THE UNIVERSITY
of EDINBURGH

Dark Matter *and* How To Find It

A search for low-mass leptophobic Dark Matter mediators
and the development of mass-decorrelated jet taggers
with the ATLAS experiment

Andreas Søgaard

Doctor of Philosophy
The University of Edinburgh
April 2019

To my father,

for always inspiring and encouraging me.

Abstract

A SEARCH FOR low-mass leptophobic Dark Matter (DM) mediator particles in 36 fb^{-1} of pp collision data at $\sqrt{s} = 13$ TeV collected by the ATLAS experiment is presented. The search is performed in final states where the mediator decay into a quark pair is reconstructed as a single, large-radius jet produced in association with a photon or a jet. No deviations from the Standard Model expectation are observed, and limits are placed on the production cross-section of leptophobic mediator particles and their coupling to quarks for mediator masses between 100 and 220 GeV. At the time of publication, this result constituted the lowest limits on leptophobic DM mediator masses for high-mass DM particles reported by ATLAS.

Adversarial neural networks (ANN) are presented as a way to train jet taggers which decorrelates them from the invariant mass of the jet. An extensive study of five different approaches to constructing mass-decorrelated jet taggers is presented. The ANN tagger is found to provide the largest QCD multijet rejection at similar levels of mass-decorrelation.

Lay Summary

The Standard Model of particle physics (SM) is the theory of all known fundamental particles and how they interact with each other. Its predictions have held up to rigorous scrutiny for more than half of a century, but there are still open questions that it leaves unanswered. One of these questions is the seeming existence of Dark Matter (DM), called so because it does not reflect light, in contrast to ordinary matter. In fact, DM has only been observed through its gravitational pull on celestial objects. However, given the success of the SM in describing almost all physical phenomena through particles and their interactions, it is reasonable to assume that DM might also have a particle nature. However, if that is the case, why has DM not already been observed to interact with ordinary matter? One explanation may be through the existence of a so-called mediator particle, Z', which is responsible for the interaction between ordinary matter and DM particles. If this particle is sufficiently massive, it will not be produced in abundance in the current, cold Universe. Nevertheless, if that is the case, it may be possible to produce it in a controlled environment *e.g.* in the high-energy proton–proton (pp) collisions at the CERN Large Hadron Collider (LHC).

If DM mediator particles can be created by fusing two particles in a high-energy collision, it is expected that it could decay back to ordinary SM particles, which can then be detected by the ATLAS experiment. The resulting spray of particles, also called a 'jet', can then be reconstructed as the experimental signature of the decay, and its mass measured. For collision events producing a DM mediator, the jets will have roughly the same mass as that of the mediator. These DM processes are expected to be extremely rare, however. It is much more likely for a pp collision to produce spurious jets not originating from the decay of a massive particle. These spurious jets will resemble the decay of a DM mediator, but will exhibit a continuum of mass values. The task is then to search for a tiny, localised DM mediator "bump" on a continuous spectrum of jet mass values.

This thesis presents such an analysis, in which the mediator particle is produced in association with either a photon or another spurious jet. Collision data collected by the ATLAS experiment in 2015 and 2016 was analysed for evidence of DM mediator particles decaying to a single jet. The analysis quantified potential deviations of the data from the SM expectation in order to identify which combinations of mediator masses and coupling strengths to SM particles were consistent with the observed data. The data was found to be in good agreement with the expectation from the SM alone, with no significant "bumps" that could point to evidence of DM mediator particles. The analysis therefore suggests that there is no evidence for DM mediators with masses in the range considered.

In order to identify the jets that could have originated from a DM mediator decay, the analysis relied on a so-called jet substructure observable. These observables describe properties of the shape of each jet, which are characteristic for DM mediator decays. This helps distinguish them against the spurious jets produced in the dominant SM processes. Furthermore, machine learning (ML) techniques may help improve these substructure observables *e.g.* for future searches for DM in ATLAS. Generally speaking, ML allows one to combine a large number of weak discriminators to a single, more powerful observable. In this case, multiple jet substructure observables can be combined to improve jet identification. The problem with standard ML algorithms is that, when they are used to select a signal-enriched sample with a simple threshold selection, they distort the shape of the jet mass spectrum. As a result, if the continuum background is not smooth any more, it becomes harder to search for localised "bumps."

To solve this problem, this thesis presents a study of five different methods to create jet identification observables that are constructed in a way that leaves the jet mass spectrum as smooth as possible. These methods range from a simple linear transform to the simultaneous, so-called adversarial training of two neural networks. It is found that, to varying degrees, these methods are effective at removing the sculpting effect mentioned above. These methods are currently being used in the next generation of DM searches at the LHC, and may therefore allow physicists to harness the power of ML to advance the discovery frontier of high energy physics.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

Parts of this work have been published in [1, 2].

(*Andreas Søgaard, April 2019*)

Acknowledgements

First and foremost, I am grateful to my supervisor, Christos Leonidopoulos, for an outstanding induction into the world of experimental particle physics; for providing me with sage guidance along the way; for having my best interests at heart; for humour, structure, and encouragement; but most importantly, for having faith in me and for allowing me to pursue my own ideas and interests, even when they weren't what you had imagined. I hope it was worth it. I hope I was. This freedom has been essential in making my PhD the enjoyable and rewarding experience it has been. I could not have asked for anything more. Thank you.

I also had the great fortune of having Flavia de Almeida Dias guide me through these formative years. Rarely have I met anyone as fierce, funny, and loyal, and it has been a pleasure working with you and growing as a physicist under your wings.

No man is an island, and I have benefited immensely from interacting with, and learning from, the other students and researchers at the University of Edinburgh. Similarly, each of my research projects in ATLAS has been carried out alongside analysis teams from all over the world. Taking even a small part in the massive scientific endeavour that is CERN has been an eye-opening and humbling experience.

I am also thankful for the support of the Scottish Universities Physics Alliance (SUPA), whose Prize Studentship provided me with the opportunity to live and study in Edinburgh in the first place.

Finally, and most importantly, all of this was only possible due to the unwavering support, patience, and kindness of my fiancée, Louise. Sharing these past years abroad with you has been the adventure of a lifetime.
Du er min klippe i et stormfuldt hav.

Contents

List of common abbreviations

SM	Standard Model of particle physics
BSM	Beyond the Standard Model
DM	Dark Matter
QCD	Quantum chromodynamics
MC	Monte Carlo
LHC	Large Hadron Collider
pp	Proton–proton
ISR	Initial-state radiation
DDT	Designed decorrelated taggers
TF	Transfer factor
GP	Gaussian process
SR	Signal region
VR	Validation region
MVA	Multivariate analysis
ML	Machine learning
NN	Neural network
BDT	Boosted decision tree
ANN	Adversarial neural network
***k*-NN**	k-nearest neighbours
CSS	Convolved substructure
JSD	Jensen-Shannon divergence

Introduction

Despite its remarkable success, the Standard Model of particle physics (SM) still faces a number of shortcomings. For instance, experimental observations in cosmology and astrophysics suggest the existence of Dark Matter (DM), making up approximately 26% of the energy in the Universe. This is a clear indication of physics beyond the Standard Model (BSM).

In simplified models for DM, the DM may interact with SM quarks through a neutral mediator boson Z'. Therefore, DM may be produced in the proton–proton (pp) collisions at the Large Hadron Collider (LHC) and detected in subsequent analyses. In particular, searches for simplified models for DM may focus on processes involving only the mediator boson. Mediator bosons produced in pp collisions couple to SM quarks, and so may decay back to a pair of quarks, reconstructed as hadronic jets. For low-mass DM mediators decaying at rest, the back-to-back jets in the resulting so-called dijet topology are not energetic enough to pass the standard jet trigger thresholds. Therefore, to probe even lower DM mediator masses, the analysis in this thesis focuses on the hadronic decay of the Z' boson produced in association with either a photon or a jet, which can be used for triggering. Triggering on such energetic initial-state radiation (ISR) objects, instead of the Z' decay products themselves, means that it is possible to circumvent the jet trigger threshold problem. The analysis focuses on mediators with low enough masses that the Z' is reconstructed as a single jet, such that a resonance search can be performed in the jet mass spectrum. In addition, the substructure within the jet may be used to distinguish the two-body Z' decay from the dominant background jets, characterised by the emission of a single energetic parton.

Part I presents the foundations upon which the work in this thesis builds. Chapter 1 presents the particles and forces of the SM, with a focus on quantum chromodynamics (QCD) and the concept of jets for reconstructing and identifying hadronic decays of *e.g.* Z' bosons. In Chapter 2, the existence of DM is motivated with experimental evidence, and a case for hadron collider searches for DM in the context of simplified

models is presented. Chapter 3 then gives an overview of the LHC and the ATLAS experiment as a general-purpose detector well suited for performing such DM searches. Lastly, Chapter 4 gives a brief introduction to machine learning (ML) techniques, which can be used to identify faint signals, such as new physics processes, in vast datasets.

