Boosted particle tagging at the LHC based on large R jet substructure using machine learning techniques.

Carlos Fernando Buitrago Cárdenas

April 17, 2023

Abstract

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Table of Contents

Introduction

Despite its huge success, the Standard Model (SM) of particle physics is regarded as an incomplete theory, as has been shown multiple times in the past. Some of the challenges the SM faces include the hierarchy problem (which involves the mass of the Higgs boson), the absence of a candidate for dark matter and the matter-antimatter asymmetry of the universe. Several extensions to the SM have been proposed in order to deal with these issues and their predictions are constantly being put to test by the Large Hadron Collider (LHC) experiments. Most of these SM extensions involve models which predict the existence of new heavy particles with decay channels involving top quarks, electroweak bosons and Higgs bosons. If these new states are heavy enough their decay products are expected to have large transverse momenta greatly exceeding their rest masses ($p_T \gg m$). These kind of objects are called *boosted objects*.

In light of the previous statements, the search for new physics at the LHC will continue to look further and further into previously unexplored kinematic regimes, and as a result of the high centre-of-mass energy that it has achieved, the LHC is already able to produce a large number of boosted particles across many final states. As the sensitivity of searches for new phenomena depends directly on them, it is of utmost importance to efficiently reconstruct and identify these kind of objects. In order to do so, many new techniques have been developed which rapidly gave birth to a new fast growing field of research.

When the boost factor is large enough, boosted objects decay into a highly collimated

spray of hadrons. Because of this, the decay products of these kind of particles would be reconstructed as a single jet by standard jet algorithms. When boosted objects decay hadronically commonly used reconstruction methods become inefficient due to the large background of ordinary QCD jets originated from light quarks or gluons.

The identification of boosted hadronically decaying objects can be achieved by using jet substructure techniques. Looking at the decay products of boosted particles as collimated (and unresolved) jets separately most likely becomes a futile effort. Instead, a single large R jet (called a fat jet) containing all of the decay products can be reconstructed. The internal structure of these fat jets can be used to distinguish between those originating from boosted electroweak bosons and those originating from light quark or gluons. In order to achieve this, different jet shape methods are used, which take advantage of the difference in the energy patterns of signal and background jets.

This work will be mainly focused on the exploration of the "N-subjettiness" jet shape, denoted by τ_N. This substructure variable gives us information about how much the radiation of a jet is aligned along N different axes within it, effectively telling us how well a jet is described as having N subjets. As will be discussed further below, the τ_N variables have discriminant capabilities which will be analysed in the present study.

Along with the jet mass, which also gives us information of the jet's substructure, τ_N will be used as discriminant variables in the tagging of boosted top quarks and boosted W bosons. In order to carry out the tagging, these variables will be used as inputs for machine learning algorithms in the form of a multivariate analysis. The performance of the tagging method will be assessed, including a comparison between the different multivariate classifiers implemented and a comparison between the input variables proposed and other variables that are commonly used in the tagging of heavy particles.

Theoretical aspects

2.1 Jets and algorithms

After being produced in a high-energy event, quarks and gluons fragment and hadronize resulting in a collimated spray of hadrons called a jet. The reason behind the process of hadronization lies in the concept of colour confinement. In quantum chromodynamics (QCD), colour confinement states that only objects with non zero colour charge can propagate as free particles, therefore quarks and gluons are only seen bound together in the form of hadrons. When particles carrying colour charge (namely quarks and gluons) are separated in a high-energy event, new colour carrying particles are spontaneously created from the vacuum in order to form colourless hadrons, thus obeying confinement.

While hadronization is not yet fully understood and a theoretical description of the process is not yet available, there is a number of phenomenological models such as the Lund String Model that do a good job of describing it [1]. The phenomenon can be understood qualitatively through these models by taking into account that the gluon field between colour charges becomes a narrow flux tube as they get separated and eventually it becomes energetically favourable for a new particle to appear rather than extending the tube further, as can be seen in figure 2.1.

The particles resulting from the hadronization of a single quark (parton) tend to travel in

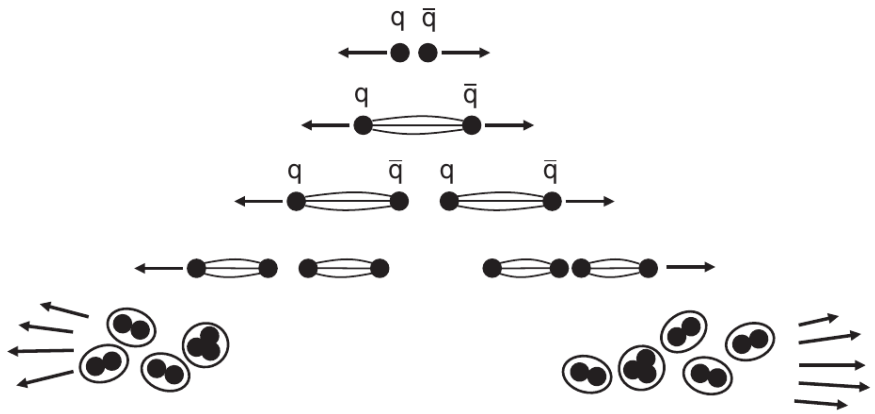


Figure 2.1: Schematic representation of the hadronization process. Image taken from [2].

the same direction as their parent forming a narrow cone, which is what is known as a jet. In particle detector experiments jets are observed instead of quarks, and their structure is quite visible when looking at the events reconstructed in the detector. Through the measurement of the different properties of a jet it is possible to obtain information about the original parton that originated it. Therefore jets are an essential part of the analyses carried out in collider experiments.

As important as they are, to be used effectively in analyses jets need to be well defined. As stated in the work of Salam [3] a jet definition is constituted by a jet algorithm with its respective parameters and recombination scheme. A jet algorithm is a set of rules that group particles into jets. These algorithms usually involve parameters that govern their behaviour, for example in defining how close two particles need to be in order for them to be considered part of the same jet. Jet algorithms are also related to a certain recombination scheme, which indicates how the momentum is assigned to the object resulting of merging two particles during a clustering process.

Jet algorithms can be usually classified into two broad categories: cone algorithms and sequential recombination algorithms. Cone algorithms originate from the initial idea of Sterman and Weinberg [4]. They are considered "top-down" algorithms, since they group together particles within specific conical angular regions so that the resulting cone is "stable",

meaning that the direction of the cone matches that of the 4-momenta sum of the particles. On the other hand, sequential recombination algorithms are considered "bottom-up", as they iteratively recombine nearby particles in accordance to a certain distance measure.

The anti-k_t algorithm [5] is a sequential recombination algorithm widely used in collider experiments, being also the preferred jet identification algorithm in ATLAS analyses. Since the present study is based on simulated data from this detector (see section 3.2) it is relevant to give a brief overview of this algorithm. The anti-k_t takes as an input a list of N objects and it returns a list of jets, which correspond to clusters of said objects grouped according to specific rules regarding distances between them. The distances used by the algorithm are calculated from the quantities k_{tX}, η_X and ϕ_X which correspond to the transverse momentum, pseudo-rapidity and azimuthal angle of the object X. These distances are d_{ij} (the distance between objects i and j) and d_{iB} (the distance between the object i and the beam), they are defined as follows:

$$d_{ij} = \min(k_{ti}^{-2}, k_{tj}^{-2}) \frac{\Delta_{ij}^2}{R^2} \quad , \tag{2.1}$$

$$d_{iB} = k_{ti}^{-2} \quad , \tag{2.2}$$

where $\Delta_{ij}^2 = (\eta_i - \eta_j)^2 + (\phi_i - \phi_j)^2$ and R is the radius parameter that sets the size scale of the jets found. The algorithm iteratively forms clusters by identifying the smallest of the two distances for all the objects in the input list. If d_{ij} is the smallest, objects i and j are recombined and replaced in the object list by the recombined object. If on the other hand d_{iB} is the smallest, object i is removed from the object list and marked as a jet.

2.2 Boosted particles and fat jets

The use of the term "boosted" in particle physics originates from the concept of a Lorentz boost, which is a type of rotation-free Lorentz transformation between frames moving with

different velocities. A "boosted object" refers then to a particle which travels at a very high speed (with very high transverse momentum p_T) after being produced in a collision. With the high centre-of-mass energy that has been achieved by the LHC (see section 2.4) large samples of top quarks, W, Z and Higgs bosons with a p_T considerably higher than their rest mass m ($p_T \gg m$) are being produced like never before [6], which might also be true for heavier particles that remain yet unknown. However, in this kinematic regime the traces left by even well-known particles differ greatly from those left by the same particles with lower p_T values.

