breaking* (SSB).

2.1.1 *EW Gauge Bosons, SSB and the Higgs Mechanism*

Assuming a complex $SU(2)$ doublet field Φ, its Lagrangian is

$$\mathcal{L} = (D^\mu \Phi)^\dagger \left(D_\mu \Phi\right) - \mu^2 \Phi^\dagger \Phi - \lambda \left(\Phi^\dagger \Phi\right)^2 \tag{2.1}$$

with

$$\Phi = \frac{1}{\sqrt{2}} \begin{pmatrix} \phi_1 + i\phi_2 \\ \phi_3 + i\phi_4 \end{pmatrix} \tag{2.2}$$

where $\phi_i, i = 1..4$ are real scalar fields and

$$D_\mu \Phi = \left(\partial_\mu + igT^i W_\mu^i + ig' Y B_\mu\right) \Phi, \tag{2.3}$$

is the *covariant derivative*. $W_\mu^i, i = 1..3, B_\mu$ are three gauge fields originating from the $SU(2)_L$ symmetry and one from $U(1)_Y$, respectively, g and g' are coupling constants and $T^i \equiv \frac{\sigma^i}{2}, Y \equiv \frac{I}{2}$ are the generators of $SU(2)_L$ and of $U(1)_Y$, respectively, where σ^i are the Pauli matrices. If $\mu^2 < 0$, the potential

$$V(\Phi) = \mu^2 \Phi^\dagger \Phi + \lambda \left(\Phi^\dagger \Phi\right)^2 \tag{2.4}$$

has a minimum at $\Phi^\dagger\Phi = \sqrt{\frac{-\mu^2}{\lambda}} \equiv v$. All fields Φ that satisfy this relation are minima of (2.4), and they're all related by transformations of $SU(2)_L \times U(1)_Y$. To break the symmetry, we select one of those vacua. We can make a $SU(2)$ rotation from (2.2) so that $\phi_1, \phi_2, \phi_4 = 0$ and $\phi_3 = v$. After selecting a vacuum the state of lowest energy is not symmetric and the $SU(2)_L \times U(1)_Y$ symmetry is said to be *spontaneously broken*. Expanding around this minimum field value, we can write

$$\Phi = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v + h(x) \end{pmatrix} \tag{2.5}$$

where $h(x)$ is a real field with zero expectation value at lowest energy and v is a real number. The W term in (2.3) can be written as

$$\begin{aligned} \sigma^i W^i_\mu &= \left(\sigma^+ + \sigma^-\right) W^1_\mu + \left(\frac{\sigma^+ - \sigma^-}{i}\right) W^2_\mu + \sigma^3 W^3_\mu \\ &= \sqrt{2}\sigma^+ \left(\frac{W^1_\mu - iW^2_\mu}{\sqrt{2}}\right) + \sqrt{2}\sigma^- \left(\frac{W^1_\mu + iW^2_\mu}{\sqrt{2}}\right) + \sigma^3 W^3_\mu. \end{aligned}$$

Defining $W^\pm \equiv \frac{W^1_\mu \mp iW^2_\mu}{\sqrt{2}}$ and using (2.5), the first term of (2.1) becomes

$$\begin{aligned} \left(D^\mu\Phi\right)^\dagger \left(D_\mu\Phi\right) = \frac{1}{2}\partial^\mu h \partial_\mu h + \frac{g^2}{4} W^{-\mu} W^+_\mu \left(v + h\right)^2 + \\ \frac{1}{2}\left(\frac{g}{2}W^{3\mu} - g'B^\mu y\right)\left(\frac{g}{2}W^3_\mu - g'B_\mu y\right)\left(v + h\right)^2, \end{aligned} \tag{2.6}$$

where y is the hypercharge of the Higgs field. For convenience, we'll set it to its final value, $y = \frac{1}{2}$. Setting $Z \equiv \frac{gW^{3\mu} - g'B^\mu}{\sqrt{g^2+g'^2}}$, we are left with the following three massive gauge bosons and an additional massless field $A \equiv \frac{g'W^{3\mu} + gB^\mu}{\sqrt{g^2+g'^2}}$ which is a combination of W^3 and B orthogonal to Z^0:

$$\begin{aligned} &W^\pm, && M_{W^\pm} = \frac{gv}{2} \\ &Z^0, && M_Z = \frac{v\sqrt{g^2 + g'^2}}{2} \\ &A, && M_A = 0. \end{aligned}$$

From (2.6) the interaction vertices between the gauge bosons and the Higgs are evident and similarly the Higgs self-interaction vertices can be obtained by substituting (2.5) into (2.4).

The existence of a massless gauge boson implies that there is a transformation under which the ground state is degenerate, i.e. it transforms the ground state into itself, or, equivalently, its generator annihilates the ground state. This generator is a combination of those of $SU(2)_L \times U(1)_Y$ and is precisely the electric charge $Q = Y + T_3$:

$$(Y + T_3)\,\Phi = \left(\frac{1}{2}I + \frac{1}{2}\sigma_3\right)\frac{1}{2}\begin{pmatrix} 0 \\ v + h(x) \end{pmatrix} = \left(\frac{1}{2} - \frac{1}{2}\right)\Phi = 0. \quad (2.7)$$

2.1.2 *Fermion Masses*

We can now examine the coupling of the fermions to the gauge and Higgs bosons by introducing a left-handed $SU(2)_L$ doublet $\begin{pmatrix} \nu_e \\ e_L \end{pmatrix}$ and a right-handed $SU(2)_L$ singlet e_R, where e_L, e_R are left- and right-handed Dirac spinors of the electron, respectively and ν_L is the left-handed Dirac spinor of the neutrino. According to the relation $Q = Y + T_3$, the values of the hypercharge Y are $-\frac{1}{2}$ for the doublet and -1 for the singlet.

Mass terms of the form $m\bar{\psi}\psi$ for a Dirac spinor ψ are not allowed in the case of the SM, since, e.g. for the electron spinor, $\bar{e}e = \bar{e}_R e_L + \bar{e}_L e_R$ and the $SU(2)_L$ transformation can rotate e_L into ν_L. Therefore, fermion masses are also acquired through coupling with the Higgs boson. After SSB, the coupling between the Higgs and the fermions becomes:

$$c\begin{pmatrix} \bar{\nu}_{eL} & \bar{e}_L \end{pmatrix}\frac{1}{\sqrt{2}}\begin{pmatrix} 0 \\ v + h(x) \end{pmatrix} e_R + \text{h.c.} = \frac{c}{\sqrt{2}}\bar{e}_L e_R\,(v + h(x)) + \text{h.c.} \quad (2.8)$$

where c is an arbitrary constant. Due to the arbitrariness of c, the masses of the fermions are also arbitrary, but it is evident that the coupling of the fermions to the Higgs is proportional to their masses.

For quarks the coupling is similar, but since it is observed that, e.g. strange and down quarks can both decay into up quarks, the quantum states of down-type quarks undergoing a weak interaction are a linear combination of their states when they propagate freely. One can transform between those bases using the Cabibbo-Kobayashi-Maskawa (CKM matrix).

Thus, to conclude, the fermions as well as the gauge bosons acquire masses by interacting with the condensate of the Higgs field and

the coupling of the latter to all fermions and gauge bosons can be summarized as follows:

$$\begin{aligned} &\text{Coupling to } Z/W && \frac{2M_{Z/W}^2}{v}, \\ &\text{Coupling to fermion } f && \frac{M_f}{v}. \end{aligned} \tag{2.9}$$

2.2 HIGGS PRODUCTION AND DECAY MODES

In hadron colliders, the main processes in which a Higgs boson is produced are shown on Figure 2.2. The Higgs production cross sections for these processes are shown on Figure 2.3. The processes are:

- **Gluon-gluon fusion** (ggF) - The Higgs is produced with no associated particles, making it a challenging production mode to identify, even though this process has the highest production cross section. It is mediated by the exchange of a heavy top quark and the contribution to the cross section from lighter quarks propagating in the loop are suppressed proportionally to the square of their mass. The dominance of this process originates from the high momentum fraction carried by the gluons in a p-p collision. Due to the mass-dependant coupling of the Higgs to fermions, the top quark loop has the biggest contribution to Higgs production.

- **Vector boson fusion** (VBF) - the second dominant production mode. In this process two quarks interact via a weak gauge boson exchange, which radiates a Higgs boson with the quarks forming two associated jets. The extra two jets facilitate the identification of this process.

- **WH and ZH associated production** (VH) - the Higgs is produced in association with a weak gauge boson primarily through *Higgs-strahlung*, in which it is radiated off a weak boson produced in the interaction of a quark and an anti-quark. For the ZH production mode there is a contribution from production via a top quark loop (diagram (d) in Figure 2.2). Due to the leptonic decays of the gauge bosons, Higgs production via this mode is easier to identify than the previous two.

- **Higgs in association with top quarks** ($t\bar{t}H$) - the Higgs is produced from a fusion of a pair of top quarks, in association with a pair top quarks. It provides a direct probe to the top-Higgs

Yukawa coupling. A similar process exists for b-quarks, but is suppressed due to the much lower b-quark mass.

- **Higgs in association with a single top** - the Higgs is produced in association with a single top quark and an additional quark. The process occurs either through the Yukawa coupling of the top quark with the Higgs or via an emission from a W boson.

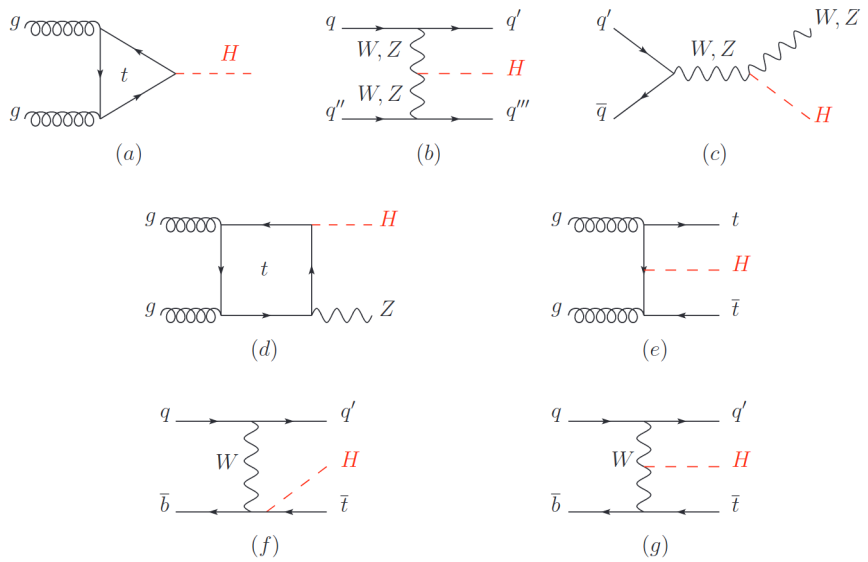


Figure 2.2: Leading order Feynman diagrams for Higgs production modes - (a) gluon fusion, (b) vector boson fusion, (c) Higgs-strahlung, (d) associated production with a gauge boson from a gluon-gluon interaction, (e) associated production with a pair of top quarks, (f-g) production in association with a single top quark.

The branching fractions for the various Higgs decay modes are summarized in Figure 2.4. Due to the dependence (2.9) of the Higgs coupling on the mass of the particle it couples with, the most dominant tree-level decays are $H \to bb$ for quarks, $H \to WW^\star$ for vector gauge bosons and $H \to \tau\tau$ for leptons. Decays into products containing massless particles, such as $H \to gg$, $H \to \gamma\gamma$ and $H \to Z\gamma$ are loop-induced.

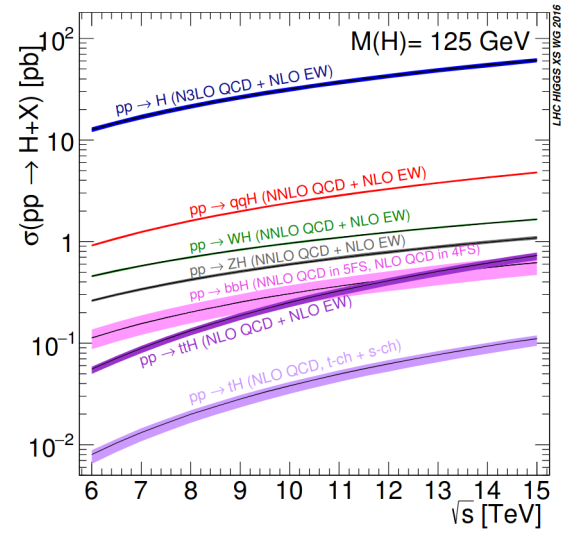


Figure 2.3: Production cross sections for a Higgs boson of mass 125 GeV at different center-of-mass energies of p-p collisions [12].

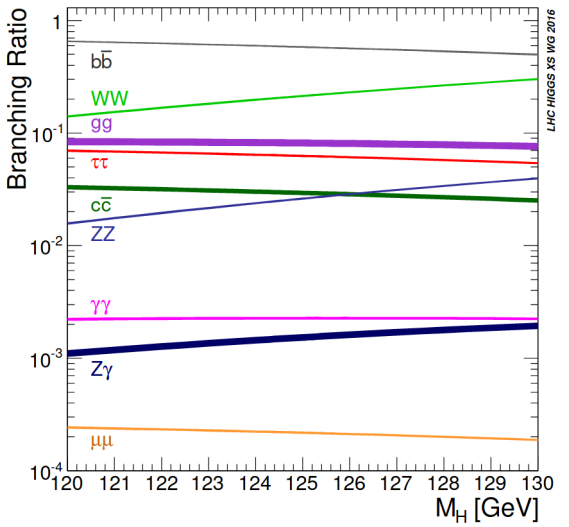


Figure 2.4: The branching fractions for the main decays of a Higgs boson of mass around 125 GeV [12].

3

TAU LEPTONS

The τ lepton is the only lepton which, due to its high mass of 1.777 GeV c^{-2} can decay both into leptons and hadrons. The former occurs in 35% of the cases and the latter - in the remaining 65%. The τ lepton has a lifetime of 2.9×10^{-13} s ($c\tau = 87$ µm) and therefore it decays before it reaches any of the ATLAS sub-detectors. Identification of the τ lepton therefore relies on the ability to identify its decay products.

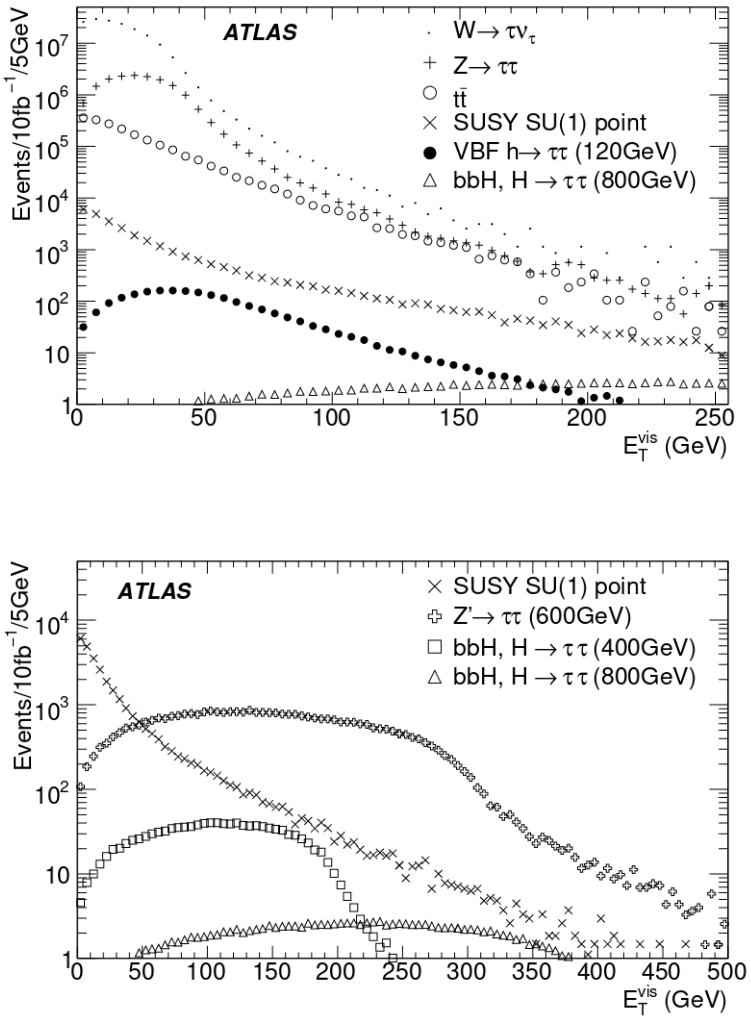


Figure 3.1: The visible transverse energy of τ leptons from different physics processes [13].

Since the leptonic decay of the τ is indistinguishable from prompt leptons, it is not considered for τ identification [14] and will not be further discussed. In the hadronic decay mode, the majority of the decays are characterized by one ("one-prong") or three ("three-prong") charged pions which may be accompanied by neutral pions. Those form a jet signature in the detectors, which is very similar to that of a quark- or gluon-initiated jet (QCD jet). Besides two identifying features - the fact that τ-originating jets are typically narrower than QCD jets and their distinct number of prongs, it is very challenging to discriminate τ jets from QCD jets [14] in a hadron collider.

The range of interest for different physics scenarios of the τ transverse momentum spans from below 10 GeV to above 500 GeV. Figure 3.1 shows some interesting processes with taus in the final state, such as top quark decays, W/Z production, SM Higgs vector boson fusion production with $H \to \tau\tau$ and BSM scenarios.

The products of the τ lepton decay that are stable enough to reach the calorimeter are charged pions and kaons with a mean lifetime of $2.6 \cdot 10^{-8}$ sec and $1.2 \cdot 10^{-8}$ sec, respectively, that travel a distance of $\mathcal{O}(10)$ m in the laboratory frame before decaying. Additional decay products of taus are neutral pions which, having a lifetime of $8.4 \cdot 10^{-17}$ sec, decay immediately into photon pairs. Table 3.1 shows a summary of the leading decay modes of the τ lepton.

3.1 KINEMATICS OF A TWO-BODY DECAY

The kinematics of a general two-body decay of a particle in its rest frame are illustrated in Figure 3.2. The two decay products are produced back-to-back. The $\hat{\mathbf{z}}$ axis is the direction of boost in the laboratory frame.

Let $p_1^{*\mu} = (E_1^*, \mathbf{p}_1^*)$ and $p_2^{*\mu} = (E_2^*, \mathbf{p}_2^*)$ be the 4-momenta of particles 1 and 2, in the decaying particle's rest frame. We then have

$$\begin{aligned}
\mathbf{p}_1^* + \mathbf{p}_2^* &= 0 \\
m &= E_1^* + E_2^* \\
m_1^2 &= E_1^{*2} - p_1^{*2} \\
m_2^2 &= E_2^{*2} - p_2^{*2}
\end{aligned}$$

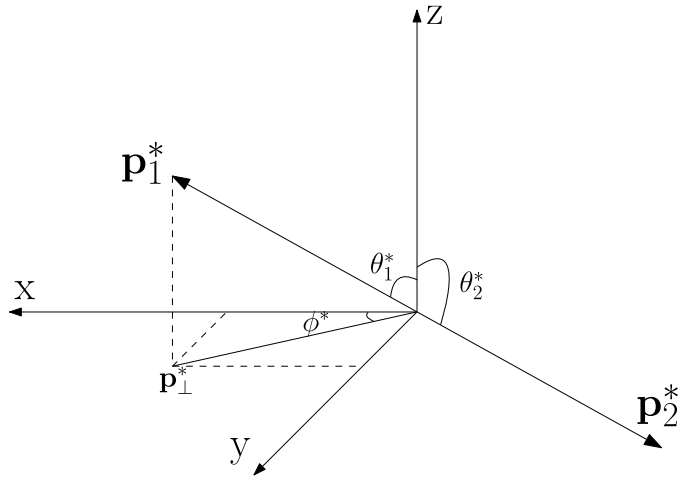


Figure 3.2: The two-body decay of a particle in its rest frame

From which we can express E_1^*, E_2^* and $p^* \equiv p_1^* = p_2^*$, as

$$\begin{aligned} E_1^* &= \frac{m^2 - m_2^2 + m_1^2}{2m} \\ E_2^* &= \frac{m^2 + m_2^2 - m_1^2}{2m} \\ p^{*2} &= \frac{\left(m^2 - m_2^2 + m_1^2\right)^2}{(2m)^2} - m_1^2. \end{aligned} \tag{3.1}$$

When boosting $p_1^{*\mu}$ and $p_2^{*\mu}$ to the laboratory frame in the $\hat{\mathbf{z}}$ direction, we get the energy and momentum of both particles in the laboratory frame:

$$\begin{aligned} E_1 &= \gamma\left(E_1^* + \beta p^* \cos\theta_1^*\right) \\ E_2 &= \gamma\left(E_2^* - \beta p^* \cos\theta_2^*\right) \\ \mathbf{p}_1 &= \mathbf{p}_\perp^* + \gamma\left(\beta E_1^* + p^* \cos\theta_1^*\right)\hat{\mathbf{z}} \\ \mathbf{p}_2 &= -\mathbf{p}_\perp^* + \gamma\left(\beta E_2^* - p^* \cos\theta_2^*\right)\hat{\mathbf{z}} \end{aligned} \tag{3.2}$$

where $\mathbf{p}_\perp^*$ is the projection of $\mathbf{p}^*$ on the xy plane, orthogonal to the boost direction.

The angle ψ_i between the trajectory of particle $i(=1,2)$ and the boost direction $\hat{\mathbf{z}}$, in the laboratory frame is

$$\sin\psi_i = \frac{p_\perp^*}{p_i} = \frac{p^* \sin\theta_i^*}{p_i} \tag{3.3}$$

It is already evident from this equation that for given quantities in the rest frame, ψ_i is smaller for higher boosts (i.e. a more energetic mother particle), so that the resulting two particles are more collimated. Squaring equation (3.3), substituting (3.2) and rearranging, we get expressions for ψ_1 and ψ_2:

$$\tan\psi_1 = \frac{p^* \sin\theta^*}{\gamma\left(\beta E_1^* + p^* \cos\theta^*\right)} \tag{3.4}$$

$$\tan\psi_2 = \frac{p^* \sin\theta^*}{\gamma\left(\beta E_2^* - p^* \cos\theta^*\right)} \tag{3.5}$$

The maximum angle between the original and daughter particle in the laboratory frame would then correspond to $d\tan\psi_{1,2}/d\theta^* = 0$, which gives

$$\cos\theta^* = \mp\frac{p^*}{\beta E_{1,2}^*}. \tag{3.6}$$

For a characteristic rest frame decay angle of $\theta^* = \pi/2$, the above maximum angle for a massive daughter particle would be

$$\psi_{\text{C, max.}} \approx \frac{p^*}{\gamma E^*}. \tag{3.7}$$

and for a massless daughter particle

$$\psi_{\text{C, max.}}^{\text{m=0}} \approx \frac{1}{\gamma}, \tag{3.8}$$

where the approximation is true for the relativistic limit.

It is also useful to derive the energy distribution of one of the two daughter particles. If N unpolarized mother particles decay, we can compute dN/dE_1 by assuming a flat distribution of $dN/d(\cos\theta^*) = 1/2$. From (3.2),

$$\frac{dN}{dE_1} = \frac{dN}{d\cos\theta^*}\frac{d\cos\theta^*}{dE_1} = \frac{1}{2}\frac{d\cos\theta^*}{dE_1} = \frac{1}{2\gamma\beta p^*} \approx \frac{1}{2\gamma p^\star}. \tag{3.9}$$

This is a flat distribution in E_1, with the values for E_1 bounded by the extreme values of equation (3.2),

$$\gamma\left(E_1^* - p^*\right) \lessapprox E_1 \lessapprox \gamma\left(E_1^* + p^*\right), \tag{3.10}$$

for which the average value is

$$\langle E_1 \rangle = \gamma E_1^*. \tag{3.11}$$

The approximations are valid for the relativistic limit.

3.2 KINEMATICS OF A HADRONIC τ DECAY

A hadronic decay of a τ lepton produces an electrically charged hadronic resonance and a neutrino, according to the Feynman diagram shown in Figure 3.3.

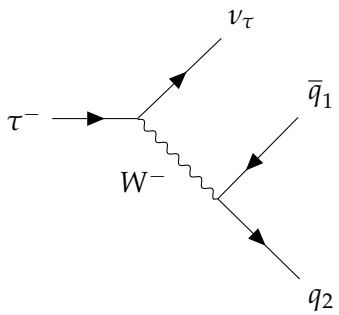


Figure 3.3: Tree level Feynman diagram of the hadronic decay of the τ lepton. In the majority of cases, $q_1 = u$ and $q_2 = d$.

We can use the results of Section 3.1 to estimate some practical values used in the identification and reconstruction of hadronically decaying τ leptons, such as the typical angle between the direction of the visible τ decay products and the original τ particle and a rough estimate of the opening angle of the τ jet. In equations (3.1), if particle 1 is the hadronic resonance and particle 2 is the neutrino, we can set $m_2 = 0$, $m_1 = m_{\text{had}}$, $m = m_\tau$ and ignore ψ_2 in (3.4). Then (3.1) becomes

$$E^* = \frac{m_\tau^2 + m_{\text{had}}^2}{2m_\tau} \qquad (3.12)$$
$$p^* = \frac{m_\tau^2 - m_{\text{had}}^2}{2m_\tau}.$$

Then, using (3.7) and (3.12), we can estimate a characteristic angle between the τ lepton and the visible hadronic jet in the case of a charged pion and ρ meson by setting, $m_{\text{had}} = m_{\pi^\pm} \approx 140$ MeV or $m_{\text{had}} = m_{\rho^\pm} \approx 770$ MeV, respectively, for the lowest energy of interest of a τ of approximately 10 GeV. That gives $\psi^{\tau,\pi}_{\text{C, max.}} \approx 0.17$ radians and $\psi^{\tau,\rho}_{\text{C, max.}} \approx 0.12$ radians. This matching angle between the true τ particle and the visible hadronic τ jet is often used to match τ calorimeter clusters and reconstructed hadronic τ jets in simulated data.

Similarly, we can also get a rough estimate of the opening angle of the visible hadronic τ jet assuming the most common, 1-prong decays $\tau \to \pi^- \nu_\tau$ and $\tau \to \pi^- \pi^0 \nu_\tau$ (see Table 3.1).

For the $\tau \to \pi^- \nu_\tau$ decay, the pion is typically highly energetic because it originates from a decay of a highly boosted τ lepton. With a lifetime of $2.6 \cdot 10^{-8}$ sec, it lives long enough to reach the EM calorimeter's main sampling layer positioned roughly 1.7 m from the interaction point around $\eta = 0$ (see Section 5.2.4.2). As shown on Figure 3.4, the pion deposits a minimum amount of energy of around 1.5

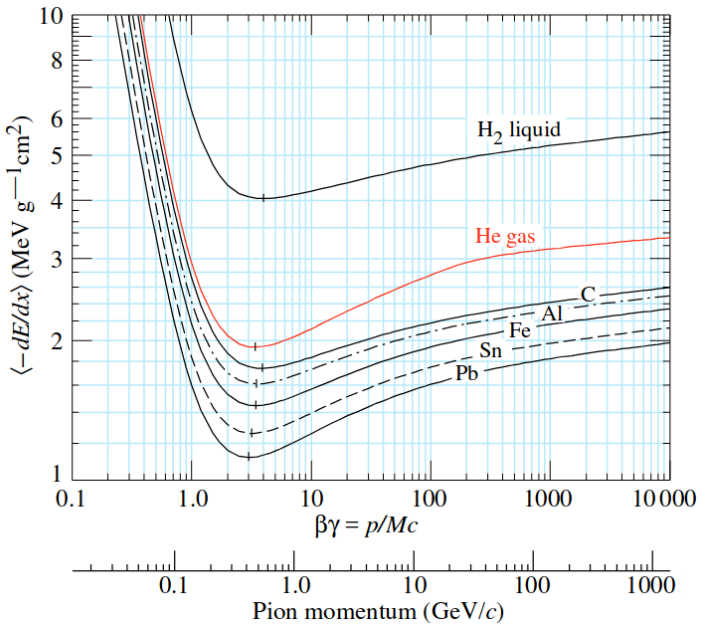


Figure 3.4: Mean energy loss for $\pi^{\pm}$ in different materials [12].

MeVg^{-1}cm^2 above a momentum of 100 MeV/c, making it a *minimum ionizing particle*. For a calorimeter with 400 mm of lead as an absorbing medium with a density of 11.35 g/cm^3, which is a very rough approximation of the ATLAS EM calorimeter absorber medium close to $\eta = 0$, it will leave roughly 700 MeV in the EM calorimeter, passing it as a single "prong".

To estimate the opening angle of the $\tau^{\pm} \to \pi^{\pm}\pi^0\nu_\tau$ decay, which occurs via an intermediate $\rho^{\pm}$ meson, we use (3.11) to get the typical value of the energy of a ρ meson for the same low τ lepton energy of 10 GeV for which $\gamma \approx 6$. Using (3.12) to compute E^* for a $\tau^{\pm} \to \rho^{\pm}\nu_\tau$ decay, we get a typical value of $E_\rho = 6$ GeV. Then, to compute the angle between the ρ meson and each of the pions in the $\rho^{\pm} \to \pi^{\pm}\pi^0$ decay, we can use (3.1), where we assume that $m_1 = m_2 = m_{\pi^{\pm}} \approx m_{\pi^0} \approx 140$ MeV and $m = m_\rho \approx 770$ MeV, and (3.7) to get a value of $\psi^{\rho,\pi}_{\text{C, max.}} \approx 0.18$ radians. Of course, for higher energies of the τ lepton, the computed opening angles will be smaller.

The immediate decay of the energetic neutral pion into two photons will result in a highly collimated photon pair, with the opening angle roughly given by (3.8).

To conclude this section, these computations indicate that, for the energies of interest of τ leptons produced in proton-proton collisions at the LHC, the angular window in which a τ lepton can be contained is of the order of 0.3 radians. Translated to ATLAS coordinates (see Section 5.2.1), and assuming we only consider the plane transverse

to the beam axis, this corresponds to an angular distance of about $\Delta R = 0.3$. In addition, the visible hadronic decay products of a τ lepton will be typically located within a cone of $\Delta R = 0.2$ with respect to the true τ lepton's direction for the low-end of the expected energies and the most common τ decays. That being said, most analyses involving τ leptons typically look at high energy regions due to high background contamination at low energies. In these regions, these angles are lower still.

Process	Branching fraction (%)
Leptonic	
$e^- \bar{\nu}_e \nu_\tau$	17.8175 ± 0.0399
$\mu^- \bar{\nu}_\mu \nu_\tau$	17.3937 ± 0.0384
Hadronic, "1-prong", pions only	
$\pi^- \pi^0 \nu_\tau$	25.4941 ± 0.0893
$\pi^- \nu_\tau$	10.8164 ± 0.0512
$\pi^- 2\pi^0 \nu_\tau$	9.2595 ± 0.0964
$\pi^- 3\pi^0 \nu_\tau$	1.0429 ± 0.0707
Hadronic, "3-prong", pions only	
$\pi^- \pi^- \pi^+ \nu_\tau$	8.9868 ± 0.0513
$\pi^- \pi^- \pi^+ \pi^0 \nu_\tau$	2.7404 ± 0.0710
Hadronic, "1-prong", involving kaons	
$\pi^- \bar{K}^0 \nu_\tau$	0.8384 ± 0.0138
$K^- \nu_\tau$	0.6964 ± 0.0096
$K^- \pi^0 \nu_\tau$	0.4328 ± 0.0148
$\pi^- \bar{K}^0 \pi^0 \nu_\tau$	0.3817 ± 0.0129
$K^- \pi^0 K^0 \nu_\tau$	0.1500 ± 0.0070
$K^- K^0 \nu_\tau$	0.1486 ± 0.0034
Hadronic, "3-prong", involving kaons	
$\pi^- K^- K^+ \nu_\tau$	0.1435 ± 0.0027

Table 3.1: Summary of the leading decay modes and corresponding experimental branching fractions for the τ lepton. The branching fractions exclude processes with intermediate resonances. Only several leading ones are shown.

4

MACHINE LEARNING

Machine learning (ML) has historically played a significant role in signal-background classification tasks in particle physics using classical techniques, such as boosted decision trees (BDT), support vector machines (SVM), simple multi-layer perceptrons (MLP) and others [15]. It is an extremely fast-growing field, where novel ML models are introduced regularly, especially in the field of Artificial Neural Networks (ANN) which show exceptional performance in learning tasks in a variety of fields.

4.1 ARTIFICIAL NEURAL NETWORKS

ANNs are ML models inspired by the functioning of networks of neurons in the brain. They consist of interconnected layers of neurons, each having multiple inputs and a single output, so that the output is a weighted, and optionally biased, sum of the inputs followed by an activation function, such as a sigmoid or a simple ramp function. The model's weights can be learned from training data, by minimizing a distance metric between the output of the ANN and the expected output, for each data point. An important property of ANNs is that they can approximate any well-behaved function given a deep or wide enough architecture [16] [17].

4.2 GRAPH NEURAL NETWORKS IN EVENT CLASSIFICATION

A recent development in the field of ANNs is an approach for learning functions on graphs, called graph neural networks (GNN). In many classification problems the data is inherently permutation invariant and is naturally representable in graph form. For example, in a collision event we have candidate leptons, jets and missing transverse energy, which can be placed on nodes of a graph with the edge connecting any two nodes being some geometric proximity measure of the two objects, such as their $\Delta R = \sqrt{\Delta\eta^2 + \Delta\phi^2}$. Such an approach has been shown to outperform other cutting-edge ML algorithms such

as Deep Neural Networks (DNNs), which are multi-layered ANNs, in the context of event classification at the LHC [18][19][20].

4.2.1 *Graph Network Formalism*

Formally, a graph can be described by $G = (\mathbf{u}, V, E)$, where $\mathbf{u}$ represent graph-level attributes, $V = \{\mathbf{v}_i\}_{i=1:N_v}$ are a set of node attributes with N_v the number of nodes and $E = \{(\mathbf{e}_k, r_k, s_k)\}_{k=1:N_e}$ where $\mathbf{e}_k$ is a collection of attributes of edge k which connects nodes with indices s_k and r_k.

A convenient framework within which a general function of a graph or, equivalently, a graph neural network, can be described is the *graph network* (GN) formalism [21]. *GN blocks* are graph-to-graph functions, whose output graphs have the same node and edge structure as the input. The processing stages of a GN block are:

$$\begin{aligned} \mathbf{e}'_k &= \phi^e\left(\mathbf{e}_k, \mathbf{v}_{s_k}, \mathbf{v}_{r_k}, \mathbf{u}\right) & \bar{\mathbf{e}}'_i &= \rho^{e\to v}\left(E'_i\right) & \text{Edge block} \\ \mathbf{v}'_i &= \phi^v\left(\bar{\mathbf{e}}'_i, \mathbf{v}_i, \mathbf{u}\right) & \bar{\mathbf{e}}' &= \rho^{e\to u}\left(E'\right) & \text{Node block} \\ \mathbf{u}' &= \phi^u\left(\bar{\mathbf{e}}', \bar{\mathbf{v}}', \mathbf{u}\right) & \bar{\mathbf{v}}' &= \rho^{v\to u}\left(V'\right) & \text{Global block} \end{aligned} \tag{4.1}$$

Note that there are two types of functions - *update functions* (ϕ), which accept a fixed-sized input and produce a fixed-sized output, and *aggregation functions* (ρ), which accept a variable-size set of inputs and produce a fixed-sized representation of it, so that the output is invariant under permutations of the input set, which is the reason the GN can approximate a function of a set of objects, invariant under permutations.

The edge block updates the attributes $\mathbf{e}_k$ of edge k to $\mathbf{e}'_k$, a representation (also called *embedding*) of the attributes, which is a function ϕ^e of the edge attributes $\mathbf{e}_k$, the attributes of its start and end nodes, $\mathbf{v}_{s_k}$ and $\mathbf{v}_{r_k}$, respectively, and the graph-level attributes $\mathbf{u}$. Similarly, the node block updates the attributes $\mathbf{v}_i$ of each node i to an embedding $\mathbf{v}'_i$ which is a function ϕ^v of the aggregation $\rho^{e\to v}$ of the embeddings E'_i of edges terminating at node i, the node attributes $\mathbf{v}_i$ and the graph-level attributes $\mathbf{u}$. The global block transforms the graph-level attributes to an embedding $\mathbf{u}'$, which is a function ϕ^u of the aggregations $\rho^{e\to u}$, $\rho^{v\to u}$ of the set of all edge (E') and node (V') embeddings, respectively, and the graph-level attributes $\mathbf{u}$. This is illustrated graphically in Figure 4.1.

In practice, the update functions ϕ are often implemented using a simple ANN per function while the aggregation functions are imple-

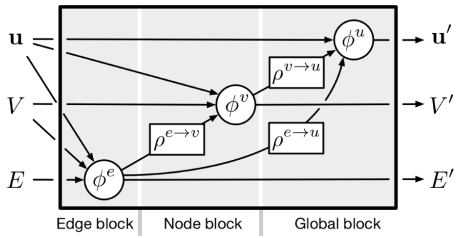


Figure 4.1: A GN block. A graph $(\mathbf{u}, V, E)$ is the input graph and the output is $(\mathbf{u}', V', E')$

mented as permutation-invariant reduction operators, such as a sum, mean or maximum.

An important thing to note is that after passing a single GN block, the embedding of each node incorporates information from nodes directly connected to it by an edge as well as from the edges themselves. That way, nodes receive information from their surroundings through edges connecting them with adjacent nodes. Also, graph-level embeddings receive information from the set of all edges and nodes and edge embeddings receive information from their originating and terminating nodes. This information sharing is referred to as *message passing* and it's an important advantage of GNNs over other ANN architectures. It allows the GNN to take into account node attributes, global attributes and pairwise relations between nodes encoded in edge attributes to produce the output. Multiple GN blocks can be chained together, allowing node embeddings to incorporate information from outside their immediate neighboring nodes.

The most general GN block given by (4.1) and Figure 4.1 describes the most general GNN architecture. However, it can be modified by removing or rearranging some of the update and aggregation functions. The GN formalism is therefore suitable for describing many GNN architectures.

4.2.2 *Deep Sets*

A deep set is a simple GNN architecture, without message passing between nodes. In the framework of the GN formalism, it can be described by a node embedding function ϕ^v, an aggregation function $\rho^{v\to u}$ and a graph-level update function ϕ^u, arranged as shown on Figure 4.2.

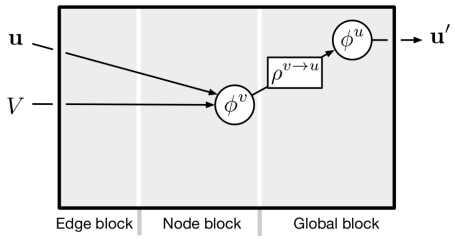


Figure 4.2: A DeepSet GN block.

4.3 BOOSTED DECISION TREES

A boosted decision tree (BDT) is a prediction model which accepts a vector of a fixed set of features as inputs and produces a single score as an output. The score can be further thresholded for the BDT to act as a classifier. For any input represented by a vector of values corresponding to the input features, a collection of decision trees, each representing a set of cuts in the input feature's space and providing a score for a given input is *boosted* by summing all scores together, producing the final BDT score. Each tree by itself is a weak, low performance classifier. Using ensembles of weak learners to boost performance was introduced in the early '90s and in many instances results in dramatic performance increase [22].

The package that is used to train the BDT for the τ algorithm is a Python implementation of XGBoost [23]. Since the choice of the different hyperparameters (HP) of the BDT model affects both the classification power and resource usage, they will be briefly discussed.

In the general case of a two-class classification problem, a BDT is trained on a training data set which is comprised of a set of features and their class labels. Given a training set of n feature vectors $\mathbf{x}_i$ each of dimension m and corresponding labels y_i, $\{(\mathbf{x}_i, y_i)\}_{i=1}^{n}, \mathbf{x}_i \in \mathbb{R}^m, y_i \in \mathbb{R}$, the BDT score of a model with K trees for sample i is computed as such:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(\mathbf{x}_i) \tag{4.2}$$

where $f_k(\mathbf{x}_i)$ is a score for tree k, and example i which is computed by applying a set of thresholds on the components of $\mathbf{x}_i$. The latter can be described by $q_k(\mathbf{x}_i) : \mathbb{R}^m \to T_k$, a function that returns the leaf index into which example $\mathbf{x}_i$ falls in tree k and a leaf weights vector $w_k \in \mathbb{R}^{T_k}$, so that $f_k(\mathbf{x}_i) = w_{q_k(\mathbf{x}_i)}$, with T_k the number of leaves in tree k. The model is trained by adjusting the tree structures q_k (and in the

process also the number of leaves T_k) and leaf weights w_k, so that the following objective function is minimized:

$$L = \sum_i l\left(\hat{y}_i, y_i\right) + \sum_k \left(\gamma T_k + \frac{1}{2}\lambda \|w_k\|^2\right), \tag{4.3}$$

where l is a differentiable convex distance measure between the class label and predicted score, and the second sum is a regularization term with two parameters, where the first term penalizes trees which have many leaves and the second penalizes large values of leaf scores. The index i runs over a set of input vectors used to evaluate the training process while k runs over the trees. The importance of the two regularization terms is controlled by the parameters γ and λ. The γ parameter controls the level of *pruning* of each tree - the higher it is, the less leaves each tree will tend to have, while the λ parameter keeps the magnitude of the leaf scores from diverging.

4.3.1 *The XGBoost Training Algorithm*

The training algorithm for XGBoost is a gradient boosting algorithm with optimizations aimed at making it exceptionally fast and resistant to overfitting. As in other gradient boosting algorithms, every training iteration a new decision tree is added. The new tree's structure is converged upon by starting with a root node and scanning different cut points for all features in the input vectors so that the split has a maximum gain to the predictive power of this tree until this gain is lower than a certain threshold.

If $\hat{y}_i^{(t-1)}$ is the prediction of the model for instance i at iteration $t-1$, then when we add a new tree $f_t(\mathbf{x}_i)$, at iteration t, we need to minimize the following objective function

$$L^{(t)} = \sum_i^n l\left(y_i, \hat{y}_i^{(t-1)} + f_t(\mathbf{x}_i)\right) + \Omega(f_t), \tag{4.4}$$

where $\Omega(f) \equiv \left(\gamma T + \frac{1}{2}\lambda \|w\|^2\right)$ is the regularization term from (4.3). Expanding in f_t,

$$L^{(t)} \simeq \sum_i^n \left[l\left(y_i, \hat{y}_i^{(t-1)}\right) + g_i f_t(\mathbf{x}_i) + \frac{1}{2} h_i f_t^2\left(\mathbf{x}_i\right)\right] + \Omega(f_t), \tag{4.5}$$

where $g_i \equiv \partial_{\hat{y}^{(t-1)}} l\left(y_i, \hat{y}_i^{(t-1)}\right)$ and $h_i \equiv \partial^2_{\hat{y}^{(t-1)}} l\left(y_i, \hat{y}_i^{(t-1)}\right)$.

The first term is constant and therefore can be dropped. This can then be re-written the following way:

$$\tilde{L}^{(t)} = \sum_{j=1}^{T} \left[\left(\sum_{i \in I_j} g_i \right) w_j + \frac{1}{2} \left(\sum_{i \in I_j} h_i + \lambda \right) w_j^2 \right] + \gamma T. \tag{4.6}$$

Where I_j is the subset of all input vectors which fall into leaf j. If the tree structure q is fixed, we can minimize $\tilde{L}^{(t)}$ to get the optimal set of leaf weights

$$w_j^* = -\frac{G_j}{H_j + \lambda} \tag{4.7}$$

for which the minimum value for $\tilde{L}^{(t)}$ is

$$\tilde{L}^{(t)}(q) = -\frac{1}{2} \sum_{j=1}^{T} \frac{G_j^2}{H_j + \lambda} + \gamma T \tag{4.8}$$

where $G_j = \sum_{i \in I_j} g_i$, $H_j = \sum_{i \in I_j} h_i$.

Using this result, we can start constructing a tree in the following manner. Starting from a root node representing a leaf into which all input vectors fall and which has some initial weight value, introduce a cut on one of the input features at a certain value of the feature, creating a split of the set of all input vectors into two subsets I_L and I_R. This split will increase the loss function by

$$L_{\text{split}} = \frac{1}{2} \left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{G^2}{H + \lambda} \right] - \gamma, \tag{4.9}$$

where G and H are evaluated over $I = I_L \cup I_R$. Repeat over all possible cuts on all features and select a cut that produces a minimal value for (4.9). Repeat the process recursively for the resulting tree until a stopping condition is reached. A common stopping condition is a requirement that some minimum number of input vectors shall fall into a leaf or that no further cut exists that reduces the loss (4.9) sufficiently.

The problem now is reduced to finding an optimal way to pick the cut values and the features on which to cut in order to reduce training time. Implementations such as scikit-learn [24] scan over every possible cut value of all features, being very powerful but at the same time resulting in long training times which are prohibitive when trying to analyse and compare multiple trainings. To shorten training times XGBoost uses an approximate algorithm which, for each feature, proposes optimal cuts which are used for the cut values in

the construction of the tree. That, along with optimal use of computer resources by distributing the training over multiple processors makes training a BDT with XGBoost exceptionally fast.

To mitigate overfitting, in addition to the regularized objective function (4.3), which helps smooth the final learnt weights, additional techniques are used. Feature sub-sampling selects a random sub-set of features to consider for each new training iteration. Shrinkage is used to assign a progressively smaller weight to subsequent trees. Early stopping is used to evaluate the trained model at every iteration on a validation set and stop the training if the performance on the latter starts to degrade.

4.3.2 *XGBoost Hyperparameters*

In the training of BDTs in general and with XGBoost in particular, several HP can be modified to achieve optimal performance, latency and resource usage. These are the number of trees K, maximum depth of a tree M and the γ pruning regularization parameter discussed previously, as well as the number of early stopping rounds and scale_pos_weight, which compensates for an unbalanced dataset.