Part II presents the search for DM mediator bosons Z' in the so-called boosted dijet + ISR final state. Chapter 5 gives an overview of the complementary DM searches performed in ATLAS and motivates the chosen final state. The Monte Carlo (MC) simulated and recorded datasets used in the analysis are described in Chapter 6. In Chapter 7, the reconstruction and calibration of the relevant physics objects — hadronic jets and photons — is described. Chapter 8 outlines the event selection used to define the signal-enhanced search region as well as the studies of jet substructure and the decorrelation from the jet mass. The process for estimating the leading SM background process contributions using Gaussian process (GP) regression is described in Chapter 9. Finally, the statistical analysis and the search results are presented in Chapter 10, and concluding remarks are given in Chapter 11. This is the first search in the boosted dijet + ISR final state in ATLAS. Furthermore, the search in the ISR γ channel, for which I was responsible, is the first of its kind in any experiment. This analysis was published in Phys. Lett. B [1] and reported the lowest-ever exclusion limits on leptophobic DM mediator masses for high-mass DM particles ($m_{\mathrm{DM}} \gg m_{Z'}$) in ATLAS.

A key challenge for this analysis is the use of jet substructure to identify the hadronic two-body decay of the Z' boson. The use of ML to improve identification is promising, but standard algorithms tend to result in tagging variables that are correlated with the jet mass. Performing a selection on such a variable sculpts the jet mass spectrum for the leading SM background, complicating resonance searches.

Part III therefore introduces and studies five different techniques for constructing mass-decorrelated jet taggers, including two ML-based methods, which may benefit future DM searches by providing better jet identification and more robust background estimation. Chapter 12 presents the existing work on ML-based jet tagging in ATLAS and introduces the problem of mass-correlation. In Chapter 13, the MC simulated samples, event and jet selection, and event weighting schemes used in this study are described. Metrics for evaluation the performance of jet taggers in terms of classification power and jet mass-decorrelation are proposed in Chapter 14. Chapter 15 describes the five mass-decorrelation techniques considered in this study, with a focus on adversarial neural networks (ANNs). Finally, the performance results are shown in Chapter 16, and Chapter 17 concludes on the study. This study, which I initiated and led as main analyser, was approved and published by the ATLAS Collaboration as a public note [2].

PART I

Foundations

CHAPTER 1

Theoretical background

In this chapter, the theoretical background for much of the work in this thesis is presented. First, the particles and forces of the Standard Model of particle physics (SM) are briefly introduced. Then, some of the central topics in hadron collider physics are described. Finally, the concept of hadronic jets is introduced along with the tools for reconstructing and identifying them. Throughout this chapter, and the rest of this thesis, natural units in which $\hbar = c = 1$ will be used for convenience.

1.1 Particles and forces

The SM [3–8] is a mathematical framework that unifies the electromagnetic, weak, and strong force in a single, coherent theory. It describes, with remarkable precision, a very large number of observed subatomic phenomena in terms of a small number of fundamental particles and their interactions.

The matter particles of the SM consist of 12 fermions and their anti-particles. The fermions, comprising leptons and quarks, are categorised in three generations of increasing masses,[1] as shown in Figure 1.1. The charged leptons are the electron (e), muon (μ), and tau (τ), all with electric charge $-1e$. Within each generation, the charged lepton has an associated electrically neutral lepton, called a neutrino (ν_e, ν_μ, and ν_τ). The six quarks in the SM are similarly grouped in generations of pairs of up-type (u, c, t) and down-type (d, s, b) quarks, with electric charges of $+2/3e$ and $-1/3e$, respectively.

The SM also comprises the four force-mediating vector bosons shown in Figure 1.1.

[1] Assuming normal ordering of the neutrino mass hierarchy.

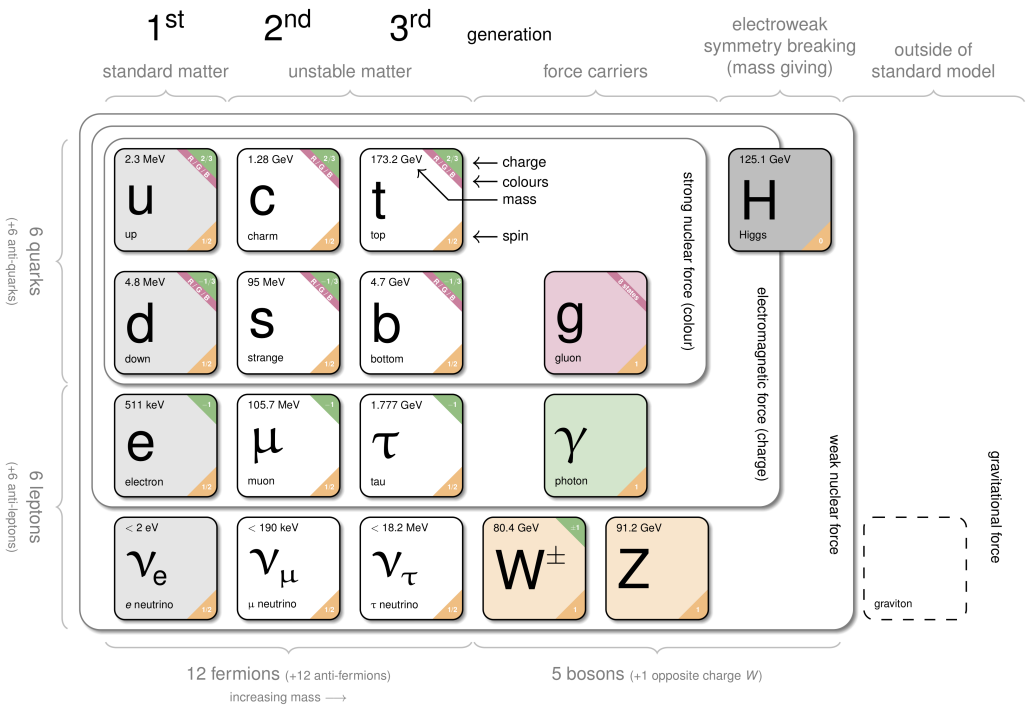


Figure 1.1 The particle content of the Standard Model of particle physics (SM). Fermions are divided into three generations with increasing particle masses, where the first generation makes up ordinary matter. Each generation contains an up-type and a down-type quark, a charged lepton, and a neutrino. The gluon, photon, and W/Z bosons are the force-carrying particles of the SM. The Higgs boson is responsible for generating the masses of the elementary particles and couples to all massive particles (except the neutrinos in the current SM). Figure reproduced from Refs. [9, 10].

The massless, electrically neutral photon (γ) is the mediator of the electromagnetic force. The weak force is mediated by the $W^{\pm}$ and Z^0 bosons, which for brevity will be referred to as simply W and Z in this thesis. These are the only massive force-mediating particles, with masses $m_W = 80.379 \pm 0.012$ GeV and $m_Z = 91.1876 \pm 0.0021$ GeV [11].

The quantum field theory of the strong interaction is called QCD, describing the interaction of the particles carrying the strong force charge, also called colour. Quarks are the only fermions carrying colour charge, which is three-valued with dimensions colloquially referred to as red (r), green (g), or blue (b). The strong force is mediated by gluons, eight massless bosons coupling to the colour charge. The gluons themselves also carry colour charge meaning that they can self-interact. The strength of the strong force is given by α_S, the strong coupling constant [12], which changes ("runs") with the momentum exchange Q^2 due to the gluon self-interaction: In the short-distance or high-energy limit, $\alpha_S(Q^2)$ goes to zero and the coloured quarks and gluons become

asymptotically free. Conversely, in the long-distance limit, where Q^2 approaches the QCD scale $\Lambda_{\text{QCD}} \approx 200$ MeV, $\alpha_S(Q^2)$ diverges. This means that colour-charged particles are effective confined to distance scales of $O(1\text{ fm})$. This leads to the formation of bound, colour-neutral states called hadrons. These may be quark–anti-quark pairs, called mesons — *e.g.* $(u\bar{u})$ with colour charges $r\bar{r}$, $g\bar{g}$, or $b\bar{b}$ in the case of the neutral pion π^0 — or a combination of three quarks, called baryons — *e.g.* (uud) with colour charges rgb in the case of the proton p. Only such colour-neutral states can exist as free particles.

In the original formulation, all particles in the SM were massless, in contrast to all observations. The 1964 theory by Higgs, Brout, and Englert [13, 14] provides a mechanism for generating the masses of the elementary particles in the SM through the coupling of massive SM particles to the Higgs field. A central prediction of this theory is the existence of a scalar Higgs boson. This predication was confirmed with the 2012 observation and subsequent measurements by the ATLAS and CMS experiments [15, 16] of a particle consistent with the SM Higgs boson. The Higgs boson is the only fundamental scalar in the SM, and its coupling to other particles is proportional to the mass of the particle (except the neutrinos in the current SM).

1.2 Hadron collider physics

Particle colliders like the Large Hadron Collider (LHC), see Chapter 3, can be used to test the validity of the SM, and to search for signs of physics beyond the Standard Model (BSM). These accelerate *e.g.* protons to high energies and collide them head-on to produce the highest-energy controlled particle interactions. Protons are abundantly available, their comparatively large masses, $m_p = 938$ MeV *cf.* $m_e = 511$ keV, reduces their synchrotron radiation power output which scales as m_p^{-4} [17], and their collisions naturally scan a large range of effective centre-of-mass energies (see below). This makes them excellent candidates for use in discovery experiments at particle colliders.