When a massive particle that has been produced with a significant boost decays hadronically, its decay products appear collimated in the momentum direction of the boosted mother particle. If the particle is boosted enough it won't decay to separated (resolved) jets but into collimated jets which will merge. As a result, what is detected are not multiple jets but a single large R jet, usually called fat jet [7]. The effect of the boost of a particle in its decay is illustrated in figure 2.2.

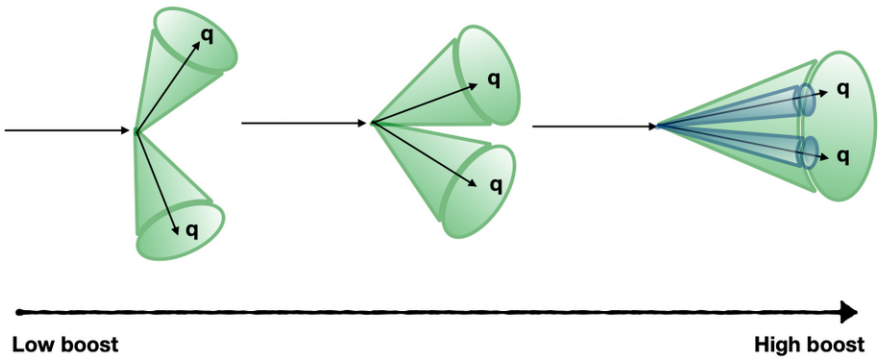


Figure 2.2: Schematic representation of jets arising from a particle that decays to two quarks with increasing Lorentz boost. Image taken from [8].

For a two-body decay, the distance between the decay products of a particle in the rapidity-azimuth plane $\left(\Delta R = \sqrt{\eta^2 + \phi^2}\right)$ is given by:

$$\Delta R \approx \frac{2m}{p_T} \quad , \tag{2.3}$$

in which m and p_T are the mass and the transverse momentum of the particle. If the p_T of the mother particle is high enough the decay products will be too close for conventional reconstruction techniques to be able to resolve them. As an example, for a W boson with $p_T = 300$ GeV the distance between its decay products is $\Delta R \approx 0.5$. Reconstructing them with jets of the conventional size used in the LHC ($R = 0.4 - 0.6$) would result in a failed attempt.

The chosen strategy to deal with the hadronic decays of these boosted particles is to use a much larger jet radius parameter in order to capture the energy of the whole decay in a single fat jet. The key to identifying and measuring boosted particles lies then in the internal structure of the reconstructed fat jets [9], whose discriminating power is the object of our study. Since it is important that the fat jets contain all of the decay products of the boosted particles, there will be a minimum possible size for the fat jet radius parameter, which will decrease as the boost of the particle increases. An example can be seen in figure 2.3, in which the distance between the products in hadronic top quark decay $t \to bqq$ define the minimum radius parameter that should be used for the reconstructed fat jet.

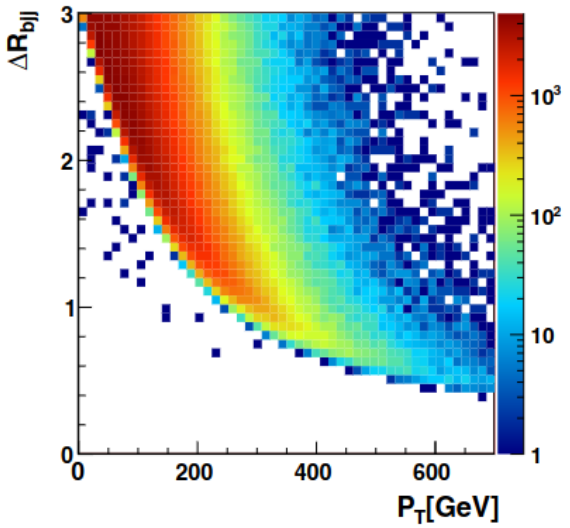


Figure 2.3: Distance between the products of the decay $t \to bqq$ as a function of the top quark p_T. Image taken from [10].

2.3 Fat jet substructure: N-subjettiness

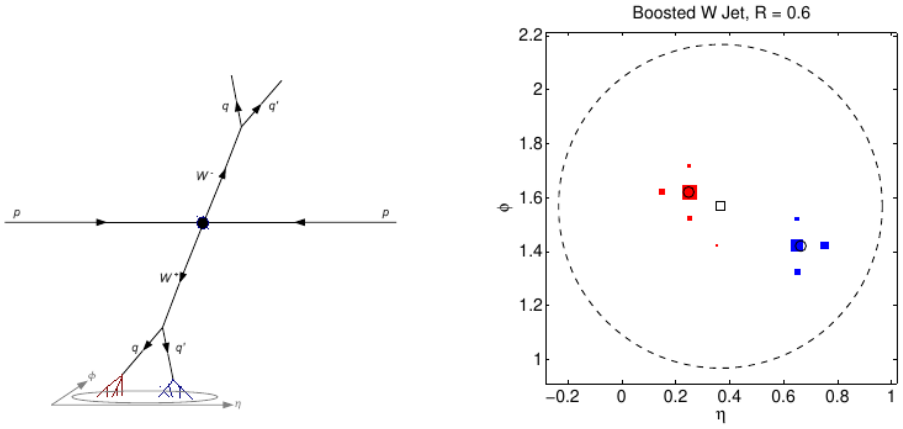
The high centre-of-mass energy achieved by the LHC together with the detection capabilities of the ATLAS calorimeter provide an excellent environment to study hadronic jets. Analyses involving hadronic decay channels are becoming increasingly more relevant after it was initially shown that these channels become usable by looking at boosted topologies [11]. Hadronic channels often have the highest branching ratios, but had been considered impractical due to the large background they present at a hadron collider.

The efficient reconstruction and identification of boosted object decays is of utmost importance in searches of new phenomena at high energies. As was stated in the end of section 2.2, the internal structure of fat jets give us crucial information for the identification of boosted particles. Therefore, since the utility of boosted signals was brought to light, several studies on jet substructure have been carried out [12] and a rich and continuously expanding field was born. The objective of these studies is to increase how well the jets resulting from boosted electroweak bosons and top quarks can be distinguished from background QCD jets (originating from hard light quarks or gluons).

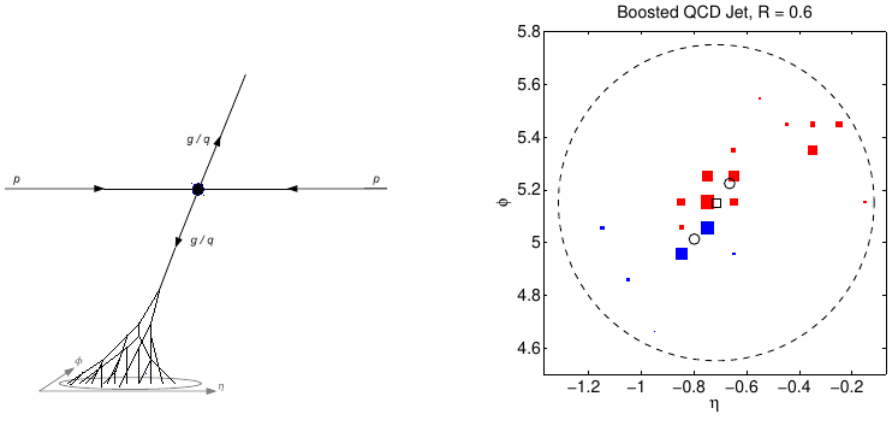
Substructure information used to distinguish between jets originating from boosted hadronically decaying objects and QCD jets is extracted from the jet clustering procedure by algorithmic methods called shape methods. Jet shape methods use certain observables that take advantage of the different energy flow in the decay pattern of signal and background jets. The method studied in this work revolves around a group of variables called N-subjettiness, a jet shape denoted by τ_N and first introduced by Thaler and Van Tilburg [13], which is based on the original N-jettiness shape [14].