The deeper the trees in the BDT, the better the BDT captures dependencies between variables during training. It is important to not have this parameter set too high, however, since then it would fail to generalize and will overfit on outliers. The benefit of higher M can somewhat be preserved and overfitting can be mitigated by using a higher value of γ, keeping the tree pruned, so that tree branches which contribute to better separation between the classes are kept, while others are pruned during the training process.

Each successive tree in a BDT corrects for the errors of the previous tree's prediction, so it's generally better to have a high value of K. However, at some point further increase in K might also lead to slight overfitting when subsequent trees start to compensate for noise.

Lastly, if the training dataset is unbalanced with e.g. more signal than background samples, the BDT's structure should be more strongly corrected for prediction errors on the minority sample than on the majority one, which is controlled by the scale_pos_weight HP.

4.4 EVALUATION OF A MODEL'S PERFORMANCE

A trained ML model assigns a score to a data point presented at the input which is preferably high for signal and low for background data. A two-class classifier model, which we'll be concerned with in this work, sets a threshold on the score to separate signal from background. There are several metrics that can be computed to describe the model's performance, but the two most critical ones for our purpose, which is trying to learn, from data, about the presence or absence of a signal, are *efficiency* and *background rejection*.

The efficiency ϵ_S, is the probability that a signal input will be tagged as such by the model. Background rejection R_B is the probability that a background input is not tagged as signal and is trivially related to the *background efficiency* ϵ_B, which is the probability to tag a background input as signal:

$$R_B = 1 - \epsilon_B \tag{4.10}$$

For example, in the case of a trigger application in high energy physics, that will be discussed in depth in the following chapters, high trigger rates are prohibitive, so R_B should be as high as possible, while we want to keep as many interesting events as we can, which translates to having a high ϵ_S.

Usually, training a model for classification is a computationally intensive task, so the model is trained once after carefully selecting the hyperparameters and then tuned by modifying the threshold above which the classifier tags an input as signal. Increasing this threshold results in less signal and background events being selected and thus in lower ϵ_S and higher R_B and vice versa.

A convenient metric for visualizing the power of a trained model in the context of signal-background classification by a model producing a thresholded score is the receiver operating characteristic (ROC) curve, as the one shown on e.g. Figure 13.10b. It is computed by varying the threshold over a signal and background distribution of the model's score and for each threshold plotting ϵ_S on the y axis and ϵ_B on the x axis. The straight blue dashed line is produced when the model has no classification power, i.e. attributing random scores to the inputs, while a concave curve implies a more powerful model. Furthermore, the area under the curve (AUC) has a convenient probabilistic interpretation - it is the probability that a randomly selected signal event is scored higher by the model than a randomly selected background event, and is often used as a metric to describe the power of the model.

Part II

EXPERIMENT

5

THE LHC AND THE ATLAS EXPERIMENT

5.1 THE LARGE HADRON COLLIDER

The Large Hadron Collider (LHC) [25] at the European Organization for Nuclear Research (CERN) located near Geneva on the Franco-Swiss border is a hadron accelerator and collider installed in the circular 26.7 km tunnel, which lies in a depth between 45 and 170 meters, originally constructed for the Large Electron-Positron (LEP) collider.

The LHC is designed to accelerate protons or heavy ions, according to a predetermined schedule, and collide them at record energies for physics studies. Particles are accelerated in a sequence of circular accelerators after being boosted in a linear accelerator, as shown in Figure 5.1 until they reach the LHC beam pipe. Two beams travel in opposite directions until they reach the target energy. There are four *interaction points* at which the beams cross, where the particles, arranged in *bunches*, cross each other and produce collisions every *bunch crossing*, which occur at 40 MHz during physics data taking periods.

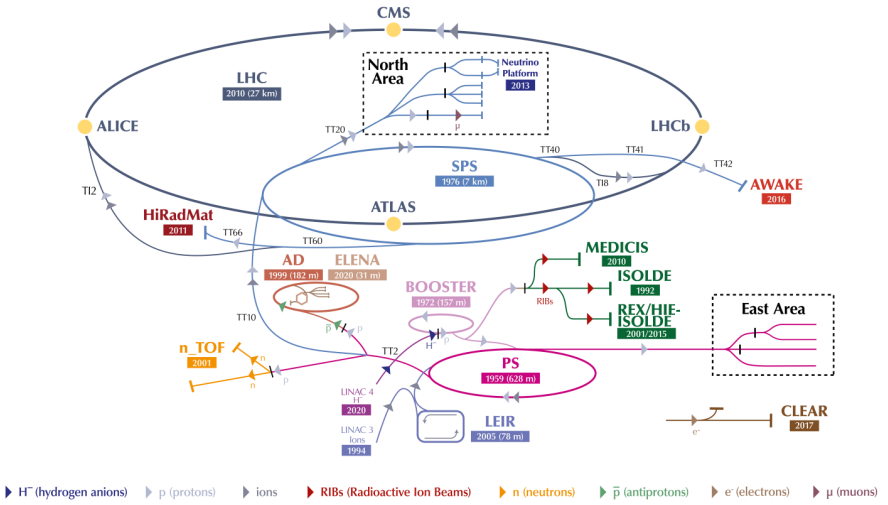


Figure 5.1: The CERN accelerator complex.

The LHC is an infrastructure that is designed to deliver hadron collisions to several detectors located around each interaction point and is maintained by large international collaborations. The physics program

of the LHC is dominated by four experiments - ATLAS (A Toroidal LHC ApparatuS) and CMS (Compact Muon Solenoid), designed to be general-purpose experiments for the detection of the Higgs boson, super-symmetry, dark matter and precision measurements of the SM, ALICE (A Large Ion Collider Experiment) which is a detector dedicated to heavy ion physics and the study of conditions which existed just after the big bang, and LHCb (Large Hadron Collider beauty), which focuses on the study of b-mesons and the matter-antimatter asymmetry. These experiments record collisions delivered by the LHC in multi-year runs with data already taken during Run 1 (2010 - 2013) with beam energies of 3.5 and 4 TeV [26], and Run 2 (2015 - 2018) with a beam energy of 6.5 TeV [27].

5.2 THE ATLAS DETECTOR

5.2.1 *Coordinates and Kinematic Variables*

The ATLAS detector [28] is a particle detector surrounding the interaction point with respect to which it is nominally forward-backward symmetric. Cylindrical coordinates are used to describe the detector and the particles emerging from collisions. The origin of the coordinate system is nominally defined to coincide with the interaction point. The beam direction defines the z-axis; the x-axis is defined as pointing from the origin toward the center of the LHC ring, and the y-axis is defined as pointing from the origin upward. The azimuthal angle ϕ is measured from the positive x-axis and takes values in the range π and $-\pi$, while the polar angle is measured from the positive z-axis and takes the values in the range 0 and π. The transverse plane is defined as the x-y plane, while the longitudinal direction is defined as being along the beam axis.

Since in a hadron collider, only some partons of two colliding hadrons interact, the initial momentum along the beam direction is essentially unknown, and therefore, all resulting particles originating from a single collision are arbitrarily boosted along the beam axis. Angles of particles as expressed by the polar angle θ would therefore depend non-trivially on the boost, which would be inconvenient for expressing, e.g., angular distances between particles. Therefore, it is convenient to introduce rapidity, defined as $y = \frac{1}{2}\ln\left(\frac{E+p_z}{E-p_z}\right)$, where $p_z = |\mathbf{p}|\cos\theta$. This quantity is additive under Lorenz boosts along the longitudinal direction, and therefore, for each collision

event, the angular distances between particles expressed in this variable are Lorenz invariant. For high energies, $p \gg m$ and therefore $y \to \frac{1}{2} \ln\left(\frac{1+\cos\theta}{1-\cos\theta}\right) \equiv \eta$, called the pseudorapidity, which is the actual quantity used for energetic particles. Thus, particles detected in the ATLAS experiment are described, in addition to their energy, by three quantities invariant under boosts along the longitudinal direction - the transverse momentum $p_T = \sqrt{p_x^2 + p_y^2}$, azimuthal angle ϕ and pseudorapidity η.

5.2.2 *Detector Layout*

The overall layout of the detector is shown in Figure 5.2. It consists of

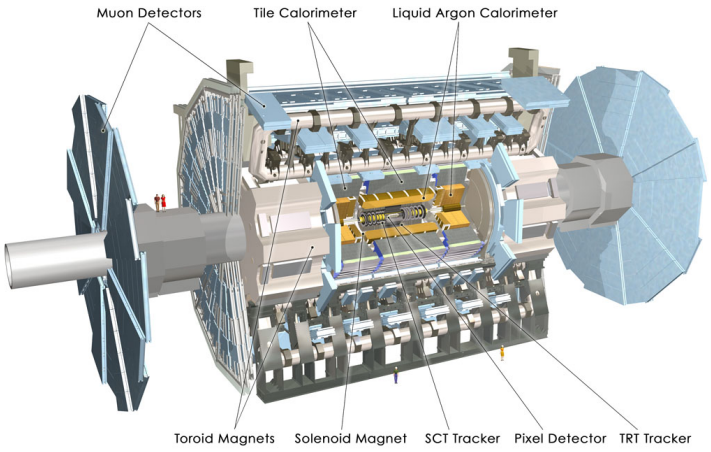


Figure 5.2: Cut-away view of the ATLAS detector.

an inner detector (ID) used for tracking, surrounded by electromagnetic and hadronic calorimeters and an external muon spectrometer. The solenoid is surrounded by a lead/liquid argon electromagnetic calorimeter, which, along with its end-cap components (EMEC), covers the region $|\eta| < 3.2$. Electromagnetic calorimetry coverage is also provided by a copper/liquid-argon module of the forward calorimeter (FCal) in the region $3.1 < |\eta| < 4.9$. The electromagnetic calorimeter is surrounded by a hadronic calorimeter comprised of steel as the absorber medium and scintillator tiles, along with a copper/liquid argon hadronic end-cap calorimeter (HEC) and two copper-tungsten/liquid-argon modules of the FCal, providing hadronic calorimetry coverage of $|\eta| < 4.9$. The muon detection system surrounds the calorimeters. It consists of a system of high-precision monitored drift tube chambers (MDT) detectors arranged in 3 concentric cylindrical layers, which provides coverage in the range $|\eta| < 2.7$ along with corresponding

MDT end-cap layers. In the region $2 < |\eta| < 2.7$, cathode-strip chambers (CSC) are used instead of MDT in the inner layer. An additional system of thin gap chambers (TGC) and resistive plate chambers (RPC) is used to trigger on muons. Barrel and end-cap toroids produce a toroidal magnetic field of strengths 1 T and 0.5 T, respectively, for the barrel and end-cap muon detector systems.

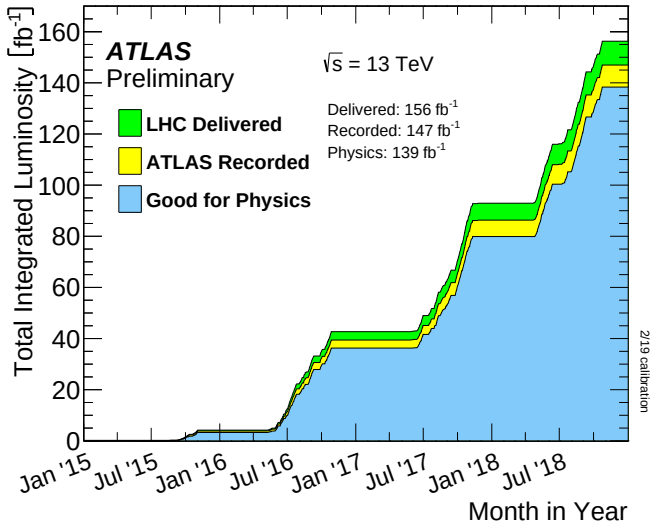
The triggering system consists of a hardware-based first-level (L1) trigger and a software-based high-level trigger (HLT), which reduces the rate of events from around 40 MHz of collisions to around 1 kHz for recording.

By the end of Run 2, an integrated luminosity of 139 fb^{-1} of collision events good for physics analysis at a center-of-mass energy of $\sqrt{s} = 13$ TeV were collected with an average of 33.7 hard interactions per event (pileup), as shown on Figure 5.3. The ongoing Run 3 is expected to collect 400 fb^{-1} of events at $\sqrt{s} = 14$ TeV with peak pileup of 55 [29].

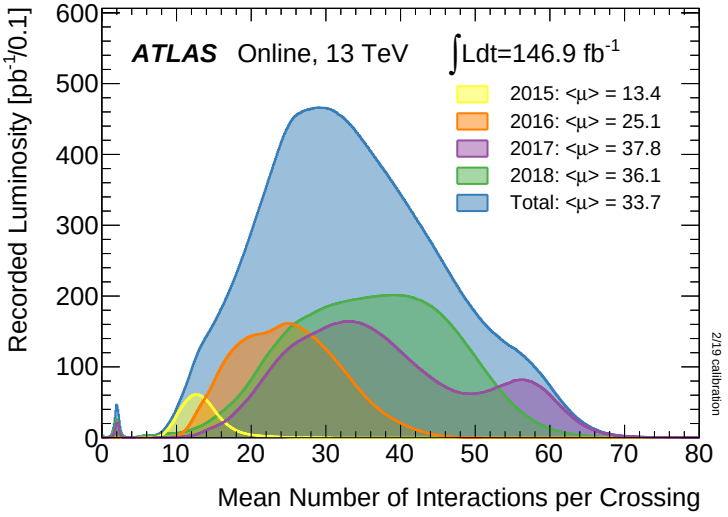
5.2.3 *Inner Detector*

The inner detector (ID) [30], being closest to the beam pipe, is responsible for constructing the tracks and vertices in a collision event with high efficiency. Along with the calorimeter and muon systems, it contributes to the electron, photon, and muon reconstruction and supplies the important extra signature for displaced vertices of short-lived particle decays. The detector provides full tracking coverage in the region $\eta < |2.5|$ and is immersed in a 2 T magnetic field, curving trajectories of charged particles and allowing the reconstruction of their momenta.

The inner detector, illustrated in Figure 5.4, consists of three subsystems relying on different tracking technologies and having different spatial resolutions. The first layer, closest to the interaction point, is comprised of fine-granularity pixel semiconductor tracking detectors, with a typical resolution of 12 μm in the $R\phi$ plane and 66 (77) μm in the z (R) direction in the barrel (end-cap) region. The second layer is comprised of several layers of semiconductor trackers based on silicon microstrip technology providing a resolution of 16 μm in the $R\phi$ plane and 580 μm in the z (R) direction in the barrel (end-cap) region, followed by a third layer comprised of transition radiation trackers made of gas-filled drift tubes which reduce cost and material density of the inner detector while still allowing efficient track reconstruction



(a)



(b)

Figure 5.3: Total integrated luminosity delivered by the LHC and recorded by ATLAS (a) and pile-up (b) during Run 2.

by providing a large number of points along the track. These provide a resolution of 170 μm per straw.

In the barrel region, the first-pixel layer receives most of the radiation and is, therefore, subject to significant degradation due to radiation damage. As displaced vertex-based tagging of b-jets is crucial for the majority of ATLAS analyses [31], between Runs 1 and 2 a new pixel detector, the Insertable B-layer (IBL) [31] was installed in the barrel region between this layer and the beam pipe to improve displaced vertex identification efficiency, mitigating the effect of accumulated radiation damage during Run 1.

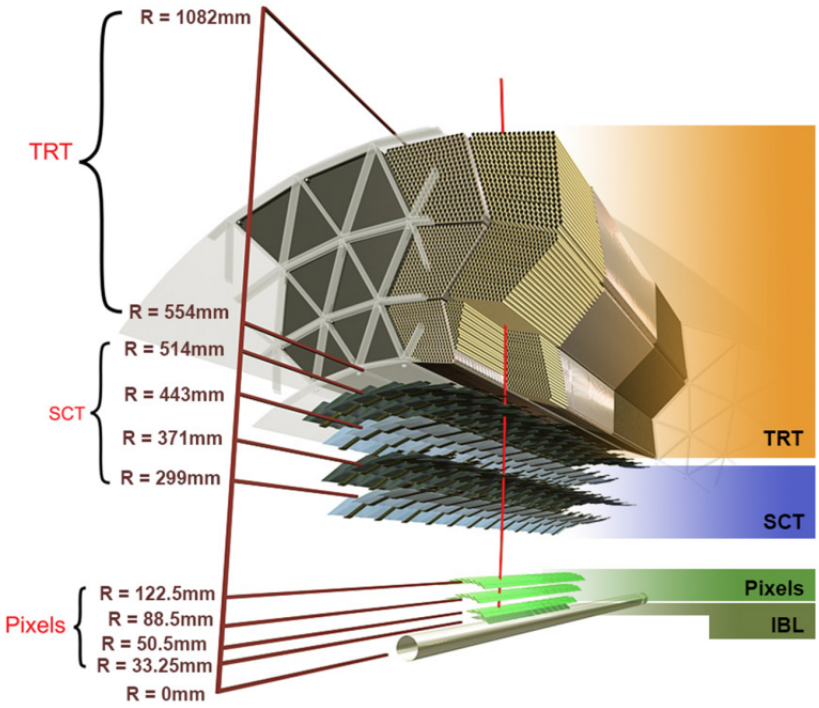
5.2.4 *Calorimeters*

The ATLAS calorimeters are designed to measure the energy and position of electromagnetically and strongly interacting particles. They are designed to absorb the energy of most of these particles produced in collisions. The ATLAS calorimeters are sampling calorimeters in that they consist of layers of "absorbing" high-density materials that stop incoming particles, interleaved with layers of "active" media that measure the particle energies [32]. The calorimeter layout of the ATLAS experiment is shown in Figure 5.5

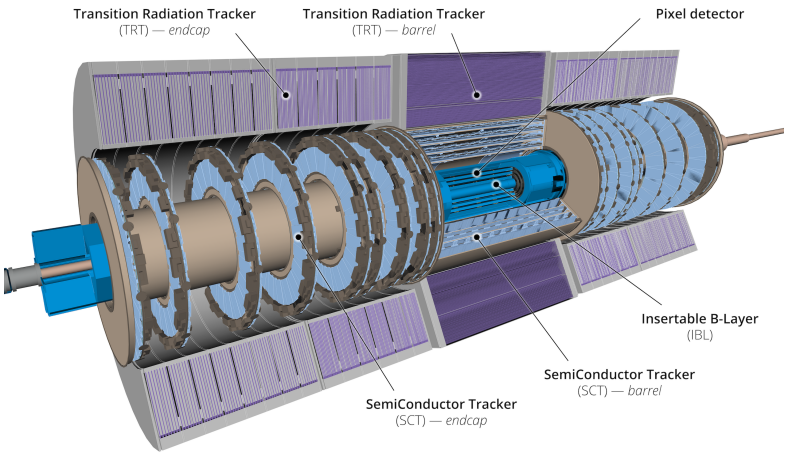
The calorimeter system is composed of a liquid argon-based EM calorimeter, which absorbs the energy of electromagnetically interacting particles in the barrel ($|\eta| < 1.475$) and end-cap ($1.375 < |\eta| < 3.2$) regions. A hadronic calorimeter surrounds the EM calorimeter, the purpose of which is to absorb any strongly interacting particles that pass through the EM calorimeter to capture the full energy of the event minus the energy carried by muons and neutrinos and minimize punch-through into the surrounding muon system. The hadronic calorimeter covers a region of $|\eta| < 4.9$.

5.2.4.1 *EM calorimetry*

An EM calorimeter is designed for optimal absorption of the energy of electromagnetically interacting particles such as photons and electrons. A charged particle loses its energy in a medium primarily either through radiative effects, such as bremsstrahlung, which is the emission of photons due to the charged particle's deceleration, or collisional losses, such as excitation and ionization of the atoms in the medium. An energetic photon loses energy primarily due to electron-positron pair production, which further interacts with the medium as



(a)



(b)

Figure 5.4: (a) Illustration of the inner detector layers in the barrel region, showing a trajectory of a traversing particle (red line). (b) Full view of the inner detector.

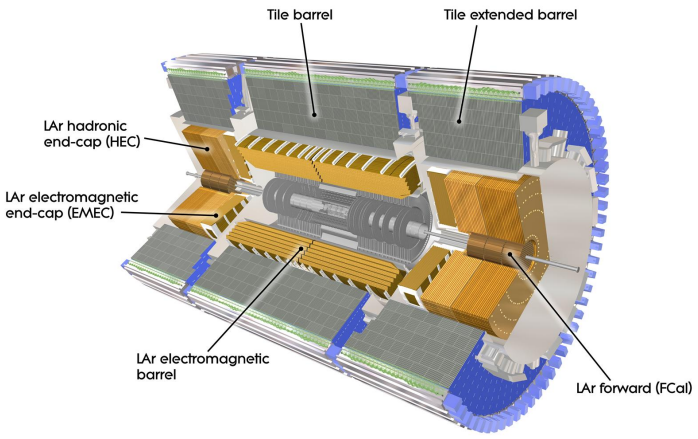


Figure 5.5: Cut-away view of the ATLAS detector.

electrically charged particles. For a given charged particle type, a *critical energy* E_c can be defined above which radiative and below which collisional losses dominate. When a highly energetic charged particle traverses the absorbing medium, it produces bremsstrahlung photons, which produce e^+e^- pairs, forming an electromagnetic shower of particles with progressively lower energies. When a charged particle's energy falls below E_c, energy loss through bremsstrahlung radiation becomes sub-dominant and eventually negligible as the charged particle slows down to rest. The dominant mechanism of energy loss at low energies is excitation and ionization. The resulting ions and free electrons are collected towards electrodes, producing electric signals allowing an estimate of the energy of the initial particle. Values of the E_c of electrons for solids and gases for different chemical elements are shown in Figure 5.6.

The typical length scale for charged, energetic particle paths in matter is the radiation length X_0, which is defined as the typical distance an energetic electron travels in the calorimeter medium before it loses all but $1/e$ of its energy, or $7/9$ of the mean free path of a en energetic photon before e^+e^- pair production. Lengths of the calorimeter are, therefore, often conveniently stated as multiples of X_0. Values of X_0 for chemical elements with $Z > 20$ are shown in Figure 5.7.

The lateral extent of an EM shower is approximated by the *Molière radius* R_M, given by

$$R_M = 21[\text{MeV}]\frac{X_0}{E_c}, \tag{5.1}$$

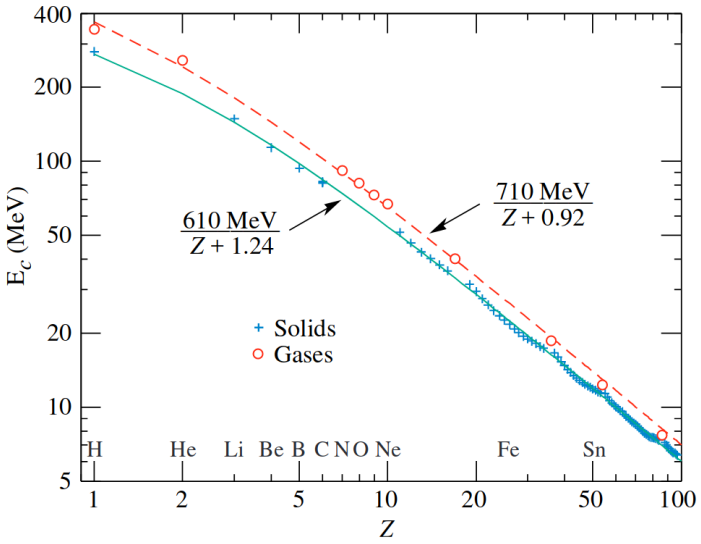


Figure 5.6: Critical energy E_c for electrons in gaseous and solid states of different chemical elements [12].

where E_c is given in MeV, and defined so that 90% of an electromagnetic shower's energy is contained within an infinite cylinder of the same radius.

For electrons, the critical energy is of the order of 10 MeV for high-Z materials (from Figure 5.6), while X_0 is of the order of 1 cm (from Figure 5.7), which results in $R_M \sim 2$ cm. This value drives the choice of cell granularity of the electromagnetic calorimeter.

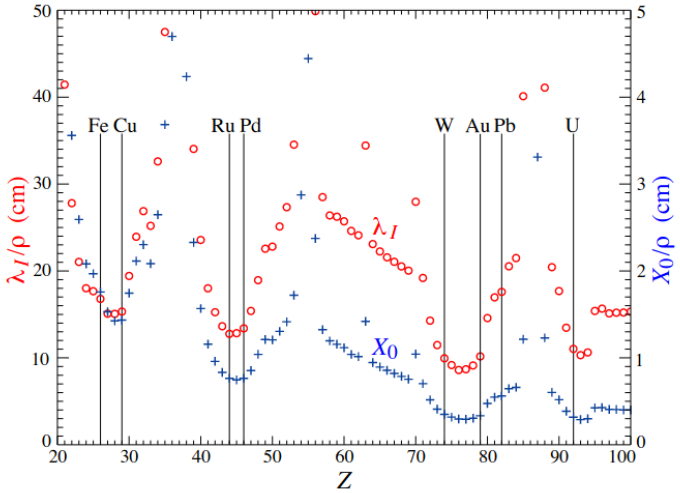


Figure 5.7: Radiation length X_0 and nuclear interaction length λ_I for chemical elements with $Z > 20$ [12].

5.2.4.2 *The ATLAS EM LAr calorimeter*

In ATLAS, the EM calorimeter is a lead/liquid argon (LAr) detector, with the former serving as the active medium and the latter as the absorber medium. It has an accordion geometry which provides complete ϕ coverage without azimuthal cracks and is divided into a barrel part ($|\eta| < 1.475$) and two end-caps ($1.375 < |\eta| < 3.2$). The thickness of the lead absorber plates changes with η and has been optimized in terms of EM calorimeter performance in energy resolution. The EM calorimeter has a thickness of $> 24X_0$ in the barrel region and $> 26X_0$ in the end-caps.

Over the region devoted to precision physics ($|\eta| < 2.5$) the EM calorimeter is segmented into three longitudinal sections, as shown in Figure 5.8. The strip section, which has a constant thickness in η of $4.3X_0$ is equipped with narrow strips of width $\Delta\eta = 0.0031$, well below the Molière radius R_M, and is used as a 'preshower' detector, enhancing identification of particles and providing a precise position measurement in η. The second section is thick and is designed to absorb most of the energy. It is transversally segmented into square towers of size $\Delta\eta \times \Delta\phi = 0.025 \times 0.025$, which translates to roughly 37×37 mm at the narrowest around $\eta = 0$, which is enough to contain more than 90% of an electromagnetic shower, based on the estimated value $R_M = 20$ mm. This layer has a varying thickness, tapering towards high η. The third section is segmented into $\Delta\eta \times \Delta\phi = 0.05 \times 0.025$ sections and has a thickness of at least $2X_0$.

Over the η range where the amount of material before the calorimeter exceeds $2X_0$, such as in the region $|\eta| < 1.8$, the EM calorimeter is preceded by a presampler layer, with granularity $\Delta\eta \times \Delta\phi = 0.05 \times 0.025$, which is necessary to correct for the energy lost in the material located between the interaction point and the calorimeter edge.

During the Phase-I upgrade that occurred between Run 2 and Run 3, the readout of the LAr calorimeter was significantly improved. Before the upgrade, each trigger tower in the EM calorimeter had only a single readout. Following the upgrade, this number increased to 10, with access to the four individual layers and four sub-divisions in η of the 0.1×0.1 regions of the first and second sampling layers. Each such region that can be individually read out is called a *supercell*. Figure 5.9 illustrates a 70 GeV electron as seen by the L1 Calorimeter trigger before and after the upgrade.

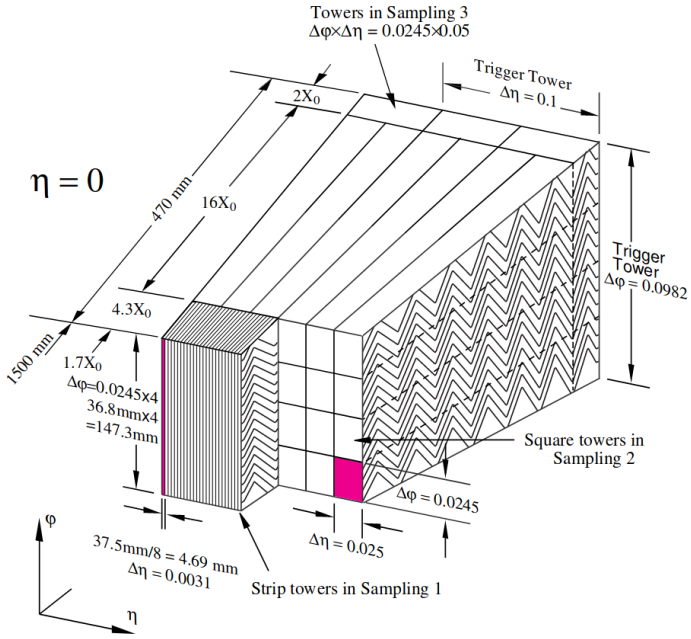


Figure 5.8: Inner structure of the EM LAr calorimeter in the barrel region.

5.2.4.3 *Hadronic calorimetry*

Hadronic calorimeters are designed to absorb the energy of strongly interacting particles, such as protons, neutrons, pions, and kaons. These particles lose their energy primarily through hadronic interactions, where secondary particles such as protons, neutrons, and pions are produced, causing a hadronic shower to be formed. At each step of the shower, about $1/3$ of the energy is converted to photons through neutral pions, which immediately decay into a pair of photons. Therefore, hadronic showers have two distinct components - electromagnetic and hadronic, composed of charged pions, heavy fragments, and excited nuclei. This gives rise to a much more complex cascade development that limits the hadronic calorimeter's energy resolution.

The typical length scale for hadronic showers is the nuclear interaction length λ_I, which is the mean distance a hadron travels in the medium before undergoing an inelastic nuclear interaction. λ_I is typically larger than X_0 and the typical length of a hadronic calorimeter is larger than that of an EM calorimeter. Values of λ_I for different chemical elements are illustrated in Figure 5.7.

5.2.4.4 *The ATLAS Hadronic Calorimeters*

The hadronic calorimeter system in ATLAS covers a wide range of $|\eta| < 4.9$ and is divided into sections utilizing different materials

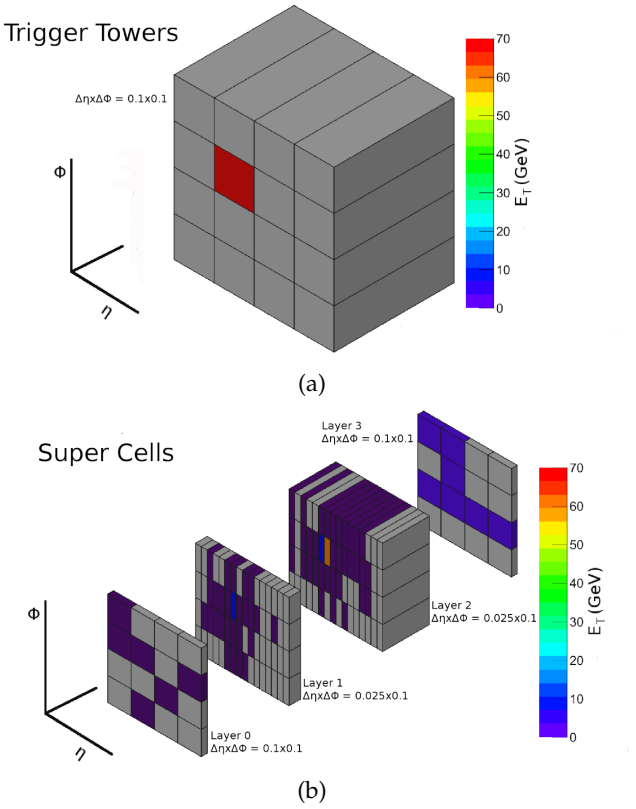


Figure 5.9: A 70 GeV electron as seen by the Level-1 calorimeter trigger before (a) and after (b) the Phase-1 upgrade.

and techniques best suited for the varying requirements and radiation environments in them. In the region $|\eta| < 1.7$, the hadronic calorimeter is built using iron as the absorber and scintillating tiles as the active medium. Over the range $1.5 < |\eta| < 4.9$, LAr calorimeters are used. The latter region is sub-divided into the hadronic end-cap (HEC) calorimeter in the range $|\eta| < 3.2$ and the high-density forward calorimeter (FCAL) in the range $3.1 < |\eta| < 4.9$.

The Hadronic calorimeter has to reduce punch through to the muon system to a minimum and provide good containment for hadronic showers. A total thickness of $11\lambda_I$ was shown to reduce the punch through to a minimum while a thickness of the active calorimeter medium of around $10\lambda_I$ was shown to be adequate to provide good resolution for high energy jets [31].

5.2.5 *Muon Spectrometer*

The Muon Spectrometer (MS) [33][32] forms the large outer part of the ATLAS detector and detects charged particles exiting the calorimeters, measuring their momentum in the pseudorapidity range $|\eta| < 2.7$. Since muons are minimum ionizing particles, depositing a minimal amount of energy of around 1.5 $\text{MeVg}^{-1}\text{cm}^2$ in any solid medium they pass through, which corresponds to $\mathcal{O}(1)$ GeV/m for lead and iron, most muons produced in collisions will escape the calorimeter system into the MS.

Since the MS volume is very large it is impossible to provide continuous tracking. Instead, tracks are reconstructed from measurements at 3 stations - the first one close to the calorimeters, upstream of the toroid magnetic field, the second one inside or immediately after the magnetic field, and a third one well outside the magnetic field of ATLAS. At each of the three stations, the measurement is multilayered, reconstructing a straight-track segment with a well-measured orientation in space.

The MS is immersed in a nonhomogenous magnetic field, causing the muon paths to curve. Through the measurement of this curvature the muon momentum is estimated. A magnetic field mostly orthogonal to the trajectories is generated by the large barrel toroid magnet in the range $|\eta| < 1.0$, smaller end-cap magnets in the range $1.4 < |\eta| < 2.7$ and a combination of the fields in the transition region of $1.0 < |\eta| < 1.4$.

The MS is designed for a momentum resolution of $\Delta p/p_T < 10^{-4}$ for $p_T > 300$ GeV and a lower resolution of a few percent for lower muon p_T. To achieve this target resolution from the measurement of tracks obtained from 3 points, each point must be measured with an accuracy better than 50 μm. This, along with economic constraints, guided the choice of design and detector technologies utilized in the MS, resulting in several sub-detector types. Over most of the η range, a precision measurement of the track coordinates in the principal bending direction is provided by Monitored Drift Tubes (MDTs). At large pseudorapidities and close to the interaction point, Cathode Strip Chambers (CSCs) are used to withstand the demanding rate and the conditions of the background. The muon triggering system covers the pseudorapidity range $|\eta| < 2.4$ and is comprised of Resistive Plate Chambers (RPCs) in the barrel region and Thin Gap Chambers (TGCs) in the end caps. The layout of the muon system is shown on Figure 5.10.

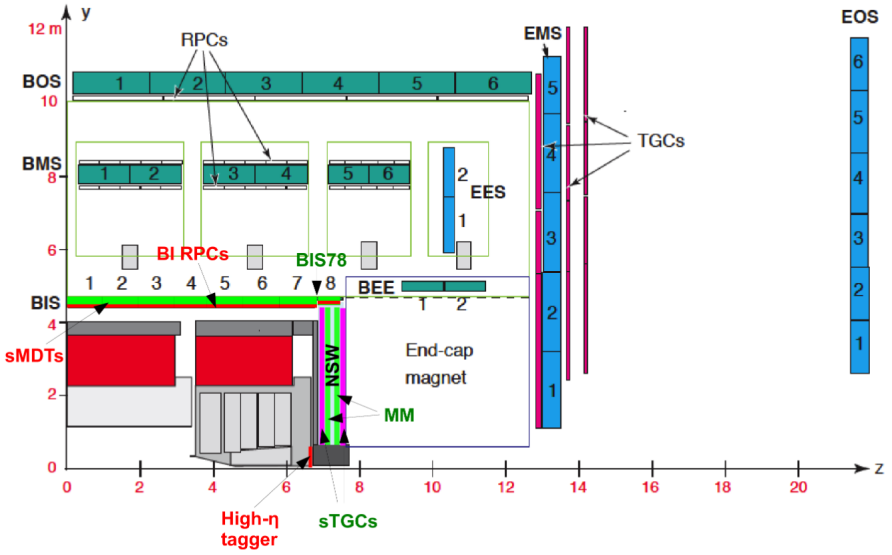


Figure 5.10: Upper right quadrant of the muon system.

During the Phase-I upgrade of the ATLAS detector, part of the end-cap TGC system was upgraded to cope with the large backgrounds expected from the planned luminosity increase. Without this upgrade, muon triggers would have to set higher p_T thresholds, losing data at low-p_T. As part of the upgrade, the New Small Wheel (NSW) was installed in place of the Endcap Inner (EI) wheel. This detector is comprised of two layers of different technologies - small strip TGC detectors (sTGC) and micro-mesh gaseous structure detectors (MM). The author of this thesis was responsible for the construction and

maintenance of a system used to coordinate the recording of measurements and hierarchy of components of the sTGC wedges shown in Figure 5.11 and eventual storage of those measurements in a central database.

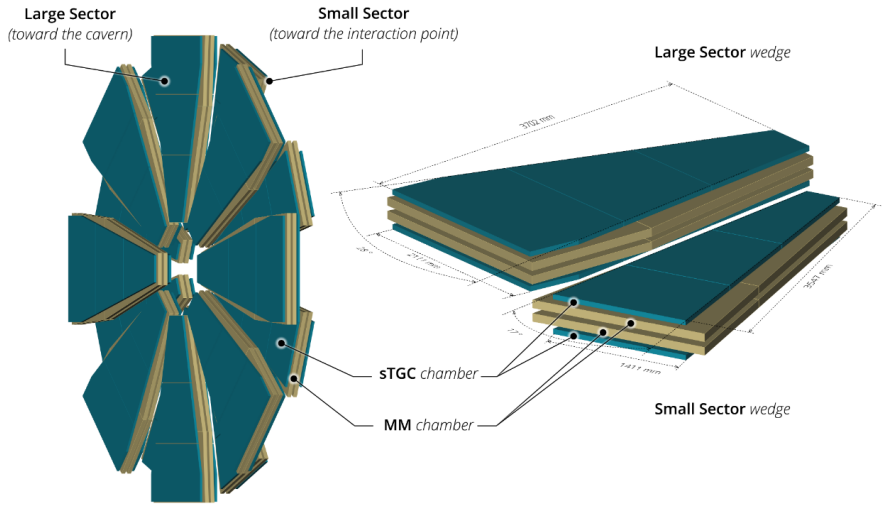


Figure 5.11: The NSW wedge (right) and arrangement of wedges into the NSW (left).

5.3 TRIGGER AND DATA ACQUISITION

The ATLAS triggering system is designed to select the most interesting events defined through physics requirements in order to reduce the vast volume of data produced by the ATLAS experiment to that manageable by the data storage system - while collisions are produced at a rate of 40 MHz, the latter can only handle a rate of 1 kHz.

The trigger system schematic for Run 3 and data rate flow comparison for Run 1 and Run 3 are shown in Figure 5.12. The system has two levels. The Level-1 trigger system (L1), which receives the unfiltered torrent of data, has to be in sync with the collision rate of 40 MHz, running simple and fast identification algorithms and therefore is implemented in hardware, driven by a hardware clock, while the latter, the High-Level Trigger (HLT), requires to run more elaborate algorithms, having less stringent latency requirements and is, therefore, software-based.

The L1 trigger system has several components that are used to identify interesting patterns in data from a collision. These include the L1 Calorimeter (L1Calo) and L1 Muon (L1Muon), which receive inputs from the calorimeters and muon detectors, respectively. The processed

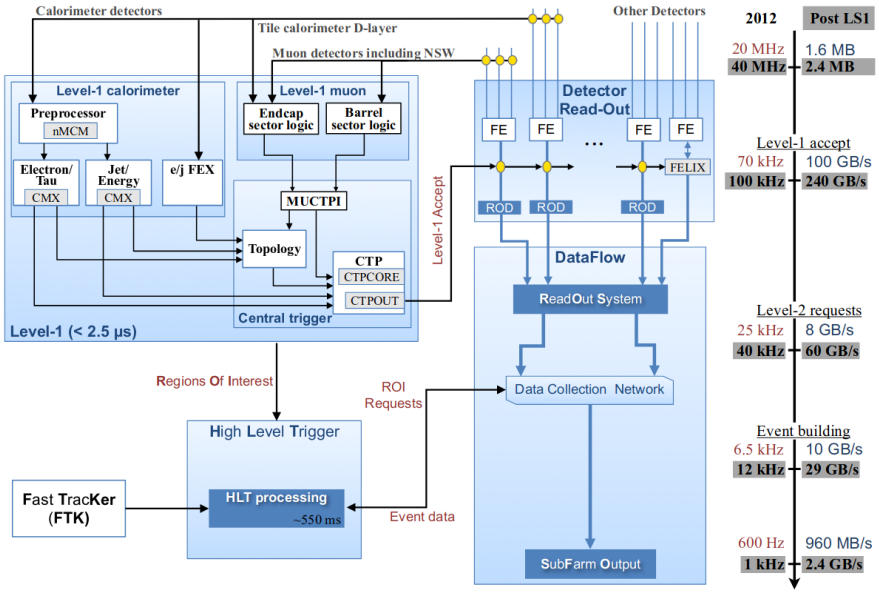


Figure 5.12: Schematic overview of the trigger system for Run 3. The data rate at each stage comparison between Run 1 and Run 3 is shown on the right.

outputs from these components are combined and processed in the L1 Central Trigger, which generates an L1 Accept (L1A) signal. While the decision is being made at L1, the signals from the detectors wait in buffers in the detector readout systems until they receive an L1A signal.

When the detector readout receives L1A, the buffered data are released and passed through the Readout Drivers (RODs), which perform various levels of feature extraction and into the Readout System (ROS), where the data await readout by the HLT. At this stage, the HLT has access to regions of interest identified by the L1 system and full tracking information provided by the Fast Tracker (FTK) system introduced during the Phase-1 upgrade. With these inputs, it is able to run essentially offline algorithms to provide a final decision on whether to store the event within about 550 ms.

5.3.1 *The Level-1 Calorimeter Trigger*

As mentioned in Section 5.2.4.2, the LAr readout granularity has increased following the Phase-I upgrade. The goal was to provide more information to the trigger system in order to improve background rejection under the new high luminosity conditions of Run 3. The L1Calo system was heavily upgraded to utilize this additional infor-

mation. Figure 5.13 describes the architecture of L1Calo during Run 3.

Several new modules were introduced in the Phase-I upgrade. The Electron Feature Extractor (eFEX) identifies $e/\tau/\gamma$, the Jet Feature Extractor (jFEX) identifies jets, computes the E_T sum and the missing E_T per event, and provides additional information for τ identification. An additional module not appearing in Figure 5.13 is the Global Feature Extractor (gFEX), which identifies large area jets. In order to perform topological cuts on objects identified by the three previous modules as well as the muon triggering system, an additional L1Topo module, external to L1Calo, was introduced.

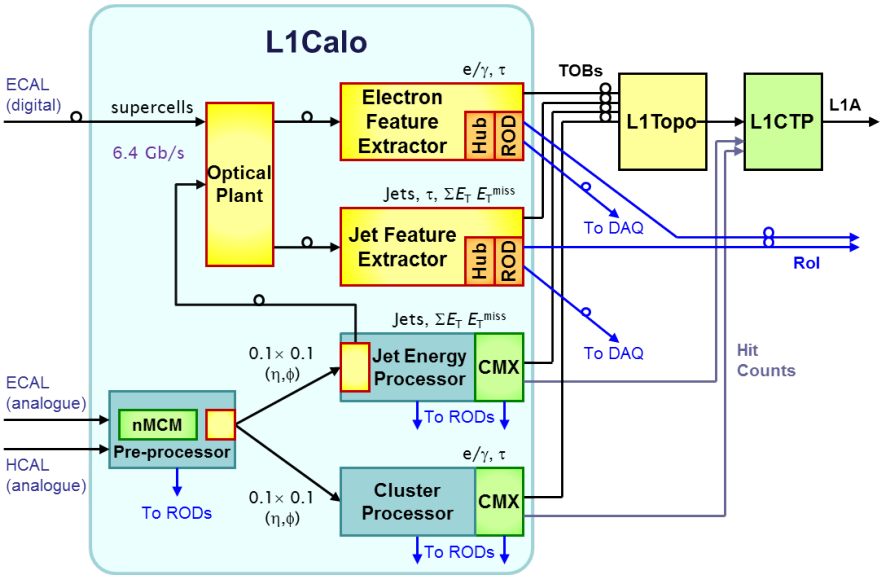


Figure 5.13: L1Calo system architecture during Run 3. Yellow components were added during the Phase-I upgrade. Components left over from Run 1 (Run 2), marked by dark (light) green, are being gradually decommissioned during Run 3.

6

DATA PROCESSING AND ANALYSIS

6.1 RECONSTRUCTION OF PHYSICS OBJECTS

Data acquisition from the ATLAS sub-detectors is followed by complex post-processing, an important part of which is the identification and reconstruction of physics objects used for physics analyses. Offline analysis should be distinguished from online triggering. Events selected by the latter are stored, and an offline process of reconstruction of physics objects is applied to them. This chapter will introduce the common physics objects and a quick overview of their reconstruction methods in ATLAS.

It is worth mentioning that despite this separation between online (trigger) and offline analysis, the HLT uses algorithms very similar to the offline ones to reconstruct the physics objects in real time for triggering.

6.1.1 *Auxiliary Objects*

The reconstruction of primary physics objects is preceded by the reconstruction of auxiliary objects used in the reconstruction of the former. These include clusters of energy in the calorimeters, tracks in the tracking detectors, and interaction and decay vertices.

6.1.1.1 *Topo-clusters*

A cluster needs to be defined to compute the energy depositions in the calorimeters. Topo-clustering [34] is an algorithm for combining calorimeter cells with significant energy depositions to form a cluster of energy with well-defined boundaries and energy content. The algorithm starts with finding a cell with an energy deposition significantly higher than the noise level resulting from electronic noise and pile-up (e.g., above 4σ). It then includes surrounding cells in layers containing progressively lower energies (e.g., 2σ followed by 0σ).