Coordinate system

Typically, proton–proton (pp) collision events at the LHC are studied in the laboratory frame coordinate system with the origin positioned at the nominal interaction point (IP) of the proton beams: The x-axis points from the IP towards the centre of the LHC, the y-axis points upwards from the IP, and the z-axis points along the direction of motion of

the incoming protons in the counter-clockwise direction when viewed from above. The x- and y-axes span the so-called transverse plane, perpendicular to the proton beams. The azimuthal angle, ϕ, is measured in the transverse plane starting from the positive x-axis. In pp collisions, the initial state of any interaction process will have a net momentum in the transverse plane which is roughly zero, but indeterminate momentum along the beam axis. Therefore, the transverse momentum of a particle, $p_{\mathrm{T}} = \sqrt{p_x^2 + p_y^2}$, is a convenient, boost-invariant variable for characterising collision events.

The polar angle, θ, is measured relative to the positive z-axis, but more commonly the pseudo-rapidity η is used

$$\eta = -\log\left(\tan\frac{\theta}{2}\right). \tag{1.1}$$

The pseudo-rapidity variable is zero for $z = 0$ and anti-symmetric in z, extending to $\eta \to \pm\infty$ as $\theta \to 0$ and π, respectively. The purely geometric pseudo-rapidity is an approximation to the rapidity of a particle

$$y = \frac{1}{2}\log\left(\frac{E + p_z}{E - p_z}\right), \tag{1.2}$$

where E is the energy of the particle and p_z is the z-component of its three-momentum. The rapidity is additive under boosts along the z-axis, meaning that differences in rapidity are boost-invariant. The pseudo-rapidity approximation is valid for relativistic particles, *i.e.* ones for which $m \ll E$, which will be the case for most of the final state objects considered in this thesis. Therefore, for describing directions of final-state particles, the $\eta - \phi$ plane will be used. The angular separation of particles i and j in this plane is given by $\Delta R_{ij} = \sqrt{(\eta_i - \eta_j)^2 + (\phi_i - \phi_j)^2}$, which is similarly boost-invariant.

Factorisation

In high-energy pp collisions, the characteristic interaction is not between the protons themselves, but rather between their constituent quarks and gluons — collectedly called partons. The main constituents of the proton, the (uud) quarks, are referred to as the valence quarks. However, at high energies, interactions involving the so-called virtual sea quarks and gluons become increasingly important. The probability for a particular type of parton to carry a fraction x of the total proton momentum, is quantified through parton distribution functions (PDFs). These functions are measured experimentally, and their values for each parton type change as a function of the momentum exchange Q^2.

For a pp collision with a centre-of-mass energy of $\sqrt{s}$, the incoming partons in a hard scatter process, each carrying fractions $x_{1,2}$ of their respective proton momenta, will have a partonic centre-of-mass energy of $\sqrt{\hat{s}} = \sqrt{x_1 x_2 s}$. This means that only a fraction of the total pp centre-of-mass energy will enter a given hard scatter process, but also that pp colliders naturally scan partonic centre-of-mass energies, in contrast to e^+e^- colliders.

While the proton constituents are confined at low energies, due to the running of the strong coupling constant α_S, the partons interacting in high-energy collisions behave as free particles. These energetic parton-level interactions are referred to as the hard scatter processes, and in the high-energy limit they can be described using perturbation theory, *i.e.* expanding the parton-level interaction cross-section in orders of $\alpha_S \ll 1$. In this perturbative regime, the phenomenology of QCD interactions can be described reasonably well by computing the matrix element for parton-level Feynman diagrams at the lowest orders in α_S [12]. The outgoing partons in a given hard scatter process are colour-charged and will exhibit colour-connections to other parts of the event. This will result in additional emissions, in the form of a shower of partons, which will gradually reduce the energy of each individual parton until the process becomes non-perturbative around the QCD scale, Λ_{QCD} [12, 18]. After this point, the outgoing quarks will form bound, colour-neutral hadronic states in a process known as hadronisation.

This factorisation of processes [19] allows for treating the perturbative hard scatter process separately from the non-perturbative physics in the interaction ($\alpha_S \approx 1$), *e.g.* contained in the PDFs, the parton showering, and the subsequent hadronisation.

Parton showers and soft QCD

While it is not possible to calculate high-multiplicity hadronic showers exactly, it is possible to simulate the process using parton showers. Given an n-parton final state with cross-section σ_n, the cross-section for the same process with an additional emission with a small opening angle is given by [18]

$$\frac{\mathrm{d}\sigma_{n+1}}{\sigma_n} \approx \frac{\alpha_S}{2\pi} \sum_{\text{partons } i,j,k} \frac{\mathrm{d}\theta^2}{\theta^2} \mathrm{d}z \, P_{i\to jk}(z). \tag{1.3}$$

Here α_S is the strong coupling constant; the labels i, j, k enumerate the configurations of quark and gluon types such that the emission $i \to jk$ is allowed (*e.g.* $q \to qg$ and $g \to gg$); θ is the opening angle for the emitted parton j with respect to the emitter i,

with associated angular phase space $\mathrm{d}\theta^2$; and z is the momentum fraction carried by the emitted parton. The Dokshitzer-Gribov-Lipatov-Altarelli-Parisi (DGLAP) splitting functions $P_{i\to jk}(z)$ [20–22] diverge for $z \to 0$ or 1 [12, 18], *i.e.* in the soft emission limit. Equation (1.3) provides a prescription for sequentially constructing a parton shower from an initial matrix element configuration. Furthermore, Equation (1.3) is approximately process-independent, meaning that the parton shower model it facilitates is universal. It is seen from Equation (1.3) that the parton branching probability diverges for soft ($z \to 0$) and collinear ($\theta \to 0$) emissions. This means that a single high-p_{T} hard scatter emission will generate a collimated spray of partons. However, due to colour-confinement, this parton shower is never observed directly. Instead, the partons hadronise to free, colour-neutral bound states of mesons and baryons.

In addition to the hadronisation, a pp collision event will also contain other sources of low-p_{T} physics. In a pp interaction, a pair of partons may interact in a high-Q^2 perturbative process. However, the remainder of the two interacting protons will be colour-connected to the interacting partons, and may produce *e.g.* initial-state radiation (ISR) and final-state radiation (FSR) of additional partons, and multiple-parton interactions (MPIs) in the same pp collision. This so-called underlying event (UE) constitutes soft radiation noise that is correlated with the hard scatter process. Finally, when colliding bunches of $\mathcal{O}(10^{11})$ protons, see Chapter 3, any hard scatter collision is likely to be accompanied by a number of additional pp collisions in the same bunch-crossing. Due to the large cross-section for QCD processes, these additional pp collisions are most likely to be low-p_{T} QCD-dominated interactions. These so-called pile-up interactions produce a diffuse flux of particles, which is not correlated with the hard scatter process, with a roughly uniform energy density per pile-up interaction of $\langle\rho\rangle/\mu \approx 0.8$ GeV [23], where $\langle\rho\rangle$ is the average energy density in $\eta - \phi$ and μ is the number of pile-up interactions.

1.3 Hadronic jets

The collimated spray of stable, colour-neutral hadrons arising from the showering of a high-p_{T} parton is called a hadronic jet. Due to energy-momentum conservation and the collinear structure of parton emissions in Equation (1.3), these jets are kinematically representative of the initiating parton. Jets are therefore useful proxies in pp collision experiments, where the aim is to reconstruct the hard scatter process. However, for a given final state, there is no way to uniquely identify which particle originated from what part of the parton-level process. It is also not clear that such a distinction would be

physically meaningful, considering the quantum nature of the interaction. Therefore, it is necessary to establish a robust, operational definition for constructing jets, which relates the observable final state to the underlying, unobservable parton-level process.

Such definitions are referred to as jet algorithms. The first such algorithm was proposed by Sterman and Weinberg in 1977 [24] in the context of hadronic final states at e^+e^- colliders. Initially, jet algorithms were cone-based, where a fixed cone in $\eta - \phi$ is seeded by a single high-p_{T} particle and its orientation iteratively updated based on the momentum sum of all final-state particles within the cone. However, such cone algorithms are generally slow to evaluate and are often not resilient to soft and collinear emissions, which is discussed further below. This means that they are not operationally robust, and have generally been abandoned for so-called sequential recombinations algorithms. Jet physics is a vast field, of which only a small portion is covered in this thesis. An excellent introduction to this field is given in Ref. [25], with more recent theoretical and experimental reviews given in Refs. [26, 27].