The energy pattern of boosted hadronically decaying particles is fundamentally different from that of QCD jets of a similar invariant mass. As an example, take the case of a boosted W boson (a similar discussion holds for the case of other boosted objects), which decays

hadronically to two quarks. The single jet containing the decay products of the boosted W should be composed of two distinct subjets with a combined invariant mass near 80 GeV. A background QCD fat jet with a similar invariant mass originates from a single hard parton and acquires mass through large angle soft splittings. N-subjettiness exploits the difference in energy flow between this two types of jet by taking into account the number of energy lobes within each one. This difference can be easily seen in the reconstructed jets for the previous example, as shown in figure 2.4.



(a)



(b)

Figure 2.4: Schematic of the hadronic decay and a typical event display for a reconstructed jet with mass $m \approx 80$ GeV in (a) W^+W^- and (b) dijet QCD events. The jets are clustered with the anti-k_T algorithm using $R = 0.6$. The marker size for each calorimeter cell is proportional to the logarithm of the energy deposition and each marker color corresponds to a subjet found by the anti-k_T. The small square indicates the total jet direction and the small circles the two subjet directions. Image taken from [13].

The N-subjettiness jet shape denoted by τ_N is defined using the N candidate subjets identified during the jet clustering process as follows:

$$\tau_N = \frac{1}{d_0} \sum_k p_{T,k} \times \min(\Delta R_{1,k}, \Delta R_{2,k}, \ldots, \Delta R_{N,k}) \quad , \tag{2.4}$$

where k runs over the constituent particles of the jet, $p_{T,k}$ is their transverse momenta, $\Delta R_{i,k}$ is the distance in the rapidity-azimuth plane between the candidate subjet i and the constituent particle k, and d_0 is a normalization factor taken as:

$$d_0 = \sum_k p_{T,k} R_0 \quad , \tag{2.5}$$

where R_0 is the jet radius parameter used in the original jet clustering algorithm.

As can be seen from its definition, τ_N gives us information about how well the jet substructure is described by N subjets, or in other words, how "N-subjetty" the jet is. This is done by assessing the degree to which constituent particles are near the subjets. A jet with $\tau_N \approx 0$ has most of its radiation aligned with the direction of the candidate subjets and therefore tends to have N (or fewer) subjets.

A jet originating from a boosted W is expected to have a small τ_2 value. However, QCD jets can also have small τ_2 values. Likewise, W jets are expected to have large τ_1, but QCD jets can also have large τ_1. The key to using N-subjettiness as a discriminating variable effectively lies in the fact that QCD jets with large τ_1, which correspond to jets with a diffuse spray of large angle radiation, typically have large τ_2 values as well. Therefore, the preferred discriminating variable will be the ratio τ_2/τ_1 usually denoted as τ_{21} [15].

It should be noted that the previous discussion holds for boosted Z bosons and Higgs bosons as well. In the case of boosted top quarks, the decay chain needs to be taken into account. Fully hadronically decaying top quarks decay to a b jet and a W boson, followed by the decay of the W boson to two quarks. Therefore a jet resulting from a boosted top

quark will have three lobes of energy, not two. Thus, the preferred discriminating variable for top quarks will be τ_{32} instead of τ_{21}.

Finally, it is important to clarify how jet mass is defined, since it will be used together with the N-subjettiness as it is a variable with discriminating power (on its own). Jet mass M is calculated from the energies and momenta of its constituents as follows:

$$M^2 = \left(\sum_i E_i\right)^2 + \left(\sum_i \vec{p}_i\right)^2 \quad , \tag{2.6}$$

where E_i and $\vec{p}_i$ are the energy and three-momentum of the i^{th} constituent.

2.4 The LHC and the ATLAS detector

The Large Hadron Collider (LHC) [16] is the world's largest particle accelerator. It was built by the European Organization for Nuclear Research (CERN) in the existing 26.7 km tunnel that was constructed originally for the Large Electron-Positron Collider (LEP), which lies between 50 m and 175 m below the surface beneath the France-Switzerland border. The 3.8 m wide tunnel contains two adjacent beamlines, which allow two high-energy particle beams to travel in opposite directions around the accelerator ring. The beams are guided by about 10.000 superconducting magnets [17], dipole magnets are used to bend the beams, quadrupole magnets are used to focus them and magnets of higher multipole orders are used for smaller corrections in the geometry of the field.

The maximum energy that can be reached by the protons in the beam is limited by the peak dipole field, which has a nominal value of 8.33 T (corresponding to a proton energy of 7 TeV) [18]. However, the actual value of the attainable field depends on external factors that cause beam losses. As such, the highest proton energy achieved as of today is 6.8 TeV, which in turn means an attained centre-of-mass energy of $\sqrt{s}$ =13 TeV. In order to achieve

said energy, before being injected to the LHC the protons are pre-accelerated in a series of steps in which their energy is successively increased (see figure 2.5). Initially, H^- ions with an energy of 160 MeV are produced in the linear accelerator LINAC4, which feeds the Proton Synchrotron Booster (PSB), where the electrons are stripped from the ions and the remaining protons are accelerated to 2 GeV and injected into the Proton Synchrotron (PS), which accelerates them to 26 GeV and injects them into the Super Proton Synchrotron (SPS) where they are accelerated to 450 GeV before being finally injected into the main ring.

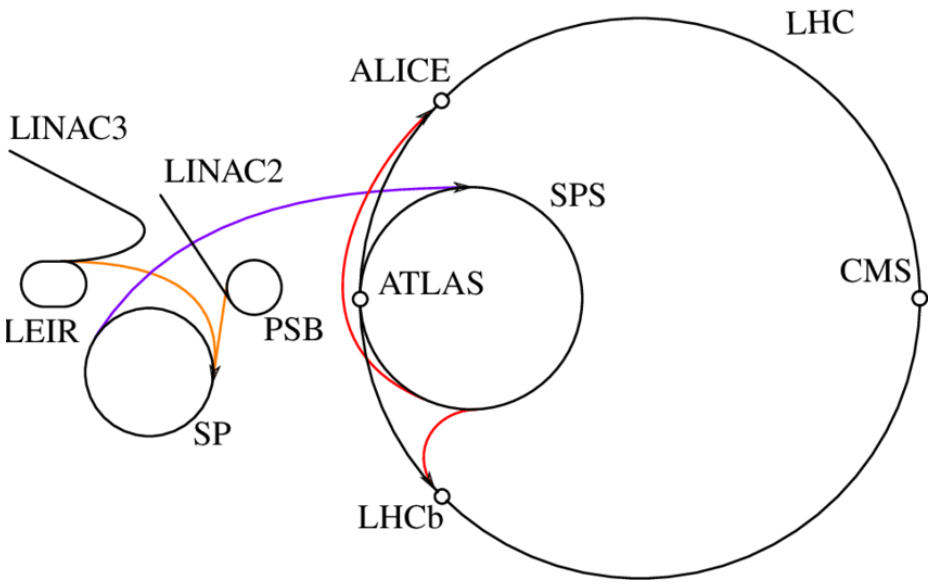


Figure 2.5: Schematic representation of the LHC and its pre-accelerators (in 2020 LINAC2 was replaced by LINAC4). Image taken from [19].

It is also worth noting that the protons do not travel in the form of continuous beams, but rather in bunches of 10^{11} protons each. Under nominal operating conditions each proton beam is composed of 2808 bunches, so that interactions take place at discrete intervals each 25 ns (collision rate of 40 Hz). Taking into account the previous parameters the design (proton-proton) luminosity of the LHC is 10^{34} $\text{cm}^{-2}\text{s}^{-1}$.

The two beams intersect at four different points around the accelerator ring, which is where the collisions occur. Specially strong magnets are used near these points in order to increase the interaction chance. Built around the collision points are seven experiments

installed in underground caverns [20]. Of these, the main four are the ATLAS, CMS, LHCb and ALICE detectors. These detectors are used to count, track and characterize all the particles that are produced in the collisions in order to reconstruct the different events as a whole. As was stated previously, this study will be carried out using simulated data from ATLAS, therefore an overview of the detector will be given below.

The ATLAS (A Toroidal LHC ApparatuS) detector [21] is a multi-purpose detector designed to cover a wide range of physics at the LHC. The detector is forward-backward symmetric with respect to the interaction point, and it has nearly full coverage in solid angle. In order to reconstruct the particles originated from the collisions, ATLAS is composed of multiple layers, each one of them sensitive to different types of particles. The detector layout is shown in figure 2.6.