6.1.1.2 *Tracks*

The ATLAS tracking system reconstructs the curved trajectories of charged particles propagating through the ATLAS magnetic field in the form of tracks [35], which are crucial for the measurement of their momentum and charge as well as subsequent association with EM clusters and interaction vertices to form a complete picture of the physics object being reconstructed.

Tracks are reconstructed from energy depositions in the layers of the ID. A track is characterized by a reference point, which is usually the average position of the interaction point between the protons, the transverse and longitudinal *impact parameters* d_0 and z_0, which are the closest points to the reference point in the transverse and longitudinal planes, the polar θ and azimuthal ϕ angles and the ratio between the measured charge and momentum of the track.

6.1.1.3 *Primary Vertices*

The points at which interactions and decays happen are called *vertices*. The *primary vertex* (PV) is the position of the primary p-p interaction. The existence of pile-up makes it challenging to identify the PV since multiple vertices are present.

The reconstruction of the PV is described in [36] and involves an iterative process of associating tracks with a reference point at the interaction region, performing a fit, removing the less relevant tracks, and repeating. This results in a set of PV candidates for which the sum of the squared momenta of the associated tracks is computed, and the vertex with the highest value is selected.

6.1.2 *Physics Objects*

The full reconstruction of physics objects includes their reconstruction, identification, and calibration. The reconstructed objects used in analyses often come with a quality ranking in the form of *working points*, often denoted as *loose*, *medium*, and *tight*, corresponding to progressively tighter requirements on the thresholds of the various discriminants that are considered in their identification and reconstruction stages and resulting in a progressively purer sample of reconstructed objects.

6.1.3 *Electrons and Photons*

Reconstruction of electrons and photons [37] is done by summing the energy of multiple adjacent topo-clusters, known as a *supercluster* and, for electrons, including an associated track of high enough quality.

Electrons and photons are identified by energy clusters in the EM calorimeter and a lack thereof in the hadronic calorimeter. An electron can be distinguished from a photon by the existence of a track detected by the ID associated with its EM cluster. Semi-leptonic decays of hadrons, $\pi^0 \to \gamma\gamma$ decays, and hadronic jets form the backgrounds for electron and photon identification. Electrons' EM shower shape and track characteristics are used for identification. For photons, the shower shape and leakage of energy into the hadronic calorimeter are used as discriminants since non-prompt photons (i.e., those that are not produced directly in the primary p-p interaction) result in higher activity in the hadronic calorimeter and the main $\pi^0 \to \gamma\gamma$ background is characterized by two energy depositions in the first fine-grained layer of the EM calorimeter, affecting the shower shape.

A further discriminant that helps reject background is the isolation of energy depositions for electrons and photons in the calorimeters - the active region in the calorimeter should be small and the surrounding region around it should contain little energy. A similar requirement is applied to associated tracks in the case of electrons.

6.1.4 *Muons*

Muons are reconstructed from tracks in the ID and the MS [38]. First, segments of tracks are formed from hits inside each muon chamber. Muon candidates are formed by fitting segments in different layers. A χ^2 fit is used, and tracks with a sufficiently low χ^2 value are kept and combined with ID tracks and information from the calorimeter.

Identification of prompt muons is performed to separate them from muons originating from π and K decays. This stage involves the requirement of isolation on the candidate muon tracks and considers variables describing muon track quality.

6.1.5 *Jets*

Jets are the most abundant objects produced in LHC collisions, forming the majority of background for physics analyses.

Partons produced in the initial p-p collision cannot exist in isolation due to color confinement and, therefore, hadronize as they propagate away from the interaction point. This forms an object that looks like a jet of particles around the direction in which the original parton was traveling. The object has clusters of energy in the calorimeter and tracks in the tracking detectors associated with it.

Jet formation generally happens in two stages, which can be factorized and treated separately, depending on the underlying physics. The initial stage immediately following the hard scattering interaction, where energetic partons constituting the colliding protons interact, is a parton shower, which is governed by perturbative QCD. At each stage of the shower, the probability of a parton emitting another parton increases with the decreasing energy of the radiated parton and its angle from the initial parton, resulting in a collimated shower of rapidly decreasing energy.

When the energy scale of the partons falls below that of around 1 GeV, the strong coupling constant α_s is too large for perturbative calculations. Here, a second, non-perturbative stage of hadronization begins, in which partons form color singlets in a process resulting in a final state of hadronic resonances, the most abundant of which are the light mesons - pions, kaons, and ρ mesons [39].

In physics analyses, jets are objects defined by the algorithm that reconstructs them, as the produced set of jets is dependent on the algorithm. Jet reconstruction in ATLAS is usually done using an iterative bottom-up calorimeter clustering algorithm called *anti-kt* [40]. The algorithm can accept as inputs either calorimeter topo-clusters, tracks (producing *track jets* as a result) or MC truth particles.

The anti-kt algorithm works by taking all topo clusters in an event and computing a measure of proximity between every pair of them and between everyone to the beamline with a parameter R scaling the ΔR distance between two topo-clusters and controlling the cone angle of the resulting jets. A minimum on this set of proximity measures is computed, and if it belongs to a pair of clusters, they are merged together into a larger cluster. If the minimum is between a cluster and the beamline, the cluster is taken as a separate jet, and the proximity measure is removed from the set. This proceeds iteratively until the set is empty. The result is a collection of clusters in the calorimeter that are clustered around the hardest object and which typically have a symmetric shape around it, as can be seen in Figure 6.1.

A recent upgrade to the jet reconstruction process known as the *particle flow (PFlow)* algorithm [41] also considers track information.

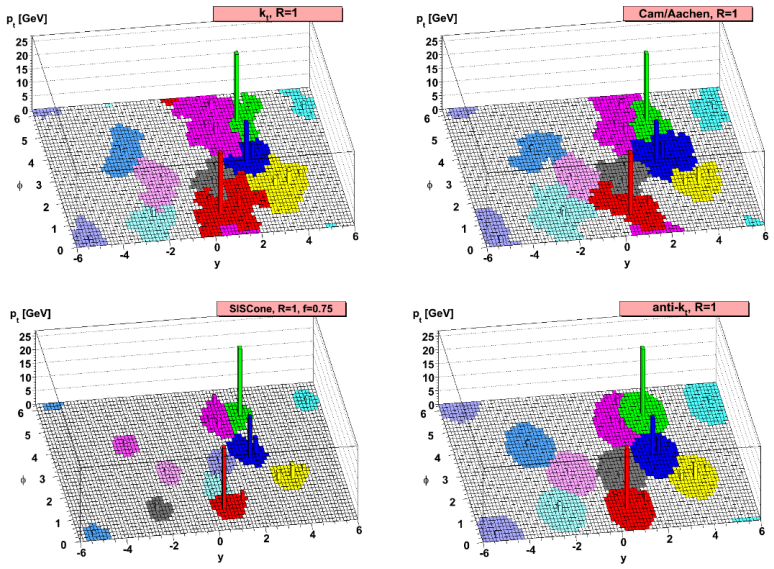


Figure 6.1: Sets of jets from applying different clustering algorithms in a single simulated event [40]. The clustering is done on stable truth particles in MC and not on calorimeter depositions.

It allows, e.g., to mitigate pile-up effects by removing calorimeter clusters associated with tracks from pile-up vertices and improving energy resolution by including soft charged particles swept away by the magnetic field.

The production of energy clusters by hadronic showers does not account for all the energy of the jet because of undetectable energy emitted during its formation. Additionally, the effects of pile-up reduce the resolution of the measured jet observables. A multi-stage calibration procedure is therefore necessary to correct the raw observables. For example, the pile-up contribution to the jet energy is corrected, and the jet four vectors are corrected to point to the primary vertex. An additional stage in the processing of jets is the removal of entire jets that likely originate from pile-up interactions.

6.1.5.1 *Flavour Tagging*

A crucial part of jet reconstruction for physics analyses is understanding what is the flavour of the parton initiating the jet. This process is referred to as *flavour tagging*.

In ATLAS, the three distinguishable categories of QCD jets are *light jets*, originating from up/down quarks or gluons, *b-jets*, and, recently, *c-jets*. The distinction between the three is possible by considering

features of the tracks and vertices associated with the jet and is based on the decay properties of the hadrons formed by these particles.

A b-hadron typically produces a detectable secondary decay vertex due to its relatively long lifetime ($\langle c\tau \rangle = 450\ \mu$m in the rest frame) [42]. The b-hadron typically carries about 70% of the jet momentum and has a high decay product multiplicity, which includes a charged lepton in around 20% of the time. Due to b-quarks decaying primarily to c-quarks as a result of the suppression from the CKM matrix elements ($|V_{cb}|^2 \gg |V_{ub}|^2$), a tertiary vertex inside the jet cone is an additional identifying sign of a b-jet. A c-hadron has characteristics similar to the b-hadron, except for having only a single displaced vertex and an intermediate decay multiplicity at the secondary vertex.

The flavour tagging strategy developed by the ATLAS collaboration is based on a two-stage approach. First, *low-level algorithms* reconstruct the features of the tracks and vertices of the heavy-flavour (i.e., b- and c-) jets and then *high-level algorithms* combine these features via multivariate methods into jet flavour classifiers. The analysis of Run 2 data uses a new generation of some of those algorithms based on recurrent and deep neural networks, which results in considerable improvements in flavour tagging performance [43].

6.1.6 *Missing Transverse Energy*

Neutrinos are particles appearing in the final states of many decays. However, they are undetectable as they interact with matter only weakly and are thus practically invisible to the detector. Various BSM particles, such as dark matter candidates, are also expected to pass undetected. Since the total transverse momentum in each collision must be zero, in order to identify such particles, one can sum all of the transverse momenta originating from the visible objects to estimate that of the invisible ones, denoted as E_T^{miss}.

The reconstruction of missing transverse energy (MET) is generally done by taking the negative of the vectorial sum of all the reconstructed physics objects and a *soft term*, which is the p_T of all the tracks associated with the PV but not associated to any of the reconstructed objects.

6.1.7 τ *Leptons*

The properties and decay modes of the τ lepton have been discussed in detail in Chapter 3. Since the leptonic decays of the τ are indistinguishable from prompt leptons, usually only hadronic τ decays are considered. The signature of these decays is one or three charged tracks due to the $\pi^{\pm}$ decay products associated with a narrow jet in the EM and hadronic calorimeters.

The reconstruction of taus starts with the anti-kt clustering algorithm with a small radius parameter of $R = 0.4$. The seeding jets must be energetic enough (typically $p_T > 10$ GeV) and located within $|\eta| < 2.5$ [44]. A recurrent neural network (RNN) is trained to separate the τ jets from QCD-originating jets based on information on the associated tracks and calorimeter clusters as well as high-level discriminating variables and is used to set a working point for τ-based analyses [45]. The latter is an improvement to a previous BDT-based discriminant.

6.2 CONSTRAINING PARAMETERS OF INTEREST IN A COLLIDER EXPERIMENT

To measure and constrain the value of a physics parameter from data collected in a particle collider experiment, one usually resorts to a counting experiment of a well-defined and sufficiently detectable process. One then parametrizes the expected result using parameters of interest (POIs), the value and uncertainty we're interested in obtaining, and nuisance parameters resulting from uncertainties in our assumptions, which, along with the POIs, must be determined from measurements. To perform the parametrization, the parameters must be set in a model describing the data being measured, which is the SM for particle physics experiments. From the latter, the expected result can be reproduced, and the parameters can, e.g., be constrained by computing the range of their values that would likely produce the observed result.

6.2.1 *Monte-Carlo Experiments*

To obtain the expected result of our counting experiment from the SM, it is not possible to use a mathematical derivation directly from the SM since the experimental setup is vastly complex. Starting from the predictions of the SM about the probabilities to produce given final states from initial states with basic given quantities such as momenta and spin, it is necessary to pass the resulting simulated particles through a computer simulation of the material surrounding the collision point and all detectors, their readout electronics and subsequent processing of the signals and data. The response of the simulated detector and the processing stages must be identical as much as possible to that of the real detector at the time of the acquisition of the analyzed data. This stage is critical for any particle physics experiment and is called a Monte-Carlo (MC) simulation.

An important concept in MC is the *truth information*. Since the event is simulated via careful bookkeeping, one knows exactly which particles were reconstructed into which final physics objects, a feature crucial for performing preliminary studies.

MC simulations are done in multiple stages. Generally, the particle interactions are simulated first, followed by the simulation of the detector response. The particle interactions are simulated starting from the probability distribution functions of the partons. Then, focusing

on a specific process, the hard scattering interaction is simulated to a fixed order (LO, NLO, NNLO, etc.), followed by the process of parton showering, hadronization, and hadronic decays to final states. Multiple MC generators specialize in the different stages of particle interactions, which are often interfaced together to produce a simulation of a given process. Each generator has a set of parameters called a *tune*.

MC-generated samples will have a *MC event weight* assigned to each event, which originates from the fact that the differential cross section with respect to phase space configuration for a particular process is not constant. When unweighted, a sample generated by a MC generator for any particular process generally does not have a natural distribution of that process like it would be found in collision events. This characteristic is restored when the MC event weights are applied.

The individual MC event weights must be further multiplied by additional weights to represent the data. These include weights that scale the number of MC events to the actual luminosity of the data and weights that are applied to correct for changes in yields due to pileup.

MC generators will introduce systematic uncertainties originating from imperfections of the assumed model. These can be estimated by varying the configuration of the generators or using different generators to produce *nominal* and *alternative* MC samples, with the former serving as the baseline sample and the latter producing a variation to estimate these uncertainties.

6.2.2 *Statistical Formalism*

In a collider experiment, we typically measure an observable that is selected to be sensitive to the signal we're interested in. For example, when searching for a resonance that decays into two particles, a common observable choice is the invariant mass of the two candidate final state particles since, in the presence of the resonance, the distribution of this observable is expected to be affected around the rest mass of the resonance as opposed to its absence. Another common choice is a multi-variate discriminant, usually a score from a machine learning model trained on MC samples to discriminate signal and background events, resulting in a clearly visible signal peak.

Once measured, the observable is binned, and at this stage, the parametrization is introduced. Each bin is treated as a counting experi-

ment and is therefore Poisson distributed with a mean of the expected number of background events and the SM expected number of signal events scaled by the *signal strength* μ, which is usually our POI. At this stage, we can define the likelihood function, parametrizing in terms of the POI, the probability of getting the measured result assuming a value of μ:

$$\mathcal{L}(\mu) = \prod_{i \in \text{bins}} \text{Pois}\left(N_i | b_i + \mu s_i\right) = \prod_{i \in \text{bins}} \frac{\left(b_i + \mu s_i\right)^{N_i}}{N_i!} e^{-(b_i + \mu s_i)}, \tag{6.1}$$

where N_i is the measured number of events in bin i and $b_i(s_i)$ are the number of background (signal) events in bin i expected from the SM, estimated from MC samples.

To take into account systematic uncertainties, they are introduced as additional parameters $\vec{\theta}$ called *nuisance parameters*. These parameters affect s_i and b_i, so that $\{s_i, b_i\} \to \left\{s_i(\vec{\theta}), b_i(\vec{\theta})\right\}$. This effect is studied in auxiliary measurements, and the parameters are multiplied into the likelihood function as Gaussian terms with zero mean and unit variance so that a value of a nuisance parameter that is far from its nominal pulls the likelihood value down:

$$\mathcal{L}_{\text{Aux}} = \prod_{\theta \in \vec{\theta}} \frac{1}{\sqrt{2\pi}} e^{-\theta^2/2}. \tag{6.2}$$

Not all nuisance parameters are constrained, and some are allowed to take any value to be determined from the final fit to the data and are not included in the above constraint. These parameters are called *free floating*.

Statistical uncertainty effects due to low simulation statistics for the background are parametrized by parameters γ_i, one per each bin, which scale the background yields ($b_i(\vec{\theta}) \to \gamma_i b_i(\vec{\theta})$):

$$\mathcal{L}_{\text{BkgStat}}(\vec{\gamma}) = \prod_{i \in \text{bins}} \text{Gauss}\left(\beta_i | \gamma_i \beta_i, \sqrt{\gamma_i \beta_i}\right), \tag{6.3}$$

with $\beta_i = 1/\sigma_{\text{rel}}^2$, where σ_{rel} is the relative statistical uncertainty on the expected total background yield.

The final likelihood function is obtained by multiplying 6.1, 6.2 and 6.3:

$$\mathcal{L}(\mu, \vec{\theta}, \vec{\gamma}, \vec{\tau}) = \prod_{i \in \text{bins}} \frac{\left(b_i(\vec{\theta}, \vec{\gamma}, \vec{\tau}) + \mu s_i(\vec{\theta})\right)^{N_i}}{N_i!} e^{-(b_i + \mu s_i)} \times \prod_{\theta \in \vec{\theta}} \frac{1}{\sqrt{2\pi}} e^{-\theta^2/2} \times \prod_{i \in \text{bins}} \text{Gauss}\left(\beta_i | \gamma_i \beta_i, \sqrt{\gamma_i \beta_i}\right), \tag{6.4}$$

where the $\vec{\tau}$ parameters are the free-floating ones. We can promote the POI μ to a vector $\vec{\mu}$ since we are often measuring the signal strength in multiple regions, which results in multiple signal strengths, one per region.

Once the likelihood function is defined and the data is measured, the POIs and the nuisance parameters can be constrained by fitting our model to the measured data, which is done by a maximization procedure for the likelihood function. At this stage, we'd like to test different hypotheses on the values of the POIs $\vec{\mu}$. To test a certain hypothesis, we define the *log likelihood ratio*:

$$\lambda(\vec{\mu}) = \frac{\mathcal{L}(\vec{\mu}, \hat{\hat{\vec{\theta}}}, \hat{\hat{\vec{\gamma}}})}{\mathcal{L}(\hat{\vec{\mu}}, \hat{\vec{\theta}}, \hat{\vec{\gamma}})}, \tag{6.5}$$

where $\hat{\hat{\vec{\theta}}}, \hat{\hat{\vec{\gamma}}}$ are the values that maximize the likelihood for a specific value of $\vec{\mu}$ and $\hat{\vec{\mu}}, \hat{\vec{\theta}}, \hat{\vec{\gamma}}$ maximize the likelihood over all the parameters. A value of $\lambda(\vec{\mu})$ close to one implies consistency of the $\vec{\mu}$ hypothesis with the data, while that close to zero implies inconsistency. A test statistic $q_{\vec{\mu}} = -2 \ln \lambda(\vec{\mu})$ is then defined, which, for values of $\vec{\mu}$ inconsistent with the data, would take high values. Thus, for a given measurement, we can compute the value of q_μ and ask what is the probability of obtaining in measurement a value at least as extreme, assuming the tested hypothesis, by computing the *p-value*:

$$p_{\vec{\mu}} = \int_{q_{\vec{\mu},\text{obs.}}}^{\infty} f(q_{\vec{\mu}} | \vec{\mu}) dq_{\vec{\mu}}, \tag{6.6}$$

where $q_{\vec{\mu},\text{obs.}}$ is the observed value of $q_{\vec{\mu}}$ and $f(q_{\vec{\mu}}|\vec{\mu})$ is the probability distribution function of $q_{\vec{\mu}}$ assuming a signal strength of $\vec{\mu}$.

From the p-value, the significance Z of the observation can be computed, which is related to it by

$$Z = \Phi^{-1}(1 - p_{\vec{\mu}}). \tag{6.7}$$

As an example, the Value of $\vec{\mu} = 0$ implies that we're testing a hypothesis of background only. A value of $p_{\vec{\mu}=0} = 2.87 \times 10^{-7}$ would

imply that the observation has a significance of 5σ relative to the background-only hypothesis, which is the accepted standard for declaring an observation in high energy physics.

The analysis is often not sensitive to specific regions of $\vec{\mu}$ values. In that case, a sufficiently large statistical fluctuation in the data could lead to a rejection of the $\vec{\mu}$ hypotheses in these regions. To mitigate that, instead of the bare p-value, a common metric used in LHC experiments to exclude a hypothesis is CL_s [46], defined as

$$CL_s = \frac{p_\mu}{1 - p_0}, \tag{6.8}$$

where p_μ (p_0) is the p-value assuming the μ ($\mu = 0$) hypothesis.

The power of CL_s vs the regular p-value for the exclusion of a given hypothesis may be illustrated by an example. If we assume that a certain μ hypothesis is true but the analysis is not sensitive to it, it means that the distributions of q_μ and q_0 strongly overlap. If there is a large upward statistical fluctuation in the data, it will result in a high observed value $q_{\mu,\text{obs.}}$. If we only consider p_μ for exclusion, we would wrongly exclude the true hypothesis. However, in the case of CL_s, and assuming the center of mass of the q_0 distribution is situated to the right of that of q_μ, p_0 would be large, thus increasing the value of CL_s and "softening" the rejection of such regions.

Part III

VHBB(CC) ANALYSIS

7

VHBB(CC) LEGACY ANALYSIS OVERVIEW

7.1 INTRODUCTION

Since the discovery of a particle consistent with the SM Higgs boson in 2012 at the LHC, measurements have been performed to study its physical properties. Perhaps the most important feature of the Higgs boson is its coupling to massive particles, which is expected to depend on the mass of the particle. Therefore, measurements of the Higgs' couplings to different particles play a very important role at the LHC.

Decays of the Higgs boson to τ leptons [47] and to a pair of b quarks [48] were recently observed at the LHC, as well as a decay to a pair of muons with a significance of 3σ [49]. A summary plot of the coupling strengths of the Higgs to various fundamental particles is given on Figure 7.1. A similar observation into lighter quarks is challenging due to the Higgs' coupling to fermions being proportional to their mass, resulting in fewer data for lighter quarks, the significant amount of background from the quark- and gluon-initiated jets in hadron colliders, and challenges in jet flavour identification.

As discussed in Section 2.2, the Higgs boson's production modes with the highest cross-section are ggF and VBF. However, the signatures of these modes do not allow for easy separation of those processes from the abundant QCD jet background of quark- and gluon-initiated jets, as the Higgs boson is either produced alone or in association with jets. The next mode with the highest cross section is the VH production mode. The existence of a vector boson can be used to suppress the background of QCD jets by selecting events with leptons from the leptonic decay modes of the Z and W bosons, which decay leptonically approximately 30% and 33% of the time, respectively [12]. The decay of the Higgs to two b- or c-quarks may thus be observed, and the relevant couplings may be measured and constrained.

The observation of the Higgs boson produced in association with a vector boson and decaying to two quarks is done differently if the Higgs is of relatively low energy as opposed to when it's energetic, since for low-energy Higgs bosons, the resulting jets can be resolved by the detector. At the same time, in the latter case, they are observed as

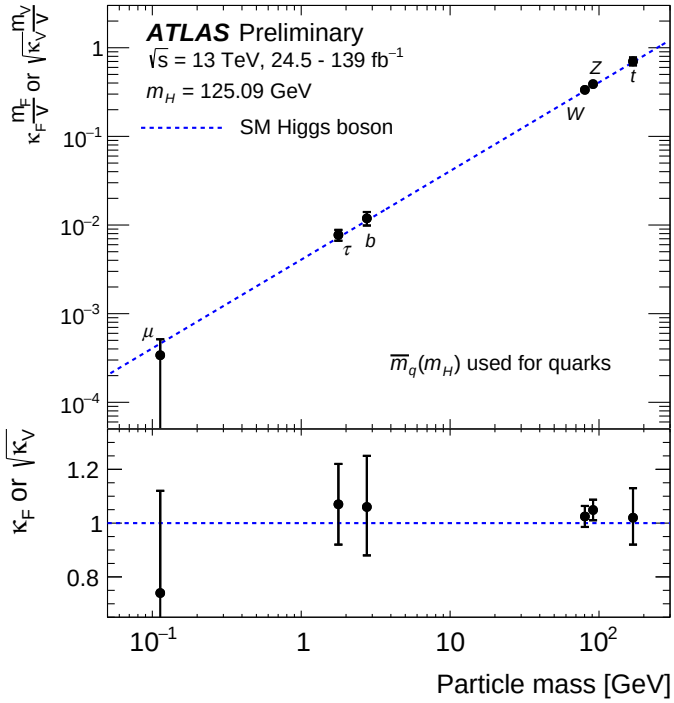


Figure 7.1: A plot summarizing the measured coupling strengths of the Higgs to various fundamental particles. κ is the ratio between the coupling measured to that of the SM. Since the couplings go as the square of the mass for gauge bosons and as the mass for fermions (see eq. (2.9)), there is a square root over v and κ_V for the vector bosons.

being merged into a single large-radius jet due to the highly boosted two-body decay of the Higgs. These observations and subsequent measurements of cross-section times branching ratio of the Higgs decaying to two b-quarks in association with a leptonically decaying vector boson have recently been performed in these two regimes and the corresponding analyses are referred to as $VH(\to b\bar{b})$ *resolved* [48] and $VH(\to b\bar{b})$ *boosted* [50], respectively. The $VH(\to b\bar{b})$ resolved analysis produced an observation of 6.7σ for this process [48].

An analysis that sets an upper limit on the signal strength of a Higgs decaying to two c-quarks in association with a leptonically decaying vector boson, referred to as $VH(\to c\bar{c})$, has also recently been published and produces an upper limit of 26 times the SM prediction for the production rate of this process [6].

The $VH(\to b\bar{b})$ resolved and boosted analyses have an overlapping region of energy at $p_T^V > 250$ GeV, with the scale of an event's energy

determined by p_T^V, the p_T of the associated vector boson. In this region, the efficiency of the Higgs reconstruction is low for the resolved analysis and high for the boosted one. Merging these analyses would increase Higgs reconstruction efficiency and better measure the Higgs coupling to b-quarks around $p_T^V = 250$ GeV. Additionally, the tagging of jets in the $VH(\to b\bar{b})$ and $VH(\to c\bar{c})$ analyses is done by assigning a probability for a jet being a b- or a c-jet. Therefore, merging the $VH(\to b\bar{b})$ analyses and the $VH(\to c\bar{c})$ one can increase the efficiency of reconstruction of the Higgs from b- and c-jets by, e.g., including jets considered for the $VH(\to c\bar{c})$ signal as events with low weight in the $VH(\to b\bar{b})$ analysis and vice-versa. In contrast, at the moment, the $VH(\to b\bar{b})$ and $VH(\to c\bar{c})$ analyses have no overlap at all in the data that passes their signal selections. It is, therefore, natural to merge these three separate analyses together into a single one, referred to as the *VHbb(cc) Legacy analysis*.

In addition to the aforementioned benefits of merging the three analyses, the VHbb(cc) Legacy analysis benefits from the harmonization of the analysis framework code and from tools used in the constituent analyses. For example, the $VH(\to c\bar{c})$ analysis was performed only in the resolved regime and relied on a fit to the reconstructed invariant mass of the Higgs candidate. The $VH(\to b\bar{b})$ boosted analysis relied on a similar fit. The $VH(\to b\bar{b})$ resolved analysis, on the other hand, is used to fit the score of a BDT model trained to classify signal and background events, and thus, its BDT framework can be used to boost the performance of the other analyses. Similarly, it is much easier to extend $VH(\to c\bar{c})$ into the boosted regime using the framework developed for the $VH(\to b\bar{b})$ boosted analysis.

Various enhancements are likewise introduced into the combined analysis, such as improvements in algorithms for reconstructing hadronic jets and moving from the older, BDT-based, MV2 jet flavor tagging algorithm to DL1r, a newer, deep neural network-based one. Additionally, a technique was utilized based on GNNs for assigning weights to events in order to be included in different analysis regions and enhance their MC statistics, referred to as *truth tagging* and briefly described in Section 7.6.2.

An additional motivation to merge the analyses is their similarity - each of the three analyses was done in the same three channels - $ZH \to \nu\nu q\bar{q}$, $WH \to \nu l q\bar{q}$ and $ZH \to l l q\bar{q}$, where q is b for $VH(\to b\bar{b})$ and c for $VH(\to c\bar{c})$. In the first and last ones, the associated Z boson decays to invisible neutrinos or two charged leptons, respectively, and in the second one, the associated W boson decays to one charged

lepton and a neutrino. The Higgs decays to two b- or c-quarks in all channels.

The VHbb(cc) Legacy analysis is naturally very complex, lasting several years and involving many participants. As a result, it will not be possible to describe all of its aspects in detail, and in those described in this chapter, the author was one of multiple contributors. In this section, the focus is mainly on providing an overview of the analysis methodology and describing the components relevant to the author's contributions to the analysis, part of which was technical in nature - improving the analysis infrastructure in terms of performance and usability. In contrast, the other part was the evaluation of BDT training on physics-motivated input variables for a final statistical fit. An additional contribution is a preliminary assessment of using a GNN-based classifier instead of BDT, described in Section 8.1.4.

7.2 SIMPLIFIED TEMPLATE CROSS SECTIONS

During Run 1, the main results of the Higgs boson coupling measurements were signal strength and multiplicative coupling modifiers. As more data becomes available, it is possible to measure the Higgs production and decay cross sections in a more granular fashion. Simplified template cross sections (STXS) [51] have been adopted by the LHC experiments as an evolution of the Run 1 measurements. They provide more fine-grained measurements of physical cross sections for individual Higgs production modes in mutually exclusive kinematic regions, known as *STXS bins*.

The different STXS bins are defined to be as close as possible to typical experimental kinematic selections or, more generally, the kinematic regions that dominate the experimental sensitivity. By looking at the *fiducial phase space*, which is a phase space restricted to a region defined by the experimental selection and the acceptance cuts of an analysis instead of the full phase space, the dependence on theory predictions is reduced, and so are the resulting theoretical uncertainties. The uncertainties due to extrapolations outside of the fiducial phase space are also reduced. An additional benefit of organizing measurements in STXS bins is that it facilitates combinations of the different production modes and phase space regions in combination analyses that aim to constrain BSM physics, e.g. [52] and [53].

As the amount of collected data increases, measurements in increasingly finer phase space bins become obtainable. With time and

experience, refinements to the bins are also expected to be introduced. The STXS binning is therefore defined in stages, with the latest stage at the time of writing being 1.2. For cross-section measurements of the Higgs produced in association with a vector boson, which is considered in the current analysis, the STXS bins are shown in Figure 7.2. Since there is currently not enough sensitivity to distinguish between the $qq \to ZH$ and $gg \to ZH$ processes, they are treated as a single ZH process.

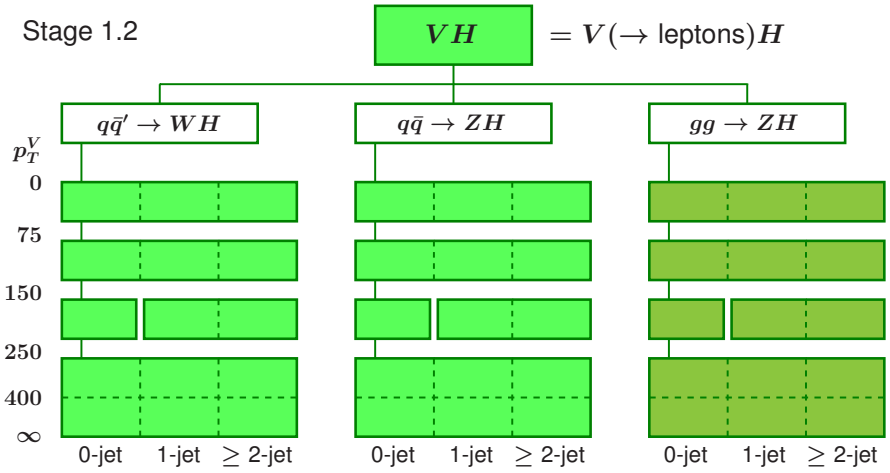


Figure 7.2: Definition of stage 1.2 STXS bins for a Higgs boson produced in association with a vector boson.

7.3 ANALYSIS STRATEGY

The goal of the analysis is a coherent measurement of $VH \to (b\bar{b})$ and $VH \to (c\bar{c})$ decays, utilizing the leptonic decay modes of the associated vector boson. The three possible leptonic decay channels for this process are illustrated in Figure 7.3. Each channel is characterized by the vector boson decay products, producing different detector signatures and, therefore, must be treated individually. The channels are further subdivided into regions of the transverse momentum of the vector boson p_T^V and the number of jets in the final state, allowing reporting of the results in STXS bins.

Three regimes are combined together in this analysis, characterized by the reconstruction technique of the Higgs. The $VH \to (b\bar{b})$ configuration, which is characterized by the presence of two b-jets, is sub-divided into a *resolved* regime, in which the Higgs has relatively low energy, resulting in well-resolved b-jets, and a *boosted* regime, in which an energetic Higgs produces two collimated b-jets resulting in

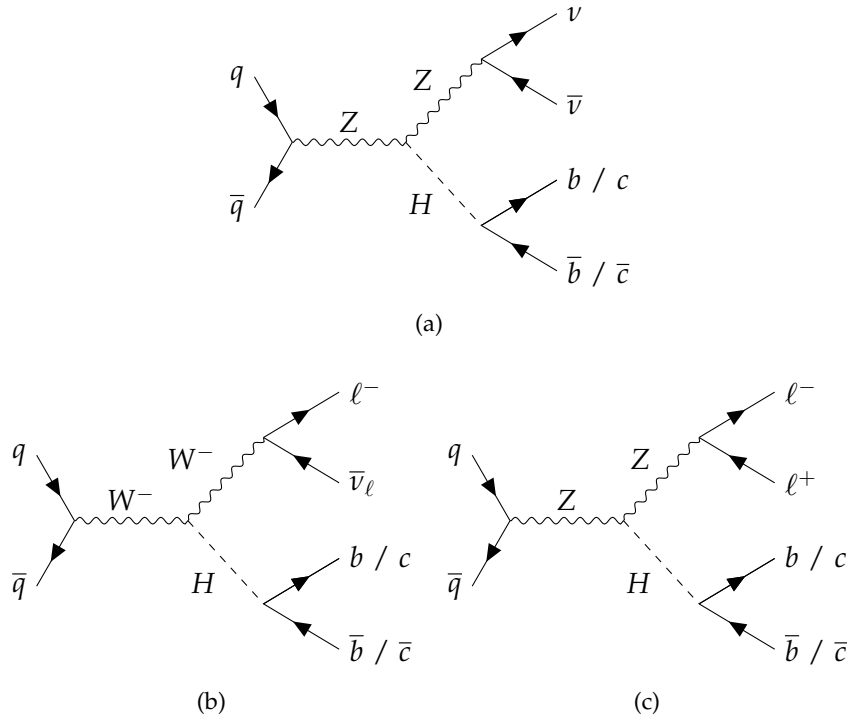


Figure 7.3: Tree level Feynman diagram of the VH production and the three Higgs decay channels targeted by the VHbb(cc) Legacy analysis- (a) 0-lepton, (b) 1-lepton and (c) 2-lepton.

a difficulty to resolve the two. The cut between these two regimes is done at $p_T^V = 400$ GeV. The third regime is $VH \to (c\bar{c})$, in which the Higgs decays to two c-jets and is characterized by the presence of c-jets. These regimes are further categorized by p_T^V, guided by the STXS bins for the VH process shown in Figure 7.2. The different regimes and categories are illustrated in Figure 7.4.

The main backgrounds in all analyses come from processes, the final products of which are indistinguishable from those of the signal processes. In this analysis, these are the production of a vector boson in association with jets (V + jets), $t\bar{t}$, single-top production, and di-boson (e.g. WW). Monte-Carlo simulated samples are used to model the background and the signal $H \to b\bar{b}$, $H \to c\bar{c}$ processes.

7.4 OBJECT RECONSTRUCTION

The physics objects used by the VHbb(cc) Legacy analysis are electrons, muons, missing transverse energy, jets, and hadronic taus. The following sections contain details on their reconstruction strategy specific to the analysis.

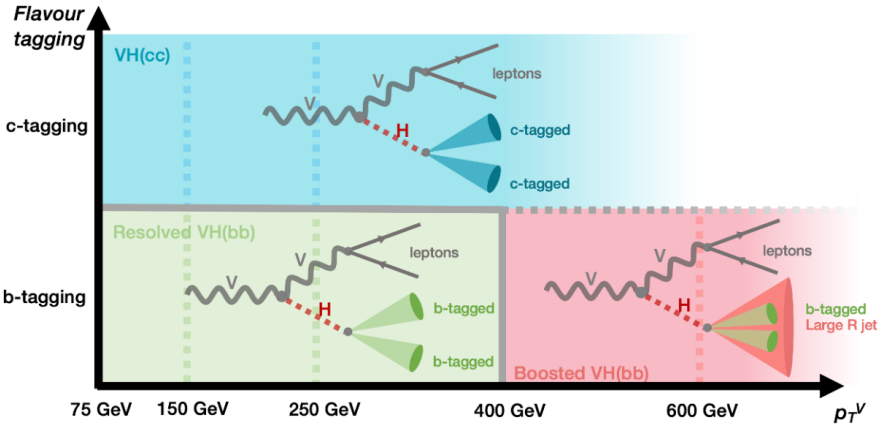


Figure 7.4: The VHbb(cc) Legacy analysis regimes and p_T^V categorization.

7.4.1 *Leptons*

Electrons and muons are the decay products of the associated vector boson in the 1- and 2-lepton channels. They are reconstructed in general as described in Sections 6.1.3 and 6.1.4, respectively. In the analysis, three selection criteria are defined for the electrons, *VH-Loose*, *ZH-Signal* and *WH-Signal*, summarized in Table 7.1. *VH*-Loose is defined to allow for the maximum electron selection efficiency for signal processes and is used as the default selection for electrons in all relevant channels. For the 2-lepton channel, two *VH*-Loose electrons are required with one of them satisfying the *ZH*-Signal criterion. In the 1-lepton channel, the tighter *WH*-Signal criterion is required to suppress the multi-jet background, which is more abundant than in the 2-lepton channel.

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

Table 7.1: Electron selection criteria. The ID is an electron identification criterion based on likelihood providing the two LH Loose and Tight working points, isolation is the isolation working point with HighPtCaloOnly being tighter. The d_0^{sig} and $\lvert\Delta z_0 \sin\theta\rvert$ columns are constraints on the transverse and longitudinal impact parameters of the electron tracks.

7.4.2 *Hadronic Taus*

Hadronically decaying taus are used as a veto to suppress $t\bar{t}$ background in the 1-lepton channel where both top quarks decay to a W boson and a b-quark and one of the W bosons decays to a hadronically decaying τ, resulting in one lepton, two b-jets and an additional jet in the final state corresponding to a 1-lepton channel signature. Hadronic taus are also identified in the 0-lepton channel to retain $W(\rightarrow \tau\nu)H$ decays where the τ decays hadronically and can be identified as a τ jet, using the latest RNN-based τ vs. QCD jet classifier mentioned in Section 6.1.7, bringing almost 5% signal gain.

The taus considered in this analysis are required to have $p_T > 20$ GeV, $|\eta| < 2.5$, excluding the calorimeter crack region $1.37 < |\eta| < 1.52$, one or three tracks corresponding to one- or three-pronged taus and a loose RNN identification criterion.

Future iterations of this analysis following the conclusion of Run 3 will use data containing taus identified using the author's τ algorithm at the L1 trigger, described in the last part of this thesis, bringing further gains via hadronic taus.

7.4.3 *Jets*

Three types of jets reconstructed with the anti-kt algorithm are considered in this analysis.

- *Small-R jets* - jets based on particle flow information with anti-kt parameter $R = 0.4$ [40]. Used to reconstruct the Higgs candidate for $VH(\rightarrow c\bar{c})$ and resolved $VH(\rightarrow b\bar{b})$ and also in the event categorization of boosted $VH(\rightarrow b\bar{b})$ to improve the signal sensitivity. Two types of small-R jets are considered - *signal jets* and *forward jets*. Only the signal jets are used to reconstruct the Higgs candidate, while the forward jets are also included when counting jet multiplicity in events. Table 7.2 summarizes the selection criteria for the two types.

- *Large-R jets* - jets with $R = 1$. Used to reconstruct the Higgs candidate for boosted $VH(\rightarrow b\bar{b})$. They are required to have $p_T > 150$ GeV and $|\eta| < 2.0$.

- *Variable-R track jets* - track jets reconstructed with a variable R parameter, depending on the jet p_T, $R(p_T) = \rho/p_T$ with maximum and minimum cutoffs $R_{\text{max}}, R_{\text{min}}$. These track jets are used to re-

construct the b-tagged objects inside the large-R jet in $VH(\to b\bar{b})$ and the parameters ρ, R_{max}, R_{min} are optimized to give the highest truth sub-jet double b-labelling efficiency [54] ($\rho = 30$ GeV, $R_{\text{max}} = 0.4, R_{\text{min}} = 0.02$). An overall requirement for VR track jets is to have $p_T > 10$ GeV and $|\eta| < 2.5$.

Jet Category	Tight Jet Cleaning	p_T	η	JVT/fJVT
Signal Jet	true	> 20 GeV	< 2.5	JVT>0.5 for $\|\eta\|$ <2.4, $p_T < 60$ GeV jets
Forward Jet	true	> 30 GeV	$2.5 < \|\eta\| < 4.5$	fJVT<0.5 for $p_T < 120$ GeV jets

Table 7.2: Small-R jet selection requirements. Jet cleaning removes events in regions corresponding to abnormally active calorimeter cells. Tight Jet Cleaning is applied to all channels to suppress non-collision background. To suppress jets arising from pileup, a TightJVT (Jet Vertex Tagger) and a LooseJVT is applied to signal or forward jets, respectively.

7.4.3.1 *Flavour Tagging*

Flavour tagging plays a crucial role in this analysis, as the signal final states are b- and c-jets. Distinguishing between these and light and τ jets plays an important role in improving the sensitivity to the signal. The high-level ATLAS flavour tagging algorithm DL1r [43] is used. It is implemented as a neural net that takes 31 inputs produced by the low-level flavour tagging algorithms and produces three outputs $p_{\text{light}}, p_c, p_b$ corresponding to the jet being a light, c or b-jet, respectively.

For b-tagging, the final DL1r discriminant is computed as

$$D_{\text{DL1r}}^{b\text{-tag}} = \ln\left(\frac{p_b}{f \cdot p_c + (1-f) \cdot p_{\text{light}}}\right). \tag{7.1}$$

The parameter f is selected to optimize the performance of the tagger. A high value will better discriminate b-jets against c-jets, while a low value will discriminate better against light jets. It should, therefore, correspond to the effective c-jet fraction in the background sample. Similarly, a $D_{\text{DL1r}}^{c\text{-tag}}$ discriminant for tagging c-jets can be defined. A working point is then defined by setting the value of f and a cut on $D_{\text{DL1r}}^{b\text{-tag}}$ and $D_{\text{DL1r}}^{c\text{-tag}}$. Working points are denoted by the efficiency of tagging the respective flavored jet by its true flavor.

In the resolved $VH(\to b\bar{b})$ and $VH(\to c\bar{c})$ regimes, since both b- and c-jets are considered, a scheme called *Pseudo Continuous Flavor*

Tagging (PCFT) is utilized. The DL1r discriminants for b-tagging and c-tagging are viewed in two dimensions, and the score is segmented in the way shown in Figure 7.5. Two b-tagging working points and two c-tagging ones are defined, dividing the plane into 5 exclusive bins. For b-tagging, the working points are 70% and 60%, and for c-tagging, they are 40% and 20% for loose and tight settings, respectively. These values stand for the efficiency of selecting true jets of the respective flavor. The bins are numbered from 0 to 4 and jets are classified into them according to their DL1r scores for b- and c-tagging. The distinction between the two b-tagging bins is only used for the multivariate analysis (MVA) training described in Section 8. The cut values of the 60% and 70% working points follow the official ATLAS recommendation from the FTAG group responsible for flavour tagging. In contrast, the 20% and 40% working points are optimized as part of the analysis to maximize the sensitivity to the $VH(\rightarrow c\bar{c})$ signal strength.

In the boosted $VH(\rightarrow b\bar{b})$ regime, only b-tagging is targeted. Here the *Pseudo Continuous b-Tagging* (PCBT) scheme is applied, where the $D_{\text{DL1r}}^{b\text{-tag}}$ discriminant of the tagged jet is binned into 4 bins corresponding to b-jet tagging efficiency of 60%, 70%, 77% and 85%. Two sub-jets inside the large R jet resulting from the boosted Higgs decay are subject to the tagging, and to maximize the signal yield, the 85% WP is chosen for this regime.

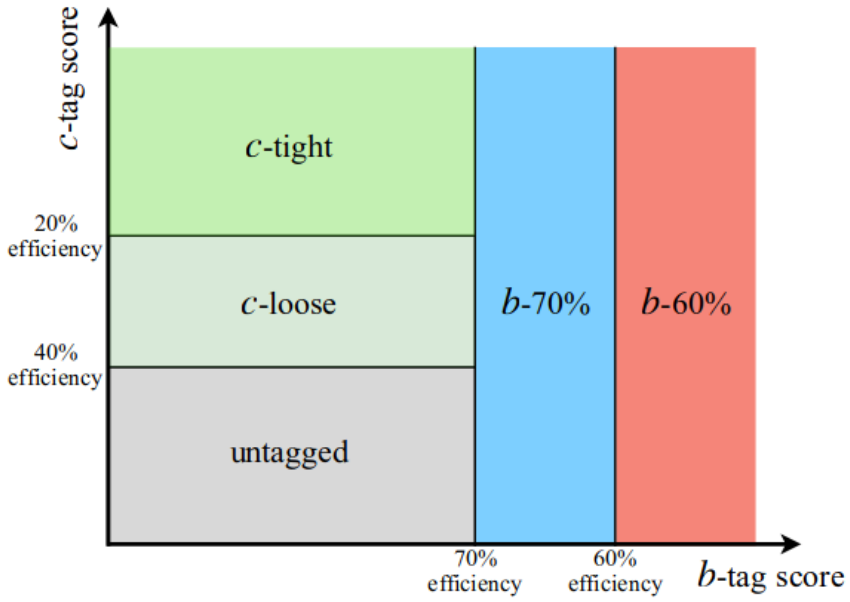


Figure 7.5: The DL1r-based PCFT bins.

7.4.4 *Missing Transverse Energy*

The presence of missing E_T (E_T^{miss}) is used as a sign of neutrinos in the 0- and 1-lepton channels in all analysis regimes. It is also used in the 2-lepton channel in MVA studies to reduce $t\bar{t}$ contamination. The general reconstruction procedure is described in 6.1.6.