Jet clustering algorithms

For a jet algorithm to be robust, it should be infra-red and collinear (IRC) safe, where "infra-red" refers to the low-energy limit. That is, the result of a jet algorithm should be unchanged if a number of arbitrarily low-energy particles are added to the final state (infra-red) as well as if any constituent is replaced with two constituents with the same direction, sharing the total energy (collinear). From the parton splitting in Equation (1.3), it is seen that the probability for increasingly soft ($z \to 0$) and collinear ($\theta \to 0$) parton emissions diverges. IRC safety therefore implies that the number of jets resulting from the application of some jet algorithm, as well as their properties, is not susceptible to such emissions.

The standard jet algorithms used in present-day general-purpose pp collision experiments are based on sequential recombination [28, 29]. This is a bottom-up approach which iteratively searches for the pair of final-state particles that are deemed most compatible with originating from the same process according to some distance measure, which are then combing. In the context of sequential recombination algorithms, particles and their four-vector combinations are often referred to as "pseudo-jets," *i.e.* possibly composite final state objects which are not yet fully clustered. These distance measures

for clustering pseudo-jets generally take the form [25]

$$d_{ij}^{(p)} = \min\left(p_{\mathrm{T},i}^{2p}, p_{\mathrm{T},j}^{2p}\right)\frac{\Delta R_{ij}}{R} \quad \text{and} \tag{1.4a}$$

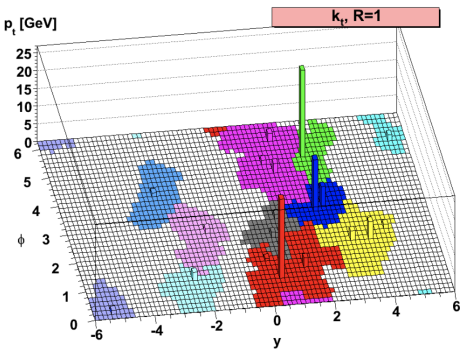
$$d_{iB}^{(p)} = p_{\mathrm{T},i}^{2p}\,, \tag{1.4b}$$

where $p_{\mathrm{T},i}$ is the transverse momentum of the i^{th} pseudo-jet in the final state, ΔR_{ij} is the distance in $\eta - \phi$ between the i^{th} and j^{th} pseudo-jet, R is a radius parameter, and p is a free parameter that characterises different algorithms, as described below. The variable $d_{iB}^{(p)}$ is historically referred to as the particle-beam distance, and is used in the termination criterion for the algorithm.

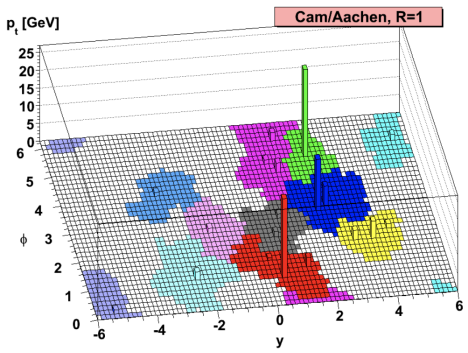
The sequential recombination algorithms start from the set of all final-state particles and searches for the minimal values of $d_{ij}^{(p)}$ and $d_{iB}^{(p)}$. These distance minima are found by iterating over all pairs of particles, or pseudo-jets, i and j in the event in the former case, and over all pseudo-jets i in the event in the second case. If $\min_{i,j}(d_{ij}^{(p)})$ is smaller, *i.e.* if the minimal distance in the event according to the measures in Equations (1.4) is between the pseudo-jets i and j, then these are removed from the final state and replaced with a new pseudo-jet, with four-momentum equal to the sum of i and j. If $\min_i(d_{iB}^{(p)})$ is smaller, *i.e.* if the minimal distance is between the beam and the pseudo-jet i, then this is deemed isolated enough to be removed from the final state and labelled a jet. This procedure continues until all final state particles have been included in a jet. The final state particles that have been clustered into a particular jet are referred to as the jet constituents.

The parameter p controls the dependence of the distance measure $d_{ij}^{(p)}$ on the transverse momentum of the final state particles, and thereby the characteristics of the jet algorithm in question. For any value of p, it is seen that collinear splittings ($\Delta R_{ij} \to 0$) lead to $d_{ij}^{(p)} \to 0$ due to the second factor in Equation (1.4a). Similarly, additional arbitrarily soft final state particles may be recombined as part of a jet at any stage in the clustering, but will have no impact on the output of the remainder of the clustering history, since their four-momentum contribution will be vanishingly small. Therefore, Equations (1.4) parametrises a class of fast, IRC-safe jet algorithms. Examples of jets resulting from each of the three common sequential recombination algorithms, discussed below, are shown in Figure 1.2.

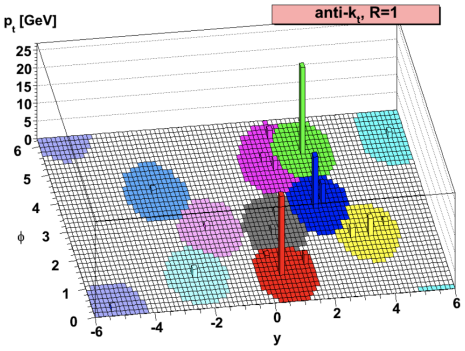
The k_t algorithm [28, 29] with $p = +1$ was the first sequential recombination algorithm. With a positive exponent, the distance measure in Equation (1.4a) favours recombining final-state particles that are soft and close in $\eta - \phi$. This behaviour mirrors the sequential



(a) k_t algorithm



(b) Cambridge/Aachen algorithm



(c) Anti-k_t algorithm

Figure 1.2 Examples of the effect of different sequential recombination algorithms applied to the same event. The coloured regions correspond to the area in $\eta - \phi$ of the different reconstructed jets. Figures from [30].

splitting model of parton showers, discussed above, by preferring soft and collinear emissions. Specifically, it attempts to reverse a parton shower evolution with an angular ordering, starting from wide-angle emissions and becoming gradually more collinear, similar to what is done in some Monte Carlo (MC) event generators [18]. However, this algorithm results in jets with irregular boundaries in $\eta - \phi$ that are susceptible to soft, wide-angle emissions. The k_t-jets are still IRC-safe, but the irregularity of the boundaries makes it harder to calibrate the jets and to mitigate the effect of pile-up radiation within the jet area.

A simpler approach is offered by the Cambridge/Aachen (CA) algorithm [31, 32], using $p = 0$. This algorithm is purely geometric, ignoring all kinematic information in the final state. This class of jet are commonly used in current-day physics analyses, but similarly to the k_t algorithm, the jet areas and boundaries are also somewhat susceptible to soft radiation.

Finally, the anti-k_t algorithm [30] uses $p = -1$. This has the implication of favouring combinations of hard particles, through the first factor in Equation (1.4a), with particles close-by in $\eta - \phi$. Therefore, anti-k_t jets tend to grow radially around high-p_T "seed-particles," resulting in generally circular jets. This regularity makes it easier to calibrate the jets and to mitigate the effect of pile-up and soft radiation, which will be described in Section 7.2. These features have made the anti-k_t algorithm the standard choice in the ATLAS Collaboration, and it is therefore used throughout both Parts II and III in this thesis.

Jet substructure

Hadronic two-body decays of massive particles at rest — such as W, Z, or Higgs bosons — typically result in two back-to-back jets with small radii. However, if these particles are produced with sufficient transverse momentum in the laboratory frame, the hadronic shower initiated by each of the two primary decay quarks may start to overlap. In this so-called boosted regime, the angular separation of the two quarks is roughly

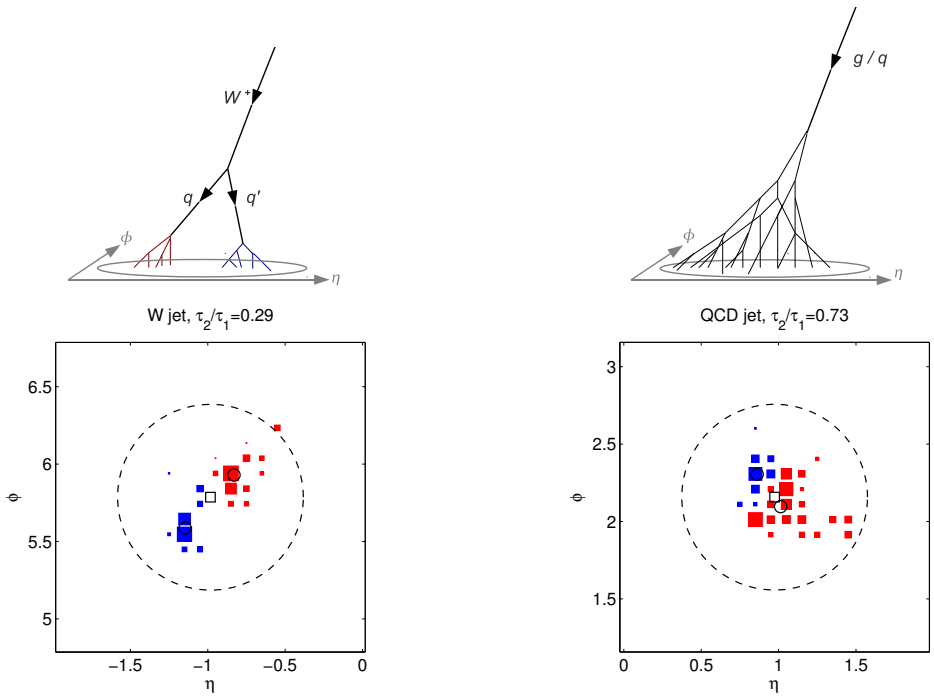
$$\Delta R_{12} \approx \frac{2m}{p_\mathrm{T}}, \tag{1.5}$$

where m is the mass of the decaying particle and p_T is its transverse momentum, see Appendix A. For $p_\mathrm{T} \gtrsim 2m/R$, the hadronic decay may become sufficiently collimated so as to be reconstructed as a single jet with radius parameter R. In these cases, radius parameters of $R \approx 1$ are typically used, and the jets are referred to as large-radius (large-R) jets. An important task is then to distinguish these hadronic two-body decays from the dominant non-resonant jet production at hadron colliders, which may also be reconstructed as large-R jets. This may be achieved by exploiting differences in the radiation patterns inside the jet, also called the substructure of the jet. This will be crucial to Parts II and III of this thesis.