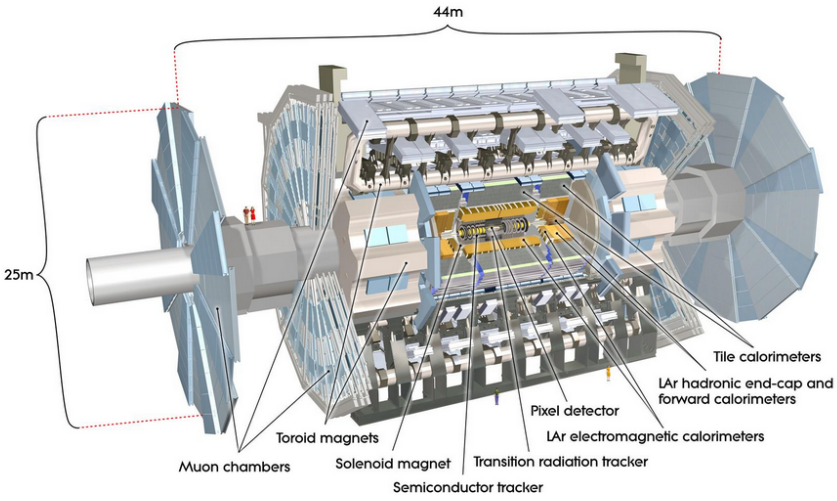


Figure 2.6: Cut-away view of the ATLAS detector. Image taken from [21].

The layers [22] that are closer to the interaction point form what is known as the inner detector, they are immersed in a 2 T magnetic field generated by a solenoid magnet. Charged particles are curved by this field and their trajectories (tracks) are measured by silicon pixel and microstrip detectors, which are surrounded by a transition radiation tracker. Surrounding the inner detector is the calorimeter system, composed by the liquid argon (LAr)

electromagnetic calorimeters and the scintillator-tile hadronic calorimeters. LAr technology is also used in the hadronic end-caps (matching the outer limits of the electromagnetic calorimeters) and forward calorimeters, which extend the detection coverage. Both the electromagnetic and hadronic calorimeters use the same principle: when a particle enters the calorimeter, it showers and deposits energy in the calorimeter cells. The difference between the two kind of calorimeters lies in the materials they are made of and the size of the cells, which determine the particles they target (electromagnetic calorimeters targeting electrons and photons and hadronic calorimeters targeting hadrons). Surrounding the calorimeter system is the muon spectrometer, which is designed to measure the muon tracks (together with neutrinos, muons are the only particles not stopped by the calorimeters). It is composed of high precision tracking chambers immersed in a magnetic field provided by toroid magnets.

Sample generation and machine learning techniques

3.1 Generation of signal and background events

The Monte Carlo samples used in the analysis are produced using a generation chain involving multiple steps involving different software. Both signal and background event generation consist of the same steps. Different samples are produced separately in order to obtain the respective signal and background datasets. Two signal processes will be considered, the first one is used to study the tagging of boosted top quarks and the second one to study the tagging of boosted W bosons. These signal samples will be used together with a single QCD background sample of light quarks and gluons, taking into account the discussion of section 2.3.

The generation chain begins with the generation of matrix elements at leading order (LO) in QCD and electroweak couplings using MadGraph5_aMC@NLO v2.9.12 [23] with the NN23LO1 PDF set [24]. The process defined to produce the signal top jets consists of top quarks decaying in the entirely hadronic channel $pp \to t\tilde{t}$, ($t \to W^+b$, $W^+ \to jj$), ($\tilde{t} \to W^-\tilde{b}$, $W^- \to jj$). On the other hand, the process used to produce the background sample is the default QCD production dijet production routine $pp \to jj$ where the j jets are composed of light quarks (b's included) and gluons.

After the amplitudes and mappings for the relevant processes have been obtained they

are passed to MADEVENT [25], which generates the unweighted tree-level Les Houches events. Both signal and background datasets consist of 500000 events generated at a centre-of-mass energy of 14 TeV. During the event generation stage, a parton-level momentum phase space cut is implemented on all partons in order to reduce computing time for the simulations. Only boosted particles are of interest for this study, and the parton-level cut allows more fat jets to be produced in the established number of events. The cut used is $p_T > 350$ GeV and it applies at parton-level to both gluons and light quarks, as well as W bosons and top quarks. **This initial p_T cut value is later increased for subsequent analyses.**

The tree-level events are showered using PYTHIA v8.306 [26] with the default settings proposed in the MADGRAPH5's implementation of PYTHIA 8 in order to obtain hadron-level events in HEPMC format [27]. Subsequently, the detector response is simulated via DELPHES v3.5.0 [28] using the default ATLAS detector configuration card with the inclusion of the fat jet reconstruction module provided by the FASTJET v3.3.4 library [29].

The fat jets used in the analysis are reconstructed via FASTJET using the anti-k_T algorithm with $R = 1.0$, **(this radius parameter is later increased for subsequent analyses)**. The τ_N values associated to each fat jet are computed from the three-momenta of N candidate subjets using the N-subjettiness package (v2.2.6) of the FASTJET contrib v1.051. In order to obtain the candidate subjets needed for the τ_N computations the fat jet is reclustered with an exclusive k_T algorithm [30].

At high luminosity hadron colliders such as the LHC, the presence of pile-up energy and the underlying event interferes in analyses involving large R jets because they lead to soft, wide-angle contaminations which lead to the dilution of the jet substructure. Grooming techniques are useful to remove these contaminations. The grooming methods applied to the jets reconstructed in this analysis are trimming [31], pruning [32] and soft-dropping [33] algorithms. These methods are implemented via FASTJET integration, using the default parameters in the ATLAS configuration card of DELPHES 3. This selection of parameters is done based on various studies that have been carried out previously in order to compare

and optimize the performance of the different methods.

It should be noted that the description of the event generation process described above also applies for the generation of the signal sample used to study the tagging of boosted W bosons. In this case the process used for the tree-level event generation is W pair production, with each W decaying in the full hadronic channel $pp \to W^+W^-$, $(W^+ \to jj)$, $(W^- \to jj)$, where the j jets are defined as in the process used to produce the background sample.

3.2 Machine learning implementation

The discrimination between signal and background events can be initially done as a cut based analysis by applying subsequent cuts on the substructure variables proposed earlier. This can be easily seen from the comparison between signal and background distributions of the jet mass and τ_{N+1}/τ_N variables (see section 4). However, in previous studies this kind of cut based selections have been found to be insufficient when dealing with data similar to that of this work [34]. With that in mind, the tagging method proposed in this study will be based on machine learning techniques, which take an increasingly important role in LHC experiments as time passes, specially in the field of jet tagging [35].

The machine learning aspect of the proposed tagging method consists on the use of multivariate analysis algorithms. Initially, multiple multivariate classifiers will be considered and their performance will be compared. The classifiers being evaluated are the k-nearest neighbour classifier (k-NN), linear discriminant analysis (LD), function discriminant analysis (FDA), 1-dimensional likelihood estimator (with and without PCA-transformed input variables), Friedman's RuleFit method (RuleFit), multilayer perceptron artificial neural network (MLP), boosted decision tree (BDT) and a deep learning neural network (DNN). For more information on each one of these methods see ref. [36].

The different multivariate methods are implemented using the Toolkit for Multivariate Data Analysis with ROOT (TMVA 4) [36]. This toolkit, integrated into the ROOT analysis framework [37] allows for the training, testing, performance evaluation and application of the different multivariate classification algorithms. The multivariate techniques used belong to the family of "supervised learning" algorithms. They use a training set of events for which the desired output is known (i.e. the algorithm knows if the event is a signal or a background event) in order to determine the mapping function that describes the decision boundary used during the classification.

The discriminating variables used as input for the classifier algorithms are the fat jet mass and N-subjettiness, which contain information on the substructure of the fat jets. As it was mentioned before, the τ_N variables do not have much discriminant power by themselves, but rather their ratios do. An advantage of the multivariate classifiers is that they can be supplied with a full set of τ_N variables so that more complex relations between them can be taken into account when defining the decision boundary that optimizes the classification. As the number of τ_N variables used as input increases, so does the information that the algorithm can use to carry out the classification. However, it is expected that for higher N values, τ_N won't contribute as much to the discrimination. Therefore the performance with different number of τ_N variables used as input will be considered.