7.5 EVENT SELECTION

All selected events in the $VH(\to b\bar{b})$ and $VH(\to c\bar{c})$ categories are required to pass a trigger criterion that is harmonized between the two but different for the different lepton channels. For the 0-lepton channel, the event must have been triggered by a E_T^{miss} trigger. For the 1-lepton channel, the trigger is required to be a low-p_T lepton for events with $p_T^V < 150$ GeV and the same E_T^{miss} trigger as in the 0-lepton channel for $p_T^V > 150$ GeV. For the 2-lepton channel, the trigger requirements are the same except that the p_T^V threshold is 250 GeV. The E_T^{miss}-based triggering was shown to increase the signal acceptance by around 5% in internal studies.

To classify into lepton channels, leptons defined in Section 7.4.1 are used. The 0-lepton channel contains events with zero VH-loose leptons, The 1-lepton channel contains those that have exactly one WH-signal lepton and no additional VH-loose ones, and events with exactly two VH-loose leptons are considered for the 2-lepton channel. Other combinations of leptons are not considered in the analysis.

In the 0-lepton channel, the signal is expected to be $ZH \to \nu\nu b\bar{b}$ or $ZH \to \nu\nu c\bar{c}$, where the neutrinos appear as E_T^{miss} and therefore events are required to have $E_T^{\text{miss}} > 150$ GeV. Additional cuts are applied to the sum of the p_T of all jets in the event. One issue in this channel is the contamination of 1-lepton events from $WH(\to b\bar{b})$ and $WH(\to c\bar{c})$ where the W decays to a τ lepton and a neutrino and the τ decays hadronically, which results in a signature consistent with the 0-lepton channel. Therefore, in events with a reconstructed hadronic τ, the W boson mass m_T^W is reconstructed and is used as a discriminant to exclude them from this channel.

In the 1-lepton channel, the signal is expected to be $WH \to l\nu b\bar{b}$ or $WH \to l\nu c\bar{c}$, where the transverse momentum of the W boson is reconstructed from the vectorial sum of the lepton and E_T^{miss} and is required to be above 75 GeV.

In the 2-lepton channel, the signal is expected to be $ZH \to llb\bar{b}$ and $ZH \to llc\bar{c}$. Here the Z boson can be reconstructed from the two leptons, which are required to have the same flavour. Due to the high charge mis-identification rate for electrons, a requirement of opposite charge for the leptons is only imposed in the muon channel. A $p_T > 25$ GeV cut for electrons and 27 GeV for muons is required, and at least one of the leptons is required to pass the ZH-Signal criterion described in Section 7.4.1. Furthermore, the invariant mass of the di-lepton system is required to be consistent with the Z mass.

7.5.1 *Higgs Candidate in Resolved Regime*

The Higgs candidate is reconstructed from jets using the jet flavour tagging information. Ideally it can be reconstructed from two jets tagged as b- or c-jets, however, there are cases where there are more than two jets. Furthermore, due to limitations in flavour tagging and possible overlap between $VH(\to b\bar{b})$ and $VH(\to c\bar{c})$, a more complicated selection is required, which is done in two steps.

After tagging the jets in the event based on the PCFT scheme, The first step avoids overlap between $VH(\to b\bar{b})$ and $VH(\to c\bar{c})$. Only events without any b-tag pass the $VH(\to c\bar{c})$ selection, while only those without any tight c-tag pass the $VH(\to b\bar{b})$ one for the 0- and 1-lepton channels. For these channels, events that have both a tight c-tag and a b-tag will not enter any signal region. In $VH(\to b\bar{b})$, for the 2-lepton channel, a tight c-tag veto is not applied.

In the second step, all signal jets are considered. In the $VH(\to b\bar{b})$ regime, an event will enter the signal region only if it has exactly two b-tagged jets, which will then be used to reconstruct the Higgs candidate. In the $VH(\to c\bar{c})$ regime, jets are prioritized as tight c-tag, then loose c-tag, then no c-tag, and the leading two are selected to reconstruct the Higgs candidate.

7.5.2 *Analysis Regions*

The analysis includes multiple signal and control regions. The events falling in the signal regions are used to reconstruct the Higgs. In contrast, those that fall in the control regions are used to constrain uncertainties and understand distributions of observables of background processes for cases in which a good description of the process is not

attainable from MC simulations. Figure 7.6 summarizes all signal and control regions in the analysis.

7.5.2.1 *Resolved* $VH(\to b\bar{b})$

In the $VH(\to b\bar{b})$ resolved regime, signal regions require exactly two b-tagged jets and are defined in three different bins of p_T^V, 75 GeV $< p_T^V <$ 150 GeV (for 1- and 2-lepton channels only), 150 GeV $< p_T^V <$ 250 GeV and 250 GeV $< p_T^V <$ 400 GeV with separation between events with two and three jets. For the 2-lepton channel a four or more jet category is included and in the 0-lepton channel a category of exactly four jets is included, due to the high $t\bar{t}$ contamination at higher jet multiplicity.

For each signal region, an orthogonal control region in which the ΔR between the two Higgs candidate jets is high is defined, referred to as *high*-ΔR. These control regions are used to determine the yield of the V + jets and $t\bar{t}$ backgrounds from data. In the 1-lepton channel, an orthogonal *low*-ΔR control region is used to better control the contribution of the $W + bb$ background by training a BDT to separate the V + jets background from the dominant $t\bar{t}$ one, referred to as $\mathrm{BDT}_{\text{Low-}\Delta R\,\mathrm{CR}}$.

A top control region is defined in the 0- and 1-lepton channels by the requirement of one b-tag and one tight c-tag and in the 2-lepton channel by requiring different lepton flavors for the two leptons, namely e and μ.

7.5.2.2 $VH(\to c\bar{c})$

In this category, the division into channels and p_T^V regions is similar to the $VH(\to b\bar{b})$ resolved one, with the highest p_T^V bin being $p_T^V >$ 400 GeV, since the impact of the overlap with the $VH(\to b\bar{b})$ boosted region of $p_T^V > 400$ GeV is small. There are two sets of signal regions, both of which require at least one tight c-tag. One set requires the presence of an additional loose or tight c-tagged jet, and the other requires the presence of additional jets, which are neither b- nor c-tagged. Each signal region has a corresponding high-ΔR control region. Additional control regions especially relevant for $VH(\to c\bar{c})$ in order to control background originating from W and Z bosons associated with light jets are the V + light control regions, in which one of the Higgs-candidate jets are required to be loose c-tagged and one not tagged. This control region is excluded in the 1-lepton

75 GeV $< p_T^V <$ 150 GeV region due to fit instabilities. The top control region is shared with $VH(\to b\bar{b})$ resolved.

7.5.2.3 *Boosted* $VH(\to b\bar{b})$

In the boosted $VH(\to b\bar{b})$ regime, the events are categorized into two p_T^V regions, 400 GeV $< p_T^V <$ 600 GeV and $p_T^V >$ 600 GeV. The signal regions are defined to have exactly two of the three leading sub-jets of the large R jet to be b-tagged. In the 0- and 1-lepton channels, events with b-tagged track jets outside of the large-R jet are moved to dedicated $t\bar{t}$ control regions.

7.6 TRUTH TAGGING

After being selected, events are categorized according to the flavor and multiplicity of the jets. There are several approaches to achieving this. In *direct tagging*, jets in the event are assigned a PCFT bin, as described in 7.4.3.1, and the event is classified by counting the number of b-jets and c-jets in it. For example, $VH(\to b\bar{b})$ resolved requires to have events with exactly two b-tags and $VH(\to c\bar{c})$ requires at least one tight c-tag. However, the low efficiency of c-tagging results in regions with very low statistics and, therefore, high statistical uncertainty.

To mitigate that, another approach called *truth tagging* is used. Instead of a binary decision of either keeping or discarding an event, it is assigned a weight

$$w = \sum_{i}^{C} \prod_{j \in T_i} \epsilon_j(\mathbf{x}|\vec{\theta}) \prod_{j \in \not{T}_i} (1 - \epsilon_j(\mathbf{x}|\vec{\theta})) \tag{7.2}$$

where $\epsilon_j(\mathbf{x}|\vec{\theta})$ is the probability to tag jet j by its flavour using a tagger trained on features $\mathbf{x}$ assuming a set of parameter values $\vec{\theta}$, T_i is the set of jets assumed tagged and $\not{T}_i$ is the complement - the jets assumed not tagged. The sum goes over all possible permutations. This weight is the event's probability of passing a selection otherwise based on its jet flavor content and multiplicity. The probability for each jet is parametrized by the components of $\vec{\theta}$, which are its p_T and η. An illustrative example is the computation of the probability of tagging two jets in a 3-jet event. The permutations, then, are all combinations of two jets tagged out of a total of three, where the third is not tagged.

After the weight is computed, it is applied to the pair of jets that are used to reconstruct the Higgs candidate. For a given event, a single permutation of tagged and untagged jets out of all possible ones is

randomly chosen with a probability proportional to the corresponding term in the sum (7.2). This way, not all permutations are considered because the likelihood fit used in the analysis assumes no correlation between bins, and using multiple permutations for the same event would spread a single event into multiple bins, increasing correlation. As an example, if we compute an observable using two tagged jets out of a total of three, using two different permutations would yield two different values of the observable for the same event, potentially placing this event in two different bins, increasing their correlation.

Truth tagging can be done either using a map-based approach or a GNN-based one. Those are discussed in more detail below. The final tagging strategy in the VHbb(cc) Legacy analysis is a combination of direct and GNN-based truth tagging based on the sizes of MC samples, tagging efficiency in different regimes and analysis regions, and the closure of truth tagging to direct tagging.

7.6.1 *Map-Based Truth Tagging*

The most straightforward way to parametrize the efficiency of a jet to be correctly tagged by a tagger is to use a *map-based* approach in which one considers the most dominant parameters contributing to the tagging efficiency, which are, as already mentioned, the η and p_T of the jet. Two-dimensional maps of the efficiency as a function of these parameters are constructed using MC samples and are then used to compute the efficiency for data. This approach was used in previous $VH(\rightarrow b\bar{b})$ [55] and $VH(\rightarrow c\bar{c})$ [6] analyses.

The map-based approach has its drawbacks. The tagging efficiency depends not only on η and p_T, but on other parameters, however, it is not practical to create multi-dimensional maps because of "the curse of dimensionality" - the statistics per bin would be too low for maps of dimensionality higher than two. Another drawback is that nearby jets in the vicinity of the analyzed jet affect its tagging efficiency. However, the map approach does not take the environment into consideration.

7.6.2 *GNN-Based Truth Tagging*

An approach that was pioneered in this analysis is a machine learning, *GNN-based* approach. A GNN is a perfect tool to address the two drawbacks of the map-based approach. It can handle high dimensionality

problems as well as consider the environment of the jets due to its internal message-passing capabilities, as described in Section 4.2.1.

The input to the network is a fully connected graph representing the jet state of the event. Each node in the GNN represents a track jet with 8 features based on its kinematics and a single event-level feature quantifying the pileup. The edges represent jet pairs with a single feature of ΔR between them. This graph is passed through several layers of message-passing networks which update the node representations while considering the information about the neighboring nodes, utilizing the message passing functionality of GNNs described in section 4.2.1. These node representations then pass through a fully-connected network which returns the flavor-tagging efficiency for each track jet. More details on this approach can be found in [56].

The GNN-based approach was shown to be more accurate than the map-based approach by exhibiting better closure of discriminating variable distributions between those produced using direct tagging and those produced with GNN-based tagging as opposed to map-based tagging. This closure is achieved after re-weighting the distributions and shows that their shapes match better with distributions produced by direct tagging when produced by GNN-based truth tagging as opposed to the map-based truth tagging.

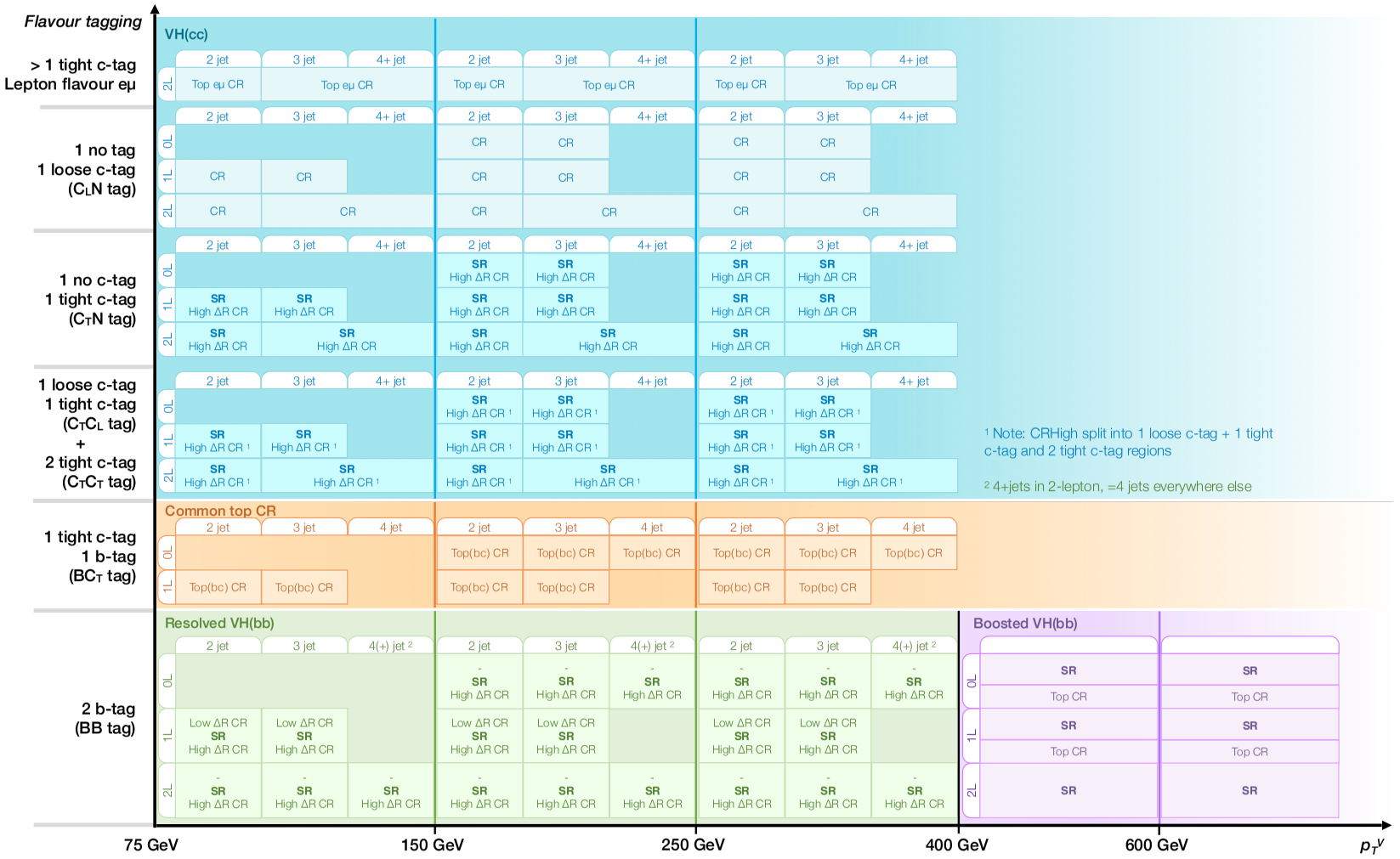


Figure 7.6: All signal and control regions used for the VHbb(cc) Legacy analysis. The green regions are used in $VH(\rightarrow b\bar{b})$, the blue ones in $VH(\rightarrow c\bar{c})$, and the orange ones are used in top CRs.

8

MULTI-VARIATE ANALYSIS

As explained in Section 6.2.2, in order to make a measurement of the POIs, a fit is performed to a distribution of an observable in which the signal can be clearly visible, increasing sensitivity. With well-chosen input variables, an ML-based multi-variate analysis (MVA) discriminant can be trained, which clearly separates signal and background. In this analysis, MVA was used in all analysis regimes, as opposed to previous iterations of the constituent analyses, where a fit on MVA was used only for $VH(\rightarrow b\bar{b})$. A BDT provided by the TMVA package of Root [22] was used for the MVA discriminant. The different regimes and lepton channels have separate input variables to the BDT, and each channel is further subdivided into regions of p_T^V and a number of jets in order to increase the sensitivity of the training.

The BDTs are trained to distinguish signal from background, so for $VH(\rightarrow b\bar{b})$ they were trained with $VH(\rightarrow b\bar{b})$ as signal while for $VH(\rightarrow c\bar{c})$ they were trained with $VH(\rightarrow c\bar{c})$ as signal. For both cases, V+jets, $t\bar{t}$, single-top, and di-boson samples are the background.

The training regions are summarized in Figure 8.1. In the resolved topology, the $p_T^V = 400$ GeV cut between resolved and boosted regimes is not applied for the training, but only considered in evaluation. Separate trainings are performed for events with different numbers of jets and in different regions of p_T^V, as shown in the figure. In the $VH(\rightarrow c\bar{c})$ regime, the trainings are split into a two c-tag region denoted in the figure as "TT+TL", where T stands for tight, and L stands for loose, which covers events that have two c-tags with at least one of them tight, and a one c-tag region "TN" where one jet has a tight c-tag, and one is untagged.

The training region of $p_T^V > 150$ GeV includes both of the analysis regions 150 GeV $< p_T^V <$ 250 GeV and $p_T^V > 250$ GeV since training on the inclusive region results in comparable performance as two dedicated trainings and since the $p_T^V > 250$ GeV region contains less statistics, which leads to overtraining.

In the boosted topology, only one BDT is trained per channel in the $p_T^V > 400$ GeV region.

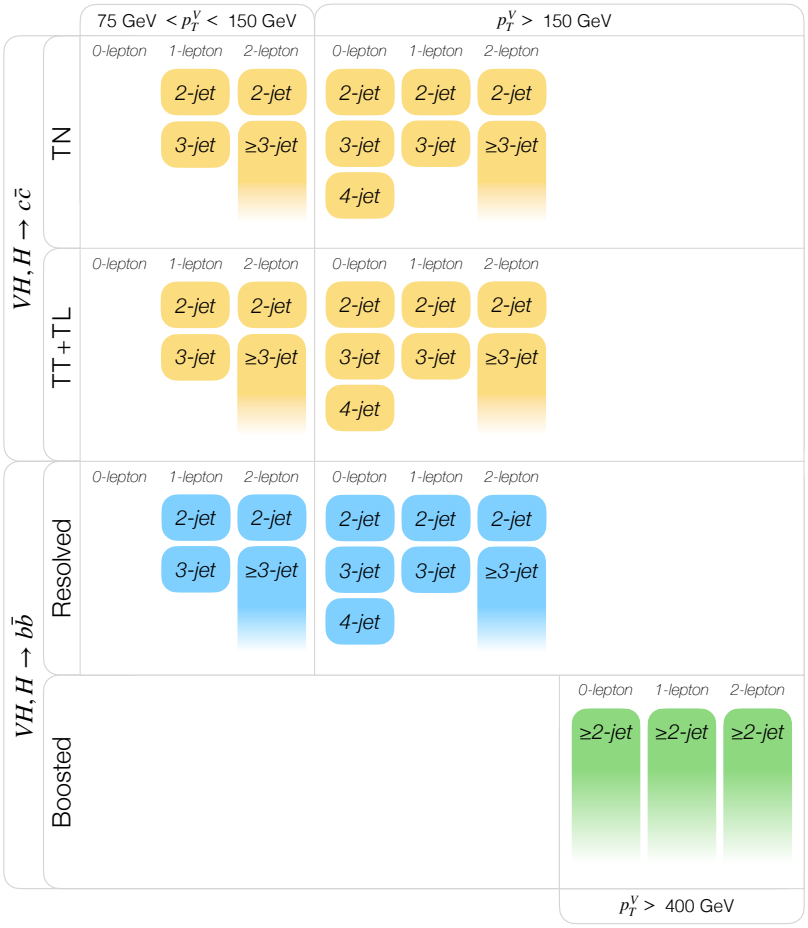


Figure 8.1: All training regions of Higgs signal vs. background in the VHbb(cc) Legacy analysis.

8.1 INPUT VARIABLES

In this section, the focus is on the variables for the 1-lepton and 2-lepton channels used in the $VH(\to b\bar{b})$ resolved and $VH(\to c\bar{c})$ analyses, in which the author made contributions to validations of the training and variable distributions.

8.1.1 $VH(\to b\bar{b})$ *Resolved and* $VH(\to c\bar{c})$ *MVA Input Variables*

In previous iterations of the constituent analyses, only $VH(\to b\bar{b})$ resolved used an MVA discriminant. The starting point for the list of variables used in the current analysis is those used in that iteration,

which were optimized by removing variables that have a negligible impact on the BDT score and adding those that bring a significant (more than 3%) improvement to the signal sensitivity in a given channel. Variables that are strongly correlated with others in the list are not included in the training, since they do not supplement the model with additional useful information. The final set of variables was harmonized between $VH(\to b\bar{b})$ resolved and $VH(\to c\bar{c})$ and is:

- p_T^V: transverse energy of the vector boson. This is computed differently per channel - it's the missing transverse energy (E_T^{miss}) in 0-lepton, vector sum of E_T^{miss} and lepton p_T in 1-lepton, and vector sum of the two leptons' p_T in 2-lepton.
- $p_T^{j_1}$, $p_T^{j_2}$: transverse momenta of the Higgs candidate jets where j_1 is the jet with the higher p_T.
- $m_{j_1 j_2}$: invariant mass of the Higgs candidate (H_{cand}) system.
- $\Delta R(j_1, j_2)$: the angular distance between the two Higgs-candidate jets forming the H_{cand} system.
- $m_{j_1 j_2 j_3}$: the invariant mass of two Higgs-candidate jets and the remaining jet with the highest p_T. When there are only two jets in an event, $m_{j_1 j_2 j_3} = m_{j_1 j_2}$.
- $\Delta\phi(V, H_{\text{cand}})$: azimuthal distance between the reconstructed vector boson and Higgs boson candidates.
- $\text{bin}_{\text{DL1r}(j1)}$, $\text{bin}_{\text{DL1r}(j2)}$: the flavour tagging bin a jet belongs to, with five possible bins - untagged, loose (70% WP) and tight (60% WP) b-tagged, and loose and tight c-tagged bins, as shown on Figure 7.5. In the MVA, the value for the two Higgs-candidate jets are used.
- $\sum_{i \neq 1,2} p_T^{j_i}$: p_T sum of non H_{cand} jets that have $p_T > 20$ GeV.
- **1-lepton channel specific variables**:
  - m_T^W: transverse mass of the W boson candidate reconstructed from the lepton and E_T^{miss}.
  - E_T^{miss}: missing transverse energy.
  - $\Delta y(V, H_{\text{cand}})$: rapidity difference between the vector boson and Higgs boson candidates.
  - $\min\{\Delta R(j_i, j)\}$, $i = 1, 2$: the distance in R between a H_{cand} jet (b- or c-tagged) and the closest additional jet.

  - $\min\{\Delta\phi(l, j_i)\}$, $i = 1, 2$: distance in ϕ between the lepton and the closest b-tagged (c-tagged) jet.
  - m_{top}: reconstructed mass of the leptonically decaying top quark to suppress top background.

- **2-lepton channel specific variables**:
  - $m_{\ell\ell}$: the invariant mass of the two leptons system.
  - $\cos\theta(\ell^-, Z)$: the angle between the Z boson and the lepton, which is sensitive to the Z boson polarization, which is distributed differently between the $Zb\bar{b}$ background and the signal [57].
  - $E_T^{\text{miss}} / \sqrt{S_{\text{T}}}$: 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.
  - $\Delta y(V, H_{\text{cand}})$: rapidity difference between the vector boson and Higgs boson candidates.

To illustrate the variable distributions, the variables of $VH(\to b\bar{b})$ 1- and 2-lepton in the evaluation region 75 GeV $< p_T^V <$ 150 GeV with exactly two and with at least four jets, respectively, are shown in Figures 8.2 and 8.3.

8.1.2 *Training*

The BDT training is done on all the nominal MC samples of this analysis. GNN truth tagging is used in order to maximize training statistics. The following BDT hyperparameters were tuned. Some of them are described in more detail in Section 4.3.2.

- Boost type - The algorithm used to enhance the classification performance of the BDT.
- Number of trees - the number of tree estimators in the BDT ensemble.
- Maximum depth - the maximum depth of any tree estimator.
- Learning rate - multiplicative factor applied to each tree to control the weight of the contribution of subsequent trees. A small value progressively "softens" the contribution of trees.
- Number of cuts - controls granularity with which to scan for the optimal cut on a feature's value while building trees.

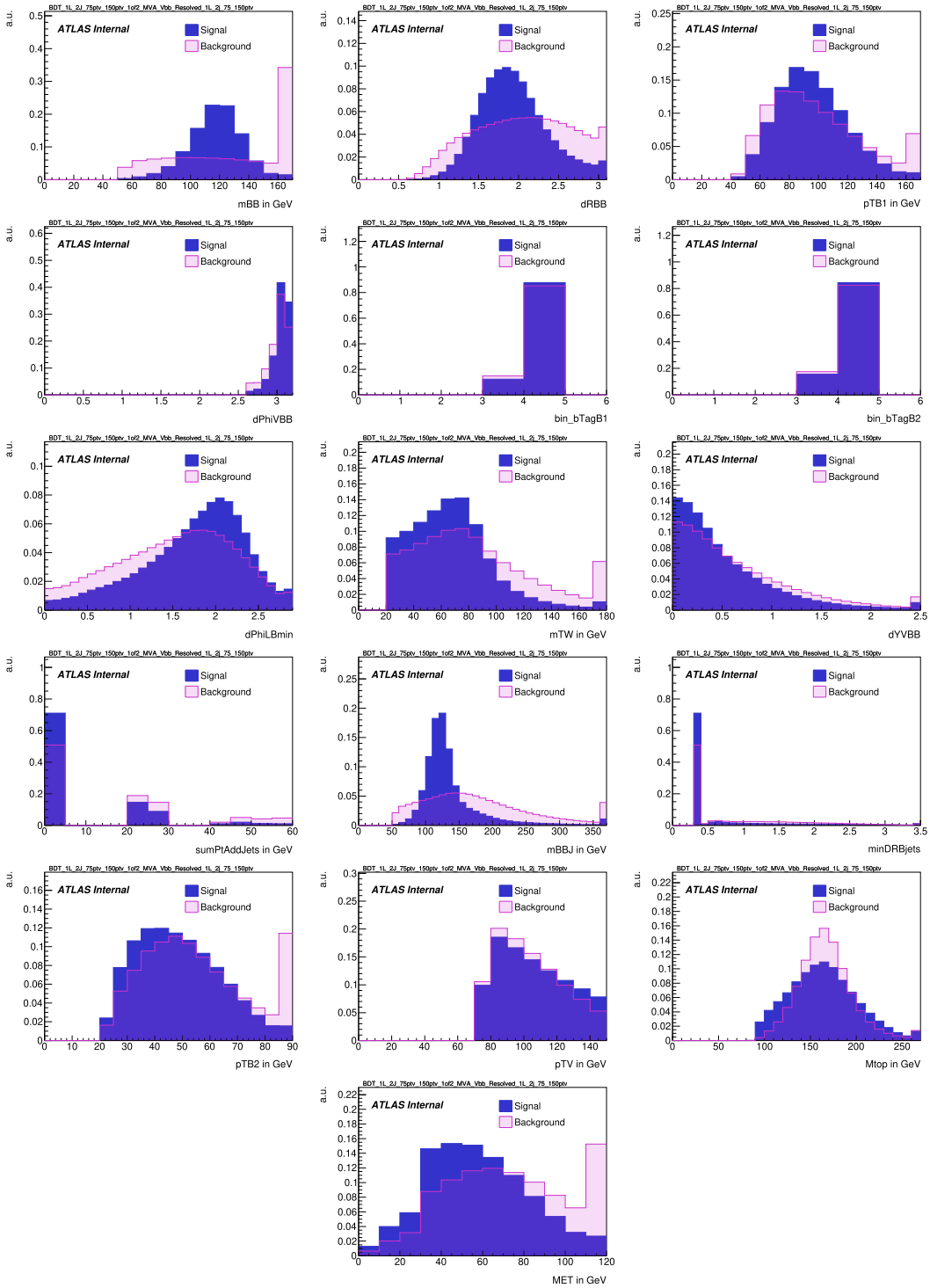


Figure 8.2: Normalised BDT input variable distributions for signal (solid blue) and background (red) in the 1-lepton channel 2-jet region with 75 GeV $< p_{\mathrm{T}}^{W} <$ 150 GeV. The upper bin is filled with the overflow content.

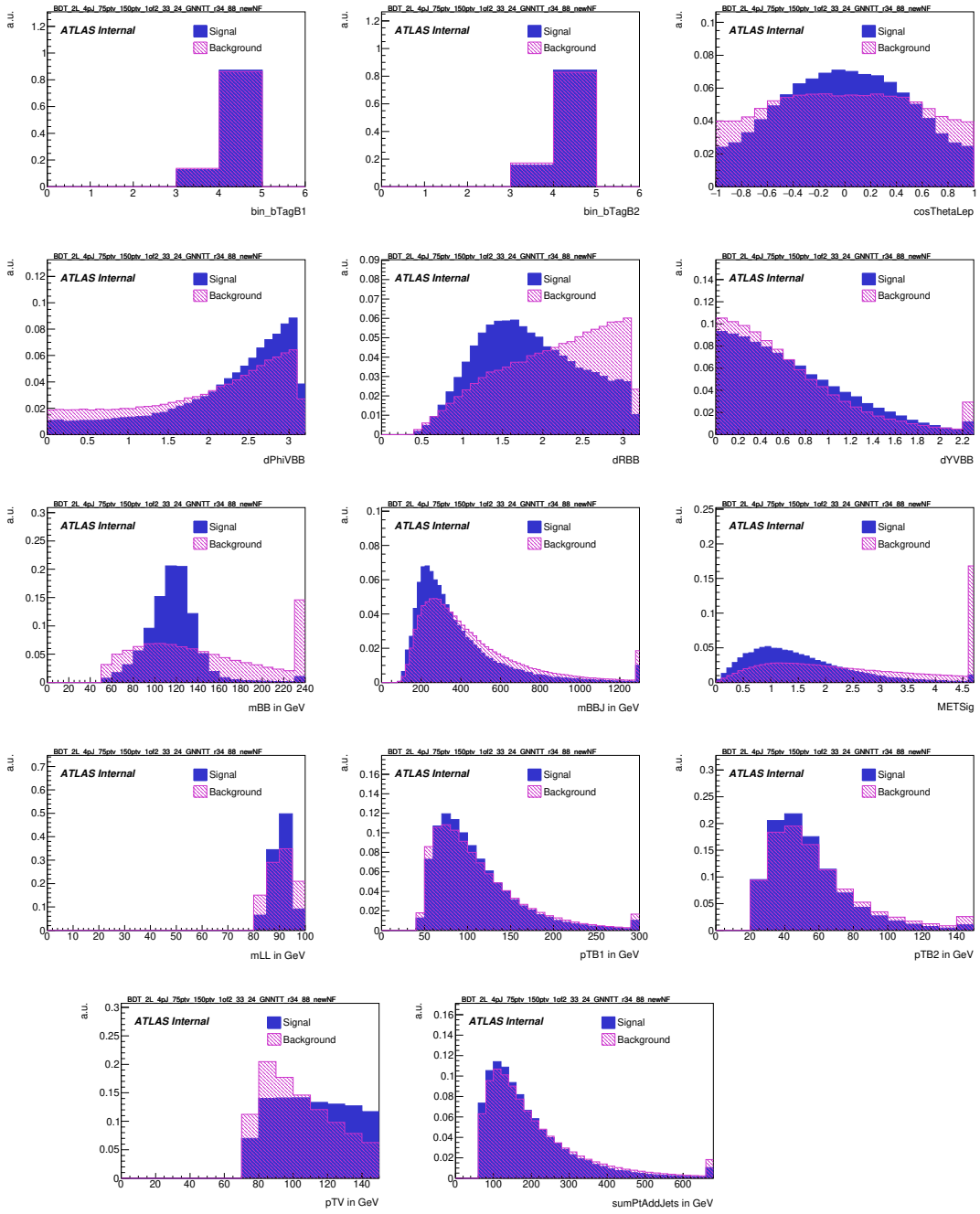


Figure 8.3: Normalised BDT input variable distributions for signal (blue) and background (red) samples in the 2-lepton channel for the $\geq$ 4-jet region with 75 GeV $< p_{\mathrm{T}}^{Z} <$ 150 GeV. Overflowing events are included in the last bin.

- Minimum node size - stop splitting a BDT node when the number of events in the node reaches this value.
- Separation method - separation criterion used to assess the signal-background separation of a specific cut on a specific variable. The criterion is maximum at the worst performance and falls off to zero. The Gini index criterion is defined as $p(1-p)$ where p is the fraction of signal events after the cut.
- Pruning method - whether pruning is applied or not.

The selected values of the hyperparameters are listed in Table 8.2 for $VH(\to b\bar{b})$ boosted and resolved and in Table 8.3 for for $VH(\to c\bar{c})$. The values were chosen for each region to maximize the BDT score-based signal significance described in Section 8.1.3.

The training of the BDT is performed using the k-fold method, where the training sample is divided into k subsamples (*folds*), and the training is done in k iterations, with each iteration trained on one subsample and evaluated on the rest. In all regions, $k = 2$ was used. For each training, the distributions of the BDT score for the different evaluation folds are summed together to produce the final distribution.

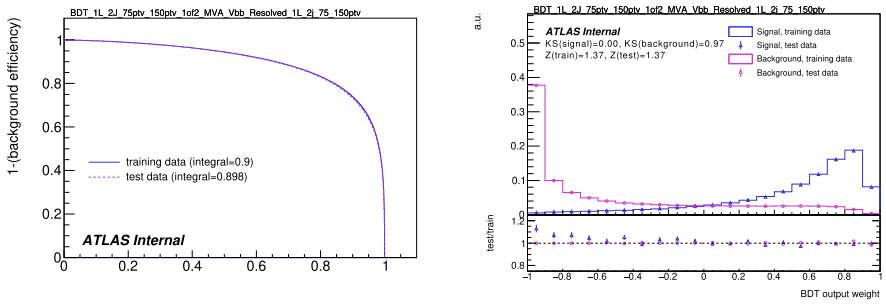
8.1.3 *Evaluation of Trainings*

An overtraining check is performed for each training by comparing the BDT score distributions and ROC curves between the training and test samples. In the absence of overtraining, these should be very close, as can be seen, for example, in the case of the $VH(\to b\bar{b})$ resolved regime's 1-lepton channel in Figure 8.4 and which is also the case in all other trainings.

The most important metric relevant for physics is how sensitive we are to the signal in the BDT distribution, or the signal significance as computed on the binned distribution, which is given for a particular binning, as

$$Z = \sqrt{\sum_i 2\left((s_i + b_i)\ln\left(1 + \frac{s_i}{b_i}\right) - s_i\right)}, \tag{8.1}$$

where i runs over the bins, s_i and b_i are the number of signal and background events in the i-th bin, respectively. Evidently, different binnings will give different signal sensitivities, so the binning must be optimized to achieve high sensitivity. In particular, regions of the



(a) (b)

Figure 8.4: Resolved $VH(\to b\bar{b})$: Overtraining checks for the 1-lepton, two b-tagged 2-jet category for BDT trainings performed with GNN truth tagging in the region 75 GeV $< p_T^V <$ 150 GeV. (a) ROC curves for the training and testing samples, (b) BDT score distributions for the training and testing samples.

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

Table 8.1: Parameter z_s and z_b settings used in Transformation D for different phase space.

BDT score with high statistics should have more bins and vice versa. A transformation called *Transformation D* is performed on the bins for that purpose. The transformation is based on the following formula:

$$Z_D = z_s \frac{n_s}{N_s} + z_b \frac{n_b}{N_b}, \tag{8.2}$$

where N_s, N_b are the total number of signal and background events in the histogram and n_s, n_b are the number of signal and background events in a certain bin. The transformation works iteratively, starting from the highest bin and merging bins into a single one, which causes n_s and n_b to increase until $Z_D > 1$ for the new bin, after which the process repeats from the next unmerged bin. The parameters z_b and z_s roughly represent the desired number of background- and signal-enriched bins, respectively. These parameters are optimized for each region to give more bins in regions with higher statistics and vice versa. The optimal values for these parameters in each training region are given in Table 8.1. An example of a BDT distribution before and after the transformation is given in Figure 8.5.

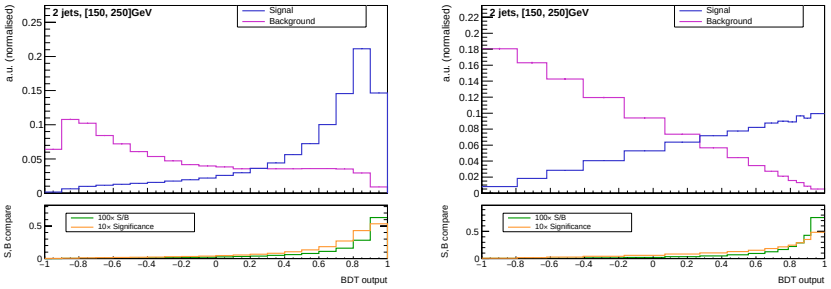


Figure 8.5: Comparison of BDT score distribution before (left) and after (right) applying Transformation D. This is evaluated in the 1-lepton channel of $VH(\to c\bar{c})$, in the region 150GeV $< p_T^V <$ 250GeV.

	Resolved $VH(\to b\bar{b})$			Boosted $VH(\to b\bar{b})$		
Settings	0-lepton	1-lepton	2-lepton	0-lepton	1-lepton	2-lepton
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%
Separation method	Gini index	Gini index	Gini index	Gini index	Gini index	Gini index
Pruning method	No pruning	No pruning	No pruning	No pruning	No pruning	No pruning

Table 8.2: BDT hyperparameters used for the 0-, 1- and 2-lepton channels for the $VH(\to b\bar{b})$ resolved and boosted analyses.

	$VH(\to c\bar{c})$	
Settings	0-, 1- and most 2-lepton regions	2-lepton, $\geq$ 3J, low p_T^V
Boost type	Gradient boost	Adaboost
Number of trees	600	200
Maximum depth	4	4
Learning rate (β)	0.5	0.15
Number of cuts	100	100
Minimum node size	5%	5%
Separation method	Gini index	Gini index
Pruning method	No pruning	No pruning

Table 8.3: BDT hyperparameters used for the 0-, 1- and 2-lepton channels for $VH(\to c\bar{c})$ analysis.

8.1.4 *GNN as Signal-Background Classifier*

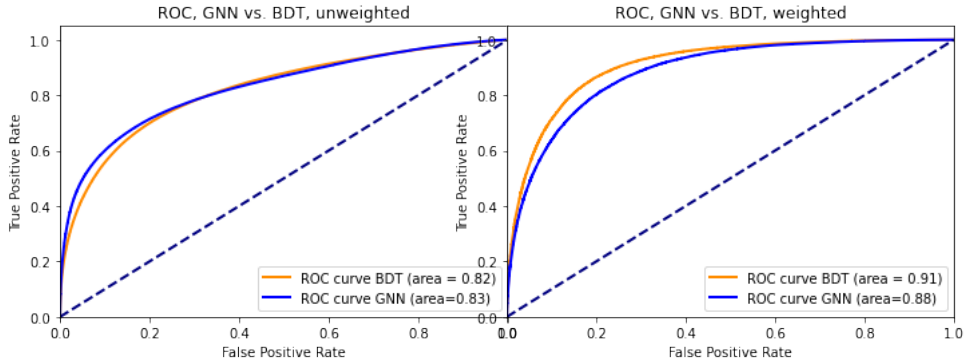
The author performed a study on the classification performance of GNN versus BDT in the context of the VHbb(cc) Legacy analysis. The data on which the study was done are MC ntuples of signal ($H \to bb$) and relevant background, the same ones used to train the BDT. The study was performed in the resolved 2-lepton channel, as that channel has kinematic information of four to five objects - two leptons and two or three jets, as opposed to, e.g., the 0-lepton channel, which has three or four objects - missing energy and two or three jets. The signal $H \to cc$ samples were omitted from the data, as in the BDT training.

As for BDT, the data were split into two folds: selecting all events with an even event number for the first and an odd number for the second. One fold is used as the training set, while the other is the test set. The events in the ntuples file are weighted so that the sum of the weights of all events belonging to a sample of a given process is proportional to the theoretical cross-section of that process.

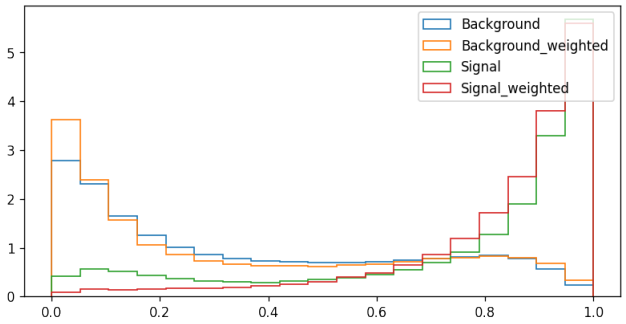
The trained proof-of-concept model was a simple DeepSet model implemented in PyTorch [58]. The model's input is an unordered set of four or five objects consisting of two leptons and two or three jets, respectively. Each object is a node in the graph, without edges connecting the nodes. The features of each node are listed in Table 8.4. For each node, there is a total of six features, the first four of them being object-specific and the other two event-wide, replicated over all nodes in the graph.

Variable	Meaning
p_T	Transverse momentum
η	Pseudorapidity
ϕ	Azimuthal angle
bin_bTag	1 to 5, corresponding to the PCBT b-tag score bin if the particle is a jet or 6 if it's a lepton
nJ	number of jets in event
FlavL	Lepton flavour, according to PDG identifiers (11 for electron, 13 for muon) [12]

Table 8.4: A list of variables used as the nodes of the input graph. The first four variables are particle-specific, while the last two are event-wide features replicated over all particles.



(a) (b)



(c)

Figure 8.6: Evaluation of a DeepSet classifier trained to separate $VH(\to b\bar{b})$ signal from background. (a) is the ROC curve of the trained classifier produced by treating each event in the test set equally, and (b) was produced by weighing each event by its weight. (c) shows the signal-background separation for the weighted and unweighted cases.

A training of the GNN model on a sample consisting of 2.5 million randomly selected events from a single region of 75 GeV $< p_T^V <$ 150 GeV and at least three jets, evaluated on a separate test set of the same region, resulted in a ROC curve which is better than that of the BDT, trained and evaluated on a much larger number of events from the same region when the ROC curve is computed without applying event weights, but worse than the BDT when weights are applied, as shown on Figure 8.6. This implies that the DeepSet model captures the internal structure of events well but, during training, emphasizes the more numerous events with low weights.

In the BDT training of the $VH(\to b\bar{b})$ resolved analysis, before the training, the event weights in the training sample are re-calculated so that the sum of signal event weights and the sum of background event

weights are each equal to the number of signal events. However, for GNN training, it is unclear how to include the weights, what kind of re-weighting should be done, and how to treat negative event weights.

Studying the effect of different re-weighting schemes on the model's classification performance required multiple trainings and pre-processing of the data. One of the challenges of training ML models is the hardware restrictions, as the amount of data in the training set often exceeds the maximum memory capacity of the workstation. ANN training also involves computationally intensive matrix multiplication operations, best performed on machines with graphics processing units (GPU), which provide orders of magnitude speed up relative to ordinary central processing units (CPU). To remedy this, a semi-automatic pipeline was implemented that leverages GPU nodes on CERN's batch computing system, HTCondor, for training. Its input is an ntuples file, and the output is several models trained on a training set with multiple event re-weighting schemes and evaluated on a test set. During evaluation, the signal significance given by (8.1) is evaluated for each training.

The framework allows the production of comparisons between different training processes and, in particular, different weighting schemes, which makes it possible to select the best data pre-processing and training scheme to increase the significance of the signal after classification. An example comparison is shown in Figure 8.7, for a different training and evaluation region of 150 GeV $< p_T^V <$ 250 GeV and two jets, in which the red line is the BDT performance and the blue bars show the performance of a GNN in different scenarios. These include mainly re-weighting schemes, such as special treatment of outlying weights by pulling them closer to the bulk of the weights' distribution ("squashing"), normalizing to different intervals, and using different loss functions (described in the next section), as summarized in Table 8.5. None of the trained models results in performance superior to the BDT.

8.1.4.1 *Modified Loss Function*

Searching for optimal re-weighting is effectively equivalent to introducing additional hyperparameters that guide the treatment of weights and optimizing them by searching for the optimal re-weighting for a specific dataset. A more general approach to the utilization of the weights is desirable. An additional problem is the appearance of very

Re-weighting	Description
abs_squash_outliers_norm_0_x_plus1	Squash outliers, normalize to interval 0-x, add 1 to all weights
norm	Normalize to interval 0-1
norm_EqualNumWeights	Average weight of 1 per event for signal, and sum events for background equal to sum events for signal (same as in BDT training)
norm_plus1	Normalize to interval 0-1, add 1 to all weights
noweights	All events have weight 1
trivial	All events have their original MC weights
trivial_bce_minus_significance	Similar to trivial, but the loss function is the binary cross-entropy minus the significance
trivial_plus1	All events have their original MC weights, add 1 to all weights

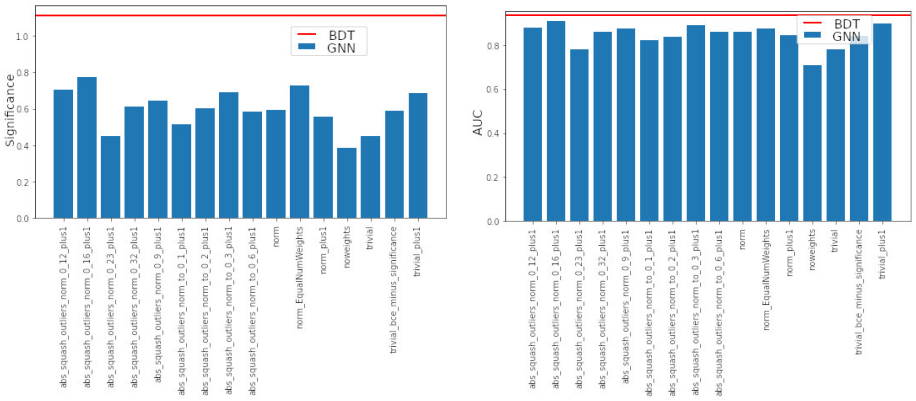
Table 8.5: A list of re-weighting schemes used to evaluate the GNN classifier.

large and negative weights, the latter being notoriously problematic for ML models. For example, the TMVA manual recommends dropping negative weights for models that don't treat them properly but keep them during evaluation [22].