A simple measure of jet substructure is the invariant mass m of a jet, defined as [33]

$$m^2 = \left(\sum_{i \in \text{jet}} p_i \right)^2 \quad \Longleftrightarrow \quad m = \sqrt{\left(\sum_{i \in \text{jet}} E_i \right)^2 - \left| \sum_{i \in \text{jet}} \mathbf{p}_i \right|^2}, \tag{1.6}$$

where E_i, $\mathbf{p}_i$, and p_i are the energy, three-, and four-momentum of the i^{th} constituent of the jet, respectively. The jet mass can be used to infer information about the mass of the decaying particle, and is therefore a useful observable for distinguishing resonant particle decays from non-resonant jet production.



(a) $W \to q\bar{q}$ **(b)** Single parton emission

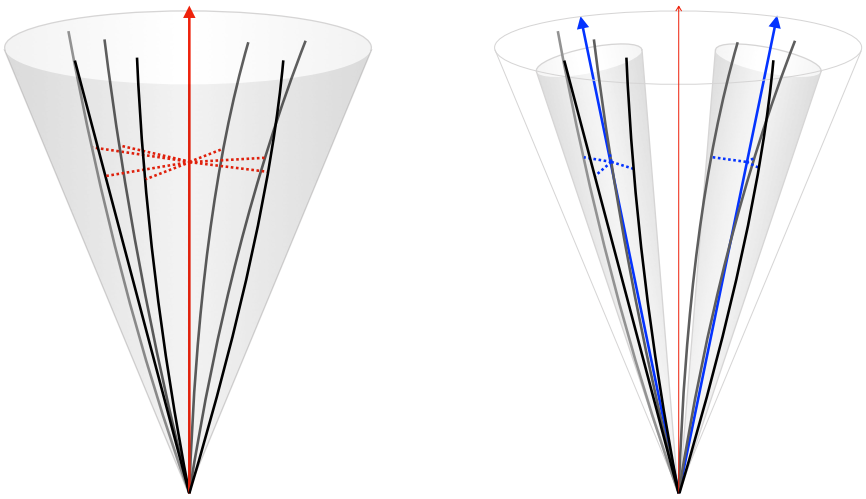
Figure 1.3 Schematic representation of **(a)** the hadronic decay of a high-p_{T} W boson and **(b)** a single energetic parton emission, as well as typical event displays for each process, with similar invariant jet masses. Cell sizes are proportional to the logarithm of the energy deposited, and colours indicate the two τ_2 subjets. The QCD jet in **(b)** has $\tau_{21} = 0.73$ while the W jet in **(a)** has $\tau_{21} = 0.29$, indicating a better match with a 2-subjet hypothesis for the latter. Figures adapted from Ref. [34].

Nevertheless, it is typically useful to have some means of distinguishing jets, originating from different hard scatter processes, which happen to have similar invariant masses. A widely used class of such substructure observables are the so-called N-subjettiness ratios [34]. These variables are based on a reclustering of the constituents of the candidate large-R jet into exactly N smaller jets, called subjets, using the k_t jet clustering algorithm [28, 29]. Examples of this are shown in Figure 1.3 in the case of a 2-subjet reclustering.

The N-subjettiness variable τ_N is then defined as

$$\tau_N = \frac{1}{R} \sum_{i=1}^{n_J} z_i \min\{\Delta R_{1i}, \Delta R_{2i}, \ldots, \Delta R_{Ni}\}, \tag{1.7}$$

where R is the jet radius parameter, N is the number of subjets, n_J is the number of jet constituents, $z_i = p_{\mathrm{T},i}/p_{\mathrm{T},J}$ is the transverse momentum fraction carried by the i^{th} jet constituent, and ΔR_{ji} is the distance between the i^{th} jet constituent and the j^{th} subjet in



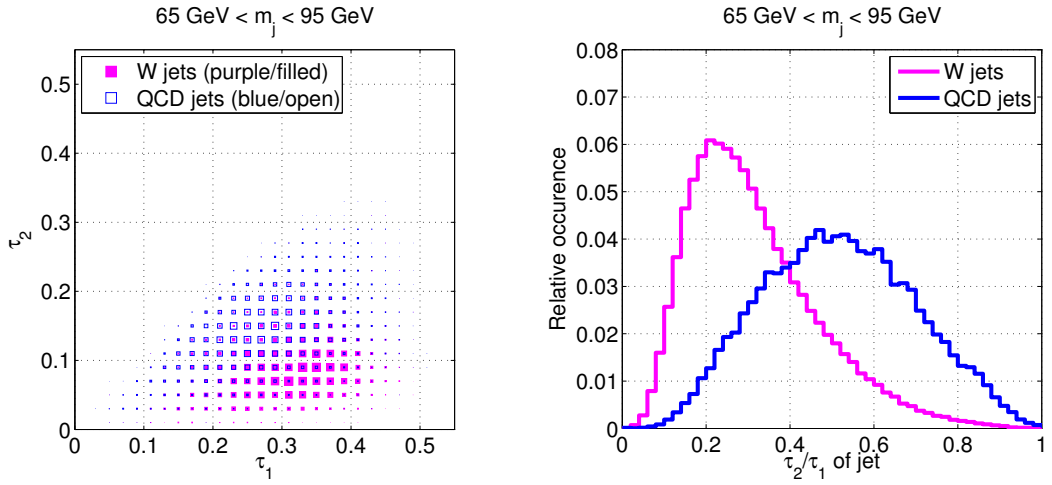
(a) 1-subjet hypothesis **(b)** 2-subjet hypothesis

Figure 1.4 Schematic representation of a **(a)** 1- and **(b)** 2-subjet hypothesis as used in the calculation of τ_{21}. See text for details.

$\eta - \phi$. In this way, τ_N quantifies the degree to which a large-R jet can be considered as composed of $N \leq n_J$ subjets: A jet with $\tau_N \approx 0$ will have all of its constituents collinear with the N subjets. Conversely, a jet with τ_N closer to 1 will contain significant energy emitted in directions which are incompatible with the N-subjet hypothesis, and will therefore be composed from at least $N + 1$ subjets. This is illustrated in the cartoon in Figure 1.4. The shaded cones indicate the reconstructed k_t subjet(s), the bold arrows indicate the corresponding subjet axes (*i.e.* first index of ΔR_{ji} in Equation (1.7)), and the dashed lines indicate the minimal distance of each constituent to a subjet axis (*i.e.* $\min\{\Delta R_{1i}, \Delta R_{2i}, \ldots, \Delta R_{Ni}\}$ for constituent i). The τ_N variable is then computed as the p_{T}-weighted sum of the dashed distances in Figure 1.4, illustrating how, in this case, the 2-subjet hypothesis offers a large relative improvement over the 1-subjet hypothesis.

However, the individual N- and $N + 1$-subjettiness values are correlated: τ_{N+1} will be strictly smaller than τ_N due to the additional degree of freedom, and a particular jet with a large value of τ_N (*i.e.* significant radiation not well described by the N-subjet hypothesis) will also be more likely to have a relatively large value of τ_{N+1}, simply because there is more non-collinear radiation to describe. This is illustrated in Figure 1.5.