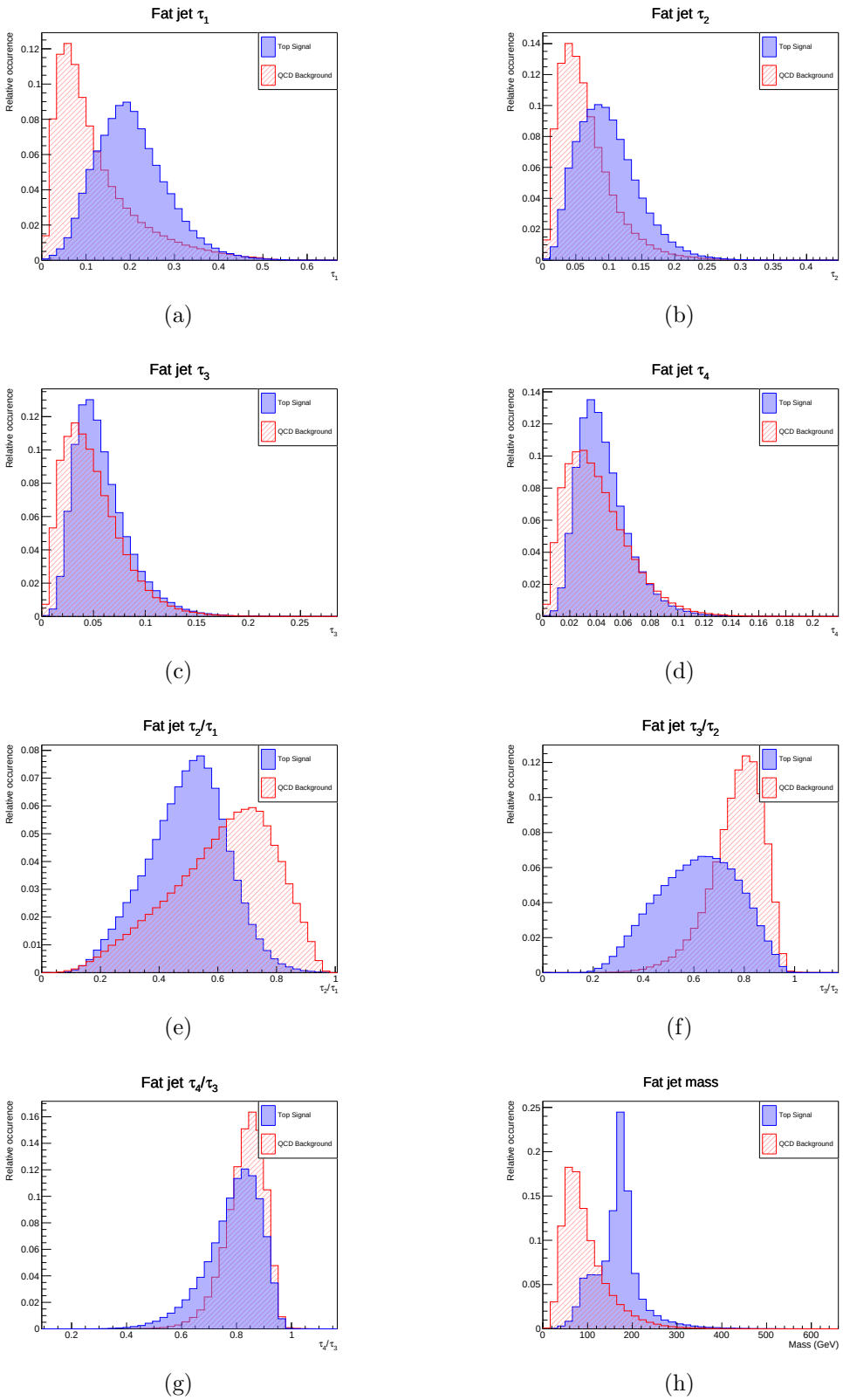
The training and testing sets of events are randomly selected from the generated signal and background samples. For both signal and background samples, half of the events are assigned to the training dataset and the other half are assigned to the testing dataset so that they both have the same size. Lastly, it is worth mentioning that for all the multivariate methods, the default configuration of the classifier algorithms in TMVA is used.

Results

An initial assessment of the discriminant capabilities of the substructure variables studied in this work can be done by looking at the difference in their distributions for signal and background events. The distributions for multiple fat jet variables in the training sample events for both the top signal and the QCD background are shown in figure 4.1.

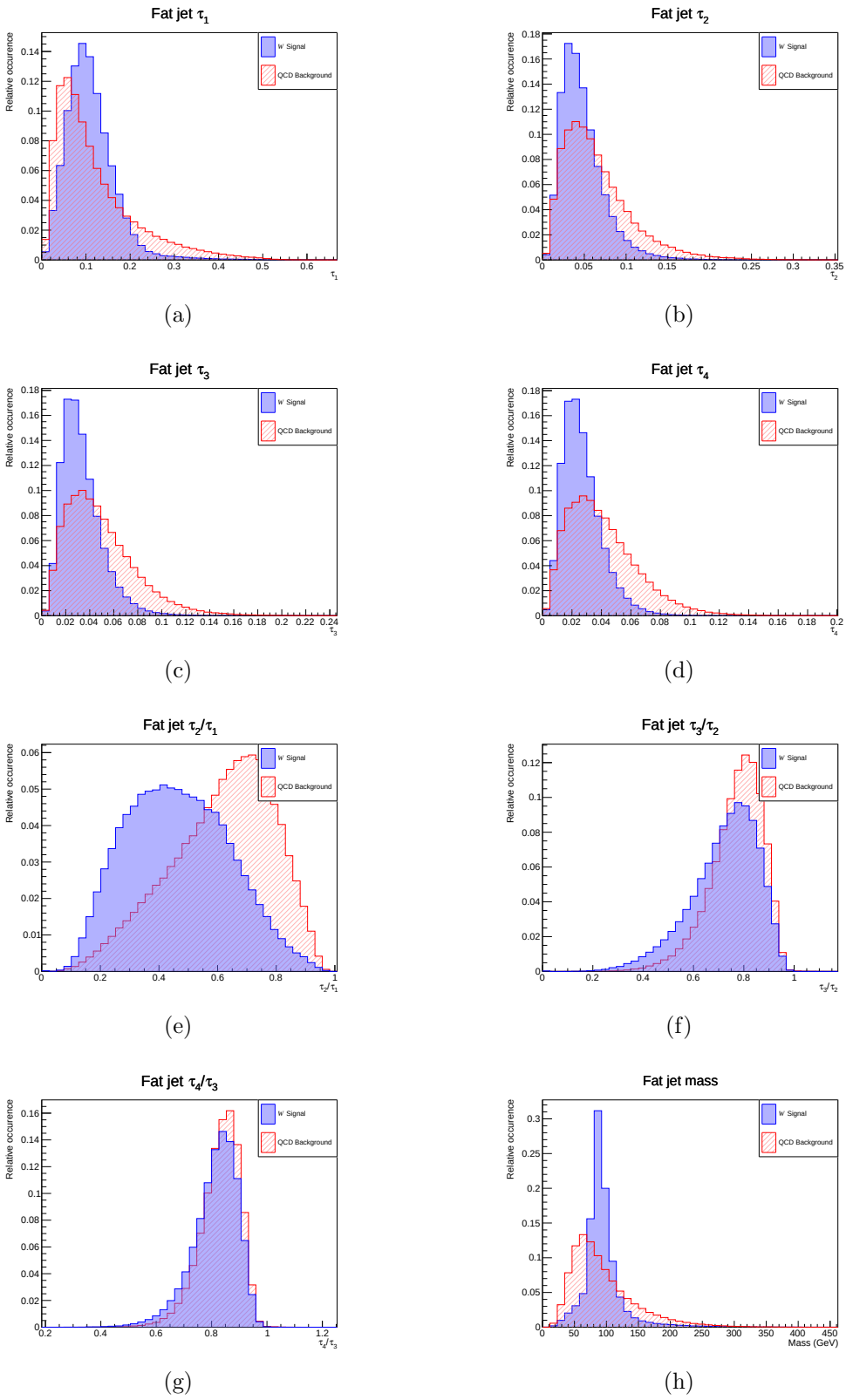
As can be seen from figures 4.1(a) - 4.1(d) the τ_N variables do not possess a significant discriminant power by themselves due to the lack of separation between the signal and background distributions. This makes a cut based selection based on these variables alone effectively impossible. On the other hand, the distributions of the τ_{N+1}/τ_N ratios exhibit a much better separation between background and signal, as depicted in figures 4.1(e) - 4.1(g). Although these ratios are not needed as input variables in the tagging method proposed, their distributions are plotted because they account for the discussion presented in section 2.3 and reinforce the advantages of a multivariate analysis.