For all re-weightings considered so far, the event weights are utilized in the training by using them to weight the binary cross-entropy loss function, which is a measure of the distance of each event's prediction score from the actual class (signal/background) of that event, the value of which the training is designed to minimize. The training is done in batches of 500 events, so the final score for a batch is the weighted sum of the values of the loss function for all events in the batch. Since it is not clear whether negative event weights can be used in this manner, as is the effect of large event weights on training, and since we're interested in the signal significance, it is natural to use a loss function that simultaneously minimizes the original, unweighted, binary cross-entropy loss function and maximizes the significance of the signal, so that the loss function for a batch is:

$$L_{\text{signif.}} = L_{\text{BCE}} - Z, \tag{8.3}$$

where L_{BCE} is the original, unweighted, binary cross-entropy loss function, and Z is given by (8.2) with 20 equal-width bins. This effectively splits the loss function into a term that is independent of weights, "pushing" the event score predictions towards their true values, and a



(a) (b)

Figure 8.7: Significance (a) and AUC (b) values for various re-weightings of the events in the training set for the DeepSet model in the two jets, $150 < p_T^V < 250$ GeV region of the 2-lepton channel. The red line is the reference BDT performance in the $VH(\to b\bar{b})$ resolved analysis in the same region.

second term that is of physical interest, which we seek to maximize and in which weights are properly treated by weighting each event that goes into the model score histogram, exactly as it's done for the final evaluation of the model.

The resulting performance for this kind of loss function can be seen in Figure 8.7, second bar from the right. The performance is not the best among the re-weighting schemes. Still, it is promising as it provides a natural treatment of weights, and the performance of this method of utilizing weights may be further improved by, e.g., increasing the training batch size.

These results are intended to inspire subsequent studies in future iterations of this analysis. Graphs are structurally suitable to represent events, and GNNs are the optimal models to e.g. classify graph representations of objects. Therefore, it is likely that with further studies on how the data should be processed and on the structure of the model, using a GNN in place of a BDT may boost the significance of the signal. That being said, the results here show that the humble BDT model is still highly performant, even with respect to the latest generation ANNs.

9

ANALYSIS RESULTS

9.1 STATISTICAL MODEL

The statistical model is based on fits to a likelihood function as described in Section 6.2.2. Multiple fits were performed, with each fit corresponding to a single set of signal-strength POIs in different regions. The effects of systematic uncertainties are included in the likelihood as nuisance parameters, and the normalizations of the largest backgrounds, top, and V+jets are determined by the fit, so they are left as free-floating parameters.

In signal regions, the BDT output discussed in Chapter 8 is used for the fits, while different observables were fitted in different control regions. Table 9.1 lists the observables used for the fit in the $VH(\to b\bar{b})$ resolved control regions. For the boosted regime, the mass of the large-R jet is fitted in the Top CRs.

The fits that were performed for VH are:

- Two-POI fit for $\{\mu^{bb}_{\mathrm{VH}}, \mu^{cc}_{\mathrm{VH}}\}$,
- Four-POI fit for $\{\mu^{bb}_{\mathrm{WH}}, \mu^{cc}_{\mathrm{WH}}\}, \{\mu^{bb}_{\mathrm{ZH}}, \mu^{cc}_{\mathrm{ZH}}\}$ to extract the signal strengths of WH and ZH production modes in $H \to bb$ and $H \to cc$,
- 13-POI fit to measure the signal cross section multiplied by the $H \to bb$ and $V \to$ leptons in 13 STXS bins,
- 10-POI fit, similar to the 13-POI fit but with no split in the number of jets.

Additional fits were performed for a similar cross-check di-boson analysis, which measures the signal strength of VZ, analogous to VH with Z replacing H and decaying either to $c\bar{c}$ or $b\bar{b}$.

9.2 RESULTS

The results are obtained by performing a combined simultaneous fit to all signal regions of $VH(\to b\bar{b})$ and $VH(\to c\bar{c})$, on the full Run 2 dataset of 140 fb^{-1}. The observed (expected) significance of the

Table 9.1: A schematic of the fit variables used in the control regions. The 'Norm. Only' label indicates that only a single bin is used in the fits and '—' indicates that the region is not used in the fits. The first row contains the flavour tagging regions for the two jets - B, C_T, C_T, N stands for *b*-tagged, tight *c*-tagged, loose *c*-tagged and not tagged, respectively.

Channel	Region	BB	C_TN	C_TC_L	C_TC_T	BC_T	C_LN
0-lepton	High-ΔR CR	Norm. Only				—	
	BC_TTop CR	—				$m_{j_1j_2}$	—
	$V+lf$ CR	—					Norm. Only
1-lepton	Low-ΔR CR	$BDT_{\text{Low-}\Delta R\text{ CR}}$	—				
	High-ΔR CR	p_T^V		$m_{j_1j_2}$		—	
	BC_TTop CR	—				$m_{j_1j_2}$	—
	$V+lf$ CR	—					p_T^V
2-lepton	High-ΔR CR	p_T^V		$m_{j_1j_2}$		—	
	Top eμ CR	—	Norm. Only			–	—
	$V+lf$ CR	—					p_T^V

$VH(\to b\bar{b})$ signal is 7.4σ (8.0σ). The signal strength and its uncertainty are:

$$\mu_{VH}^{bb} = 0.92^{+0.16}_{-0.15} = 0.92 \pm -0.10(\text{stat.})^{+0.13}_{-0.11}(\text{syst.}). \tag{9.1}$$

For $VH(\to c\bar{c})$, the fit resulted in an observed (expected) 95% confidence level (CL) upper limit on the signal strength of 11.5 (10.6) times the SM prediction. The expected and observed 95% CL upper limits on this signal strength for the individual lepton channels and the combination are summarized in Figure 9.1. The corresponding signal strength is

$$\mu_{VH}^{cc} = 1.0^{+5.4}_{-5.2} = 1.0^{+4.0}_{-3.9}(\text{stat.})^{+3.7}_{-3.5}(\text{syst.}). \tag{9.2}$$

This is a significant improvement over the previous result of 26 times the SM prediction for the previous $VH(\to c\bar{c})$ analysis [6]. The signal strength expected and observed 95% and 68% CL contours for $VH(\to b\bar{b})$ and $VH(\to c\bar{c})$ are shown on Figure 9.2.

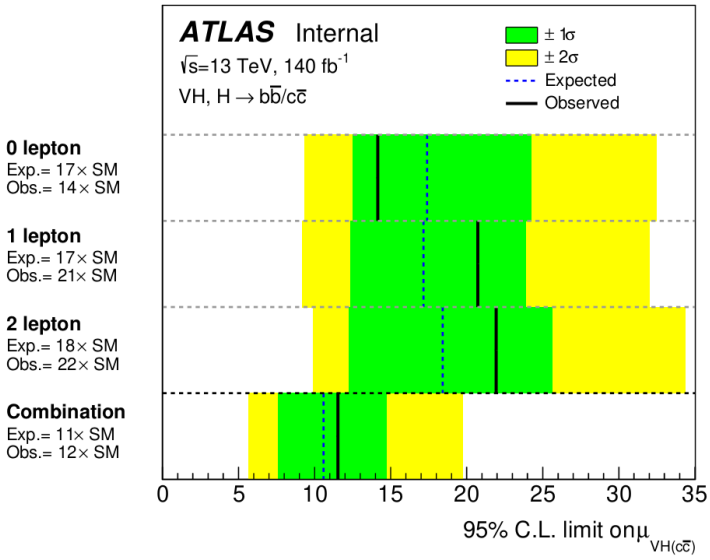


Figure 9.1: Expected and observed 95% CL upper limits for the $VH(\to c\bar{c})$ signal strength for individual lepton channels and their combination.

The result of the four-POI fit for the signal strength of WH and ZH with H decaying to $b\bar{b}$ is summarized in Figure 9.3. The observed (expected) significance of the ZH, $H \to b\bar{b}$ measurement is 4.9 (5.6) standard deviations, and for WH, $H \to b\bar{b}$ it is 5.3 (5.5) standard deviations, making this the first observation of the WH, $H \to b\bar{b}$ process.

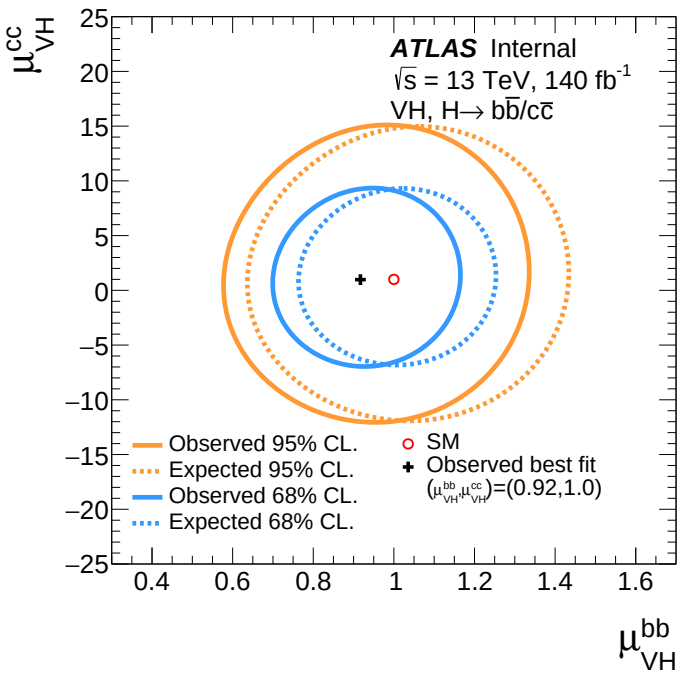


Figure 9.2: Expected and observed 95% and 68% CL contours for the $VH(\to b\bar{b})$ and $VH(\to c\bar{c})$ signal strengths.

The observed and expected values for the 13-POI and 10-POI STXS fits are summarized in Figures 9.4, 9.5, respectively, showing no significant deviations from the SM predictions.

9.2.1 *Measurement Uncertainties*

Table 9.2 shows the breakdown of the sources of statistical and systematic uncertainties contributing to the total uncertainties in the observed and expected signal strength measurements of $VH(\to b\bar{b})$, $WH(\to b\bar{b})$, $ZH(\to b\bar{b})$, and $VH(\to c\bar{c})$. The statistical and systematic uncertainties are similar in size. The largest contribution to the systematic uncertainties comes from theoretical modeling, mainly from the signal and the W + jets and Z + jets backgrounds, followed by experimental uncertainties from jet reconstruction and flavor tagging. A significant reduction in uncertainty relative to the previous iterations of the $VH(\to b\bar{b})$ resolved, $VH(\to b\bar{b})$ boosted, and $VH(\to c\bar{c})$ analyses is obtained due to several factors, some of which are the increase of 25% in the c-jet rejection of the DL1r flavor tagging algorithm at the same b-tagging efficiency, the introduction of dedicated CRs to constrain the top background and the re-optimization of the MVA.

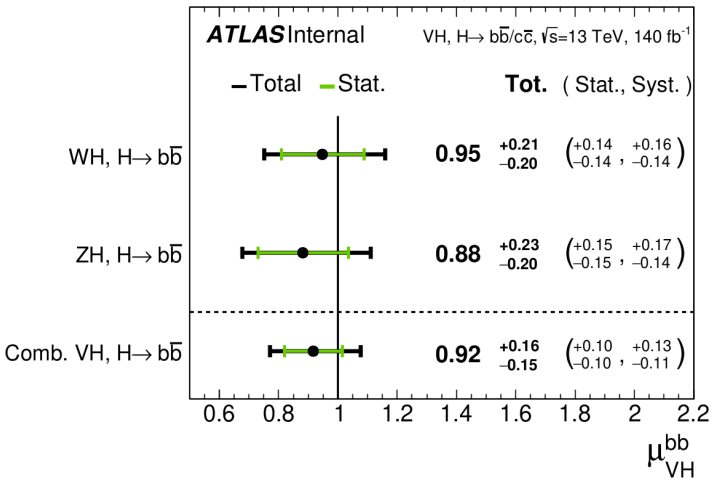


Figure 9.3: Signal strength

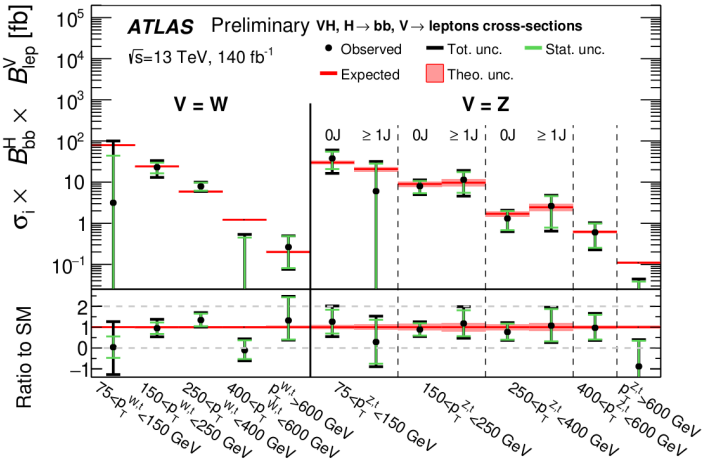


Figure 9.4: Result of the 13-POI fit for the measurement of the VH, $V \to$ leptons cross sections times the $H \to b\bar{b}$ branching ratio in 13 STXS bins.

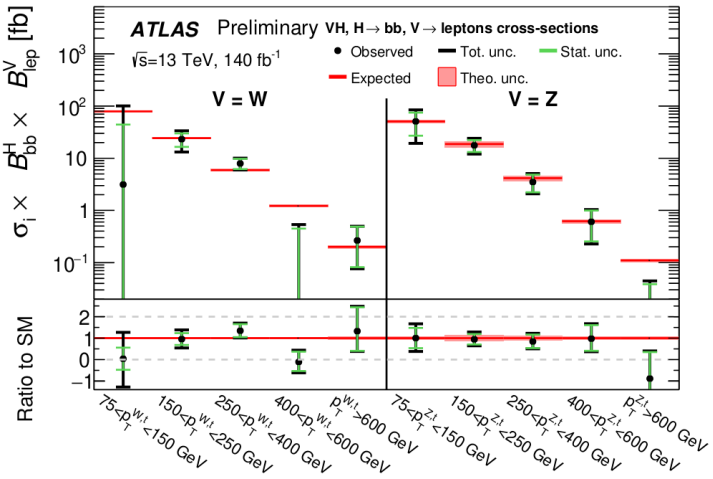


Figure 9.5: Result of the 10-POI fit for the measurement of the VH, $V \to$ leptons cross sections times the $H \to b\bar{b}$ branching ratio in STXS bins with no split in the number of jets.

Table 9.2: The breakdown of contributions to the uncertainty in the fitted value of the signal strengths for the $VH(\to b\bar{b})$, $WH(\to b\bar{b})$, $ZH(\to b\bar{b})$, and $VH(\to c\bar{c})$ processes.

Source of uncertainty		σ_μ			
		$VH(\to b\bar{b})$	$WH(\to b\bar{b})$	$ZH(\to b\bar{b})$	$VH(\to c\bar{c})$
Total		0.153	0.204	0.216	5.31
Statistical		0.097	0.139	0.153	3.94
Systematic		0.118	0.149	0.153	3.57
Statistical uncertainties					
Data statistical		0.090	0.129	0.139	3.67
$t\bar{t}$ $e\mu$ control region		0.009	0.014	0.027	0.08
Background floating normalisations		0.034	0.049	0.042	1.24
Other VH floating normalisation		0.007	0.018	0.014	0.33
Simulation samples size		0.023	0.033	0.030	1.62
Experimental uncertainties					
Jets		0.027	0.035	0.030	1.02
E_T^{miss}		0.010	0.005	0.021	0.23
Leptons		0.003	0.002	0.010	0.25
Flavour-tagging	b-jets	0.020	0.018	0.026	0.29
	c-jets	0.013	0.017	0.012	0.73
	light-flavour jets	0.005	0.008	0.008	0.66
Pile-up		0.008	0.017	0.002	0.23
Luminosity		0.006	0.007	0.006	0.08
Theoretical and modelling uncertainties					
Signal		0.076	0.074	0.101	0.72
Z + jets		0.042	0.018	0.081	1.77
W + jets		0.054	0.087	0.026	1.42
$t\bar{t}$ and Wt		0.018	0.033	0.018	1.02
Single top quark (s-, t-ch.)		0.010	0.018	0.002	0.16
Di-boson		0.033	0.039	0.049	0.52
Multi-jet		0.005	0.010	0.005	0.55

10

CONCLUSION

The VHbb(cc) Legacy analysis aims to combine together previous iterations of the measurements of Higgs couplings to b- and c-quarks, known as the $VH(\to b\bar{b})$ Resolved, $VH(\to b\bar{b})$ Boosted, and $VH(\to c\bar{c})$ analyses, by focusing on the Higgs production mode in association with a vector boson, which, though not the leading production mode, produces a clear signature which improves sensitivity to signal.

Looking at the full Run 2 data, the analysis takes the strongest features of the aforementioned analyses and combines the phase spaces to increase the sensitivity. One of the main upgrades is the usage of a BDT discriminant that was trained to separate signal and background in different phase space regions. This, along with the increase in statistics and improvements in flavor tagging techniques, among others, allowed us to perform the first 5.3σ observation of the WH, $H \to b\bar{b}$ process and tighten the constraint on the coupling constant of the Higgs boson to c-quarks from 26 to 11.5 times the SM prediction. All measurements produced results in agreement with those expected from the SM.

The paper summarizing this analysis was sent to publication [59], with future iterations of the analysis to come that would look at Run 3 data. Using further improvements based on machine learning models such as GNNs for the fit discriminant instead of the currently used BDT may potentially improve the sensitivity of the fit to the signal, especially because a GNN is designed to operate on graphs, which are highly suitable for representing collider events, taking into account the interactions between particles using the GNN's internal message passing mechanism. While the rough preliminary study presented in this chapter showed no clear advantage over the BDT discriminant, future careful studies of the GNN as a signal-background classifier for physics analyses are required. From the study presented here, two suggestions for future research direction would be the proper treatment of MC weights in the loss function when training an ML model and selecting a more suitable differentiable loss function that would more clearly reflect the sensitivity to the physics signal of interest.

Part IV

THE L1CALO EFEX BDT τ ALGORITHM

11

TAU IDENTIFICATION IN THE ATLAS LEVEL-1 CALORIMETER TRIGGER

11.1 INTRODUCTION

The τ lepton plays an essential role in the physics program of the ATLAS experiment. Being the heaviest lepton, it is an excellent probe of the coupling of the SM Higgs boson to leptons [44] as well as for searches for BSM effects in the Higgs sector [60–62], dark matter candidates [63] and super-symmetry [64, 65]. Final states with τ leptons are also good probes for anomaly searches [60] and precision measurements [66] in the electroweak sector. It is therefore crucial to implement a mechanism for a highly efficient selection of τ lepton-containing events, producing a sample of as high a purity as possible to reduce non-τ backgrounds and, therefore, maximize the sensitivity of τ-based physics analyses, both in Run 3 and onwards.

Searches for BSM and precision measurements of SM processes require collecting more and more data. Run 3 of the LHC started the commissioning process in 2022, following the Phase-I upgrade with an expected instantaneous luminosity of $\mathcal{L} = 3 \times 10^{34}$ cm^{-2}s^{-1} and pile-up values of $\langle\mu\rangle = 80$ [67], compared to $\mathcal{L} = 1.9 \times 10^{34}$ cm^{-2}s^{-1} and $\langle\mu\rangle = 33$, respectively during Run 2 [68]. During the upgrade, several systems of the ATLAS detector were upgraded, among which are the LAr calorimeter [67] and the TDAQ system [69], allowing a higher granularity readout of calorimeter energy depositions. The purpose of this upgrade was to enhance the physics reach of the experiment during Run 3 and beyond at increasing LHC luminosities.

The high mass of the τ causes it to have a very short lifetime, resulting in its decay before entering the inner detector. When a τ decays hadronically, which occurs 66% of the time, it results in a signature almost indistinguishable from QCD jets. During Run 3, this discrimination between τ particles and jets will be even more complicated due to the significant increase in pile-up. This is especially critical for the L1Calo trigger subsystem, that uses only calorimetric information and is limited in available resource and latency for algorithms. The planned increase in pile-up degrades the calorimeter resolution, requiring an increased granularity readout of the calorimeters along

with improved algorithms, all of which are addressed by the Phase-I upgrade. The upgraded L1Calo makes full use of the finer granularity, using new electron/photon, jet, and global feature extraction modules referred to as eFEX, jFEX, and gFEX, respectively [69].

The first stage of τ identification in the ATLAS trigger in Run 3 is implemented in the eFEX module. The baseline algorithm used for this runs on FPGAs and is based on several heuristics, such as an estimation of E_T of a predefined calorimeter cluster and measuring the isolation in a cluster around a local maximum of energy. These values are very quick to compute in hardware and set a threshold, eliminating the vast majority of QCD jets. Still, the finer granularity opens up a much larger space of observables to explore for similar thresholding, which is a natural setting for a machine-learning (ML) based approach for triggering.

Due to the planned increase in luminosity of the LHC and the increase in background, the ML approach will likely be required in the future to bring the trigger rates down while maintaining high signal efficiency. However, using ML algorithms in ultra-low-latency settings is a novel technology and requires experimentation to extract its full potential. The first implementation of an L1 ML-based algorithm at the LHC was done in the CMS experiment, where a BDT was trained as a regression model for muon p_T estimation from 25 input variables, improving the background rejection by a factor of three [70]. However, the implementation involved encoding the BDT as a 1.2 GB look-up table and storing it on special hardware, an impossible approach using the current ATLAS trigger hardware, where resources are limited. A new approach is required, where the ML model, instead of being encoded in a look-up table, is fully evaluated.

There is a multitude of details that need to be addressed in such a workflow by experimenting with a simple yet powerful ML model in order to get crucial experience in such a setting. As an example, from the perspective of firmware implementation, these algorithms often require a different approach, in which the model's design is generated using automated tools and incorporated into a larger, manually written design. Unlike a fixed, heuristic-based algorithm, an ML-based one may need to be re-trained at different points in time during data taking. In order to understand the effect and real-world costs of maintenance and re-training, it is essential to experiment with a simple ML model in the L1Calo trigger during ATLAS data taking.

This part describes the development, implementation, and performance studies of a BDT-based algorithm used for data taking in Run 3

during 2024 for τ triggering. The algorithm has a superior performance for resource usage over the previous heuristics-based τ algorithm in the τ E_T range relevant for physics analyses.

11.2 TRIGGER CHAINS

From the initial torrent of data resulting in the full ATLAS detector being read out at a rate of 40 MHz, the triggering system must reduce the rate to around 1 kHz. This is done in ATLAS by focusing on specific physics objects and combinations of them and forming *trigger chains*, which specialize in their identification and are motivated by the data required for offline physics analyses, where the trigger chain that fired the event is used in the initial selection of the events of interest.

The initial identification stage, which receives the full 40 MHz flow, is the level-1 trigger, which itself is separated into sub-stages for identification, with several algorithms specializing in the identification of particular physics objects. The very first stage, which is the context of the algorithms discussed herein, is the efficient identification of single objects, referred to as regions of interest (RoIs), such as isolated calorimeter clusters with particular features corresponding to a particular physics object of interest. These RoIs are then used as seeds for more elaborate algorithms within the L1 trigger. For example, two τ calorimeter clusters and one jet calorimeter cluster can be combined to form a trigger requiring a topological cut on the angle between the τ clusters in addition to a jet.

If an L1 trigger fires, the RoIs are then used as seeds for offline-style algorithms at the HLT described in Section 5.3. These can, for example, fully reconstruct a τ jet from tracking and calorimeter information and cut on the fully reconstructed p_T of the τ jet to reach a final decision on whether to fire this particular trigger chain. If one trigger chain fires, the data from this particular bunch crossing is considered for storage.

11.3 TURN-ON CURVES AND TRIGGER RATES

To evaluate the efficiency of a triggering algorithm for signal selection, a *turn-on curve* is generated, such as one of those shown in Figure 13.1, which plots the efficiency for the selection of events containing signal objects, as determined by the evaluated algorithm, as a function of the p_T of the fully reconstructed object that the algorithm is meant to

identify. Turn-on curves are generated on signal MC samples or data samples that had undergone a selection for events containing signal physics objects. They can be produced with respect to events, in which case the number of events containing at least one RoI that passes the target algorithm's threshold serves as the numerator, and the total number of signal events serves as the denominator. They can also be produced for physics objects, in which case, after reconstruction, the physics objects are matched to RoIs. The numerator is then the number of RoIs that pass the target algorithm's thresholds and match a physics signal object, and the denominator is the total number of signal objects.

The turn-on curve shows only part of the picture since a very crucial quantity for the proper operation of the ATLAS triggering system and subsequent data storage is the rate produced by a given trigger chain. The algorithm under evaluation is tuned so that the rate does not exceed that which can be handled by the ATLAS systems downstream of the trigger. To estimate the background rate, the number of events that pass selection, given a specific tuning, is evaluated on a sample containing the primary background, on which the thresholds of the algorithm are set to produce the maximum allowed rate for a particular trigger.

A data sample suitable for rate counting that represents the "background hum" inside the detector is usually a *zero bias* sample. It is collected by a purely random trigger, thus having no selection bias and hence its name. Alternatively, an MC background sample of QCD jets can be used. Still, if data is available, it is much preferable to MC because of multiple corrections that need to be applied to MC samples to reflect the real detector.

11.4 TAU TRIGGERING ALGORITHMS IN L1CALO

The RoIs that serve as inputs to calorimeter-based τ identification algorithms are *trigger objects* (TOBs), fixed-form calorimeter cell clusters. These are designed to be large enough to contain jets originating from hadronic τ decays. Based on the approximations of opening angles for such jets done in Section 3.2 and the order of the Molière radius mentioned in Section 5.2.4.1, a sufficient size would be of the order of 0.3×0.3 in $\eta \times \phi$ space.

For every bunch crossing, all possible calorimeter TOBs are analyzed by the L1Calo system using dedicated processors that look at a specific

region of the calorimeter. Logically, the algorithms work in three stages:

- **Seeding** - TOBs that have a maximum of E_T deposition in their center are considered, and the rest are ignored.
- **Discriminant computation** - compute an estimate of the E_T deposited in the full TOB as well as any additional discriminants.
- **Comparison against thresholds** - Comparison of the discriminants against thresholds passed as parameters. For the outputs of the algorithms, due to tight constraints on bandwidth, the full discriminants cannot be written out, and therefore, a few discrete working points are defined by sequences of thresholds. Usually, there are three working points for τ identification - loose, medium, and tight.

For each TOB containing a seed, an estimate of the E_T deposition within the TOB is used as the primary feature for its selection for further processing, with a configurable threshold. Additional discriminants that help separate a particular physics object from the background are then thresholded to reject background-originating TOBs.

Taus have a similar signature to QCD jets in the calorimeter, and distinguishing between the two is very challenging. To make things more difficult, the demanding latency and resource requirements on the L1Calo algorithms allow only the most basic features to be used as discriminants. The primary discriminant for identifying τ jets in L1Calo is based on the fact that τ jets are typically more collimated than QCD jets, with the energy deposition dropping off more sharply from the center of the TOB. Therefore, to improve the distinction of τ jets from QCD jets, only a single feature can be used, which is a measure of isolation around the highest E_T region.

Due to the change in readout granularity of the LAr calorimeter after the Phase-I upgrade, there is a difference in the definition of the TOB and the isolation between Runs 1 and 2 and Run 3. Consequently, the algorithms have changed as well.

11.4.1 *Tau Triggering in Runs 1 and 2*

During Runs 1 and 2 the TOB was a 4x4 cluster of 0.1×0.1 trigger towers, as illustrated on Figure 11.1. Each tower provides two values - the energy deposition in the EM part and the hadronic part,

originating from the EM and hadronic calorimeters, respectively. In the EM calorimeter, E_T values are summed for the towers in each of the four possible 1×2 and 2×1 sections within the central region. Additionally, the E_T within the towers in the central 4×4 region of the hadronic calorimeter is computed and added to each of the four sums. At least one of the resulting four values is required to pass a threshold.

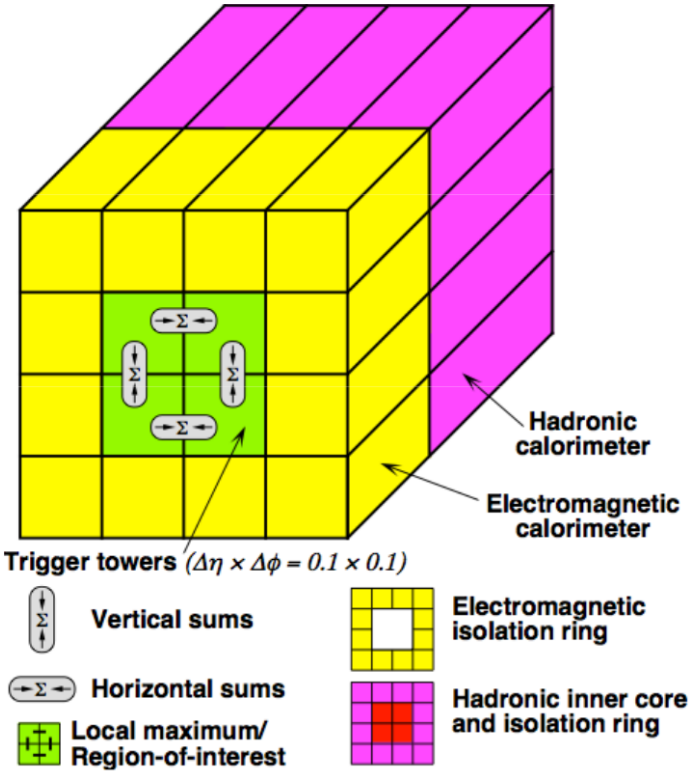


Figure 11.1: Trigger object of the calorimeter trigger for τ during Runs 1 and 2. [71]

For the isolation requirements, the E_T values in the 12 towers of the isolation ring surrounding the central 2×2 region are summed separately in the EM and hadronic calorimeters to produce the values $E_T^{\text{EM isol}}$ and $E_T^{\text{HAD isol}}$, respectively, and each of these values is required to be less than a programmable value. The isolation requirements for Runs 1 and 2 are detailed in Table 11.1.

11.4.2 *Tau Triggering in Run 3*

After the Phase-I upgrade, the amount of E_T values provided by each tower increased from 2 to 11, paving the way to finer algorithms. A baseline algorithm following the same principles as the one used in

	Isolation
Run 1	$E_T^{\text{EM isol}} < 4$ GeV
Run 2	$E_T^{\text{EM isol}}[\text{GeV}] \leq (E_T[\text{GeV}]/10 + 2)$ for $E_T < 60$

Table 11.1: Isolation requirements for τ triggering at L1 in Runs 1 and 2.

Runs 1 and 2 was designed and initially planned to be used. It will be referred to as the *Heuristic algorithm*. The second algorithm that is eventually used for Run 3, fully designed and implemented by the author of this thesis, is the *BDT algorithm*.

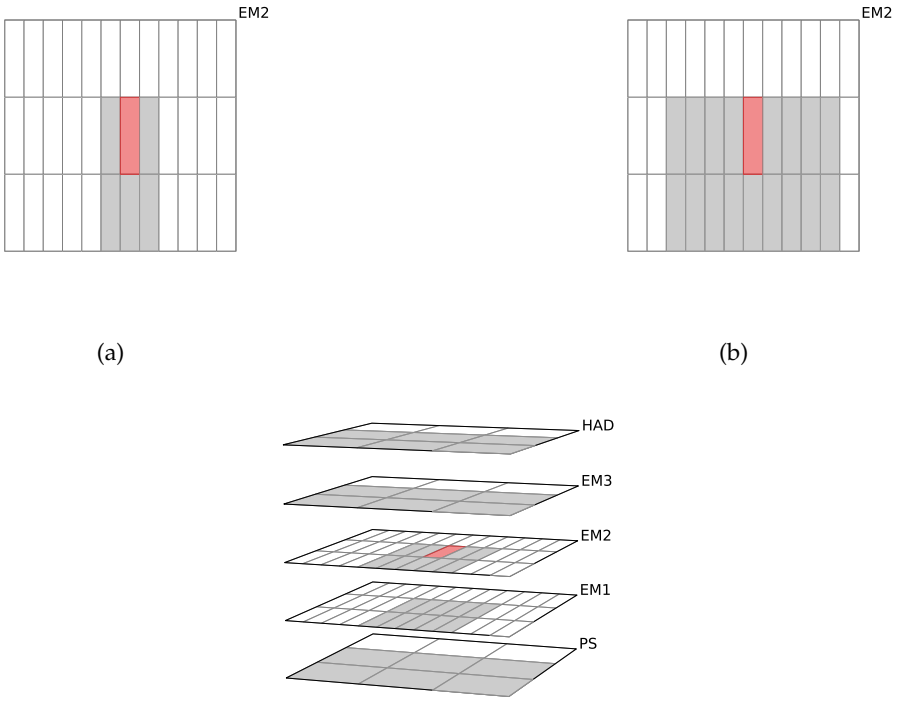
The Phase-I upgrade introduced several sub-systems to the L1Calo system in order to analyze the more granular picture provided by the LAr calorimeter, described in Section 5.2.4.2. Tau trigger chains in ATLAS start from two systems - eFEX and jFEX.

The eFEX system looks at TOBs comprised of 3×3 trigger towers, each with a size 0.1×0.1 and 11 supercells, as illustrated in Figure 11.2. It looks for *seeded TOBs*, which have a deposition of E_T in the central tower that is higher than the surrounding ones, computes a sum over a subset of the supercells to estimate the TOB E_T as well as an isolation measure internal to the TOB.

11.4.2.1 *jFEX Isolation*

The jFEX system, which is responsible for identifying jets, independently of the eFEX system and, among others, implements a *small-radius (SR) jet algorithm*, which scans the calorimeter and looks for larger regions of 5×5 trigger towers. It computes a sum over supercell E_T within a seed of 3×3 towers at the center and selects those 5×5 regions for which the seed is higher than the neighboring seeds. For each region, it computes two sums - $E_{T,\text{jFEX}}^{3\times3}$, the full sum over the 3×3 towers seed and $E_{T,\text{jFEX}}^{\text{ring}}$, a sum over the towers in a ring of $0.2 \leq R < 0.4$ around the seed, as illustrated on Figure 11.3.

Since the jFEX SR-jet seed is the same size as the eFEX TOB, the sum over the ring around the TOB can be used to compute an additional isolation measure to reject the less collimated QCD jets. Thus, in the L1Topo system, the eFEX TOBs and the jFEX 5×5 regions are matched, and a cut on this isolation is performed.



(a) (b)

(c)

Figure 11.2: Illustration of the Run 3 τ TOB structure and the supercells (colored) participating in E_T discriminant sums for the Heuristic algorithm. The red supercell is the small seed. (a) EM core sum, (b) EM environment sum, (c) E_T sum.

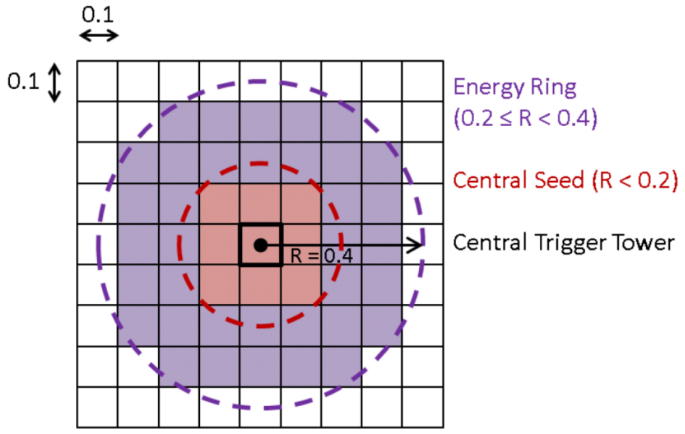


Figure 11.3: A small-radius jet, as defined by the jFEX. [32]

The jFEX isolation condition was initially computed and evaluated as

$$1024 \cdot E_{T,\text{jFEX}}^{\text{ring}} < R \cdot E_{T,\text{eFEX}}, \tag{11.1}$$

where $E_{T,\text{eFEX}}$ is the E_T estimate in the eFEX TOB as computed by eFEX, as will be described in Section 11.4.2.2, 1024 is a factor that facilitates the implementation in hardware, and R is the isolation threshold - the smaller it is, the more collimated a jet is. This computation, however, misses some isolation energy. As discussed in the following sections, the Run 3 eFEX algorithms compute $E_{T,\text{eFEX}}$ over a small cluster of supercells within a 3×3 TOB instead of the full sum over the TOB. Therefore, the E_T between this cluster and the jFEX isolation ring is lost when using the above definition. An improved scheme was proposed just prior to the beginning of data taking for the year 2024 in which the E_T in this lost region is considered by modifying the isolation condition:

$$1024 \cdot \left(E_{T,\text{jFEX}}^{\text{ring}} + \frac{A}{1024} E_{T,\text{jFEX}}^{3\times3} - \frac{B}{1024} E_{T,\text{eFEX}} \right) < R \cdot E_{T,\text{eFEX}} \tag{11.2}$$

Where A, B, and R are parameters. As an example, if $A = B = 1024$, (11.2) becomes (11.1), with $E_{T,\text{jFEX}}^{\text{ring}}$ replaced by $E_{T,\text{jFEX}}^{\text{ring}} + \left(E_{T,\text{jFEX}}^{3\times3} - E_{T,\text{eFEX}} \right)$, which is the E_T in the isolation ring around the small cluster in eFEX. Eq. (11.2) can further be re-written as

$$1024 \cdot E_{T,\text{jFEX}}^{\text{ring}} + A E_{T,\text{jFEX}}^{3\times3} < R' \cdot E_{T,\text{eFEX}} \tag{11.3}$$

where $R' = R + B$, reducing the tunable parameters to A and R'.

11.4.2.2 *Run 3 Heuristic Algorithm*

The Heuristic τ algorithm works by finding the supercell in the central tower's EM2 layer with the maximal energy deposition as illustrated in Figure 11.2. This is referred to as the *small seed*. A direction in ϕ is determined by looking at the energy deposition directly above or below the small seed. The higher value determines the ϕ direction. A 5×2 energy cluster in the ϕ direction and around the small seed is defined in the EM2 and EM1 layers, and a 3×2 cluster in the PS, EM3, and HAD layers. The E_T of the TOB is estimated by summing over these clusters. Additionally, an intra-TOB isolation measure is computed by taking in a similar fashion a 3×2 region and a 9×2 region around the small seed and in the ϕ direction determined previously, referred to as the *EMCore* and *EMEnv* sums, and illustrated on Figures 11.2a and 11.2b, respectively. An equivalent of an EMCore/EMEnv ratio between the two is then computed, forming the R_{core} discriminant [32].

11.4.2.3 *ML Model for τ Triggering in Run 3*

The granular picture of the TOB illustrated in Figure 11.2 allows the computation of discriminants that may give better separation of τ signal and background TOBs. Studying those variables by manually applying cuts is tedious and time-consuming. A natural approach in this case would be to use them as inputs to an ML model and to allow it to train on signal vs. background samples.

The score of the model needs to be used as a discriminant. Keeping in mind that any algorithm suggested at this late stage when the firmware is already written must be compliant with the firmware and the way this algorithm is expected to be utilized during data taking as well as the fact that the primary feature of interest used to select interesting TOBs is the E_T deposition in the TOB, the score can only replace the Heuristic algorithm's RCore discriminant, representing isolation. Therefore, the ML model should effectively provide a more performant isolation discriminant based on the fine-grained picture of the E_T depositions in the TOB.

There is a large variety of ML models to choose from. The two main constraints in the choice of models are the amount of resources available on the processors running the firmware and the strict latency requirements. Different models have different strengths and weaknesses. While some varieties of ANNs are potentially more performant, their power lies in automatically learning unspecified, high-level features

from low-level ones, which requires resource-heavy deep neural nets that take a long time to evaluate. Since these types of models are not feasible, it is necessary to engineer input variables, compute them, and then use them as inputs to a model. For this kind of problem, BDTs are often as performant as ANNs while being much less resource-heavy and much faster to evaluate, as will be explained in more detail in Section 12.1.2.

As the ML model of choice, a BDT model was hence chosen. In the 2024 configuration, the BDT has a maximum depth of three with nine trees and has as inputs 10 variables that represent sums at different distances around the center of the TOB and an additional variable that provides the scale of the total E_T deposited in it. During the training, the BDT learns cuts on the first ten variables at different E_T scales, providing a more sensitive isolation discriminant compared to the Heuristic algorithm's RCore. The algorithm's implementation details will be discussed further in the upcoming sections.

12

BOOSTED DECISION TREES ON HARDWARE

12.1 FIELD PROGRAMMABLE GATE ARRAYS

The strict LHC bunch crossing frequency of 40 MHz requires that all experiments are synchronized with it in order to collect data optimally. For that reason, the first stage of the triggering systems of general-purpose experiments such as ATLAS and CMS, which are required to trigger on a large number of signatures, resulting in very low latency requirements, are implemented in hardware and are controlled by a strict clock signal which must be synchronized with the bunch crossing frequency.

There are generally two kinds of integrated circuits that are capable of running clocked logic - *application specific integrated circuits* (ASICs) and *field-programmable gate arrays* (FPGAs). The main difference between the two is that the former is manufactured with the logic already present and immutable, while the latter contains a grid of logic elements that can be connected "in the field" to form logic circuits and is thus programmable and provides the best compromise between design flexibility and performance required by high energy physics experiments.

12.1.1 *FPGA Firmware Design*

Two essential components in an FPGA are *lookup tables* (LUT) and *registers* or *flip-flops* (FF). Ideally, any function can be implemented by combining multiple LUTs. However, connecting an arbitrary amount of LUTs to produce an immediate output assumes that the travel time of signals over the wires connecting the LUTs is zero. In reality, one must take care that logic signals at the inputs of the LUTs are stable long enough to get a reliable output. For that reason, the logic flow is fragmented into a sequence of stages by introducing FFs, which are storage elements controlled by a square clock signal between the stages. When the clock changes from a low to a high logic value, the FF releases the stored value and takes in the next value at its input.

This makes the flow of logic resemble a production line, where a logic operation is performed every clock cycle.

A digital signal processor (DSP) is an additional component that helps signal processing on an FPGA. These components are used instead of implementing signal processing operations such as multiplications using LUTs.

The description of the transfer of logic signals between the FPGA's components is a design abstraction called *register-transfer level* (RTL) ans is written using a *hardware description language* (HDL) such as VHDL or Verilog. A design implemented in HDL is synthesized into a file describing the exact wiring and locations of the physical components on the FPGA using a toolchain provided by the FPGA vendor, which can then be used to program the FPGA. The process usually consists of two stages, the first is the synthesis of the design, which converts the HDL code into a lower-level description of logic blocks and their connections and the second is the implementation of the design which selects the physical elements on the FPGA and determines the wires that connect them together.

Several real-world constraints need to be taken into account when designing a logic circuit on FPGAs. The most important and obvious is the amount of resources utilized by the design.

A further complication arises when a design is large enough and utilizes a sizeable fraction of the FPGA resources. In this case the vendor's toolchain may be unable to place some components of the design close enough together so that signals between them propagate within the expected time along the logic paths that connect them. Essentially, once a design is synthesized and placed on the FPGA, a minimum is computed over the estimated *slack* value of each logic path in worst-case working conditions, where the slack of a path is the required time it should take a signal to propagate along it by design and the actual time it takes in reality. A complex design that uses a lot of resources may have paths with negative slack, in which case the design is said to *fail timing*. Such a design cannot be used in reality as it may behave unpredictably.

Further constraints are energy consumption and heat dissipation, which this thesis will not discuss.

12.1.1.1 *Latency and Throughput*

An algorithm operating in an extremely low-latency domain, such as the ATLAS L1 trigger, must be fast enough to process the large

volume of incoming data and produce trigger decisions in synchrony with the LHC's 40 MHz collision rate. The temporal performance of such an algorithm is typically evaluated using two metrics: *latency* and *throughput*.

Latency is the amount of time it takes for an algorithm to process a single input. However, when processing a long sequence of inputs, the total processing time is not significantly affected by latency - if the latency is sufficiently low. Throughput, on the other hand, refers to how many inputs the algorithm can process per unit time. The maximum throughput at a given clock frequency corresponds to the algorithm consuming a new input every clock cycle.

The number of clock cycles between two successive input consumptions is called the *initiation interval* (II). Therefore, maximum throughput is achieved when $\mathrm{II} = 1$. For a long input sequence, increasing the II from, for example, 1 to 2 will approximately double the total processing time. As such, a key guideline for algorithm development in extremely low-latency applications - such as the ATLAS L1 trigger - is to maximize throughput as much as possible.

Increasing throughput is often referred to as *pipelining* the algorithm and takes advantage of the parallelism enabled by FPGAs. Certain operations that would typically be considered sequential in CPU-targeted algorithms can be parallelized on FPGAs. For example, in the context of a loop, if the iterations are independent, they can all be computed in parallel. In other cases, a subsequent iteration can begin execution before the previous one has completed. Such optimizations are generally ineffective on CPUs, which are inherently sequential in nature.

While beneficial for speed, increasing throughput often results in higher resource utilization on the FPGA. To understand why, one can imagine a logic path on the FPGA as a production line, where multiple processing stations handle incoming data packets, with each station representing a hardware resource. If the slowest station requires two clock cycles to process a packet, new packets must be introduced every two cycles, corresponding to an initiation interval of $\mathrm{II} = 2$. To reduce the II to 1, or fully pipeline the algorithm, the slowest station would need to be duplicated, thereby increasing resource usage on the FPGA.