Figure 1.5a shows the correlation between τ_1 and τ_2, including the fact that $\tau_2 < \tau_1$. For the purposes of identifying *e.g.* hadronic two-body decays, this correlation suggests a problem for using τ_2 as a substructure variable, because a very 1-subjet–like background jet ($\tau_1 \ll 1$) will also be a good match for the 2-subjet hypothesis due to the added degree of freedom ($\tau_2 < \tau_1 \ll 1$), which in this case will be redundant. Instead,



(a) $W \to q\bar{q}$ **(b)** Single parton emission

Figure 1.5 N-subjettiness distributions for hadronically decaying W bosons and non-resonant QCD jets with similar invariant masses, showing **(a)** the correlation between τ_2 and τ_1 for both classes of jets and **(b)** the distribution of the τ_{21} ratio, breaking the correlation and isolating the improvement of the 2-subjet hypothesis over the 1-subjet one. Figures from Ref. [34].

to distinguish *e.g.* two-body decays from single emissions, the N-subjettiness ratio $\tau_{21} = \tau_2/\tau_1$ is typically used, instead of the individual N-subjettiness variables. Taking the ratio of τ_2 and τ_1 isolates the improvement of the 2-subjet hypothesis relative to the 1-subjet hypothesis, and cancels out the correlation between the two subjettiness variables described above. This has the effect of cancelling out the correlation in Figure 1.5a. Since τ_2 is strictly smaller than τ_1, the N-subjettiness ratio satisfies $0 \leq \tau_{21} < 1$. Therefore, jets with $\tau_{21} \approx 0$ are more consistent with a two-subjet hypothesis, while ones with $\tau_{21} \approx 1$ are more consistent with a one-subjet hypothesis. This is illustrated in the τ_{21} distributions in Figure 1.5b, which shows the separation of 2-subjet–like W jets from 1-subjet–like single-emission QCD jets. The separation provided by the N-subjettiness ratio τ_{21} is better for than for either of the base variables individually. Additional jet substructure observables used in ATLAS, and in Part III of this analysis in particular, are detailed in Appendix A.

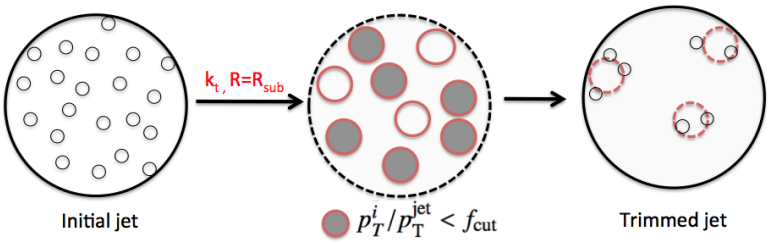
Jet grooming

Sequential recombination is an IRC-safe and robust algorithm for reconstructing jets with a characteristic radius of R. The motivation provided above suggests that these jets are kinematically representative of the initiating particle and that jet substructure observables can be used to distinguish between initiating processes. However, in hadron

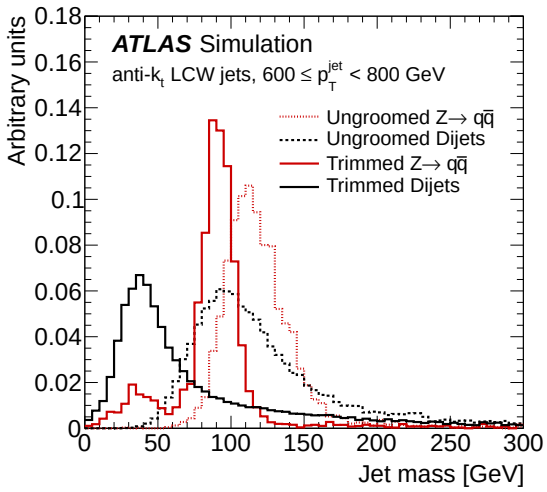
collisions, the UE may result in radiation leaking into the jet, thereby degrading the jet mass and momentum resolution. In particular, in the case of resonance decays this tends to shift the jet mass towards values larger than the mass of the initiating particle. Additionally, the roughly uniform, low-p_{T} radiation from additional pile-up interactions is accrued during jet reconstruction. Such soft radiation can also obfuscate the jet substructure, rendering the jet identification less effective. To mitigate the effect of pile-up and to remove soft radiation from the UE, so-called jet grooming algorithms may be employed [35–38]. These jet grooming techniques try, in various ways, to discard the jet constituents which are likely to have originated from soft, wide-angle radiation rather than from the hard scatter process.

A technique commonly used in ATLAS and throughout this thesis is called jet trimming [36]. Here, the constituents of a jet clustered with some radius parameter R are re-clustered using a smaller subjet radius parameter R_{sub}. In ATLAS, for large-R jets with $R = 1.0$, the k_t algorithm with $R_{\mathrm{sub}} = 0.2$ is commonly used [39]. The characteristics of the k_t algorithm discussed above means that energetic re-clustered subjets are taken as proxies for hard scatter partons. Low-energy subjets, by contrast, are considered to be spurious accumulations of soft radiation. Therefore, all subjets i which carry less than f_{cut} of the energy of the entire jet, *i.e.* $p_{\mathrm{T},i} < f_{\mathrm{cut}}\, p_{\mathrm{T},J}$, are discarded. A typical value for f_{cut} used in ATLAS is 5%. The constituents of the high-p_{T} subjets that are not discarded by this requirement are considered the constituents of the trimmed jet. The trimming procedure is illustrated in Figure 1.6a.

Grooming procedures, such as trimming, result in jets with improved mass and momentum resolution and which are also more robust against the presence of pile-up. For the hadronic decay of *e.g.* a Z boson, this leads to a jet mass peak which is more closely centred around the known Z mass. For non-resonant jet production, which constitutes the dominant background process in many searches for BSM physics, trimming mitigates the wide-angle emissions that otherwise generate significant masses for a jet initiated by the emission of a practically massless high-p_{T} parton. This reduces the number of background jets under a resonant mass peak, as illustrated in Figure 1.6b, which shows simulated jet mass distributions with and without the application of jet trimming.



(a) Trimming procedure



(b) Jet mass distributions

Figure 1.6 Figures showing **(a)** the schematic jet trimming procedure and **(b)** the effect of jet trimming on the jet mass distributions for a hadronic resonance decay ($Z \to q\bar{q}$) and non-resonant jet production (dijets). Figures from [40].

CHAPTER 2

Dark Matter

The SM, presented in Chapter 1, is a remarkable scientific success, providing a coherent framework for understanding and predicting most physical phenomena in the Universe with high precision. Nevertheless, it still has a number of shortcomings, including the fact that approx. 26% of the energy in the Universe appears to be made up of a type of non-luminous matter not described by the SM. This so-called Dark Matter (DM) is one of the clearest indications of BSM physics, and is therefore the focus of the physics analysis presented in this thesis.

2.1 Experimental evidence

The existence of DM has only been inferred through gravitational effects. Early evidence was reported by Fritz Zwicky in 1933 [41, 42], who observed a larger spread in apparent velocities of galaxies in the Coma cluster than expected based on the amount of luminous matter in the cluster. This indicated the presence of additional, non-luminous matter in the cluster.

This observation was corroborated by Vera Rubin *et al*. [43, 44] starting in 1970 with their measurement of the rotational speed of galaxies. From Newtonian mechanics, the acceleration required to sustain a circular motion with velocity v at a fixed radius r is $a = v^2/r$. Additionally, the acceleration arising from Newtonian gravitational force at radius r is $a = GM(r)/r^2$ where $M(r)$ is the total mass enclosed within the sphere of radius r and G is the Newtonian gravitational constant. That is, the dependence of the

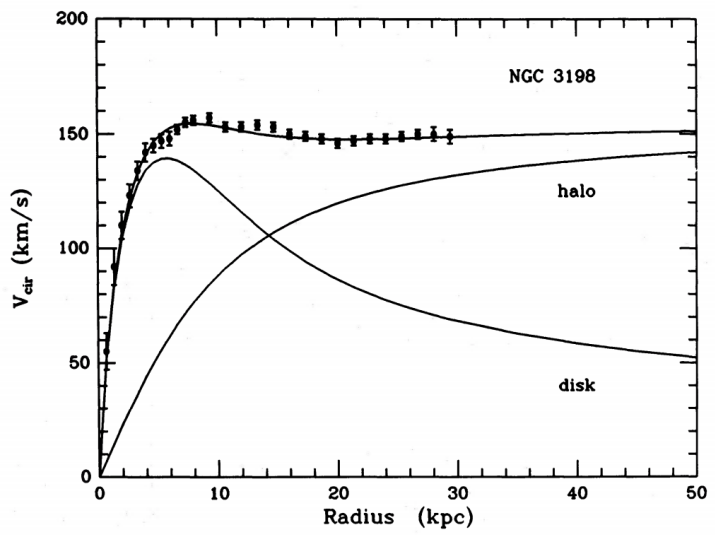


Figure 2.1 Rotation cruve for spiral galaxy NGC-3198. Expected constribution from the luminous disc, overlaid with best-fit Dark Matter (DM) halo. Figure from Ref. [45].

rotational velocity of *e.g.* spiral galaxies at radius r is given by

$$v(r) = \sqrt{\frac{GM(r)}{r}}. \tag{2.1}$$

The visible matter in galaxies is typically dominated by some combination of a localised spheroid bulge and an exponential disc with some characteristic radius [45]. This relative compactness means that at sufficiently large radii, the enclosed visible mass $M(r)$ stays roughly constant. Therefore, the expected behaviour of the rotation curve at large radii r is as $v \propto 1/\sqrt{r}$. However, the observed galactic rotation curves exhibit a roughly constant plateau at these large radii, see Figure 2.1, in clear contrast to the expected behaviour.