The statements made above also hold true when the variable distributions for the W signal are taken into account. The variable distributions for the training sample events of the W signal are plotted against those of the QCD background in figure 4.2. However one remark can be made regarding the difference with the top signal τ_{N+1}/τ_N distributions. As can be seen in figures 4.2(e) - 4.2(g), for the W signal the ratio τ_2/τ_1 exhibits the best separation between the signal and background distributions. That is not the case for the



(a) (b) (c) (d) (e) (f) (g) (h)

Figure 4.1: Distribution of (a) - (d) τ_N, (e) - (g) τ_{N+1}/τ_N ratios and (h) mass for the top and QCD fat jets in the training sample events.



(a) (b) (c) (d) (e) (f) (g) (h)

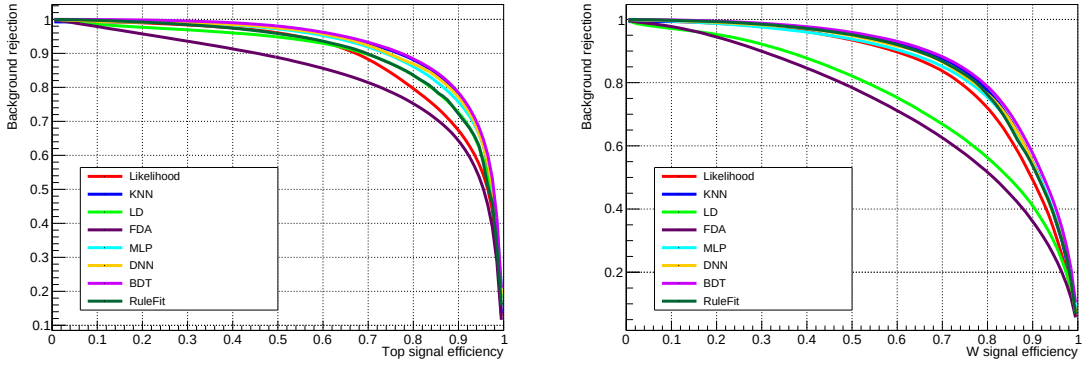
Figure 4.2: Distribution of (a) - (d) τ_N, (e) - (g) τ_{N+1}/τ_N ratios and (h) mass for the W and QCD fat jets in the training sample events.

top signal distributions, in which τ_3/τ_2 presents a better discrimination. This is because fat jets originating from W bosons are made up of two subjets while those originating from top quarks are made up of three. One common feature shared by the top and W signals is that for higher N values, the separation between signal and background for τ_{N+1}/τ_N distributions vanishes. Because of this, the contribution of τ_N variables with high N values to a classification would be negligible. Therefore, the set of τ_N variables used as input in the subsequent multivariate analysis consists of $\{\tau_1, \tau_2, \tau_3, \tau_4\}$.

For both the top and the W signals, there is a clear distinction between the signal and background fat jet mass distributions, which are shown in figures 4.1(h) and 4.2(h). The signal distribution consists of a narrow peak around the top quark mass (the W mass in the case of the W signal) while the background one covers a wider range of mass values. This accounts for the good discriminant capabilities of this variable. For this reason, the jet mass is continuously used in jet tagging algorithms, and it is employed alongside the N-subjettiness variables in the multivariate analysis presented here.

From this point forward, the results of the machine learning implementation will be addressed. The performance of the different multivariate classifiers trained using TMVA will be analysed by looking at their respective Receiver Operating Characteristic (ROC) curve. The ROC curves are created by plotting the background rejection (the fraction of background events that are correctly identified as such by the classifier) versus the signal efficiency (the fraction of true signal events that are correctly identified by the classifier) as the discrimination threshold is varied.

The ROC curves for the different MV classifiers are presented in figure 4.3. As has been already mentioned, the variables used as input for the classifier training are the fat jet τ_N (with $N = 1, \ldots, 4$) and the fat jet mass.



(a) (b)

Figure 4.3: ROC curves of the different multivariate classifiers trained for (a) boosted top tagging (b) boosted W tagging.

The area under the ROC curve (AUC) is a commonly used metric for evaluating the performance of a classifier algorithm. A higher AUC value indicates a better classifier performance, with a value of 1.0 denoting a perfect classification (total background rejection for a maximum signal efficiency). The AUC values obtained for both boosted top tagging and boosted W tagging are presented in table 4.1.

MVA Method	AUC (Top signal)	AUC (W Signal)
BDT	0.925	0.871
KNN	0.920	0.862
DNN	0.917	0.858
MLP	0.915	0.851
RuleFit	0.899	0.858
LD	0.892	0.751
Likelihood	0.887	0.838
FDA	0.840	0.723

Table 4.1: Area under the ROC curve of the different multivariate classifiers used.

As evidenced by the AUC values obtained, the classifiers perform better overall in the tagging of top quarks than in that of W bosons. This was expected, since the substructure of fat jets originating from top quarks is somewhat more distinguishable. It is worth pointing out that the best tagging performance was achieved by the boosted decision tree in both cases. Therefore, it is chosen as the preferred classifier in subsequent analyses. Despite the

fact that its performance is slightly worse, the DNN also continues to be considered as a deep learning alternative alongside the BDT method.

Before moving forward it is important to verify that the chosen classifiers are not being overtrained. In order to do so, the classifier response distributions for signal and background are plotted for both training and test data in figure 4.4. As can be seen from the figure, the training and test distributions are very similar in all cases, indicating that overtraining has been avoided.