12.1.2 *FPGA Implementation*

The structure of a BDT model allows it to be very efficiently implemented in a highly parallel design on an FPGA. Since the model is a collection of threshold comparisons and eventual summation of independent tree scores, the evaluation of thresholds and individual trees can be parallelized, producing a low-latency, fully pipelined design with the help of modern high-level synthesis (HLS) tools that convert logic written in human-friendly languages such as C++ to RTL.

The tool that is utilized in order to produce such a design for the τ algorithm is Conifer [72]. It has several back-ends to synthesize a VHDL design for a BDT, given a trained XGBoost model, of which the XilinxHLS back-end is used. It implements a two-stage process in which it first synthesizes C++ code implementing the BDT decision function from the trained model and then utilizes Xilinx's Vitis HLS tool to produce a highly parallelized, fully pipelined, low-latency implementation of the BDT.

Several parameters, such as the bit widths of the thresholds and scores, clock period, and clock period uncertainty, steer the design's synthesis. Vitis HLS takes care to produce an optimal design in terms of both resources and latency. Details of the generated RTL, such as the latency of the resulting design, which resources to use, and its interface, can be fine-tuned using special C++ pragma directives.

As an example, the resulting design of a single $M = 2$ tree is illustrated in Figure 12.1 for a simple case of an input feature vector of length three and three corresponding thresholds. All nodes in the tree are evaluated in parallel, and the tree is traversed, resulting in 2^M bits, each representing a path. Only one bit is 1, corresponding to the path taken in the tree, while the rest are 0. These bits are used to index a small lookup table which stores the scores of that tree. The thresholds and leaf scores are fixed in the logic of the FPGA firmware rather than being fetched from an external memory. This process is done for all K trees in parallel and the resulting scores are added using a balanced adder reducing the scores to their sum in a pair-wise tree structure [72].

The resulting score of the BDT, as implemented in hardware, does not have the usual sigmoid function applied to it, so its value is not constrained. Since the sigmoid is a monotonous function of its input and our purpose is to set simple thresholds on the score, this is an unnecessary step.

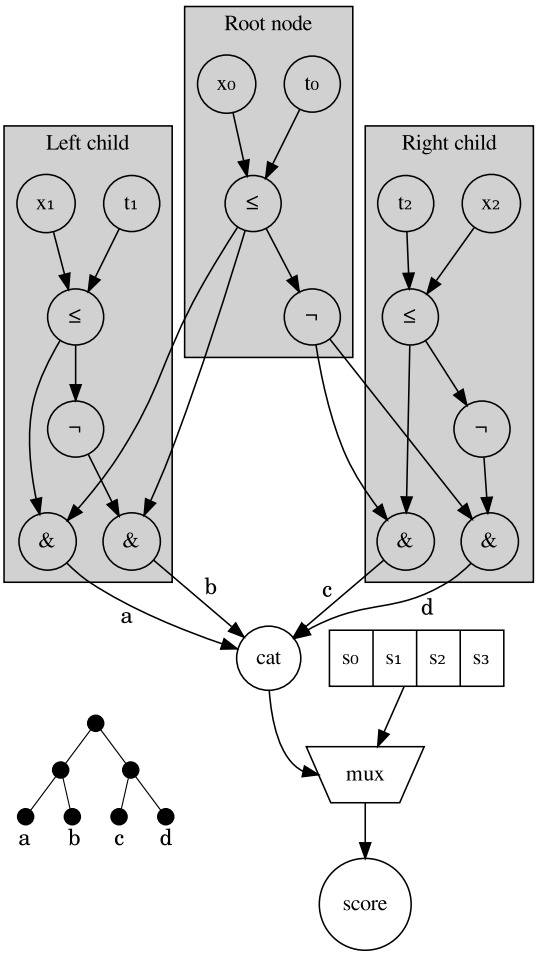


Figure 12.1: Schematic of the implementation of decision trees, showing a single tree with depth of 2. The x are the node features, and t the thresholds. The '¬' is the unary 'not' operator, and '&' is the binary 'and'. The Boolean leaf activations are concatenated and used to address a look-up table of output scores. The labels 'a', 'b', 'c', and 'd' on the schematic correspond to the respective labeled leaf nodes of the tree represented at the bottom left. [72]

12.1.3 *The Effect of Hyperparameters on the Synthesized Design*

The choice of HP affects the complexity of the trained model and, therefore, the two critical metrics of its synthesized design on hardware - resource usage and latency.

Timing closure is an additional important quantity that is difficult to estimate a priori from HP values and is only available after running full implementation and routing of the full firmware design, with the BDT model integrated, on the target FPGA.

Resource usage and timing closure heavily depend on the number of nodes N in the BDT, where a node is either a single comparison operation of a feature with a threshold or a leaf of a tree.

The number of trees K, the tree depth M, and the pruning regularization parameter γ, in turn, have an effect on N. For an unpruned tree,

$$N = K \cdot \left(2^{M+1} - 1\right), \tag{12.1}$$

and the effect of increasing the γ parameter is to reduce the number of nodes in an a priori unpredictable way.

The latency of the BDT model is affected by both M and K, and this effect is somewhat predictable. During the evaluation of the input, the tree must be traversed from the root to the leaf node, which, for an unpruned tree, takes M sequential stages and can be done in parallel for all trees. Then, all leaf scores must be added, which takes $\log_2 K$ sequential stages of addition. This results in a total latency of approximately $M + \log_2 K$ sequential stages. The HLS tool often optimizes the design to reduce the overall latency by placing multiple stages in a single clock cycle, making the actual latency of the model's design difficult to predict a priori, but this approximation allows for estimating an upper bound.

13

THE EFEX BDT τ ALGORITHM

The rationale for the choice of a BDT as the model of choice and its usage for triggering was briefly discussed in Section 11.4.2.3 and is, in hindsight, fairly trivial. However, converging to the final algorithm used in 2024 data taking, referred to as "BDT v16", involved a long process of trial and error with many successes, failures, and direction changes, involving different ML models, training methods, and input variables.

The multi-year process of developing the algorithm can, in retrospect, be divided into two stages. The first one was an exploratory *development stage*, during which the tools and data used for exploring the various possibilities and approaches to BDT training, evaluation, and firmware implementation were adequate but did not fully represent the real-world systems and methodologies in ATLAS, partly because Run 3 has not yet started and a lot was still under development. The second was a *commissioning stage*, during which the algorithm was integrated within ATLAS simulation and firmware and proper methodologies were applied for its evaluation and tuning, with the involvement of ATLAS physicists and engineers from the L1Calo system and the ATLAS τ trigger group.

This chapter describes both the progression of the algorithm during the development stage, summarizing the main directions that were studied before converging to the current state of the algorithm until Section 13.4.2 and continues to describe the algorithm in its current state.

13.1 CHOICE OF MODEL

Due to its simplicity relative to other models, the idea of using a BDT as the model of choice was fairly clear from the beginning, though some attempts at training ANNs were made. The initial plan was to use a deep BDT with many highly performant variables in order to improve the low-pt efficiency of the tau identification algorithm in L1Calo (Figure 13.1), where the identification of tau leptons is difficult due to the overwhelming amount of QCD jet background, as can be seen on Figure 13.2, and where it was believed that a ML approach

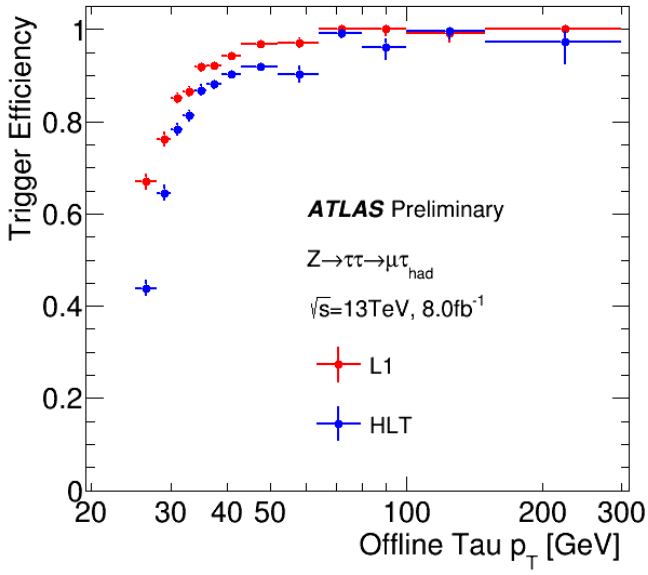


Figure 13.1: Efficiency as a function of the p_T of the reconstructed tau jet in Run 2 for L1 and HLT [73].

would increase the efficiency of identifying taus. In parallel, feasibility studies of running different models on hardware were performed.

Initially resource and latency considerations did not play a big role and the goal was to see how well a BDT can separate signal from background. The first BDTs tested were very deep and wide - up to max. depth of eight and 200 trees, with good performance relative to that of the Run 2 algorithm on the same dataset.

13.2 TRAINING AND PERFORMANCE EVALUATION

The training during the development stage was done on MC16 Monte-Carlo samples with $Z \rightarrow \tau\tau$ as signal and di-jet samples as background, described in more detail in Section 13.8. The performance during this stage was measured by producing a turn-on curve, assuming that the only tunable threshold for the algorithm is the BDT score, while the TOB E_T was not thresholded.

This way of producing the turn-on curves used during the development stage is simplified and does not fully reflect how they were produced during the commissioning stage. It was done this way because the initial idea was to use the BDT score as the primary discriminant instead of the TOB E_T.

For comparison, the Run 2 (also referred to as the Legacy) turn on was also computed and was based on the thresholds on the sub-

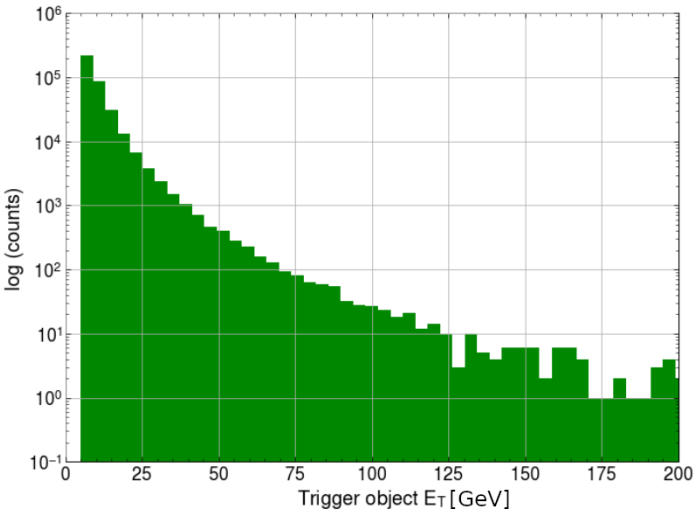


Figure 13.2: Distribution of p_T for TOBs in a di-jet MC sample.

leading tau TOB of the primary di-tau trigger used during Run 2, requiring a minimum TOB E_T of 12 GeV and isolation of $E_T^{\text{EM isol}}[\text{GeV}] \leq (E_T[\text{GeV}]/10 + 2)$ [74]. A fixed early estimate of $E_{T,\text{eFEX}}/E_{T,\text{jFEX}}^{\text{ring}} = 0.275$ for the jFEX isolation, described in Section 11.4.2.1, was applied for the BDT turn-on curves. This value was taken from earlier performance studies done on Monte-Carlo samples with a center-of-mass energy of 13 TeV. During the commissioning stage, the methodology was to switch to the TOB E_T as the primary discriminant, and the jFEX isolation, along with the BDT score applied as additional cuts to reduce the rate while keeping the same efficiency.

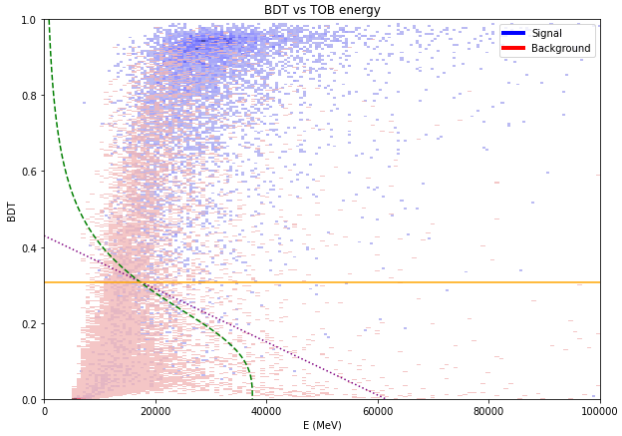
When comparing several turn-on curves meant to represent the efficiency of several algorithms (or the same algorithm with different threshold settings), the algorithms are first tuned so that they produce the same background rate, which is required to be equal to the background rate during Run 2, since Run 3 data was not yet available during the development stage. During this stage, the background rate was defined as the number of events that contain at least one TOB that passed the algorithm's threshold. As discussed in Section 13.7.3, during the commissioning stage, a more realistic requirement of events containing several TOBs and other objects is required for background rate estimation.

13.3 COMBINED CUTS

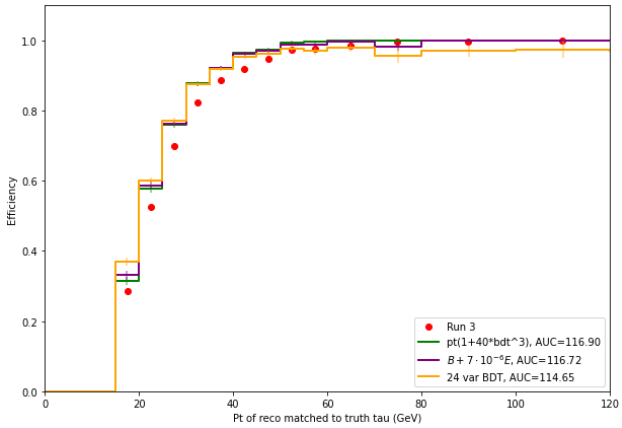
After a BDT score is produced, the most trivial way to apply a threshold is to apply it directly to the BDT score. However, some boost in performance can be gained from combining the BDT score with other variables. For example, those that take too long to compute and therefore cannot be used directly as inputs to the BDT, or the fact that shallow BDTs are used, so cuts on specific variables are omitted by the training algorithm in favor of the more powerful variables, so combining other variables may be an interesting exercise.

The effect of combining the BDT score with the TOB E_T computed by summing over all the supercells in the TOB was studied by performing linear and non-linear combinations of the two and using the combination as the discriminant instead of the BDT score to produce a turn-on curve. An example of a comparison in performance of a direct cut on BDT, as well as linear and non-linear combinations, is shown in Figure 13.3. Figure 13.3a shows the effect of the combination on the turn-on curves. The red dots are an approximation of the turn-on curve of the Run 3 Heuristic algorithm, used as a baseline here, which was computed in an independent study. All three models result in the same background rate. The effect is mostly on the low- and high-p_T regions, where a combination cut of this type loses signal TOBs at low-p_T, reducing the low-p_T performance while improving it at high-p_T by accepting more signal events in that region.

The study demonstrates the potential benefits of using a combination. For example, a model that performs much better at low p_T but worse at high p_T relative to a baseline algorithm can be combined with the TOB E_T in this manner, sacrificing some performance at low p_T but bringing it up to that of the baseline algorithm for high-p_T to produce an overall better model relative to the baseline. This effect is, however, smaller when the TOB E_T and the BDT score are correlated, which is always the case since the BDT picks up the strong dependence of the likelihood of a TOB being signal on its E_T, even though the E_T is not directly used as the input variable to the BDT. Applying a combination cut also complicates implementation in firmware and introduces additional parameters to be tuned - the coefficients of the combination. It was therefore decided not to investigate this direction further, though future studies could prove beneficial.



(a)



(b)

Figure 13.3: The effect of combining a BDT score with the full TOB E_T. (a) Distribution of signal and background TOBs in the BDT score-TOB E_T plane. The horizontal orange line is a cut on the BDT score only while the purple dotted line is a cut on a linear combination of the BDT score and TOB E_T. The green curve is a non-linear combination that attempts to "surround" the background. The resulting same-rate turn-on curves are shown in (b). The red Run 3 curve is an approximation of the Run 3 Heuristic algorithm.

13.4 INPUT VARIABLES

There were initially a total of 24 variables, which were highly performant but very expensive to compute since they were based on divisions and summations over large areas. Some example variables that illustrate their complexity are listed in Table 13.1. The corresponding distributions are shown on Figure 13.4.

Variable	Definition	Description
E^{TOB}	$\sum_{l,\eta,\phi} E_{l,\eta,\phi}$	Full TOB E_T.
Width	$\frac{\sum_{l,\eta,\phi} E_{l,\eta,\phi} \cdot R(\eta,\phi)}{E^{\text{TOB}}}$	Measures how focused the energy is in the center of the TOB. $R(\eta,\phi)$ is the distance of cell (η,ϕ) from the TOB axis.
Density	$\ln\left(\frac{\sum_{l,\eta,\phi}\left(10^3 \cdot E_{l,\eta,\phi}\right)^2 / V_l}{E^{\text{TOB}}}\right)$	Density measure of the energy in the TOB. V_l is the volume of a supercell in layer l.
MaxRatio	$\frac{\max_2 E_{2,\eta,\phi}}{\max_1 E_{2,\eta,\phi}}$	Ratio between the highest and second highest energy depositions in the EM2 layer.
MaxF2	$\frac{\max_2 E_{2,\eta,\phi}}{E_2}$	Ratio between the second highest energy deposition and the energy in EM2.

Table 13.1: Some variables used as initial inputs to the BDT model. The index l runs over all layers while η and ϕ take values of zero up to the maximum index of supercell in the respective direction, $E_{l,\eta,\phi}$ is the E_T in layer l and position (η,ϕ).

In addition to the 24 variables, other variables were evaluated based on summations over permutations of supercells and weighted sums of the supercells, described in the following sections.

13.4.1 *Energy Flow Polynomials as Variables*

One of the ideas for input variables was inspired by Energy Flow Polynomials (EFPs) [75] and motivated by the assumption that sums

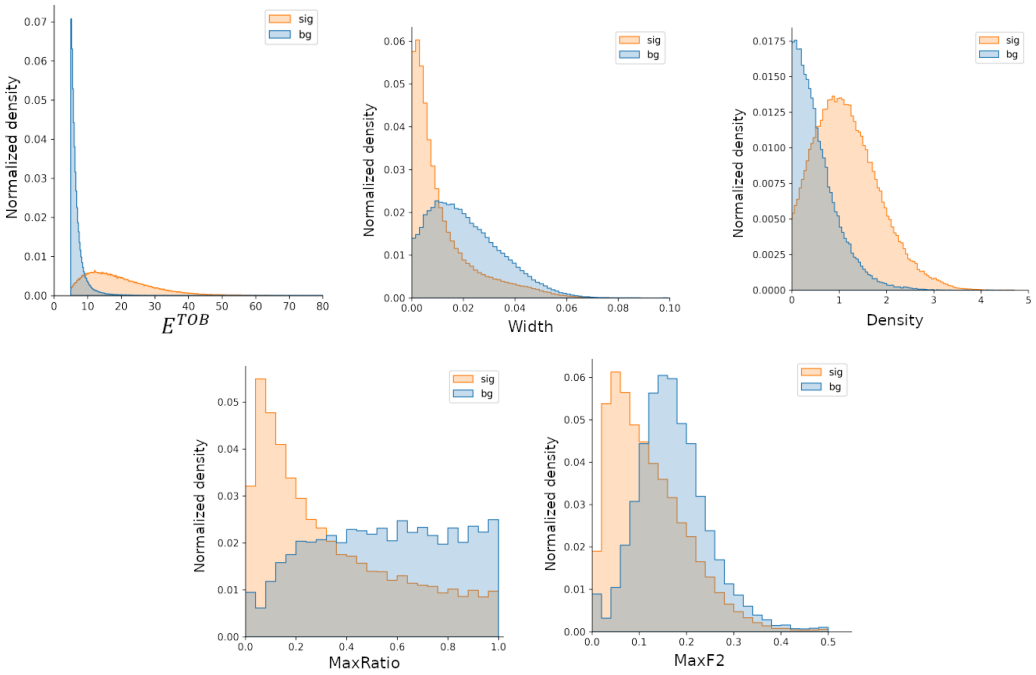


Figure 13.4: Distributions of the variables described in Table 13.1.

over specific permutations of cells within specific TOB layers might carry information about the τ jet's internal structure, such as the number of prongs or the photon decay products from neutral pions in the EM1 layer.

EFPs are functions of the form

$$\mathrm{EFP}_G = \sum_{i_1=1}^{M} \cdots \sum_{i_N=1}^{M} z_{i_1} \cdots z_{i_N} \prod_{(k,l)\in G} \theta_{i_k i_l} \tag{13.1}$$

where G is a multigraph (i.e. a graph with at least zero edges between any two vertices) with N vertices and edges $(k,l) \in G$,

$$\begin{aligned} z_a &\equiv \left(\frac{p_{Ta}}{\sum_{b=1}^{M} p_{Tb}} \right)^{\kappa} \\ \theta_{ab} &\equiv \left(\Delta\eta_{ab}^2 + \Delta\phi_{ab}^2 \right)^{\beta/2}, \end{aligned} \tag{13.2}$$

$\Delta\eta_{ab}, \Delta\phi_{ab}$ are the distances in η, ϕ, respectively, between supercell a and b, p_{Ta} is the energy deposit in supercell a, M is the total number of supercells, and κ, β are parameters. These functions can be graphically represented by multigraphs. For example:

$$= \sum_{a=1}^{M} \sum_{b=1}^{M} \sum_{c}^{M} z_a z_b z_c \theta_{ab} \theta_{ac} \theta_{bc}^2 \tag{13.3}$$

Two things should be noted. First, these functions are invariant under permutations of the vertices due to the summation of all possible

permutations, which reflects the nature of depositions of energy in a TOB resulting from a jet with multiple constituents. Second, these functions encode a relation between the vertices. For instance, in (13.3), the angular proximity between vertices b, c is more critical than that between a, b or a, c. These two features potentially allow probing various configurations of the energy depositions within the TOB.

In one study, five arbitrary EFPs were selected and generated, and a BDT was trained on the 18 most highly ranked of the 24 original variables in addition to the five EFPs, resulting in an improvement of the turn-on curve in the low-p_T region, as shown on Figure 13.5. The signal-background distribution for each EFP, their multi-graphs, and their κ and β coefficients are shown in figure 13.6.

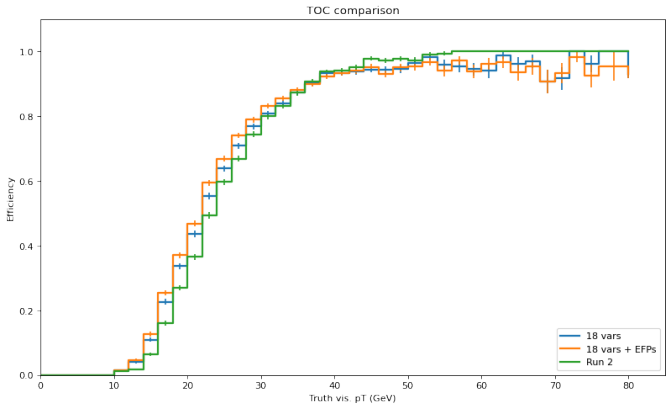


Figure 13.5: Improvement in low-p_T of the BDT performance due to adding five EFP variables. The green curve is the Run 2 algorithm performance, blue is the original BDT performance and orange is the performance of the BDT trained on the original variables plus the five EFPs.

The EFP variables, as many of the other variables, while boosting performance, were dropped at later stages because of their prohibitive resource and latency requirements when implemented on hardware.

13.4.2 *Weighted Sum as a Variable*

A weighted sum of the supercells in a TOB was another variable that was shown to boost performance. The idea for assigning weights to sections of the TOB and summing to produce a discriminant originated from a coding error. During the computation of the total E_T in a TOB to use it as the primary discriminant to produce a turn-on curve, one of the layers was accidentally counted twice, and the resulting turn-on curve was slightly better.

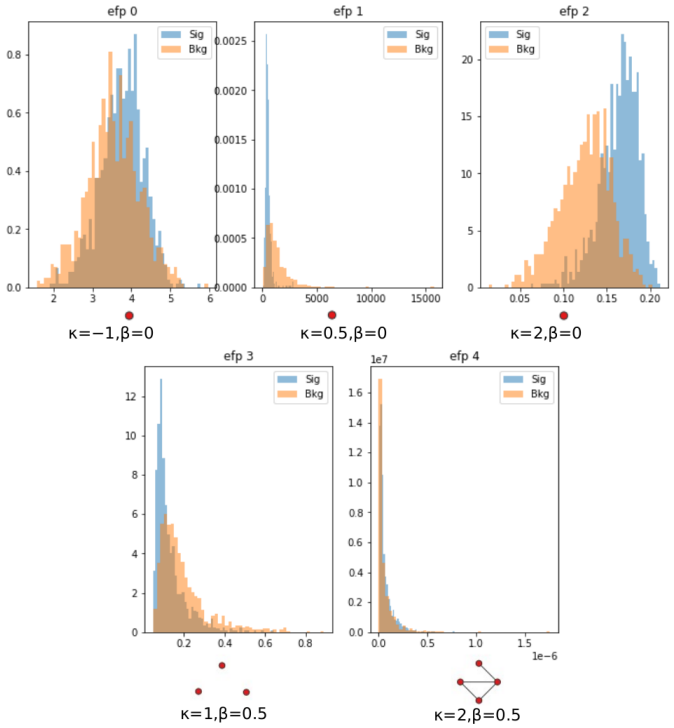


Figure 13.6: Signal-background distribution of five arbitrary EFP variables. The multi-graphs and coefficients describing them are shown under each plot.

In order to converge to the optimal weights of the weighted sum, a turn-on curve was produced, assuming a weighted sum model that produces the weighted sum discriminant as the output. The area under the turn-on curve at low-p_T for this model was used as a performance metric, and the weights randomly varied. The weights that produced the maximum area were selected. The scanning for optimal weights was done in iterations since it was not computationally possible to scan over the full space of 99 weights (one per supercell). At each iteration, the TOB was divided into a manageable number of three groups of cells in a symmetric way around its axis, assigning a weight to each group and selecting the triplet of weights that maximized the area metric. In the next iteration, a different segmentation into three groups was done, and the optimal weights of the first iteration were multiplied by those of the second iteration, and so on. The weights in the various optimal sets that resulted from these scans were usually falling off radially from the TOB axis, as seen in figure 13.7. Supercells at the periphery of the TOB had low and even negative weights, while central supercells had high weights, with depositions far from the center reducing the score, making the TOB more background-like.

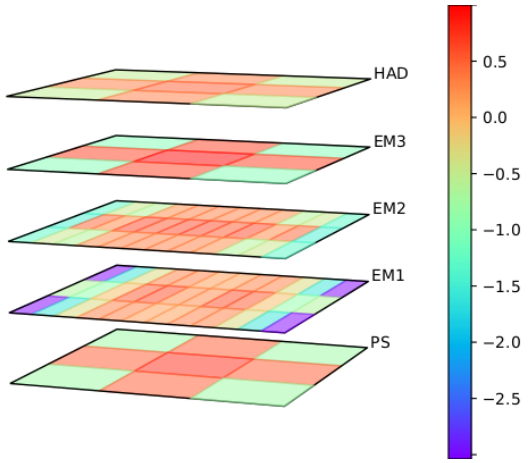


Figure 13.7: Optimal weights that maximize the area under a turn-on curve produced by assuming a weighted sum as the τ identifying discriminant.

Using this variable with the optimal weights as an input to the BDT slightly improved its performance. However, it was soon discarded, as producing a weighted sum in hardware was not feasible within the available resource and latency constraints. HLS-generated designs for evaluating a weighted sum of 99 supercells far exceeded the latency requirement of the τ algorithm in the eFEX system, which was limited to 12 cycles at 200 MHz, as will be mentioned in Section 14.1.2.1.

A possible future direction for the weighted sum could be to investigate the impact of weight quantization on latency - an aspect not explored by the author. Nonetheless, the image in Figure 13.7 sparked the idea for the final set of input variables, which are described in detail in Section 13.4.3.

13.4.3 *Symmetric Sums*

As previously discussed, many choices of variables are available that demonstrate good performance, however the latency and resource constraints render these variables unusable in reality. As a result, an alternative set of variables that capture internal TOB structure while being quick and less resource-intensive to compute was required.

Since the TOB is essentially symmetric around its axis, in the sense that, given a TOB with E_T depositions, the score of an algorithm meant to detect a τ-lepton signature in the TOB should not change if the TOB is arbitrarily rotated around its axis, it is reasonable to estimate E_T depositions at different radii around the TOB axis at different layers

and then train a BDT to classify signal vs. background with these values as inputs.

The finest granularity for probing the E_T deposition as a function of the distance from the TOB's axis, given the current readout granularity of the EM LAr and hadronic calorimeters, can be accessed by producing symmetric sums - sums over supercells located at the same distance from the TOB axis at a specified layer. There is a total of 33 such sums, as illustrated on Figure 13.8. The supercells used to compute each symmetric sum are summarized in Table 13.2, while the indexing convention of the supercells is described in Figure 13.8. The name of each variable has two parts separated by an underscore. The first part, preceded by 'l', is the layer index, and the second part, preceded by 'd', is the $\Delta R = \sqrt{\Delta\eta^2 + \Delta\phi^2}$ between the center of the layer and the center of any supercell in the sum, rounded to four digits after the decimal point and multiplied by 10000.

As each variable is constrained to a single layer, an additional variable estimating the total E_T in the TOB allows the BDT to automatically optimize the learned cuts on the symmetric sums as a function of the energy scale. To minimize resource usage, the E_T deposition in the core tower was selected and is henceforth referred to as CORE. With this variable and the symmetric sum variables as inputs, the BDT score essentially encodes an elaborate measure of isolation of the E_T deposition in the TOB as a function of its energy scale.

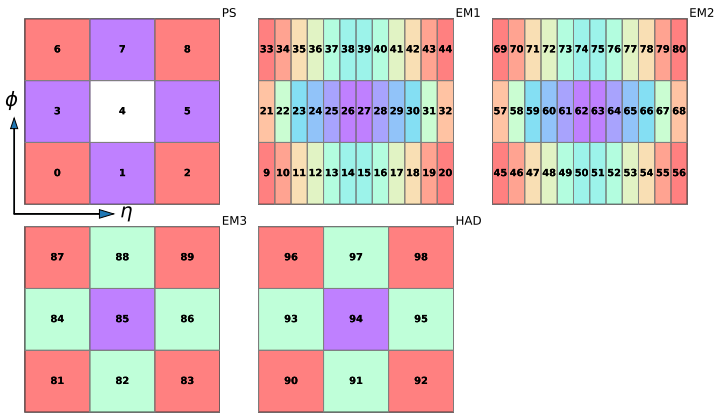


Figure 13.8: An illustration of possible variables for BDT inputs based on sums over groups of supercells equidistant from the TOB axis. For each layer, every such group is marked by the same shade. The numbering convention of all supercells is shown, as well as the increasing η and ϕ directions.

13.5 TRAINING OF THE BDT

BDT v16 is trained on individual TOBs from signal and background samples extracted from $Z \to \tau\tau$ MC and a zero bias data sample from a 2023 run, respectively. A summary table of the dataset and the hyperparameters is given in Table 13.3. The datasets used are described in more depth in Section 13.8.

The signal + background dataset was randomly split in two - 80% for training, 20% for test, preserving the integrity of the events. The train set was further split into 80% for training and 20% for validation used to evaluate the training for early stopping by XGBoost. To converge on a minimal set of the strongest variables, the BDT was trained in iterations, starting with the full set of 34 variables, performing variable ranking, removing the weakest variables, and re-training on the trimmed variable set, as further detailed below.

Following the discussion in Section 4.3.2 and given that the full evaluation of the BDT working point must be done within 12 clock cycles at 200 MHz (see Figure 14.2 for a timing diagram), during the development of the algorithm it was decided that the space of possible HPs shall be limited to a domain such that the latency of the resulting algorithm does not exceed seven clock cycles, leaving enough time for input variable computation and BDT condition estimation. In addition, the resource usage on the FPGA shall be low enough for the full firmware, with the BDT algorithm integrated, to meet timing requirements, as described in Section 12.1.1.

Within these constraints, an XGBoost model of depth three and nine trees was trained using the Python XGBoost package version 1.5.2 with five early stopping rounds. The values for max. depth and number of trees were selected to give the best ROC curve while constraining the total node count to at most 200, but lower values of the latter were preferred due to issues of the final design's difficulty in meeting timing. Since the dataset is unbalanced, XGBoost's scale_pos_weight hyperparameter with a value of 3.43 was used to compensate. It is the ratio between the size of the majority signal class and the minority background class, following the recommendation in XGBoost's online reference.

The training is visualized in Figure 13.9. The plots were obtained by training a model while allowing the number of trees to go up to 50. When training a BDT using XGBoost, each tree compensates for the errors of the sum of the preceding trees, so the number of trees is equivalent to the number of training iterations.

A low number of total nodes in the BDT of around 135 was selected to facilitate its implementation on the FPGA with no timing violations. At this value, the error function does not plateau yet, and the AUC is still increasing, so, in principle, selecting a model with more nodes would result in better performance. However, it will be shown in Section 13.7.4 that compared to the heuristic algorithm's performance, this model is already superior while being less resource intensive, as shown in Table 14.3. Potentially better hyperparameters for the BDT are explored in Section 13.5.2.

The ROC curve of the final model with the number of trees constrained to 9 is shown on Figure 13.10b, evaluated on subsamples of the test set with different minimum cuts on the TOB E_T applied while the BDT distribution for signal and background is shown on Figure 13.10a. The AUC of the final model on the full test set is around 0.77.

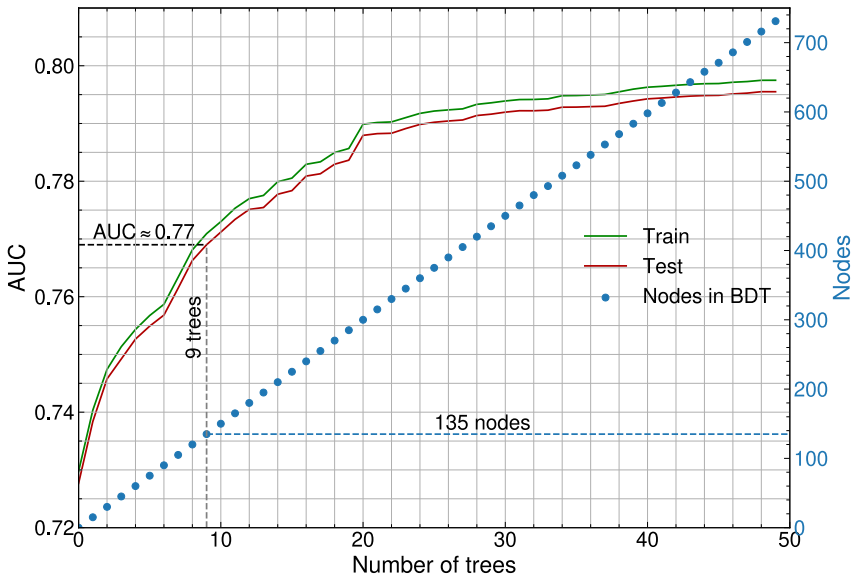
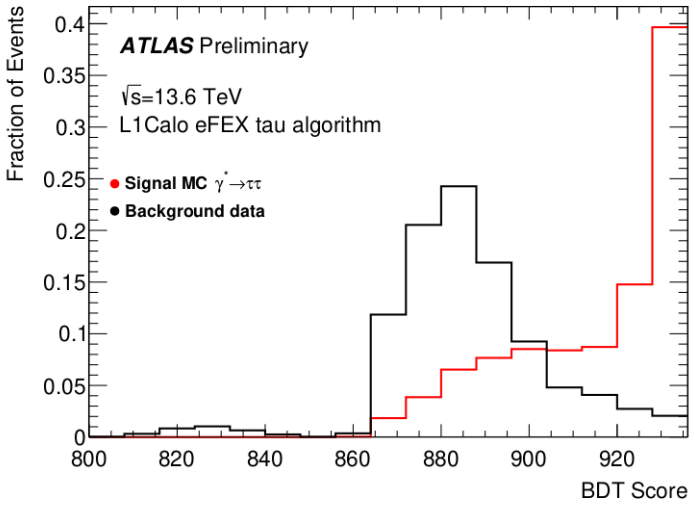
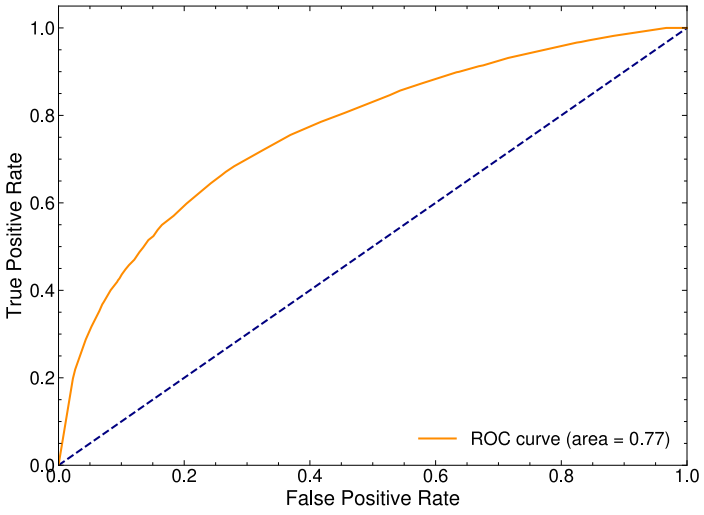


Figure 13.9: Training of the BDT algorithm. The AUC metric is on the left y-axis, and the number of nodes in the BDT is on the right y-axis as a function of the number of trees. The selected number of nodes is shown along with the number of trees and the resulting AUC of the trained model at that point.



(a)



(b)

Figure 13.10: Evaluation of the final BDT model. 13.10a Separation between signal and background samples in the test dataset [76], 13.10b ROC curve of the final model.

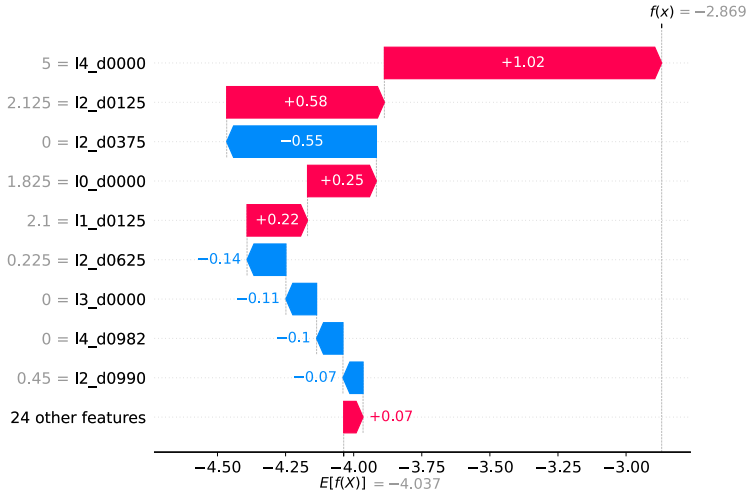
13.5.1 *Variable Ranking*

A selection of the strongest variables is made using computation of SHAP values [77], which are useful to understand how strongly different values of the variables contribute to the final score and simultaneously to explain what a model had learned.

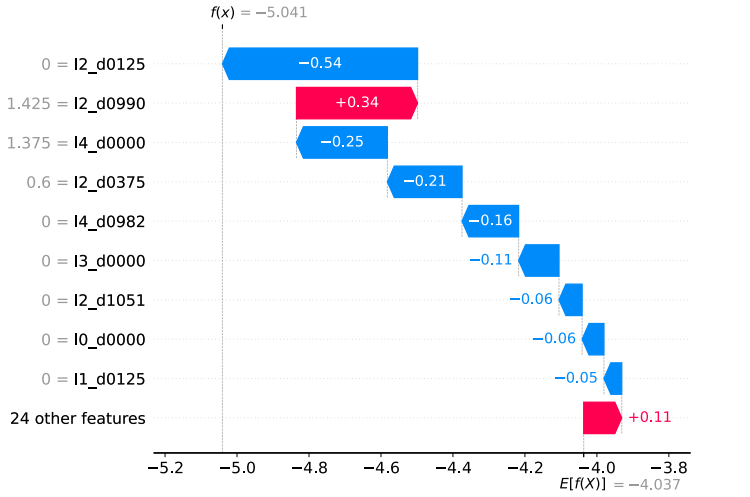
Given a single input vector of BDT variables, the computed SHAP values of its components quantify the contribution of each variable to the BDT score of the vector, which can be expressed as the sum of the SHAP values of the variables, as shown in Figure 13.11. In 13.11a, high E_T depositions in the EM2 layer and the hadronic layer give a significant contribution to the TOB's high score of -2.869 while the lack of E_T deposition in the cells comprising the symmetric sum l2_d0375 reduces the score. In 13.11b, the lack of E_T deposition in the center of the EM2 layer and a low E_T deposition in the hadronic layer push towards a lower score while the presence of a significant E_T deposition in EM2 pushes the score higher. The negative score of the BDT in this example is the consequence of not applying the sigmoid function to the BDT score.

Once the SHAP values are computed, a feature ranking can be constructed by computing the median of its SHAP values over a set of input vectors for each variable. An explanation of the model can be visualized by plotting all SHAP values for all input vectors in a dataset as shown in Figure 13.12a.

The variable selection is done iteratively. The result is a set of 11 variables shown in Figure 13.12 along with their ranking and SHAP value distributions. A BDT model is first trained on the full set of $33 + 1$ variables, where "+1" stands for the additional CORE variable. At each step, a simplified turn-on curve (used during the development stage, as described in Section 13.2) is produced on a validation set along with a Run 2 reference curve. The variables with SHAP scores close to zero are dropped, and the BDT is re-trained on the remaining variables. This process is repeated several times, until a further removal of variables either has a significant reduction in efficiency or if the BDT curve falls below the Run 2 reference curve. The resulting SHAP scores and the position of the variables in the TOB are shown in Figure 13.12.

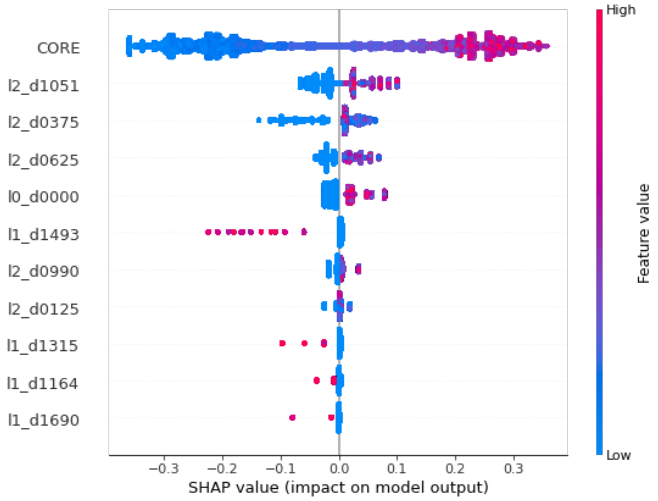


(a)

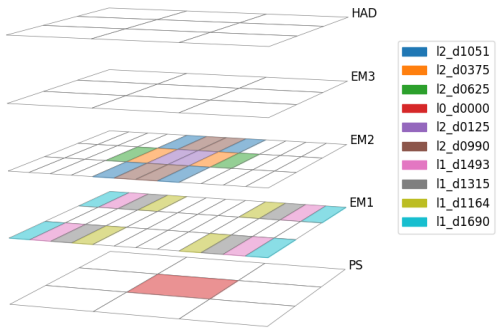


(b)

Figure 13.11: Example of TOB explanation with SHAP values for two TOBs - (a) matched to a truth tau from a $Z \to \tau\tau$ MC sample and (b) one from a di-jet MC sample, computed using a separate training. The names and values of the strongest variables are listed on the left side of each plot and the contribution of each variable to the final BDT score $f(x)$ is given by the arrows and their annotations.



(a)



(b)

Figure 13.12: (a) A bee swarm plot explaining the trained BDT model in terms of SHAP value distributions of the final set of 11 input variables. The variables are ordered by their classification power. The range of values of each variable is colored according to a blue-to-red gradient color map, with blue corresponding to low and red-to-high values. Negative SHAP values reduce the BDT score, while positive ones increase it. (b) Illustration of the 11 BDT variables in a TOB.

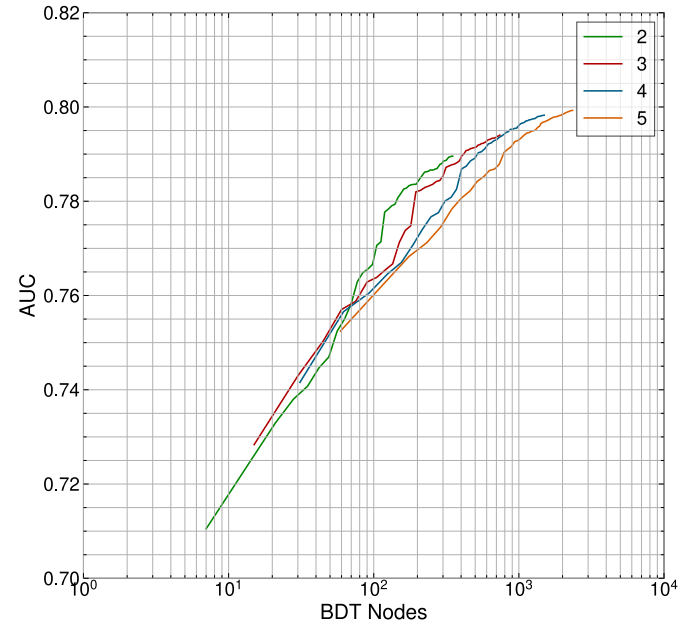
13.5.2 *Effect of Hyperparameter Choice on BDT Performance*

During training and hyperparameter selection, the main focus was on the feasibility of the project. In the late development and early commissioning stages, the primary concerns were focused on technical implementation details and the ability to outperform the Heuristic algorithm while having a lower resource footprint. At this experimental stage of using ML algorithms in the eFEX firmware, the process of re-training, integrating the BDT in the firmware, and subsequent performance evaluation and hardware tests required substantial time and effort. This, along with the BDT becoming a viable option very close to the start of Run 3, resulted in selecting a BDT with a performance slightly below optimal to be improved in future re-trainings. This section presents a recent study showing how the choice of hyperparameters affects the performance and where the selected BDT is situated relative to the optimal one.