This disagreement cannot be remedied by varying the different components of luminous matter. Therefore, a new component of matter is required for which $M(r)$ increases with r even at large radii, such that flatness is achieved in Equation (2.1). A so-called halo of gravitational DM with density profile [46]

$$\rho_{\mathrm{DM}}(r) = \rho_0 \left[1 + \left(\frac{r}{r_c}\right)^2\right]^{-1} \tag{2.2}$$

satisfies this criterion [47]. Here, ρ_0 is an overall density scale and r_c is the characteristic radius of the DM halo.

Finally, the Cosmic Microwave Background (CMB) exhibits minute temperature

anisotropies [48], originating from density differences at the time of recombination [49], *i.e.* the time when free electrons and protons started forming electrically neutral hydrogen atoms, making the Universe transparent to photons. The structure of the power spectrum of the CMB temperature is sensitive to the energy density of DM, and can therefore be used to measure its abundance in the Universe at the time of recombination, which is expected to be constant to this day. The relative relic DM energy density is measured to be $\Omega_{\mathrm{DM}} = 0.261 \pm 0.004$ [50], where the density parameter for each class of matter is defined as $\Omega_i = \rho_i/\rho_{\mathrm{crit}}$, where ρ_i is the energy density and ρ_{crit} is the critical total energy density yielding a flat Universe. The measured combined relative energy density is $\Omega = \Omega_b + \Omega_{\mathrm{DM}} + \Omega_\Lambda = 1.001 \pm 0.002$, consistent with unity, where Ω_b is the relative energy density of baryonic (luminous) matter (approx. 5%) and Ω_Λ is the relative energy density of the so-called Dark Energy (approx. 69%) [50]. This suggests that the Universe is spatially flat, and that DM constitutes approx. 26% of all energy in it.

2.2 Weakly Interacting Massive Particles

The SM explains natural phenomena through a system of fundamental particles and their interactions. Therefore, it is natural to propose particle candidates to explain the experimental evidence for the existence of DM. In order to conform to observation, such a particle (or particles) must be:

electrically neutral, as it would otherwise scatter light and be detectable using telescope experiments,

stable, or at least long-lived on the scale of the lifetime of the Universe, to still be abundant at present times [47],

massive, to interact gravitationally and be bound in halos; specifically massive enough to be non-relativistic (cold) at the time of cosmological structure formation [51],

non-baryonic, since the total energy density of matter in the Universe, as determined using the CMB, is inconsistent with the observed baryonic energy density [50, 52], implying that the DM particle candidate must be non-baryonic, and specifically not colour-charged; and

weakly interacting, *e.g.* due to the observed DM relic density (see below) and the sphericity of DM halos and observations of colliding galaxy clusters, *e.g.* the

'Bullet Cluster' where measurements using X-ray imaging and gravitational lensing suggest that DM halos are virtually collisionless [53, 54].

The simplest particle candidates are the SM neutrinos, which are known to exist and at least two of which are massive [11], thereby satisfying all of the requirements listed above. In fact, neutrinos will constitute a non-zero but insufficient part of the energy density attributed to DM. The energy density parameter for neutrinos is given by [55]

$$\Omega_\nu \approx \frac{\sum_\nu m_\nu}{93.14 \text{ eV} h^2}, \tag{2.3}$$

where $\sum_\nu m_\nu \lesssim 0.2$ eV is the sum over the masses of the SM neutrino species and $h = 0.677$ is the standardised Hubble parameter [11, 50]. This limit on the sum of neutrino masses is driven by cosmological observations of CMB anisotropies [11]. The SM neutrinos therefore contribute at most 2% of the energy associated with DM, because of their low masses. Therefore, the SM does not provide a suitable candidate for particle DM.

This motivates the search for new weakly interacting massive particles (WIMPs) [56]. These hypothetical particles must satisfy the above criteria and couple weakly to SM particles, allowing for their detection in an experimental setting.

In the early universe, at temperatures much greater than the mass m_{DM} of the WIMP DM particles χ_{DM}, these are assumed to have been pair-produced in collisions of SM particles [47]. Initially, this process is in equilibrium with its inverse — DM particles annihilating in pairs to produce SM particles — with the common rate $\Gamma_A = n\langle\sigma_A v\rangle$, where n is the DM number density and v is the DM particle velocity such that $\langle\sigma_A v\rangle$ is the thermally averaged annihilation cross-section. This equilibrium is maintained until the WIMPs can no longer efficiently self-annihilate. This occurs when the rate of expansion exceeds the WIMP annihilation rate, *i.e.* $H \gtrsim \Gamma_A$ where H is the Hubble parameter. After this point, known as freeze-out, the DM number density stays constant. This is what leads to the current-day DM relic abundance Ω_{DM}, which relates to the DM cross-section as [47]

$$\Omega_{\text{DM}} \approx \frac{3 \times 10^{-27} \text{ cm}^3 \text{ s}^{-1}}{\langle\sigma_A v\rangle h^2}, \tag{2.4}$$

where h is the standardised Hubble parameter. Due to this dependence, limits on the observed DM abundance can be used to infer limits on the WIMP annihilation cross-section. Assuming the WIMPs are pair-produced in the s-channel exchange of an electroweak boson, the thermally averaged annihilation cross-section is approximately

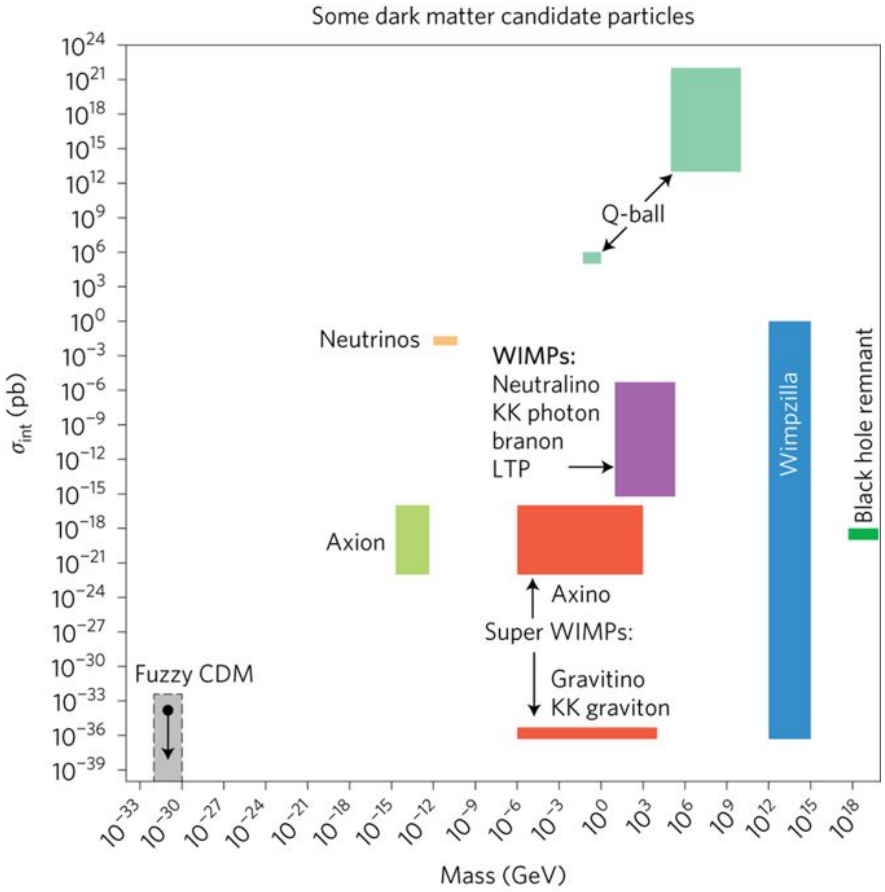
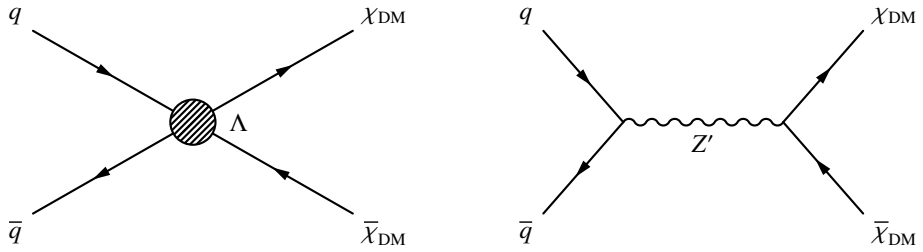


Figure 2.2 Overview of some particle Dark Matter (DM) candidates, spanning more than 50 orders of magnitude in mass. Weakly interacting massive particle (WIMP) DM is hypothesised to have masses at the TeV-scale, and is therefore particularly well suited for searches at the Large Hadron Collider (LHC). Figure from Ref. [59].

$\langle \sigma_A v \rangle \sim G_F^2 m_{\text{DM}}^2$ [57], where G_F is the Fermi constant. Inserting this into Equation (2.4) yields the order-of-magnitude relation $\Omega_{\text{DM}} \sim (m_{\text{DM}}/\text{GeV})^{-2}$. The fact that a WIMP with a mass at the GeV-scale and SM-like couplings around the electroweak scale yields a relic abundance which is consistent with experimental observation is colloquially called the "WIMP miracle" [58]. Since electroweak energies are accessible at the LHC, the above motivates the search for WIMP DM at collider experiments.