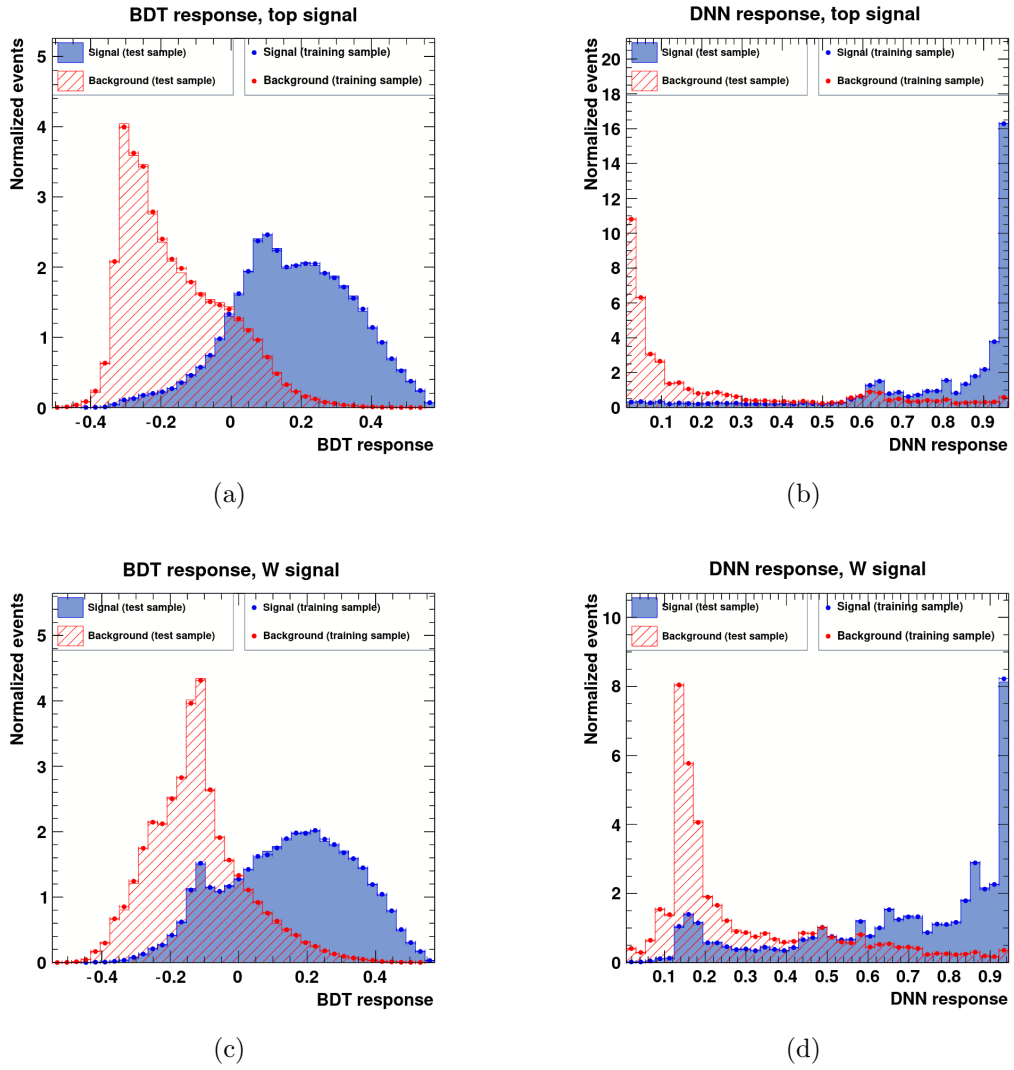
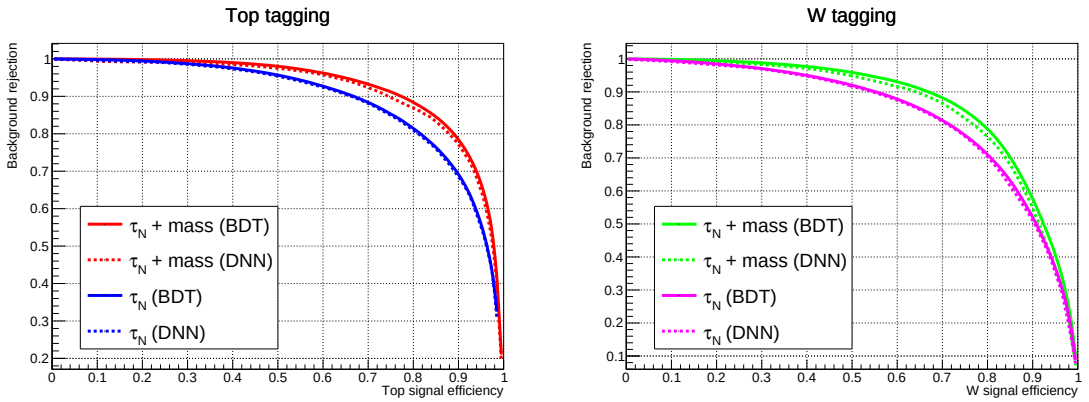


Figure 4.4: Distribution of the BDT and DNN classifier responses for training and test data in (a) - (b) boosted top tagging (c) - (d) boosted W tagging.

Up until now the fat jet mass has been used alongside the N-subjettiness variables as an input variable in the classifier training. This is because the mass works as a great discriminant variable, as has been shown already. However, in order to further examine the tagging performance of the N-subjettiness jet shape, a classification based on the τ_N variables alone is carried out. The ROC curves obtained both with and without using fat jet mass information for the classification are plotted for each case in figure 4.5.

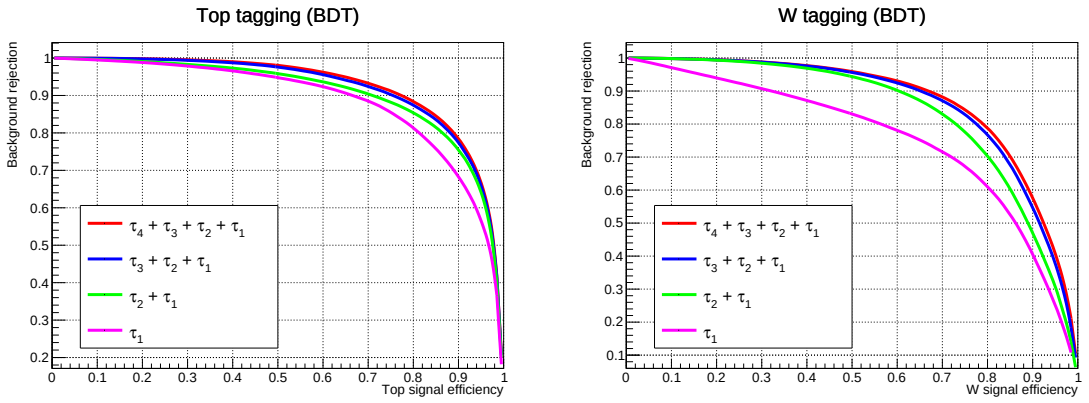


(a) (b)

Figure 4.5: ROC curves for (a) boosted top tagging (b) boosted W tagging with and without including the fat jet mass as an input variable.

As evidenced by the respective ROC curves, the fat jet mass has a noticeable contribution to the tagging performance. This corroborates that using both substructure variables in conjunction is optimal. However, it is also worth noticing the remarkable discriminating job that the τ_N can do on their own, which confirms that they are not dependant on the mass in order to carry out a classification.

So far the full set of τ_N values with $N = 1, \ldots, 4$ have been used. The impact of the number of τ_N variables used as input for the classifier training will now be addressed. In order to do so, the model is trained first using only the fat jet τ_1 value (plus the fat jet mass) then the following τ_N variables are added, one at a time until τ_4 is reached and the original set is achieved. The ROC curves obtained through this process are presented in figure 4.6.



(a) (b)

Figure 4.6: ROC curves for (a) boosted top tagging (b) boosted W tagging using different number of fat jet τ_N variables (and fat jet mass) as input.

This time the tagging performance was plotted using the BDT method only for simplicity. However, the same behaviour is observed when the trained algorithm is the DNN. As the number of τ_N inputs is increased, the gain in performance decreases notably to the point where the model gets saturated with information. This supports the initial decision of selecting the full set of τ_N input variables as those up to $N = 4$.

Conclusions

We can look at m-body subjettiness $(\tau_1^{(1)}, \tau_1^{(0.5)}, \tau_1^{(2)}, \tau_2^{(1)}, \tau_2^{(0.5)}, \tau_2^{(2)})$.

A better study can be done by analysing the ideal parameters in the grooming techniques used before generating the samples.

Following the hand rule for the radius, a better analysis could be done by studying the effect of decreasing the R of the reconstructed fat jets as the p_T of the boosted particles increases.

For simplicity the cut used in the W analysis was the same as for the top analysis, a better study could use a lower p_T cut for the W, because the Lorentz boost is achieved at lower p_T values for this particle.

The performance of the tagging algorithm could be further studied by delving into the architecture of the different TMVA analyses, specially into those that seemed to work the best (BDT and DNN).

Bibliography

[1] B. Andersson, G. Gustafson, G. Ingelman, and T. Sjöstrand. Parton fragmentation and string dynamics. *Physics Reports*, 97(2-3):31–145, July 1983. ISSN 03701573. doi: 10.1016/0370-1573(83)90080-7.

[2] Mark Thomson. *Modern Particle Physics*. Cambridge University Press, 2013.

[3] Gavin P. Salam. Towards jetography. *The European Physical Journal C*, 67(3):637–686, June 2010. ISSN 1434-6052. doi: 10.1140/epjc/s10052-010-1314-6.

[4] George Sterman and Steven Weinberg. Jets from Quantum Chromodynamics. *Physical Review Letters*, 39(23):1436–1439, December 1977. doi: 10.1103/PhysRevLett.39.1436.

[5] Matteo Cacciari, Gavin P. Salam, and Gregory Soyez. The anti-k_t jet clustering algorithm. *Journal of High Energy Physics*, 2008(04):063–063, April 2008. ISSN 1029-8479. doi: 10.1088/1126-6708/2008/04/063.

[6] BOOST2012 participants: A. Altheimer et al. Boosted objects and jet substructure at the LHC. *The European Physical Journal C*, 74(3):2792, March 2014. ISSN 1434-6044, 1434-6052. doi: 10.1140/epjc/s10052-014-2792-8.

[7] Sebastian Schätzel. Boosted Top Quarks and Jet Structure. *The European Physical Journal C*, 75(9):415, September 2015. ISSN 1434-6044, 1434-6052. doi: 10.1140/epjc/s10052-015-3636-x.