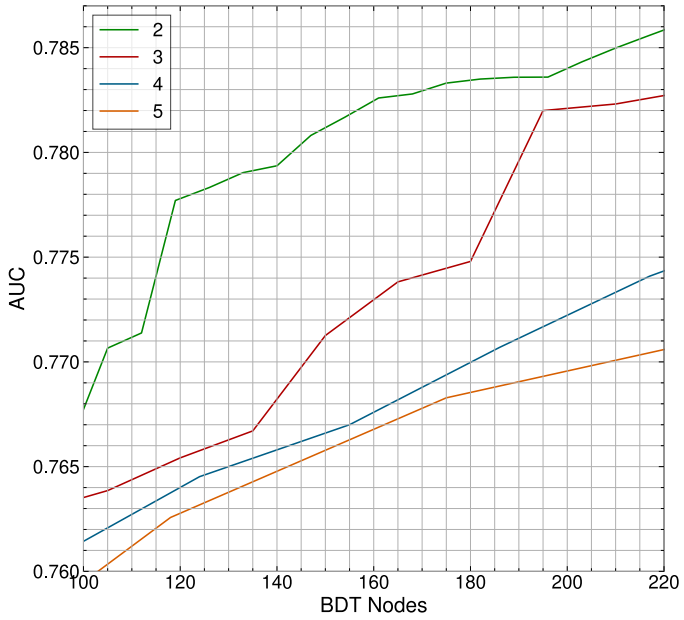
The primary hardware constraint for the BDT hyperparameters comes from resources. As discussed in Section 4.3.2, the resource usage heavily depends on the number of nodes in the BDT, and the higher the resource usage, the higher the likelihood of the design to fail timing. From experimentation, the cutoff value for the number of nodes above which the full firmware with the BDT is likely to fail timing is around 200.

In this study, several hyperparameters are varied independently - max. depth, the pruning parameter γ and scale_pos_weight. The effect of their different values on the model's performance, quantified by the AUC metric, is observed as a function of the training iteration, which is roughly proportional to the number of nodes, since a new decision tree is added every iteration.

The strongest effect was observed by varying the max. depth hyperparameter. The AUC vs. number of nodes for this variation is shown in Figure 13.22. Other hyperparameters showed negligible effect on the AUC metric. From 13.13a it is evident that the deeper the tree, the more performant it can get given a large enough number of iterations translating to a large enough number of nodes. However, 13.13b shows that in the region where the number of nodes is within our range, i.e., below the cut-off point for timing failure of $\sim$ 200 nodes, the shallower model is more performant.



(a)



(b)

Figure 13.13: (a) AUC of different models with different values of the max. depth hyperparameter over a large range of nodes. (b) The same AUCs, zoomed into the relevant range of nodes, from 100 to slightly above the timing failure cutoff value of 200.

13.6 ESTIMATION OF TOB E_T

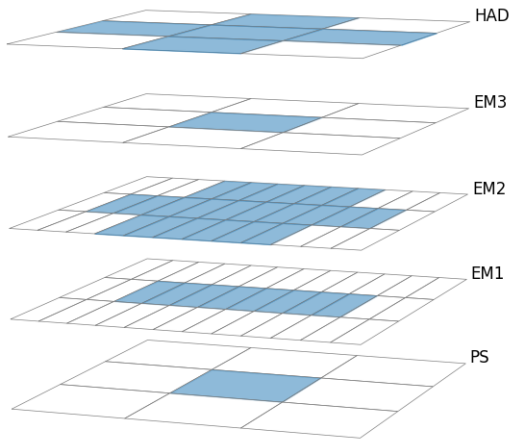
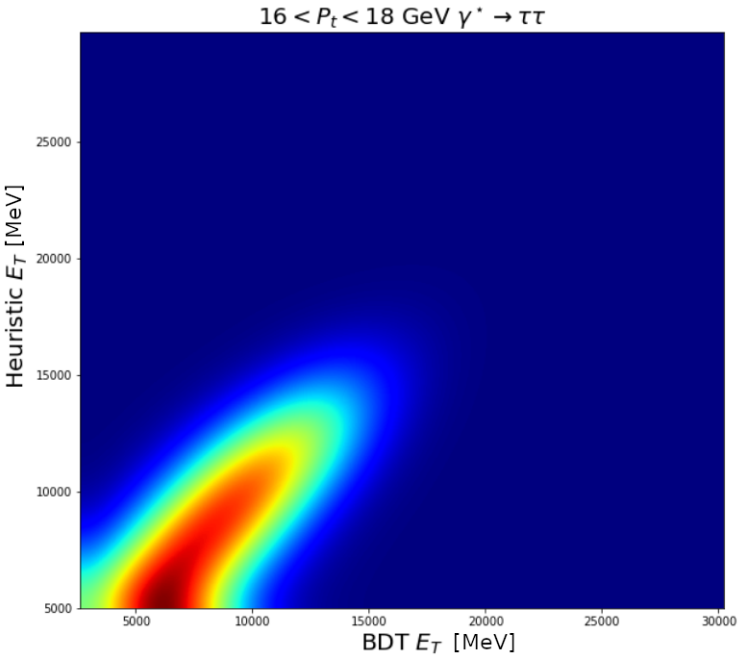


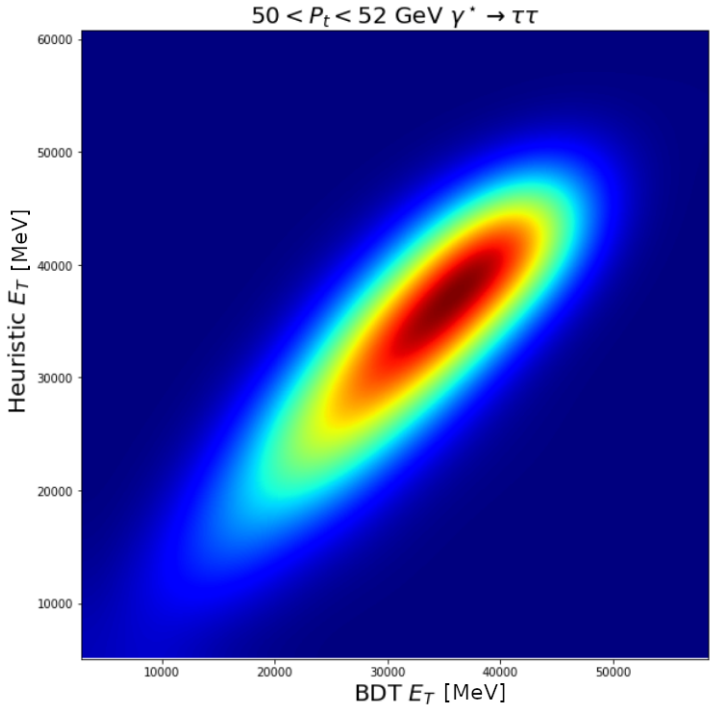
Figure 13.14: The cluster of supercells over which the BDT v16 algorithm sums to produce the TOB's E_T estimate.

The BDT τ algorithm, unlike the Heuristic, estimates the sum of the E_T within the TOB by summing over a fixed set of supercells. The Heuristic algorithm's computation of E_T is summarized in Section 11.4.2.2 and requires logic to search for the small seed in the central region of the EM2 layer and to compute the direction in ϕ in which to construct the cluster. This requires a fair amount of resources, and it was decided to drop this functionality in the BDT algorithm to make more room for a more complex BDT model and instead to sum in a symmetric fashion over a region centered on the TOB axis in all layers, with a particularly wide sum in the EM2 layer which is the deepest sampling layer of the calorimeter, absorbing most of the energy. The E_T sum of the BDT v16 algorithm is illustrated in Figure 13.14.

The fact that the Heuristic algorithm uses an asymmetric cluster while the BDT one uses a wide symmetric sum for E_T estimation results in some fundamental differences between the two. One consequence is a difference in granularity with which the algorithm reports the TOB in the η direction. The Heuristic algorithm relies on the position of the small seed, which is the supercell in the central region of the EM2 layer having the highest energy deposition, in order to refine the reporting of the η coordinate of the TOB. It can thus report the η position of the TOB to the downstream algorithms with a precision of 0.025. The BDT algorithm, having omitted the small seed, can only report the center of the TOB, which has a precision in η of 0.1. That being said, this does not affect any downstream trigger algorithms



(a)



(b)

Figure 13.15: The E_T reported by the Heuristic vs that reported by the BDT algorithm for a reconstructed τ jet at low (a) and high (b) p_T, evaluated on a MC sample of off-shell γ decaying to $\tau\tau$.

(e.g., in L1Topo) as they do not rely on the fine-grained reporting of the Heuristic algorithm.

Another fundamental difference is a skew of E_T of the BDT algorithm relative to the Heuristic one at low-p_T, as can be seen in Figure 13.15. To produce the plot, the E_T was read off the TOBs produced by the Heuristic and BDT algorithms that match reconstructed τ jets. Figure 13.15a contains only τ jets between 16 and 18 GeV while Figure 13.15b contains those between 50 and 52 GeV. At high-p_T there is no significant skew in E_T, but in the low-p_T region, there is a skew in how much the BDT algorithm over-reports the E_T relative to the Heuristic one. This skew, however, is present only at energies well below the range of TOB E_T thresholds used for triggering, the lowest being above 11 GeV, and is therefore not of concern.

Symmetric sum name	Constituent supercells
l0_d0000	4
l0_d0982	1, 3, 5, 7
l0_d1402	0, 2, 6, 8
l1_d0125	26, 27
l1_d0375	25, 28
l1_d0625	24, 29
l1_d0875	23, 30
l1_d0990	14, 15, 38, 39
l1_d1051	13, 16, 37, 40
l1_d1125	22, 31
l1_d1164	12, 17, 36, 41
l1_d1315	11, 18, 35, 42
l1_d1375	21, 32
l1_d1493	10, 19, 34, 43
l1_d1690	9, 20, 33, 44
l2_d0125	62, 63
l2_d0375	61, 64
l2_d0625	60, 65
l2_d0875	59, 66
l2_d0990	50, 51, 74, 75
l2_d1051	49, 52, 73, 76
l2_d1125	58, 67
l2_d1164	48, 53, 72, 77
l2_d1315	47, 54, 71, 78
l2_d1375	57, 68
l2_d1493	46, 55, 70, 79
l2_d1690	45, 56, 69, 80
l3_d0000	85
l3_d0982	82, 84, 86, 88
l3_d1402	81, 83, 87, 89
l4_d0000	94
l4_d0982	91, 93, 95, 97
l4_d1402	90, 92, 96, 98

Table 13.2: All symmetric sums around the TOB axis and the constituent supercells. These sums serve as the building blocks of the BDT input variables. The cell indices correspond to those in Figure 13.8.

Dataset				
Sample	Class	Size		
		Train	Validation	Test
$Z \to \tau\tau$	Signal	948867 (720436)	237225 (179926)	296589 (225081)
Run 452843 ZeroBias	Signal	276584 (170996)	69441 (42933)	86092 (53492)

XGBoost Hyperparameters	
Parameter	Value
scale_pos_weight	3.43
max_depth	3
n_estimators	9
early_stopping_rounds	5

Table 13.3: Training inputs summary for the BDT algorithm. The dataset size is the number of TOBs, and the numbers in parentheses are the number of events.

13.7 ALGORITHM TUNING AND PERFORMANCE

Before an ATLAS run starts, the parameters for all the trigger algorithms are set in a *trigger menu*. Before physics data taking, these parameters must be tuned to give the optimal performance in terms of efficiency, background rejection, and trigger rate.

Besides optimizing the BDT algorithm's parameters, it is crucial to compare the performance of the baseline Heuristic algorithm with the BDT algorithm to see which candidate algorithm has a better performance. For that, the heuristic and BDT algorithms must be tuned individually. This section describes the tuning of the thresholds of the two algorithms, based on utilizing them inside primary Run 3 τ triggers, counting the resulting rates, and changing the thresholds until the rates are similar to the equivalent triggers based on the Legacy (Run 2) τ algorithm. Efficiency plots using the tuned algorithms are presented. The tuning is performed using a combination of MC and data samples produced during Run 3 described in Section 13.8.

13.7.1 *Parameters of the Tau Algorithms*

There are several tunable parameters internal to the BDT and Heuristic algorithms. These are either triplets or single numbers. The triplets correspond to three thresholds, setting three working points - loose, medium, and tight. These can be either multipliers or direct thresholds. Multipliers are used to avoid applying direct thresholds on ratios, which are hard to compute in hardware, and instead are used to multiply the denominator of the ratio and compare with the numerator. In addition, there are several external parameters. Following is a list of all parameters:

- **Parameters common to the BDT and Heuristic algorithms**
  - **Hadronic fraction multipliers** M_1^{had},M_2^{had},M_3^{had} - a triplet of multipliers to set working points on the hadronic fraction, logically the ratio between the energy depositions in the hadronic and the EM calorimeters. The hadronic fraction working point is not used for τ triggering and is implemented as part of the τ algorithm due to a requirement for potential future triggering on long-lived particles. It was not considered in the tuning.

  - **Minimum** E_T $E_{T,\text{min.}}$- a threshold on the TOB E_T below which the TOB is not written out from the eFEX system. It is set to 5 GeV and is not tuned.
  - **Maximum** E_T $E_{T,\text{max.}}$- a threshold on the TOB E_T above which the isolation discriminant is ignored and is always considered as passed (i.e., "tight").

- **Parameters specific to the BDT algorithm**
  - **BDT thresholds** B_1, B_2, B_3- a triplet of BDT score thresholds, defining three working points, depending on where the score lands.
  - **Hadronic fraction multipliers** M_1^{had},M_2^{had},M_3^{had} - in the BDT algorithm, the Hadronic and EM energies, due to practical reasons, are computed by summing over several BDT input variables in the HAD and EM2 layers, respectively.
  - **Low** E_T **BDT cutoff** $E_{T,\text{min.}}^{\text{BDT}}$- a minimum threshold for E_T below which the BDT score is ignored, discussed in Section 13.7.1.1.

- **Parameters specific to the Heuristic Algorithm**
  - R_{core} **multipliers** M_1^{core},M_2^{core},M_3^{core}- a triplet of multipliers to set working points on the R_{core} discriminant, described in Section 11.4.2.2.
  - **Hadronic fraction multipliers** M_1^{had},M_2^{had},M_3^{had} - in the Heuristic algorithm, the EM and hadronic energies are computed around the small seed, in the ϕ direction of highest energy in the EM2 layer and over two ϕ rows in the same ϕ direction in the hadronic layer.

- **Parameters external to the eFEX algorithms**
  - **Trigger** E_T **threshold** - the primary cut on the E_T of the TOB, which, for a given trigger chain, is largely fixed at a certain energy value, while other parameters are varied during tuning.
  - **jFEX isolation multipliers** - a triplet of multipliers to set a working point for the jFEX isolation, as described by (11.1).

13.7.1.1 *BDT vs. Heuristic Isolation*

As already discussed in Section 11.4, the major difference between the Heuristic and BDT algorithms is the isolation discriminant, which in

the former case is provided by the R_{core} ratio, while in the latter - by the BDT score.

One fundamental difference in the nature of isolation between the BDT and the Heuristic algorithm is the fact that the R_{core} discriminant is computed as a ratio of the energy deposition in a core over an environment region that is not very sensitive to the energy scale, while the BDT is required to pick up the isolation's scale dependence in training, through the provided CORE variable, as described in Section 13.4.3. This can lead to sensitivity of the performance to the training dataset and is a disadvantage of the BDT algorithm in its current form. Such a dependence was indeed observed during 2024 data taking, as will be mentioned in Chapter 15.

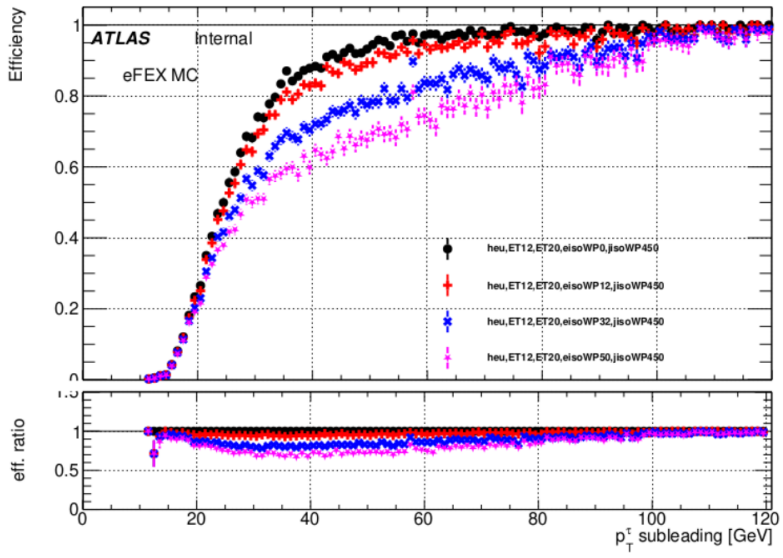
Another interesting difference is how tightening the threshold on the isolation discriminant affects the signal turn-on curve's shape. This effect is demonstrated in Figure 13.16. Tightening the R_{core} isolation results in a "dent" in the mid-p_T region of the turn-on curve, while tightening the BDT threshold results in a later turn-on. This is not surprising since the strongest BDT input variable is the CORE variable (see Figure 13.12a), which is highly correlated with the TOB's primary E_T discriminant. Therefore, tightening the BDT score has a similar effect as tightening the E_T cut, which lowers the efficiency at low-p_T values of reconstructed τ jets.

In order to mitigate this effect of low efficiency at low-p_T, an additional parameter, $E_{T,\text{min.}}^{\text{BDT}}$, was proposed and implemented by the author and accepted by the ATLAS L1Calo collaboration, of a minimal E_T value below which the BDT score is ignored.

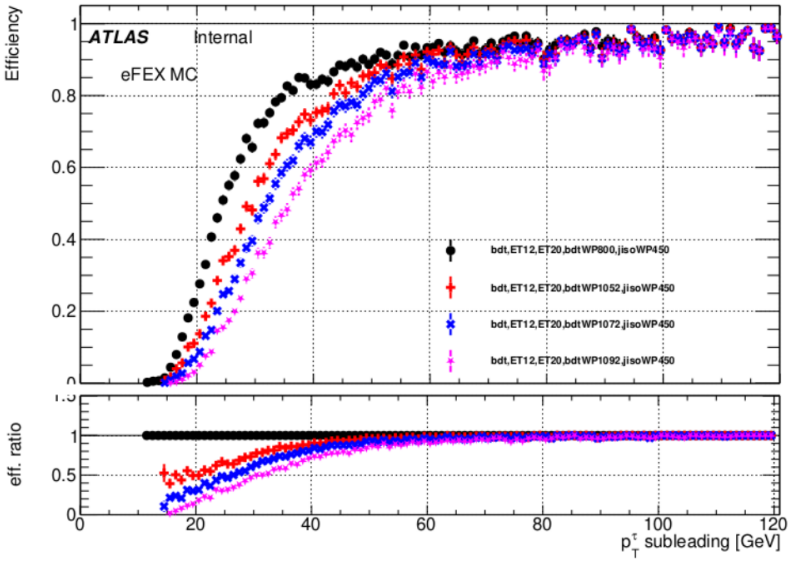
13.7.2 *Choice of Triggers for Tuning*

The challenging signature of the τ lepton does not allow using low-threshold single τ triggers for data taking since they produce unacceptable rates at L1. Therefore, to reach low thresholds which are essential for sensitive measurements of the Higgs boson's coupling to fermions, the τ triggers used for physics data taking are topological triggers that require a combination of two low-E_T taus originating from the eFEX τ algorithm with an isolation requirement on both, additional jets and topological cuts between them.

The effect of adding some example topological requirements at the L1 level on the L1 rates, as evaluated on Run 2 data, is shown in Figure 13.17. Starting with a bare di-τ trigger of two unisolated τ TOBs with



(a)



(b)

Figure 13.16: Efficiency curves for a single-τ trigger produced by progressively tightening the isolation discriminant, as a function of the p_T of the offline reconstructed τ. (a) Heuristic R_{core} discriminant, (b) BDT score discriminant. The jFEX isolation working point is kept constant and loose.

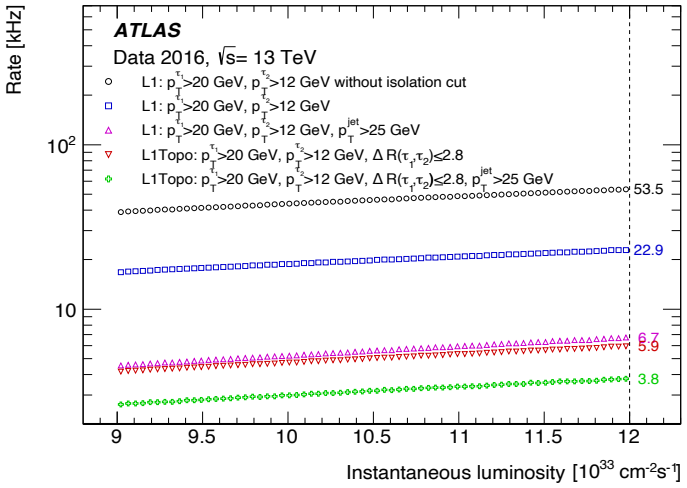


Figure 13.17: L1 rate of a di-τ trigger as a function of instantaneous luminosity for different additional topological requirements evaluated on Run 2 data using the Legacy L1Calo τ algorithm. [78]

E_T thresholds of 12 GeV and 20 GeV (black circles), the addition of isolation reduces the rate by a factor of 2.3. The requirement of an additional jet or a topological requirement of a $\Delta R(\tau_1, \tau_2) \leq 2.8$ cut between the taus as well as the requirement of an additional jet above 25 GeV reduces the rate further.

13.7.3 *Rate Counting*

In Run 3 the highest rate τ triggers are expected to be low-threshold di-τ triggers with at least two extra jets followed by a di-τ trigger with an additional jet, topological cut of $\Delta R(\tau_1, \tau_2) \leq 2.8$ between the two taus and a disambiguation cut of $\Delta R(\tau, j) > 0.1$ between each τ and an additional jet. For that reason, the thresholds for the BDT and Heuristic algorithms are computed on emulations of the di-τ+ ≥ 2 jets trigger. The jFEX group performed a separate, similar tuning on jet trigger parameters in the jFEX system, which was responsible for producing the jet TOBs used in the emulation of this trigger.

The optimal working points for the leading and sub-leading components of the di-τ trigger following tuning are given in Table 13.4. Table 13.5 summarizes the resulting event counts, as computed on a zero bias sample, for the pure di-τ trigger and the di-τ+ ≥ 2 jets one, for which the rates are roughly the same for all algorithms, as required. These results are not the final results and are not up to date,

as the process of tuning and algorithm improvement is still ongoing at the time of writing due to changes in ATLAS detector running conditions. The rates are shown for a given set of run conditions mainly to compare the primary trigger rates for legacy, heuristic, and BDT algorithms used to select individual τ TOBs, as was computed prior to the approval of the BDT algorithm.

Type	Algorithm	E_T cut [GeV]	Isolation	jFEX R'	jFEX A	Max. E_T [GeV]
Sub-leading	Heuristic	12	$M_2^{\text{core}} = 18$	1784	1024	50
	BDT		B_2=896	1704	1024	100
	Legacy		Run 2	-	-	-
Leading	Heuristic	20	$M_2^{\text{core}} = 21$	1632	1024	50
	BDT		B_2=908	1634	1024	100
	Legacy		Run 2	-	-	-

Table 13.4: Summary of a set of thresholds on discriminants of the leading and sub-leading TOBs in a di-τ trigger producing the same rate for the di-τ+ $\geq$ 2 jets trigger. Run 2 isolation is defined in Table 11.1.

Trigger	Legacy	Heuristic	BDT
Di-τ	3088	2949	2876
Di-τ+ $\geq$ 2 jets	202	210	201

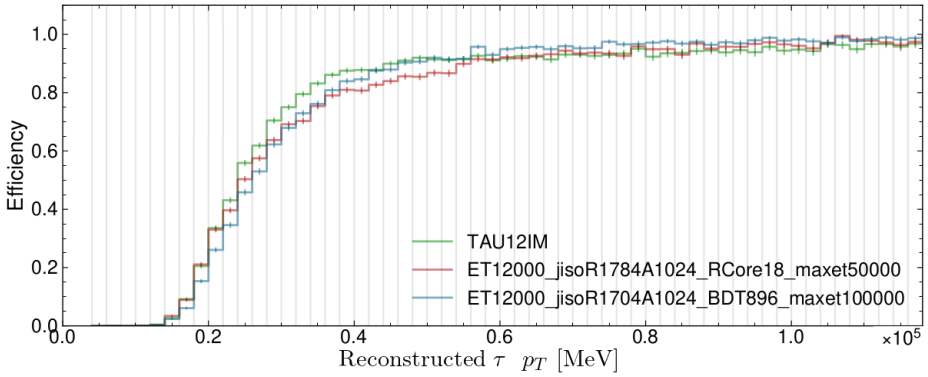
Table 13.5: Summary of the number of zero bias events that pass the di-τ and di-τ + jets triggers for the Legacy, Heuristic, and BDT algorithms.

13.7.4 *Turn-on Curves*

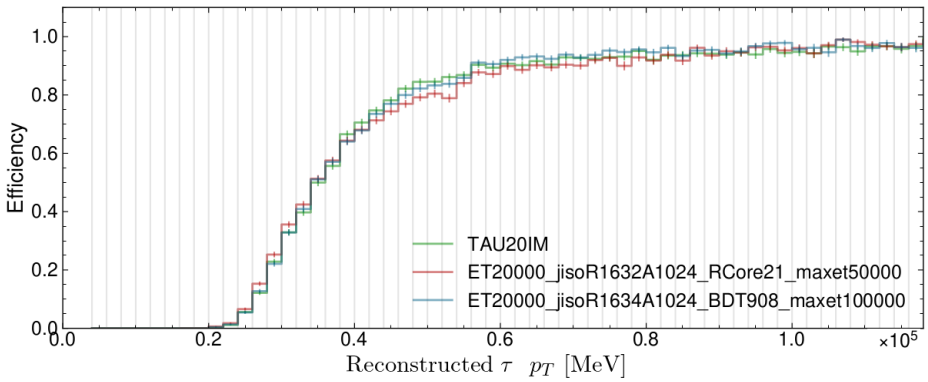
Efficiency curves for reconstructed τ jets are computed for the Legacy, BDT, and Heuristic algorithms on a $\gamma^* \rightarrow \tau(\text{had.})\tau(\text{had.})$ MC sample for the leading and sub-leading reconstructed τ jets.

All τ jets that match truth τ jets within a radius of $\Delta R = 0.2$ are selected and participate in the efficiency denominator. Those that have a matching TOB within $\Delta R = 0.2$ that passes the algorithm requirements participate in the numerator.

Figure 13.18 shows the efficiency of the subleading and leading taus for the two algorithms tuned as described in Table 13.4. The BDT algorithm performs slightly better for the identification of TOBs



(a)



(b)

Figure 13.18: Efficiency curves for a single-τ trigger produced as a function of the p_T of the offline reconstructed τ, for the tuning summarized in Table 13.4, for sub-leading (a) and leading (b). The three curves are Legacy (green), Heuristic (red), and BDT (blue).

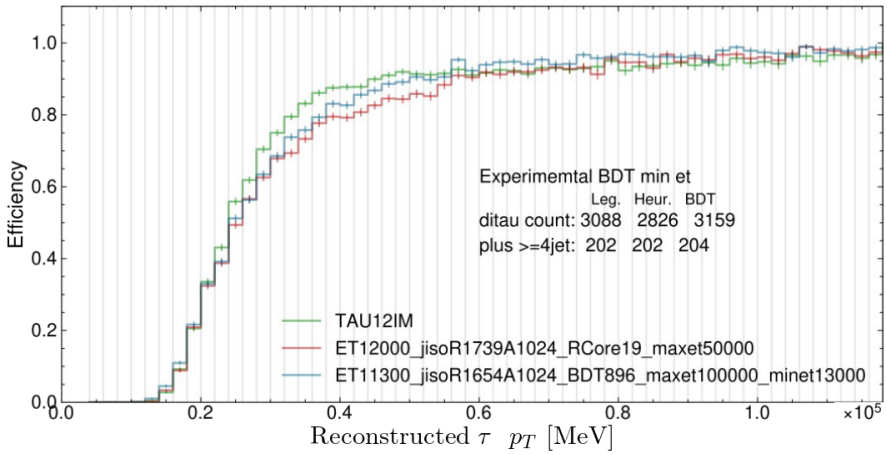


Figure 13.19: The effect of adding a minimum BDT cut below which the TOB's BDT score is ignored on the sub-leading TOB identification efficiency. The turn-on for the BDT outperforms the Heuristic over the entire p_T range. The event counts for the Legacy, Heuristic, and BDT algorithms are shown for di-τ and di-τ+ $\geq$ 2 jets (written as $\geq$ 4 jets) trigger emulations.

matched to reconstructed taus in the mid- to high-p_T range of the curve, especially for the lower 12 GeV cut on Figure 13.18a. The Heuristic algorithm performs better at low-p_T. However a recent addition of the parameter $E_{T,\text{min.}}^{\text{BDT}}$ discussed in 13.7.1.1 improves the performance, as seen in Figure 13.19, which shows a re-tuning done with this additional parameter.

13.7.5 *Performance Relative to Run 2 Algorithm*

As evident from Figure 13.18a and Table 13.5, when requiring the same rate for di-τ triggers as the Run 2 algorithm, the Run 3 algorithms are less performant in the mid-p_T region.

One reason for that is that the Run 3 algorithms, especially when considering isolated objects, are more fine-grained than the Run 2 one and are capable of better resolving close-by energy depositions as separate TOBs, triggering the di-τ triggers on the background. This artifact makes background rejection worse, while not affecting the signal efficiency since the τ jets in signal di-τ signatures usually appear back-to-back and are well separated in η and ϕ.

Figure 13.21 shows the multiplicity per event of sub-leading Run 3 BDT and Run 2 TOBs that are selected by the respective algorithms with and without the application of the isolation requirement listed in Table 13.4, evaluated on a zero bias sample. While without isolation,

there are more events with at least two TOBs, this is reversed when the isolation requirement is applied. This supports the aforementioned hypothesis - while the isolation is required to reject individual background TOBs, it also increases di-τ rate counts, which requires an increase in thresholds to lower the rate to an acceptable level, which in turn hurts the efficiency. Therefore, the better Run 2 performance at mid-p_T stems from worse background rejection for the primary di-τ triggers and not from a worse efficiency for signal detection.

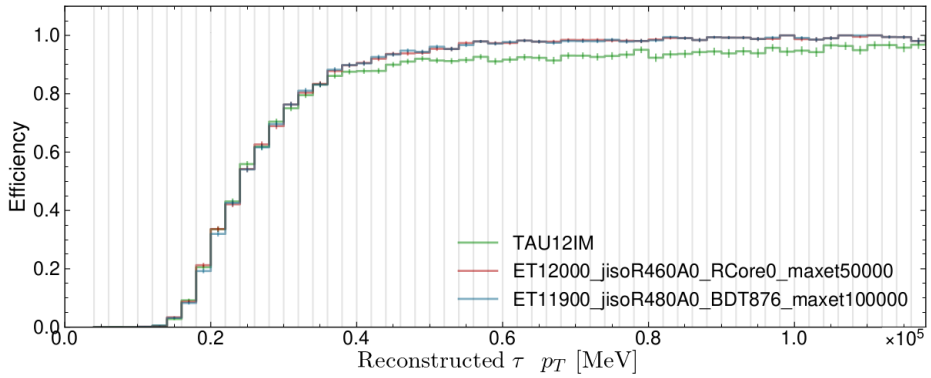
Type	Algorithm	E_T cut [GeV]	Isolation	jFEX R'	jFEX A	Max. E_T [GeV]
Sub-leading	Heuristic	12	$M_2^{\text{core}} = 0$	460	0	50
	BDT	11.9	B_2=876	480	0	100
	Legacy	12	Run 2	-	-	-
Leading	Heuristic	20	$M_2^{\text{core}} = 19$	408	0	50
	BDT		B_2=888	350	0	100
	Legacy		Run 2	-	-	-

Table 13.6: Summary of a set of thresholds on discriminants of the leading and sub-leading TOBs in a di-τ trigger producing roughly the same single-τ leading and subleading efficiencies for all algorithms. Run 2 isolation is defined in Table 11.1.

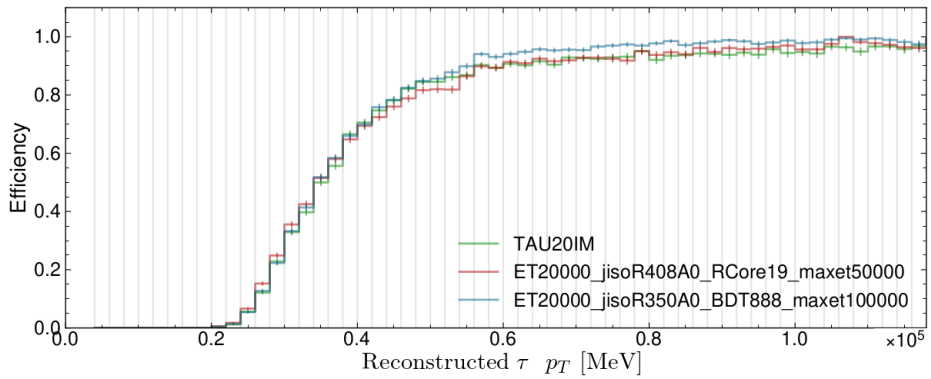
Trigger	Legacy	Heuristic	BDT
Di-τ	3088	4159	4136
Di-τ+ $\geq$ 2 jets	202	247	244

Table 13.7: Summary of the number of zero bias events that pass the di-τ and di-τ + jets triggers for the Legacy, Heuristic, and BDT algorithms with the latter two tuned to give the same efficiency as the Legacy algorithm.

To see that the same single-τ efficiency as Run 2 is attainable by the Run 3 algorithms, a separate tuning, described in table 13.6 was done to give the same efficiency between all algorithms, as seen on Figure 13.20. The resulting counts are listed in Table 13.7 and, as expected, are much higher for Run 3 vs. Run 2 than those listed in Table 13.5.



(a)



(b)

Figure 13.20: Same-efficiency curves for a single-τ trigger produced as a function of the p_T of the offline reconstructed τ, for the tuning summarized in Table 13.6, for sub-leading (a) and leading (b). The three curves are Legacy (green), Heuristic (red), and BDT (blue).

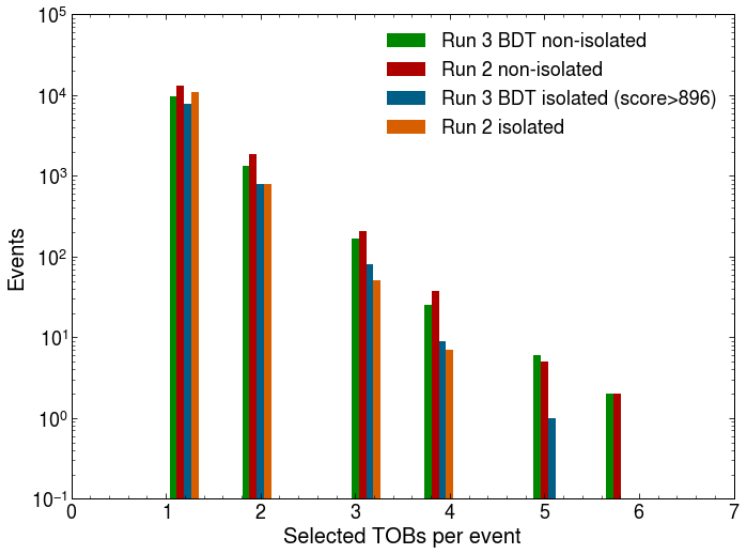
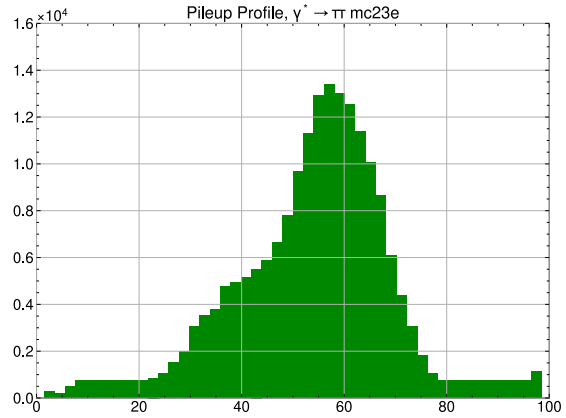


Figure 13.21: Multiplicity of Run 3 BDT and Run 2 TOBs that pass the corresponding algorithm thresholds of the sub-leading TOB with and without the isolation requirement in a zero bias sample.

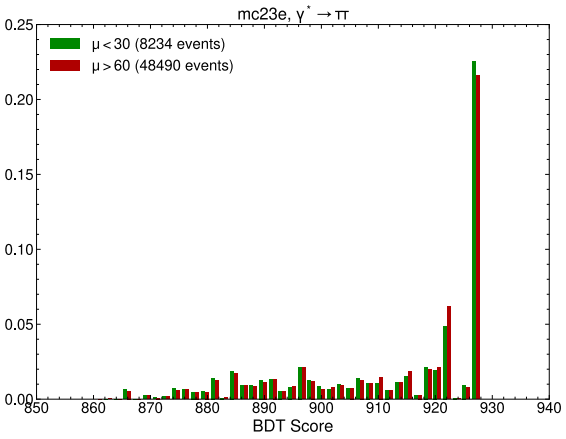
13.7.6 *Dependence on Pileup*

A MC simulation should ideally produce distributions of all observables identical to those of the data it is meant to describe. However, some observables, in particular the pileup, can only be determined after the data is taken. Therefore, any MC sample that is generated before a dataset is taken will have a deviation in this observable. Other observables may be significantly affected by varying levels of pileup. Following this logic, the fact that we used data for background and MC for signal in training the BDT means that the pileup distribution is different in the background and signal samples. The BDT could, in principle, pick up differences in observables due to pileup, which could be dominant over differences between the signal and background signatures.

In order to ensure that the dependence of the score distribution is not strongly dependent on pileup, a histogram shown in Figure 13.22b is produced on one of the signal MC samples used to tune the BDT algorithm, where the BDT score of signal TOBs taken from events falling within two extreme ranges of pileup is plotted, $\mu < 30$ and $\mu > 60$. A pileup distribution of all events in the same sample is shown in 13.22a. No change in the shape of the distribution is observed for the two ranges.



(a)



(b)

Figure 13.22: (a) Pileup profile of a $\gamma^* \to \tau\tau$ signal MC sample. (b) Distribution of the BDT score in the sample for different pileup ranges.

13.8 DATA AND MC SAMPLES

Different MC and data samples were used during the development and commissioning stages for the training and evaluation of the BDT algorithm. During the development stage, only MC samples were used. For training, a $Z \to \tau\tau$ sample was used as the signal and a di-jet sample as the background, both at a center-of-mass energy of $\sqrt{s} = 13$ TeV.

During the commissioning stage, the model was re-trained on $Z \to \tau\tau$ at $\sqrt{s} = 13.6$ TeV for signal and a zero bias run for background. For tuning, a $\gamma^* \to \tau\tau$ sample, which has a wider p_T range and in which all taus are restricted to decay hadronically, was used to compute efficiencies and compare different BDT models and working points. Zero bias data from multiple runs, which were the latest at the time, were used to count the expected L1 rates for the different BDT models. Table 13.8 summarizes all the MC and data samples used.

Process	$\sqrt{s}$ [TeV]	Generators
$Z \to \tau\tau$	13	Powheg(v.2_r3781), Pythia8(v.308)+AZNLO+CTEQ6L1, EvtGen(v.2.1.1)+Photos++(v.3.64)
Di-jet	13	Pythia8(v.308)+A14 NNPDF23LO, EvtGen(v.2.1.1)
$\gamma^* \to \tau\tau$	13.6	Pythia8(v.308), EvtGen(v.2.1.1)
$Z \to \tau\tau$	13.6	Powheg(v.2_r3781), Pythia8(v.308)+AZNLO+CTEQ6L1, EvtGen(v.2.1.1)+Photos++(v.3.64)

Table 13.8: Summary of the MC datasets used for training and evaluation of the BDT algorithm.

To extract the training data from the MC datasets, reconstructed τ jets were matched to truth taus within a radius of $\Delta R = 0.2$. For each reconstructed τ jet, simulated Run 2 and Run 3 TOBs were matched, also within a radius of $\Delta R = 0.2$. The supercell content of the TOBs was used to construct input variables for the BDT in the way described in Section 13.4.3.

For counting rates, zero bias streams from ATLAS physics data taking runs recorded at the beginning of Run 3 were used, pre-selected on events that pass the lowest Legacy isolated τ trigger. This trigger requires a single TOB with its energy computed as described in Section 11.4.1 with an additional isolation requirement of $E_T^{\text{EM isol}}[\text{GeV}] \leq (E_T[\text{GeV}]/10 + 2)$ for $E_T < 60$GeV.

14

HARDWARE AND FIRMWARE

Since the BDT τ algorithm is an alternative algorithm to the Heuristic one, which was already implemented and ready for operation as a baseline τ triggering algorithm for Run 3, the guiding principle in producing the firmware design was to make the least amount of changes to any surrounding components, whether in hardware, firmware or software and to have the same interface, expecting the same input and producing the same output. The latency of the algorithms had to be the same, and the resource usage of the Heuristic algorithm served as an upper limit to that of the BDT one. The fact that the firmware and hardware environment were well defined both facilitated the development on the one hand and constrained the physics performance on the other.

14.1 THE EFEX SYSTEM

The BDT algorithm runs within the electron Feature Extractor (eFEX) system introduced as part of the Phase-I upgrade. Its purpose is to identify isolated e/γ and τ TOBs.

14.1.1 *The eFEX Hardware*

The eFEX system is comprised of 24 eFEX modules, which together cover the calorimeter range of $|\eta| < 2.5$ [79]. Each eFEX module contains four Xilinx Virtex 7 FPGAs for data processing (part number xc7vx550tffg1927-2) and an additional FPGA for control. Each FPGA accepts a fixed region of 0.6×1.0 in $\eta \times \phi$ space as input and processes a core region of 0.4 ×0.8, except in the border regions of η, where the FPGAs accept a core region of 0.7× 1.0 and process a region of 0.5× 0.8. The regions are illustrated in Figure 14.1, where the core regions are colored orange, and the full region is the core region and the surrounding towers.

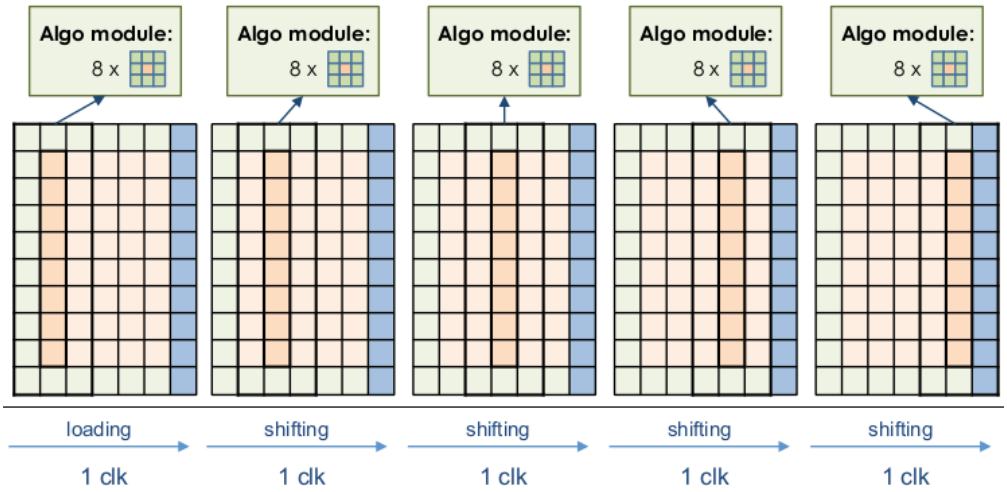


Figure 14.1: EFEX multiplexing at 200 MHz. The η direction is horizontal. The five columns illustrate the state of the system at each of the five 200 MHZ clock cycles fitting into a single 40 MHz cycle. At each clock cycle, eight algorithm cores simultaneously process eight 3×3 TOBs, partially overlapping in ϕ, positioned inside the black box in each column in the upper part, where each small rectangle is a single 0.1×0.1 trigger tower. The blue columns are filled with dummy data (e.g., all zeroes) in central regions away from $|\eta| = 2.5$ and only contain valid calorimeter data in the border regions touching $|\eta| = 2.5$.

14.1.2 *The eFEX Firmware*

Focusing here on functionality relevant to τ only, each eFEX FPGA's firmware implements eight identical copies of the τ identification algorithm, referred to as algorithm cores. While data arrive at the FPGAs at the LHC bunch crossing rate of 40 MHz, the algorithms run inside a 200 MHz clock domain, allowing the eight algorithm cores to process the 0.6×0.1 region, or a grid of 6×10 trigger towers each spanning 0.1×0.1 in $\eta \times \phi$ space, by taking five clock cycles to scan the grid in the η direction as illustrated in Figure 14.1. The design of the algorithm must be fully pipelined, which means that each algorithm core is required to produce output every clock cycle at 200 MHz.

A single algorithm core takes as an input a TOB comprised of 3×3 trigger towers, containing information about the energy deposition in a 0.3×0.3 region in $\eta \times \phi$ space of the EM and hadronic calorimeters. Each tower is comprised of 11 16-bit words, resulting in an input of 99 16-bit words for the τ algorithm. For each input TOB, the output of the

algorithm is a collection of conditions and discriminants computed internally.