[8] CMS Collaboration. W, Z, and Higgs bosons as portals to exotic physics. https://cms.cern/news/w-z-and-higgs-bosons-portals-exotic-physics, May 2022.

[9] ATLAS Collaboration. ATLAS measurements of the properties of jets for boosted particle searches. *Physical Review D*, 86(7):072006, October 2012. ISSN 1550-7998, 1550-2368. doi: 10.1103/PhysRevD.86.072006.

[10] Tilman Plehn, Michael Spannowsky, Michihisa Takeuchi, and Dirk Zerwas. Stop Reconstruction with Tagged Tops. *Journal of High Energy Physics*, 2010(10):78, October 2010. ISSN 1029-8479. doi: 10.1007/JHEP10(2010)078.

[11] Jonathan M. Butterworth, Adam R. Davison, Mathieu Rubin, and Gavin P. Salam. Jet substructure as a new Higgs search channel at the LHC. *Physical Review Letters*, 100(24):242001, June 2008. ISSN 0031-9007, 1079-7114. doi: 10.1103/PhysRevLett.100.242001.

[12] Janna Katharina Behr. Searches with Boosted Objects. In *34th International Symposium on Physics in Collision*, November 2014.

[13] Jesse Thaler and Ken Van Tilburg. Identifying boosted objects with N-subjettiness. *Journal of High Energy Physics*, 2011(3):15, March 2011. ISSN 1029-8479. doi: 10.1007/JHEP03(2011)015.

[14] Iain W. Stewart, Frank J. Tackmann, and Wouter J. Waalewijn. N-Jettiness: An Inclusive Event Shape to Veto Jets. *Physical Review Letters*, 105(9):092002, August 2010. ISSN 0031-9007, 1079-7114. doi: 10.1103/PhysRevLett.105.092002.

[15] Jesse Thaler and Ken Van Tilburg. Maximizing boosted top identification by minimizing N-subjettiness. *Journal of High Energy Physics*, 2012(2):93, February 2012. ISSN 1029-8479. doi: 10.1007/JHEP02(2012)093.

[16] Lyndon Evans and Philip Bryant. LHC Machine. *Journal of Instrumentation*, 3(08):S08001, August 2008. ISSN 1748-0221. doi: 10.1088/1748-0221/3/08/S08001.

[17] Stephen Myers. The Large Hadron Collider 2008–2013. *International Journal of Modern Physics A*, October 2013. doi: 10.1142/S0217751X13300354.

[18] Oliver S. Bruning, P. Collier, P. Lebrun, S. Myers, R. Ostojic, J. Poole, and P. Proudlock. LHC Design Report Vol.1: The LHC Main Ring. August 2004. doi: 10.5170/CERN-2004-003-V-1.

[19] Tino Michael. Determination of muon reconstruction efficiencies in the ATLAS detector using a tag & probe approach in Z to μ μ events. Master's thesis, Dresden, Technische Universität Dresden, 2011.

[20] Ana Lopes and Melissa Loyse Perrey. FAQ-LHC The guide. Technical report, 2022.

[21] G. Aad et al. The ATLAS Experiment at the CERN Large Hadron Collider. *JINST*, 3:S08003, 2008. doi: 10.1088/1748-0221/3/08/S08003.

[22] A. Airapetian et al. ATLAS: Detector and physics performance technical design report. Volume 1. May 1999.

[23] J. Alwall, R. Frederix, S. Frixione, V. Hirschi, F. Maltoni, O. Mattelaer, H.-S. Shao, T. Stelzer, P. Torrielli, and M. Zaro. The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations. *Journal of High Energy Physics*, 2014(7):79, July 2014. ISSN 1029-8479. doi: 10.1007/JHEP07(2014)079.

[24] Richard D. Ball, Valerio Bertone, Stefano Carrazza, Luigi Del Debbio, Stefano Forte, Patrick Groth-Merrild, Alberto Guffanti, Nathan P. Hartland, Zahari Kassabov, José I. Latorre, Emanuele R. Nocera, Juan Rojo, Luca Rottoli, Emma Slade, and Maria Ubiali. Parton distributions from high-precision collider data. *The European Physical Journal C*, 77(10):663, October 2017. ISSN 1434-6052. doi: 10.1140/epjc/s10052-017-5199-5.

[25] Fabio Maltoni and Tim Stelzer. MadEvent: Automatic Event Generation with MadGraph. *Journal of High Energy Physics*, 2003(02):027–027, February 2003. ISSN 1029-8479. doi: 10.1088/1126-6708/2003/02/027.

[26] Torbjörn Sjöstrand, Stefan Ask, Jesper R. Christiansen, Richard Corke, Nishita Desai, Philip Ilten, Stephen Mrenna, Stefan Prestel, Christine O. Rasmussen, and Peter Z. Skands. An Introduction to PYTHIA 8.2. *Computer Physics Communications*, 191:159–177, June 2015. ISSN 00104655. doi: 10.1016/j.cpc.2015.01.024.

[27] Matt Dobbs and Jørgen Beck Hansen. The HepMC C++ Monte Carlo event record for High Energy Physics. *Computer Physics Communications*, 134(1):41–46, February 2001. ISSN 0010-4655. doi: 10.1016/S0010-4655(00)00189-2.

[28] J. de Favereau, C. Delaere, P. Demin, A. Giammanco, V. Lemaître, A. Mertens, and M. Selvaggi. DELPHES 3, A modular framework for fast simulation of a generic collider experiment. *Journal of High Energy Physics*, 2014(2):57, February 2014. ISSN 1029-8479. doi: 10.1007/JHEP02(2014)057.

[29] Matteo Cacciari, Gavin P. Salam, and Gregory Soyez. FastJet user manual. *The European Physical Journal C*, 72(3):1896, March 2012. ISSN 1434-6044, 1434-6052. doi: 10.1140/epjc/s10052-012-1896-2.

[30] K. Khelifa-Kerfa, Y. Delenda, and N. Ziani. Jet shapes in H/V boson + jet with k_t clustering at hadron colliders, July 2022.

[31] David Krohn, Jesse Thaler, and Lian-Tao Wang. Jet Trimming. *Journal of High Energy Physics*, 2010(2):84, February 2010. ISSN 1029-8479. doi: 10.1007/JHEP02(2010)084.

[32] Stephen D. Ellis, Christopher K. Vermilion, and Jonathan R. Walsh. Techniques for improved heavy particle searches with jet substructure. *Physical Review D*, 80(5):051501, September 2009. ISSN 1550-7998, 1550-2368. doi: 10.1103/PhysRevD.80.051501.

[33] Andrew J. Larkoski, Simone Marzani, Gregory Soyez, and Jesse Thaler. Soft Drop. *Journal of High Energy Physics*, 2014(5):146, May 2014. ISSN 1029-8479. doi: 10.1007/JHEP05(2014)146.

[34] Jinmian Li, Riley Patrick, Pankaj Sharma, and Anthony G. Williams. Boosting the charged Higgs search prospects using jet substructure at the LHC. *Journal of High Energy Physics*, 2016 (11):164, November 2016. ISSN 1029-8479. doi: 10.1007/JHEP11(2016)164.

[35] Antimo Cagnotta, Francesco Carnevali, and Agostino De Iorio. Machine Learning Applications for Jet Tagging in the CMS Experiment. *Applied Sciences*, 12(20):10574, January 2022. ISSN 2076-3417. doi: 10.3390/app122010574.

[36] A. Hoecker, P. Speckmayer, J. Stelzer, J. Therhaag, E. von Toerne, H. Voss, M. Backes, T. Carli, O. Cohen, A. Christov, D. Dannheim, K. Danielowski, S. Henrot-Versille, M. Jachowski, K. Kraszewski, A. Krasznahorkay Jr., M. Kruk, Y. Mahalalel, R. Ospanov, X. Prudent, A. Robert, D. Schouten, F. Tegenfeldt, A. Voigt, K. Voss, M. Wolter, and A. Zemla. TMVA - Toolkit for Multivariate Data Analysis, July 2009.

[37] Rene Brun and Fons Rademakers. ROOT — An object oriented data analysis framework. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 389(1):81–86, April 1997. ISSN 0168-9002. doi: 10.1016/S0168-9002(97)00048-X.