14.1.2.1 *BDT τ Algorithm*

The following schematics and description of the algorithm structure will represent the current iteration of the BDT algorithm used in 2024, BDT v16.

One of the guiding principles in the BDT algorithm's firmware implementation was to avoid introducing significant modifications to an already existing system. Therefore, from the design perspective, all significant modifications take place within a VHDL module implementing all details of the algorithm, meant to serve as a drop-in substitute for the Heuristic algorithm.

The algorithm's implementation re-uses VHDL code and modules already present in the firmware as much as possible to conform with the already established coding standards. The algorithm is implemented under the assumption that the supercells participating in summations of the TOB E_T and BDT input variables, as well as the BDT model itself, may change in the future, which requires an additional external tool to automatically generate and test delay and summation VHDL code to avoid painstaking manual changes otherwise required when these modifications are made. The input to the algorithm is identical to the input for the Heuristic algorithm, and the output is a set of computed discriminants and conditions that are set downstream in data structures that are structurally identical for both τ algorithms.

The two algorithms accept several parameters as inputs, which are summarized in Table 14.1. As discussed in Section 13.7.1, parameters are either arrays of three 8-bit elements for multipliers or single 16-bit values. Initially, only parameters $0-3$ were used for the BDT algorithm to conform with the Heuristic one, and the fourth parameter was added to the BDT algorithm close to its approval.

The output structure is summarized in Table 14.2. For the Heuristic algorithm, the fields UpNotDown, Seed, and JetOrBDT contain values that are specific to that algorithm and correspond to the direction in ϕ in which the clustering is done, which of the four central supercells in EM2 to cluster around (small seed), and an isolation working point, respectively. The BDT algorithm stores the BDT score in the BDT field, the JetOrBDT field is set to the BDT working point and the UpNotDown and Seed fields are both set to zero. The E_T sum over the cluster with an energy overflow bit is reported for both algorithms in

Index	Elements	Size	Parameter description	
			BDT	Heuristic
0	3	8	B_1, B_2, B_3	M_1^{core},M_2^{core},M_3^{core}
1	3	8	M_1^{had},M_2^{had},M_3^{had}	
2	1	16	$E_{T,\text{min.}}$	
3	1	16	$E_{T,\text{max.}}$	
4	1	16	$E_{T,\text{min.}}^{\text{BDT}}$	-

Table 14.1: Summary of the parameters of the τ algorithms. The size column lists the size in bits per element. The parameters are explained in Section 13.7.1.

the fields Energy and EnergyOF, respectively. The IsMax field signals that the TOB has a maximum energy deposition in the central tower, in which case it is said to be seeded. The Frac field is the hadronic fraction working point. The Version field is set to 0 in the Heuristic algorithm and to 1 in the BDT one.

Name	Size	Description	
		BDT	Heuristic
Version	1	Heuristic/BDT algorithm indicator	
Energy	16	Estimate of TOB E_T	
EnergyOF	1	Energy overflow indicator	
UpNotDown	1	0	Direction of slowest energy decrease in ϕ
Seed	2	0	Small seed
IsMax	1	Maximum energy in core tower	
JetOrBDT	2	BDT condition	R_{core} isolation condition
Frac	2	Hadronic Fraction	
BDT	10	BDT score	0

Table 14.2: Summary of the output fields produced by the two τ algorithms. The size columns is in bits.

The full structure of the BDT algorithm is illustrated in Figure 14.2. Logically, the implementation consists of a sequence of three operations - computation of sums, computation of discriminants and evaluation of conditions. The total latency of the algorithm is precisely

equal to that of the Heuristic algorithm, which is 12 clock cycles at 200 MHz, or 60 ns.

14.1.2.2 *Resource Usage*

One of the benefits of the BDT algorithm, besides being more performant than the Heuristic one, is that it is lighter in resources.

As discussed in Section 12.1.1, there are three main components of an FPGA - LUTs, FFs, and DSPs. The exact number of each of these utilized in a design is only known after the entire FPGA's firmware is fully implemented, with all components placed and all connections between them routed. Table 14.3 shows a comparison of the amount of resources used by the full eFEX firmware design with the Heuristic algorithm and that used by the same design with the Heuristic algorithm replaced by the BDT one as the only difference.

The usage of all types of resources is markedly reduced, with the DSP count, in particular, dropping significantly. The reason for that is the R_{core} discriminant in the Heuristic algorithm, which requires multiplication for which DSPs are utilized. In the BDT algorithm, however, this discriminant is replaced by the BDT score, which only requires comparisons and summations to evaluate.

Resource	Heuristic	BDT	Reduction
LUT	195314	189596	3%
FF	295727	271187	8.3%
DSP	120	96	20%

Table 14.3: Comparison of resource usage between the Heuristic and BDT algorithms. The numbers are reported for the full eFEX firmware, the only difference being the τ algorithm.

14.2 FRAMEWORK AND TOOLING

Since ML-based algorithms have never before been implemented at the L1 trigger in ATLAS, a variety of tools had to be written to facilitate the development, including tools for data preparation, training, evaluation, simulation, tuning, and firmware generation of the BDT algorithm. Most were developed by the author at different stages of the project and are briefly described in this section.

14.2.1 *Data Preparation, Training and Evaluation*

Data and MC in ATLAS are usually stored in Root [80] files, which pass multiple derivation stages in a pipeline specific to each analysis in which unnecessary information is removed and additional useful information is added and the final lightweight Root files containing only the relevant information for a particular need are called the *ntuples*.

While Root files are optimized to be used for physics analyses implemented in C++, it is much simpler and faster to use a tabular format based on Python Pandas [81] data frames to store and manipulate data for training XGBoost models in Python. For that, a tool that extracts the supercell data from Root files into comma-separated values (CSV) files, implements bookkeeping of different data file versions and automatic download of files for training and performs data preparation such as train-test splitting was developed.

For the training itself, a special tool was developed to e.g., have the functionality to compute, in a configurable way, the BDT variables from supercells, be able to systematically keep track of the trained models to properly do parameter scans, and also manipulate already trained models in order to test e.g., the effect of digitization of the BDT leaf scores on the performance of the algorithm. The tool can also produce basic development-stage turn-on curves for model comparison, described in Section 13.2.

14.2.2 *Simulation*

The ATLAS detector, including all components and algorithms is fully simulated as part of the Athena framework written mostly in C++. A necessary requirement was to implement the BDT algorithm in Athena before the commissioning stage for performance studies, real-time monitoring of the algorithm during Run 3 and derivation of analysis data.

Two more places require the evaluation of the BDT algorithm - the *eFEX online simulation*, which is the full simulation of the eFEX system's mock-up test rig used to test any changes to the full eFEX firmware on hardware identical to the eFEX before being deployed to ATLAS. The goal of the simulation is to accept the same inputs as the test rig and verify that the outputs from the test rig produced by the firmware match the output of the simulation. The second location is

in local Python code for purposes of further algorithm development, where the functionality of evaluating the BDT algorithm exactly as implemented in Athena is greatly helpful.

One potential issue in implementing the algorithm in three different locations was the workload required to manage three different versions of the BDT algorithm's code. To avoid it, and since this management was meant to be done fully by the author, at least in the beginning, it was necessary to avoid code duplication as much as possible. The approach to implement the simulation was, therefore, to write an immutable C++ core implementation, fully configurable by an external configuration file, that evaluates all output discriminants given input supercells and parameters. A thin interface layer between either Athena, the online simulation code, or the local Python code was then implemented to bridge the core and the environment.

14.2.3 *Tuning*

During commissioning, it was critical to be able to quickly and conveniently produce turn-on curves and compute rates for different combinations of τ triggers while being highly flexible in changing the tunable parameters of both the eFEX BDT and Heuristic τ algorithms as well as the jFEX isolation, on the tabular data mentioned in Section 14.2.1. It was also necessary to be able to quickly introduce new parameters, such as $E_{T,\text{min.}}^{\text{BDT}}$ discussed in Section 13.7.1 to see how they affect performance. Prior to that, C++ analysis code based on the Root framework was used for tuning, but it required long periods of time from the idea to change or add a parameter to the production of the desired plots. Therefore, a special tuning framework was developed in Python which significantly simplified and accelerated this process. This framework was used e.g., to produce the turn-ons in Figure 13.20 and the rate counts in Table 13.5.

14.2.4 *Automatic Firmware and Simulation Generation*

Due to the experimental nature of the BDT algorithm, several critical components must be configurable:

- BDT model - it must be simple to switch with a re-trained model in case the performance is not optimal.

- BDT variables - although constrained to be summations over groups of supercells, in principle, every re-training can have a different set of highly ranked variables. Therefore, the groups of supercells over which to sum in order to produce the output must be configurable.

- E_T sums - there are three besides the BDT variables that are computed and used - the hadronic and EM sums used to compute the hadronic fraction working point (see Section 13.7.1) and the E_T sum of the TOB. These were implemented as configurable sums due to the initial concern that a sum over a large number of supercells would require too many resources to compute.

While in software, these changes are trivial, in the firmware, a change of a single sum or of the BDT model would require re-computation of delays of signals entering and exiting the summations and the BDT module and updating them in the design since, for example, a sum of two numbers requires one clock cycle to evaluate. In comparison, a sum of four numbers may require two, and, as discussed in Section 12.1.3, the depth and number of the trees in a BDT affect its latency.

To facilitate the switching of the BDT model and the summation schemas, a pipeline was developed by the author, which takes as an input a trained BDT model along with a description of which supercells to use to compute each energy sum and produces the finished VHDL code with all delays and summations automatically computed and set, failing if the resulting design exceeds the target latency of 12 cycles. All summations and delays are implemented in a separate module so that the resulting VHDL code is clean and readable. In addition, a configuration file for the simulation that fully determines its behavior is generated.

A central component of the pipeline is the automatic computation of delays and summations, for which a special tool was developed that accepts as input the summands, a list of expressions of the required sums to perform, one per each output sum, and any additional delays to add to each sum after they are computed. It then uses the NetworkX [82] Python package to analyze the sums as a directed acyclic graph (DAG) and compute the delays of each sum. The tool can also produce diagrams of the DAGs, as shown in Figure 14.3, which demonstrates the complexity and amount of sums that need to be computed within the BDT algorithm.

Testing is a time-consuming yet crucial part of firmware generation. This is also handled by the pipeline automatically, testing the resulting

VHDL implementation using Vivado Simulator on a set of inputs taken from real ATLAS data as well as random inputs by comparing the simulation outputs to those expected for each input. Once finished, the tested VHDL code is ready to be dropped to a special folder in the existing eFEX firmware repository without adding any further modifications and without disturbing any components outside of the folder.

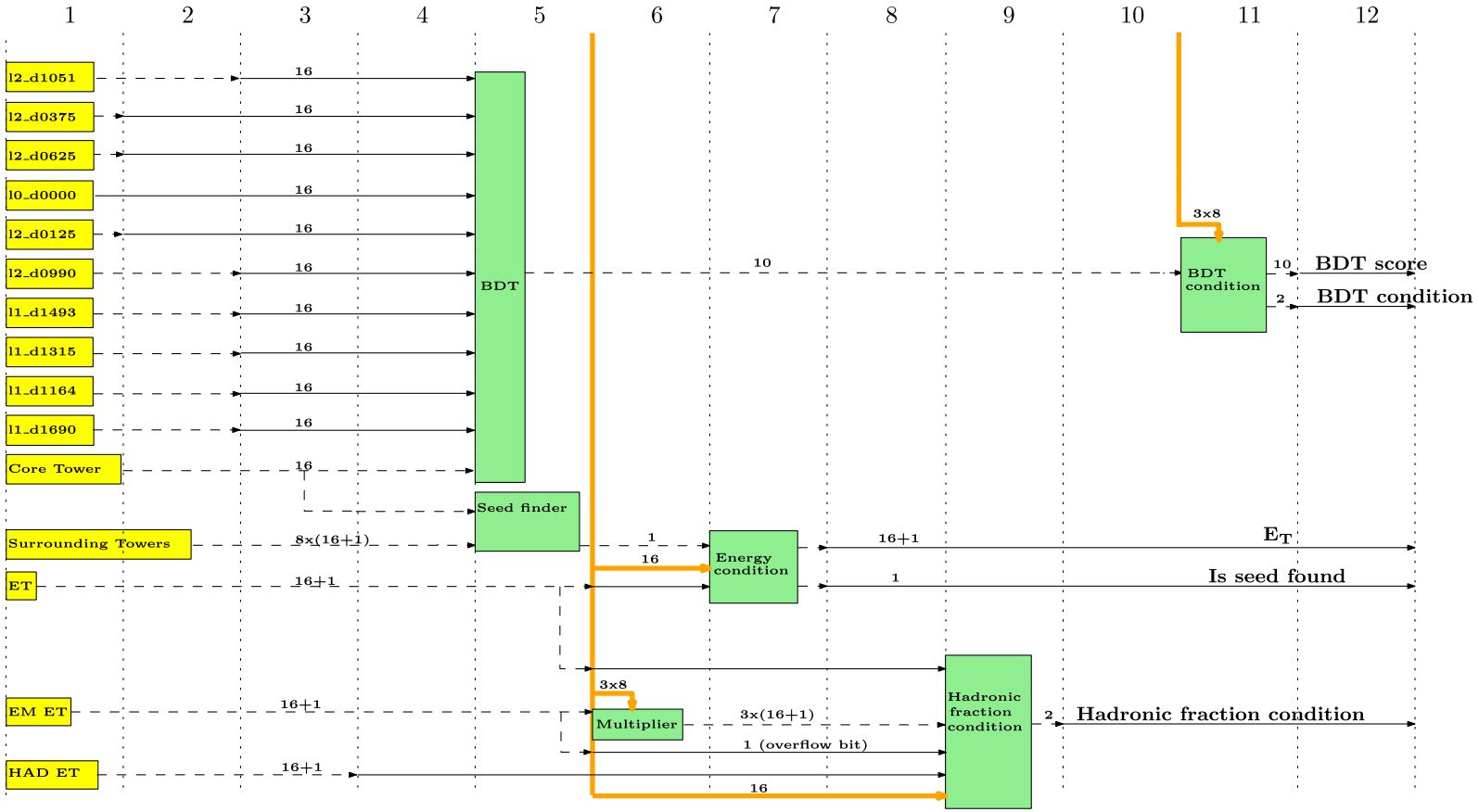


Figure 14.2: A diagram of the full algorithm. A region between any two vertical dotted lines is one clock cycle, and the cycles are numbered. Each green box stands for a VHDL module, which computes it. The Yellow boxes represent the sums used by the algorithm and are all computed in a separate module. Outgoing (ingoing) arrows symbolize output (input) signals of a module, labeled by the number of bits in the signal. When the number of bits is $16 + 1$, it means there are 16 bits and one overflow bit. A dashed arrow symbolizes computation latency and extends until the border before the clock cycle at which the result of the computation of its source module is ready, while solid arrows symbolize delayed signals. All delays are implemented in a separate module. The 99 supercells are used as inputs to the summing module in the first clock cycle. The thick orange line symbolizes the parameters that are ready at clock cycles 6 and 11. The outputs are ready to be used at clock cycle 13 (not visible).

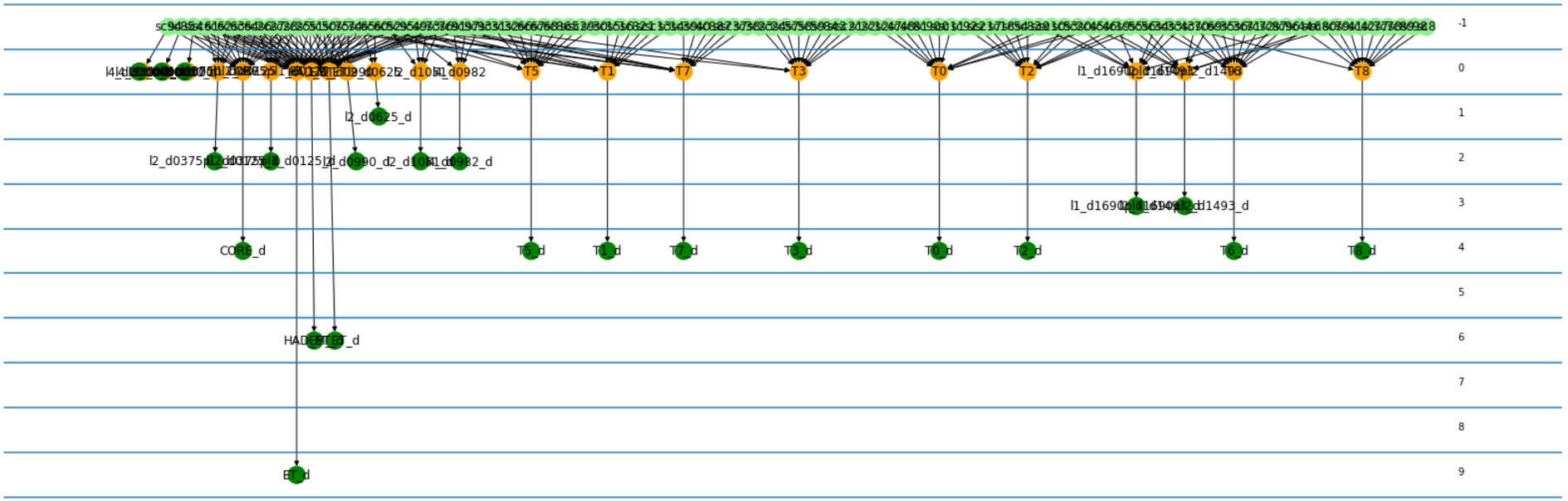


Figure 14.3: A diagram of a directed acyclic graph describing all the sums that need to be produced within the BDT algorithm. Light green circles are input supercells, orange circles are addition operations, and dark green circles are output sums. On the right is the clock cycle number at which an input is ready to be read, an operation is performed, or a sum is ready to be used. The output sums marked 'T' are sums over each of the nine towers in the TOB, used to identify if the TOB is seeded (see Section 11.4.2).

15

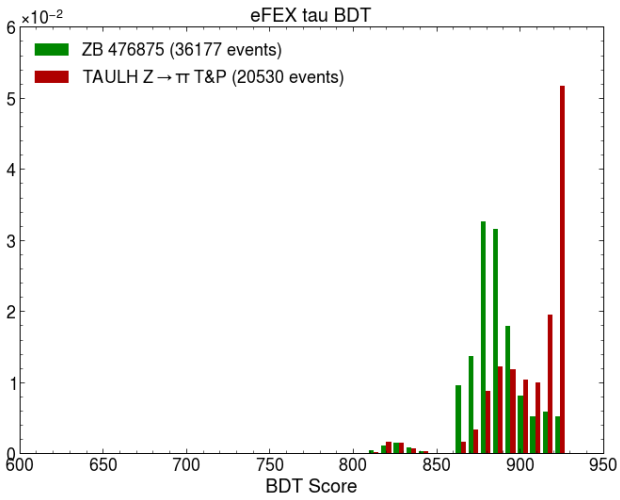
CONCLUSION AND OUTLOOK

A BDT-based algorithm for the identification of τ leptons in the L1 calorimeter-based ATLAS trigger was introduced as the default algorithm for this purpose during Run 3, demonstrating improved performance with respect to resource usage over the baseline heuristics-based algorithm. This is the first time a trained machine-learning model has been fully evaluated in the nanosecond domain in the ATLAS experiment.

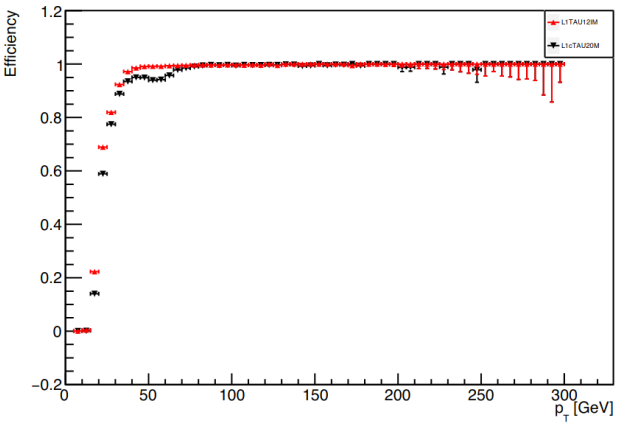
The performance gain of the BDT algorithm over the Heuristic one is not very significant due to the similarity of hadronic τ and QCD jets, especially when considering only calorimeter information. Nevertheless, perhaps the most important goal of the algorithm is to provide experience to ATLAS with using ML algorithms for L1 triggering and open the gate for additional ML algorithms to be potentially included before the end of Run 3. In addition, it provides a technical baseline for those planned to be introduced in the Phase-II upgrade in several ATLAS trigger systems. Since this work was submitted well into Run 3 as an example of the lessons learned on utilizing an ML-based algorithm during physics data taking, it is interesting to follow up with some recent developments.

Following the tuning done by the author and reported in this work, the ATLAS τ signature group performed an independent tuning to set the algorithm's thresholds. While initially, it showed the same performance, subsequent analysis following a fix in the ATLAS simulation code resulted in signal events migrating into the low-score bumps visible in the BDT score distributions, which previously only contained background data points (see Figure 13.10a), as shown on Figure 15.1a. This caused a significant drop in any τ trigger's efficiency when the eFEX BDT isolation working point was anything but the loosest, as can be seen in Figure 15.1b. Here, the red curve is based on the Legacy algorithm, while the black curve is based on the BDT algorithm, with matching thresholds.

This effect, after investigation, resulted in the realization that noise cuts that are applied to the supercells before they enter the algorithm to reduce random electronic noise have a strong effect on the algorithm's performance. As mentioned, during training, a zero-bias data sample



(a)



(b)

Figure 15.1: (a) Signal vs background distribution showing a "bump" between BDT score 800 and 850 which contains both signal and background events. (b) Drop in efficiency relative to the legacy algorithm due to changes in the MC signal sample's noise cut.

was used for the background, and an MC sample was used for the signal. The data sample had realistic, η-dependent noise cuts which were lower than the ones set in the MC sample and which were flat in η. This resulted in TOBs at high η values having high values of the peripheral input variables, which reduced their BDT score since, during training, these variables had high values only in the background sample.

This realization requires retraining the model, and a lesson was learned that retraining is required. This, in turn, requires an efficient retraining framework, which the author is developing for that purpose at the time these lines are written.

Although the BDT-based τ algorithm is already being used for data collection during Run 3, further work remains in evaluating the collected data. For instance, a data–MC comparison of the BDT score is currently underway. This is an important step in assessing the algorithm's performance, particularly in addressing concerns related to overfitting.

With all the infrastructure for tuning, training, and firmware generation in place, further experimentation with machine learning algorithms in the Level-1 trigger is now possible. Potential future directions include exploring different types of input variables, using a BDT as a regression model to estimate the E_T of TOBs, training models targeted at specific regions of phase space, leveraging newly available tools for anomaly detection using BDTs as autoencoders [83], and studying the application of neural networks for calorimeter-based τ triggering within the hardware trigger system. All of these avenues are now open for investigation and development during Run 3.

BIBLIOGRAPHY

[1] Douglas R. Hofstadter. *Gödel, Escher, Bach: an Eternal Golden Braid*. Basic Books Inc., 1979.

[2] ATLAS Collaboration. "Observation of a New Particle in the Search for the Standard Model Higgs Boson with the ATLAS Detector at the LHC." In: *Physics Letters B* 716.1 (Sept. 2012), pp. 1–29. ISSN: 0370-2693. DOI: 10.1016/j.physletb.2012.08.020.

[3] The CMS Collaboration. "Observation of a New Boson at a Mass of 125 GeV with the CMS Experiment at the LHC." In: *Physics Letters B* 716.1 (Sept. 2012), pp. 30–61. ISSN: 03702693. DOI: 10.1016/j.physletb.2012.08.021. arXiv: 1207.7235 [hep-ex].

[4] 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 (Aug. 2022), p. 175. ISSN: 1029-8479. DOI: 10.1007/JHEP08(2022)175. arXiv: 2201.08269 [hep-ex].

[5] ATLAS Collaboration. "Cross-Section Measurements of the Higgs Boson Decaying into a Pair of Tau-Leptons in Proton-Proton Collisions at $\sqrt{s} = 13$ TeV with the ATLAS Detector." In: *Physical Review D* 99.7 (Apr. 2019), p. 072001. ISSN: 2470-0010, 2470-0029. DOI: 10.1103/PhysRevD.99.072001. arXiv: 1811.08856 [hep-ex].

[6] ATLAS Collaboration. "Direct Constraint on the Higgs-charm Coupling from a Search for Higgs Boson Decays into Charm Quarks with the ATLAS Detector." Oct. 2022. DOI: 10.1140/epjc/s10052-022-10588-3. arXiv: 2201.11428 [hep-ex].

[7] Abdus Salam. "Weak and Electromagnetic Interactions." In: *Selected Papers of Abdus Salam*. Vol. Volume 5. World Scientific Series in 20th Century Physics. WORLD SCIENTIFIC, May 1994, pp. 244–254. ISBN: 978-981-02-1662-7. DOI: 10.1142/9789812795915_0034.

[8] Steven Weinberg. "A Model of Leptons." In: *Physical Review Letters* 19.21 (Nov. 1967), pp. 1264–1266. DOI: 10.1103/PhysRevLett.19.1264.

[9] Sheldon L. Glashow. "Partial-Symmetries of Weak Interactions." In: *Nuclear Physics* 22.4 (Feb. 1961), pp. 579–588. ISSN: 0029-5582. DOI: 10.1016/0029-5582(61)90469-2.

[10] G. 't Hooft and M. Veltman. "Regularization and Renormalization of Gauge Fields." In: *Nuclear Physics B* 44.1 (July 1972), pp. 189–213. ISSN: 0550-3213. DOI: 10.1016/0550-3213(72)90279-9.

[11] LHCb Collaboration. "Observation of a $J/\psi\Lambda$ Resonance Consistent with a Strange Pentaquark Candidate in $B^- \to J/\psi\Lambda\bar{p}$ Decays." Oct. 2022. DOI: 10.1103/PhysRevLett.131.031901.

[12] Particle Data Group. "Review of Particle Physics." In: *Progress of Theoretical and Experimental Physics* 2022.8 (Aug. 2022), p. 083C01. ISSN: 2050-3911. DOI: 10.1093/ptep/ptac097.

[13] ATLAS Collaboration. *Expected performance of the ATLAS experiment: detector, trigger and physics*. Geneva: CERN, 2009.

[14] ATLAS Collaboration. "Identification of Hadronic Decays of Tau Leptons in 2012 Data with the ATLAS Detector."

[15] Jonathan Shlomi, Peter Battaglia, and Jean-Roch Vlimant. *Graph Neural Networks in Particle Physics*. Tech. rep. 2020. arXiv: 2007.13681v2.

[16] Zhou Lu, Hongming Pu, Feicheng Wang, Zhiqiang Hu, and Liwei Wang. "The Expressive Power of Neural Networks: A View from the Width." In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Vol. 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper/2017/file/32cbf687880eb1674a07bf717761dd3a-Paper.pdf.

[17] 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: https://doi.org/10.1016/0893-6080(89)90020-8. URL: https://www.sciencedirect.com/science/article/pii/0893608089900208.

[18] Murat Abdughani, Daohan Wang, Lei Wu, Jin Min Yang, and Jun Zhao. *Probing triple Higgs coupling with machine learning at the LHC*. Tech. rep. 2020. arXiv: 2005.11086v3.

[19] Murat Abdughani, Jie Ren, Lei Wu, and Jin Min Yang. *Probing stop pair production at the LHC with graph neural networks*. Tech. rep. 2019. arXiv: 1807.09088v2.

[20] Jie Ren, Lei Wu, and Jin Min Yang. *Unveiling CP property of top-Higgs coupling with graph neural networks at the LHC*. Tech. rep. arXiv: 1901.05627v1.

[21] Peter W Battaglia et al. *Relational inductive biases, deep learning, and graph networks*. Tech. rep. arXiv: 1806.01261v3.

[22] A. Hoecker et al. "TMVA - Toolkit for Multivariate Data Analysis." July 2009. DOI: 10.48550/arXiv.physics/0703039. arXiv: physics/0703039.

[23] Tianqi Chen and Carlos Guestrin. "XGBoost: A Scalable Tree Boosting System." In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Aug. 2016, pp. 785–794. DOI: 10.1145/2939672.2939785. arXiv: 1603.02754 [cs].

[24] Fabian Pedregosa et al. "Scikit-Learn: Machine Learning in Python." June 2018. DOI: 10.48550/arXiv.1201.0490. arXiv: 1201.0490 [cs].

[25] Lyndon Evans and Philip Bryant. "LHC Machine." In: *Journal of Instrumentation* 3.08 (2008), S08001–S08001. DOI: 10.1088/1748-0221/3/08/s08001. URL: https://doi.org/10.1088/1748-0221/3/08/s08001.

[26] R Alemany-Fernandez et al. *Operation and Configuration of the LHC in Run 1*. Tech. rep. 2014.

[27] J Wenninger. *Operation and Configuration of the LHC in Run 2*. Tech. rep.

[28] ATLAS Collaboration. *The ATLAS Experiment at the CERN Large Hadron Collider*. Tech. rep. DOI: 10.1088/1748-0221/3/08/S08003.

[29] J. T. Boyd. "LHC Run-2 and Future Prospects." 2020. arXiv: 2001.04370 [hep-ex].

[30] ATLAS Collaboration. "ATLAS Inner Detector: Technical Design Report, 1." Geneva, 1997.

[31] Yosuke Takubo and ATLAS Collaboration. "ATLAS IBL Operational Experience." In: *Proceedings of The 25th International Workshop on Vertex Detectors — PoS(Vertex 2016)*. La Biodola, Isola d'Elba, ITALY: Sissa Medialab, Feb. 2017, p. 004. DOI: 10.22323/1.287.0004.

[32] ATLAS Collaboration. "The ATLAS Experiment at the CERN Large Hadron Collider: A Description of the Detector Configuration for Run 3." May 2023. DOI: 10.48550/arXiv.2305.16623. arXiv: 2305.16623 [hep-ex, physics:physics].

[33] ATLAS Collaboration. "ATLAS Muon Spectrometer: Technical Design Report." June 1997.

[34] ATLAS Collaboration. "Topological Cell Clustering in the ATLAS Calorimeters and Its Performance in LHC Run 1." In: *The European Physical Journal C* 77.7 (July 2017), p. 490. ISSN: 1434-6044, 1434-6052. DOI: 10.1140/epjc/s10052-017-5004-5. arXiv: 1603.02934 [hep-ex].

[35] ATLAS Collaboration. "Performance of the ATLAS Track Reconstruction Algorithms in Dense Environments in LHC Run 2." In: *The European Physical Journal C* 77.10 (Oct. 2017), p. 673. ISSN: 1434-6044, 1434-6052. DOI: 10.1140/epjc/s10052-017-5225-7. arXiv: 1704.07983 [hep-ex].

[36] ATLAS Collaboration. "Reconstruction of Primary Vertices at the ATLAS Experiment in Run 1 Proton–Proton Collisions at the LHC." In: *The European Physical Journal C* 77.5 (May 2017), p. 332. ISSN: 1434-6052. DOI: 10.1140/epjc/s10052-017-4887-5.

[37] ATLAS Collaboration. "Electron and Photon Performance Measurements with the ATLAS Detector Using the 2015-2017 LHC Proton-Proton Collision Data." In: *JINST* 14.12 (2019), P12006. DOI: 10.1088/1748-0221/14/12/P12006.

[38] ATLAS Collaboration. "Muon Reconstruction and Identification Efficiency in ATLAS Using the Full Run 2 Pp Collision Data Set at \sqrts=13 TeV." In: *Eur. Phys. J., C* 81 (2021), p. 578. DOI: 10.1140/epjc/s10052-021-09233-2.

[39] M Mangano. "QCD and the Physics of Hadronic Collisions." 2018. DOI: 10.23730/CYRSP-2018-004.27.

[40] Matteo Cacciari, Gavin P. Salam, and Gregory Soyez. "The Anti-Kt Jet Clustering Algorithm." In: *Journal of High Energy Physics* 2008.04 (Apr. 2008), p. 063. ISSN: 1126-6708. DOI: 10.1088/1126-6708/2008/04/063.

[41] ATLAS Collaboration. "Jet Reconstruction and Performance Using Particle Flow with the ATLAS Detector." In: *The European Physical Journal C* 77.7 (July 2017), p. 466. ISSN: 1434-6044, 1434-6052. DOI: 10.1140/epjc/s10052-017-5031-2. arXiv: 1703.10485 [hep-ex].

[42] ATLAS Collaboration. "Performance of b-Jet Identification in the ATLAS Experiment." In: *Journal of Instrumentation* 11.04 (Apr. 2016), P04008–P04008. ISSN: 1748-0221. DOI: 10.1088/1748-0221/11/04/P04008. arXiv: 1512.01094 [hep-ex].

[43] ATLAS Collaboration. "ATLAS Flavour-Tagging Algorithms for the LHC Run 2 pp Collision Dataset." In: *The European Physical Journal C* 83.7 (July 2023), p. 681. ISSN: 1434-6052. DOI: 10.1140/epjc/s10052-023-11699-1. arXiv: 2211.16345 [hep-ex, physics:physics].

[44] ATLAS Collaboration. "Evidence for the $VH, H \to \tau\tau$ process with the ATLAS detector in Run 2." Dec. 2023.

[45] ATLAS Collaboration. "Identification of Hadronic Tau Lepton Decays Using Neural Networks in the ATLAS Experiment." Geneva, 2019.

[46] A. L. Read. "Presentation of Search Results: The CLs Technique." In: *Journal of Physics G: Nuclear and Particle Physics* 28.10 (Sept. 2002), p. 2693. ISSN: 0954-3899. DOI: 10.1088/0954-3899/28/10/313.

[47] ATLAS Collaboration. "Cross-section measurements of the Higgs boson decaying into a pair of τ-leptons in proton-proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector." In: *Phys. Rev. D* 99 (7 2019), p. 072001. DOI: 10.1103/PhysRevD.99.072001. URL: https://link.aps.org/doi/10.1103/PhysRevD.99.072001.

[48] ATLAS Collaboration. "Measurements of WH and ZH production in the $H \to b\bar{b}$ decay channel in pp collisions at 13TeV with the ATLAS detector." In: (2021). DOI: 10.1140/epjc/s10052-020-08677-2. arXiv: 2007.02873v2.

[49] The CMS Collaboration. "Evidence for Higgs boson decay to a pair of muons." In: *Journal of High Energy Physics* 2021.1 (2021). ISSN: 1029-8479. DOI: 10.1007/jhep01(2021)148. URL: http://dx.doi.org/10.1007/JHEP01(2021)148.

[50] 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 s=13 TeV with the ATLAS detector." In: *Physics Letters B* 816 (2021), p. 136204. ISSN: 0370-2693. DOI: https://doi.org/10.1016/j.physletb.2021.136204. URL: https://www.sciencedirect.com/science/article/pii/S0370269321001441.

[51] Nicolas Berger et al. "Simplified Template Cross Sections - Stage 1.1." June 2019. DOI: 10.48550/arXiv.1906.02754. arXiv: 1906.02754 [hep-ex, physics:hep-ph].

[52] ATLAS Collaboration. "Combined Effective Field Theory Interpretation of Higgs Boson and Weak Boson Production and Decay with ATLAS Data and Electroweak Precision Observables." Geneva, 2022.

[53] ATLAS Collaboration. "Combined Measurements of Higgs Boson Production and Decay Using up to 139fb^{-1} of Proton–Proton Collision Data at $\sqrt{s} = 13$ TeV Collected with the ATLAS Experiment."

[54] ATLAS Collaboration. "Variable Radius, Exclusive-k$_T$, and Center-of-Mass Subjet Reconstruction for Higgs$\to b\bar{b}$ Tagging in ATLAS." Geneva, 2017.

[55] 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: *The European Physical Journal C* 81.2 (Feb. 2021), p. 178. ISSN: 1434-6044, 1434-6052. DOI: 10.1140/epjc/s10052-020-08677-2. arXiv: 2007.02873 [hep-ex].

[56] ATLAS Collaboration. "Flavor Tagging Efficiency Parametrisations with Graph Neural Networks." Geneva, 2022.

[57] Dorival Goncalves and Junya Nakamura. "Role of the Z Polarization in the $H \to b\bar{b}$ Measurement." In: *Physical Review D* 98.9 (Nov. 2018), p. 093005. ISSN: 2470-0010, 2470-0029. DOI: 10.1103/PhysRevD.98.093005. arXiv: 1805.06385 [hep-ex, physics:hep-ph].

[58] Adam Paszke et al. "PyTorch: An Imperative Style, High-Performance Deep Learning Library." In: *Advances in Neural Information Processing Systems 32*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett. Curran Associates, Inc., 2019, pp. 8024–8035. URL: http://papers.neurips.cc/

paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

[59] ATLAS Collaboration. "Measurements of WH and ZH Production with Higgs Boson Decays into Bottom Quarks and Direct Constraints on the Charm Yukawa Coupling in 13 TeV pp Collisions with the ATLAS Detector." Oct. 2024. DOI: 10.48550/arXiv.2410.19611. arXiv: 2410.19611 [hep-ex].

[60] ATLAS Collaboration. "Searches for lepton-flavour-violating decays of the Higgs boson into $e\tau$ and $\mu\tau$ in $\sqrt{s} = 13$ TeV pp collisions with the ATLAS detector." In: *Journal of High Energy Physics* 2023.7 (July 2023), p. 166. ISSN: 1029-8479. DOI: 10.1007/JHEP07(2023)166. arXiv: 2302.05225 [hep-ex].

[61] ATLAS Collaboration. "Search for resonant and non-resonant Higgs boson pair production in the $b\bar{b}\tau^+\tau^-$ decay channel using 13 TeV pp collision data from the ATLAS detector." In: *Journal of High Energy Physics* 2023.7 (July 2023), p. 40. ISSN: 1029-8479. DOI: 10.1007/JHEP07(2023)040. arXiv: 2209.10910 [hep-ex].

[62] ATLAS Collaboration. "Search for heavy Higgs bosons decaying into two tau leptons with the ATLAS detector using pp collisions at $\sqrt{s} = 13$ TeV." In: *Physical Review Letters* 125.5 (July 2020), p. 051801. ISSN: 0031-9007, 1079-7114. DOI: 10.1103/PhysRevLett.125.051801. arXiv: 2002.12223 [hep-ex].

[63] ATLAS Collaboration. "Search for dark matter produced in association with a Higgs boson decaying to tau leptons at $\sqrt{s} = 13$ TeV with the ATLAS detector." In: *Journal of High Energy Physics* 2023.9 (Sept. 2023), p. 189. ISSN: 1029-8479. DOI: 10.1007/JHEP09(2023)189. arXiv: 2305.12938 [hep-ex].

[64] ATLAS Collaboration. "Search for bottom-squark pair production in pp collision events at $\sqrt{s} = 13$ TeV with hadronically decaying τ-leptons, b-jets and missing transverse momentum using the ATLAS detector." Mar. 2021. DOI: 10.1103/PhysRevD.104.032014.

[65] ATLAS Collaboration. "Search for new phenomena in pp collisions in final states with tau leptons, b-jets, and missing transverse momentum with the ATLAS detector." In: *Physical Review D* 104.11 (Dec. 2021), p. 112005. ISSN: 2470-0010, 2470-0029. DOI: 10.1103/PhysRevD.104.112005. arXiv: 2108.07665 [hep-ex].

[66] ATLAS Collaboration. "Measurement of τ polarisation in $Z/\gamma^* \to \tau\tau$ decays in proton–proton collisions at $\sqrt{s} = 8$ TeV with the ATLAS detector." In: *The European Physical Journal C* 78.2 (Feb. 2018), p. 163. ISSN: 1434-6052. DOI: 10.1140/epjc/s10052-018-5619-1.

[67] ATLAS Collaboration. *ATLAS Liquid Argon Calorimeter Phase-I Upgrade: Technical Design Report*. Technical Design Report. ATLAS 22. 2013.

[68] ATLAS Collaboration. "Luminosity Determination in pp Collisions at $\sqrt{s} = 13$ TeV Using the ATLAS Detector at the LHC." Geneva, 2019.

[69] ATLAS Collaboration. *Technical Design Report for the Phase-I Upgrade of the ATLAS TDAQ System*. Tech. rep. 2013.

[70] Darin Acosta et al. "Boosted Decision Trees in the Level-1 Muon Endcap Trigger at CMS." In: *Journal of Physics: Conference Series* 1085.4 (Sept. 2018), p. 042042. ISSN: 1742-6596. DOI: 10.1088/1742-6596/1085/4/042042.

[71] ATLAS Collaboration. "The ATLAS Level-1 Calorimeter Trigger." Geneva, 2008. DOI: 10.1088/1748-0221/3/03/P03001.

[72] S. Summers et al. "Fast Inference of Boosted Decision Trees in FPGAs for Particle Physics." In: *Journal of Instrumentation* 15.05 (May 2020), P05026. ISSN: 1748-0221. DOI: 10.1088/1748-0221/15/05/P05026.

[73] ATLAS Collaboration. "Performance of the ATLAS Tau Trigger in 2016 Data." Geneva, 2016.

[74] ATLAS collaboration. "The ATLAS Tau Trigger in Run 2." 2017.

[75] Patrick T. Komiske, Eric M. Metodiev, and Jesse Thaler. "Energy Flow Polynomials: A Complete Linear Basis for Jet Substructure." In: *Journal of High Energy Physics* 2018.4 (Apr. 2018), p. 13. ISSN: 1029-8479. DOI: 10.1007/JHEP04(2018)013. arXiv: 1712.07124 [hep-ex, physics:hep-ph].

[76] ATLAS Collaboration. "L1Calo eFEX Tau BDT Plots." Geneva, 2024.

[77] 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, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Curran Associates, Inc., 2017, pp. 4765–4774. URL: http:

//papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

[78] ATLAS Collaboration. "Performance of the ATLAS Level-1 Topological Trigger in Run 2." Jan. 2022. DOI: 10.1140/epjc/s10052-021-09807-0. arXiv: 2105.01416 [hep-ex].

[79] *Technical Design Report for the Phase-I Upgrade of the ATLAS TDAQ System*. Tech. rep. Final version presented to December 2013 LHCC. 2013. URL: https://cds.cern.ch/record/1602235.

[80] Rene Brun et al. *root-project/root: v6.18/02*. Version v6-18-02. June 2020. DOI: 10.5281/zenodo.3895860. URL: https://doi.org/10.5281/zenodo.3895860.

[81] The pandas development team. *pandas-dev/pandas: Pandas*. Version latest. Feb. 2020. DOI: 10.5281/zenodo.3509134. URL: https://doi.org/10.5281/zenodo.3509134.

[82] A. Hagberg, D. Schult, P. Swart, and J. M. Hagberg. "Exploring Network Structure, Dynamics, and Function Using NetworkX." In: 2008.

[83] S. T. Roche, Q. Bayer, B. T. Carlson, W. C. Ouligian, P. Serhiayenka, J. Stelzer, and T. M. Hong. "Nanosecond Anomaly Detection with Decision Trees and Real-Time Application to Exotic Higgs Decays." In: *Nature Communications* 15.1 (Apr. 2024), p. 3527. ISSN: 2041-1723. DOI: 10.1038/s41467-024-47704-8.

תקציר

ההתאמה יוצאת הדופן בין התחזיות של המודל הסטנדרטי של פיזיקת החלקיקים לבין המדידות שבוצעו בניסויים במאיץ ההדרונים הגדול (LHC) ב–CERN דורשת איסוף כמויות הולכות וגדלות של נתונים כדי לקבוע גבולות עליונים מחמירים יותר ויותר על ערכי תצפית, במסגרת החיפוש אחר פיזיקה מעבר למודל הסטנדרטי. שדרוגים שמטרתם להגדיל את כמות הנתונים המסופקים לניסויים דוחפים את גבולות האלגוריתמים המתמחים בהפרדה בין אות לרקע, על מנת לייצר דגימות נקיות לניתוחים פיזיקליים. התקדמויות של השנים האחרונות בטכנולוגיית למידת מכונה מאפשרות פיתוח של דור חדש של אלגוריתמים אלה, אשר במקרים רבים עולים בביצועיהם על מקביליהם מבוססי ההיוריסטיקות, הן בתרחישים של טריגר בזמן אמת והן בניתוח נתונים אגורים.

עבודת מחקר זו בוחנת את האלגוריתמים מהדור החדש בהקשרים של ניתוח מקוון ולא מקוון. הניתוח המקוון מתמקד בזיהוי לפטונים מסוג טאו בטריגר מבוסס החומרה של ניסוי ATLAS תוך שימוש בעצי החלטות, ואילו הניתוח הלא מקוון בוחן את השימוש ברשתות עצביות ובעצי החלטות באנליזה בה מתבצע חיפוש של דעיכת בוזון ההיגס לשני קווארקים קסומים במקביל למדידה מדוייקת של דעיכתו לשני קווארקים תחתונים, אשר מציבה את המגבלות המחמירות ביותר עד כה על הצימוד של בוזון היגס לקווארקים אלו.

העבודה המוצגת בתיזה זו מביאה לפיתוח המודל מבוסס למידת מכונה הראשון המשמש לאיסוף נתונים בטריגר החומרתי של ניסוי ATLAS. המודל משמש לזיהוי ראשוני של לפטונים מסוג טאו ומבוסס על עץ החלטות ללא דרישות של חומרה ייעודית. טכנולוגיה זו פותחת את הדלת לדור הבא של אלגוריתמי טריגר בניסוי ATLAS.

School of Physics and Astronomy
The Raymond and Beverly Sackler
Faculty of Exact Sciences
Tel Aviv University

בית הספר לפיזיקה ולאסטרונומיה
הפקולטה למדעים מדויקים
ע"ש ריימונד ובברלי סאקלר
אוניברסיטת תל אביב

תזת מחקר עבור התואר "דוקטור לפילוסופיה"

שימוש בלימוד מכונה לזיהוי מיידי של לפטונים מסוג טאו בחומרה ולאנליזה של חלקיקי בוזון היגס בדעיכה לקווארקים מסוג תחתון וקסם בניסוי אטלס

הוגש לסנאט אוניברסיטת תל אביב על ידי

דוד רייכר

עבודה זו בוצעה בהנחייתו של

פרופ. ארז עציון