This thesis focuses on WIMP DM since its natural energy range overlaps with the energy reach of the LHC and since it may be produced in pp collisions. However, several other models for DM exist, covering a vast range in potential particle masses, see Figure 2.2. These models all have different theoretical appeals and may, in addition to being DM candidates, also solve other problems facing particles physics (*e.g.* axions were proposed as a solution to the strong CP problem and sterile neutrinos might restore left-right symmetry to the SM and provide a mass generation mechanism neutrinos).



(a) Effective field theory **(b)** Simplified model

Figure 2.3 Example diagrams for pair-production of Dark Matter (DM) particles χ_{DM} from a Standard Model (SM) quark–anti-quark pair in **(a)** effective field theory (EFT) with interaction scale Λ and **(b)** a simplified model with mediator particle Z'.

Finally, there also exist non-particle models for DM, *e.g.* modified gravity where the gravitational force acting on galaxies deviates from the Newtonian expectation in the large-distance, low-acceleration limit. These theories, however, are not suitable for tests at particle colliders. For a review of these and other models for DM, as well as cosmological evidence and experimental aspects, see *e.g.* Ref. [47].

2.3 Simplified models

In order to explain the relic abundance and the stability of WIMP candidates, DM particles should in general be pair-produced in pp collisions [60]. The WIMP properties mentioned above imply that the simplest way in which DM particles may interact with ordinary matter through leading order SM processes is through Z or H boson exchange. This is proposed in Higgs-portal models [61, 62]. However, the Z partial width to invisible particles is already constrained at the 10^{-3} level [11]. Similarly, the branching fraction of H to weakly interacting BSM particles is already constrained by searches for its invisible decays, with observed limits of $\mathcal{B}(H \to \text{inv.}) < 0.24$ [63, 64]. This suggests that a new process is required to mediate the interaction between DM and SM particles.

The simplest approach is to treat this process in the framework of effective field theory (EFT), where the DM particle pair is produced in a contact interaction with *e.g.* a pair of SM quarks, see Figure 2.3a [60, 65]. This way, the details of the mediating process need not be specified explicitly. Given a particular coupling operator, the only free parameter in the EFT, apart from the mass of the DM particle, is the overall production rate as controlled by the contact interaction scale Λ. However, such an EFT

is only valid in cases where the mass of the mediating particle is much greater than the momentum transfer in the s-channel process. Reference [66] found that in a simplified model, where the WIMP DM pair production is mediated by a BSM Z' boson, the EFT approximation is valid only for $m_{Z'} > 2.5$ TeV at $\sqrt{s} = 8$ TeV. Therefore, a more complete model of the DM-SM interaction is required. The analysis in this thesis uses benchmark signal processes generated in a simplified model for DM, comprising a DM particle χ_{DM} as well as a mediating boson Z'. An example process is shown in Figure 2.3b. This simplified model has five free parameters: the DM mass m_{DM}, the mediator mass $m_{Z'}$, and the coupling of the mediator to the DM particles (g_{DM}), SM quarks (g_q), and SM leptons (g_l) [67]. The EFT interaction scale for the process shown in Figure 2.3a is related to the parameters of the simplified model as $\Lambda = m_{Z'} / \sqrt{g_q g_{\mathrm{DM}}}$ [66]. This type of simplified model also allows for comparison of DM search results from collider experiments with those from direct detection experiments [66].

Direct detection experiments such as LUX [68] and XENON1T [69] search for WIMPs elastically scattering off heavy nuclei [47]. Assuming the Milky Way has a DM halo composed of WIMPs, their large flux through the Earth could allow for such direct interaction with ordinary matter even considering the low expected WIMP-quark interaction cross-section. These experiments search for nuclear recoils with O(keV) energies from single scatter events in a low-background environment. In the case of the above experiments, this is done using large-volume underground time-projection chambers with ultra-pure liquid xenon as detector medium. This way, direct detection experiments are sensitive to WIMP DM with masses from a few GeV and up to O(10 TeV). At low DM masses, direct detection experiments are limited by the lowest energy recoils that can be reconstructed; at high masses they are rate limited. These are complemented by collider searches, which can perform searches with no lower bound on the mass of the probed DM particles. In direct detection experiments, WIMPs with axial-vector couplings result in a dependence on the angular momentum of the target nucleus [47]. The associated WIMP-nucleus cross-section is therefore referred to as spin-dependent. By contrast, for scalar or vector WIMP couplings the cross-section is spin-independent. The spin-independent cross-section dominates for the high-mass target nuclei such as xenon used in most direct detection experiments. Therefore, it is common for collider searches to focus on the DM particles with axial-vector couplings, to offer complementarity to direct detection experiments.

For production of DM in hadron colliders, a non-zero value for g_q is necessary. Conversely, g_l may be taken to be zero, which will be assumed in the analysis presented in Part II. The extension from an EFT to a simplified model also results in a richer

phenomenology, in which it is possible *e.g.* to search for the Z' mediator itself in processes not including the DM particles. This approach will be adopted in the analysis in this thesis.

The choice of spin for the DM particle only has a minor impact on collider searches in general, and the DM particle does not enter into the signal processes targeted by this analysis. Therfore, χ_{DM} is assumed to be a Dirac fermion for concreteness [65]. The mediator is assumed to be spin-1 and to have an axial-vector coupling to both the DM and SM quarks, with flavour-universal couplings to SM quarks. For scalar and pseudo-scalar mediators, Minimal Flavour Violation (MFV) [70] is typically assumed, which implies mass-dependent, Yukawa-type couplings of the mediator to fermions. Therefore, such models are better targeted by in final-states with heavy-flavour objects, such as Ref. [71]. Similarly, spin-1 mediators might be assumed to have vector couplings to fermions; however, this has minimal impact on collider signatures *cf.* axial-vector couplings [72]. This is confirmed by results from the CMS Collaboration, which reports identical limits on vector and axial-vector couplings in the final state targeted by this analysis [73]. However, axial-vector couplings allow LHC searches to probe different parts of the simplified DM model parameter space than direct detection searches [67]. This is because the spin-independent DM-nucleon cross-section scales with the square of the number of nucleons, whereas the spin-dependent cross-section is suppressed due to the partial cancellation of amplitudies from opposite-spin nucleons [47]. Therefore, direct detection experiments are generally less sensitive to axial-vector mediators, meaning that collider experiments can offer complementarity. The relevant additional Lagrangian interaction terms are therefore [65, 67]

$$\mathcal{L} \supset \sum_{q} g_q Z'_\mu \bar{q} \gamma^\mu \gamma_5 q + g_{\mathrm{DM}} Z'_\mu \bar{\chi} \gamma^\mu \gamma_5 \chi. \tag{2.5}$$

For a given set of couplings (g_{DM}, g_q) this simplified model then provides a parameter space $(m_{\mathrm{DM}}, m_{Z'})$ which may be probed by collider searches using different experimental signatures. For each parameter configuration it is possible to compute the corresponding WIMP DM relic abundance [74], which allows for a comparison of collider search results and cosmological observations. Finally, certain parameter configurations for axial-vector mediator couplings lead to the violation of perturbative unitarity, *i.e.* cross-sections that diverge for large momentum transfers [75]. However, this is strictly only a problem of the minimal simplified model considered as the benchmark process, and may point to additional BSM physics to restore unitary similar to the Higgs boson restoring unitarity in W^+W^- scattering [76].

CHAPTER 3

The ATLAS Experiment

The European Organisation for Nuclear Research ("*Conseil Européen pour la Recherche Nucléaire*," or CERN) was established in 1954, and has since hosted a number of particle physics experiments. The Large Electron-Positron collider (LEP), in operation at CERN from 1989 to 2000, was the largest particle accelerator ever built, with a circumference of 26.7 km. The LEP tunnel, stretching below the French-Swiss border near Geneva, now houses the LHC, see Figure 3.1.

3.1 The Large Hadron Collider

The LHC [78] is a hadron accelerator, storage ring, and collider designed for pp collisions in addition to heavy ion collisions such as lead–lead (Pb-Pb). The main motivation for the construction of the LHC was the study of the mechanism for electroweak symmetry breaking, one possible manifestation of which could be the Higgs boson, as discussed in Chapter 1. With the discovery of a particle consistent with the SM Higgs boson in 2012 [15, 16], the LHC transitioned to focusing on searches for other BSM physics. To do so, the LHC needs large centre-of-mass energies and instantaneous luminosities to search for increasingly high-mass and rare BSM physics processes.

The LHC proton beams are supplied by an injector chain, which during its first two periods of data-taking — the so-called Runs 1 and 2 — started from the CERN LINAC2 linear accelerator [79]. Using an array of three radiofrequency (RF) accelerator tanks, the LINAC2 pre-accelerates a hydrogen gas from rest, strips the hydrogen atoms of their electrons, and accelerates the resulting protons to energies of 50 MeV over a distance of 33 m [79]. From LINAC2, the proton beam is passed through as series of

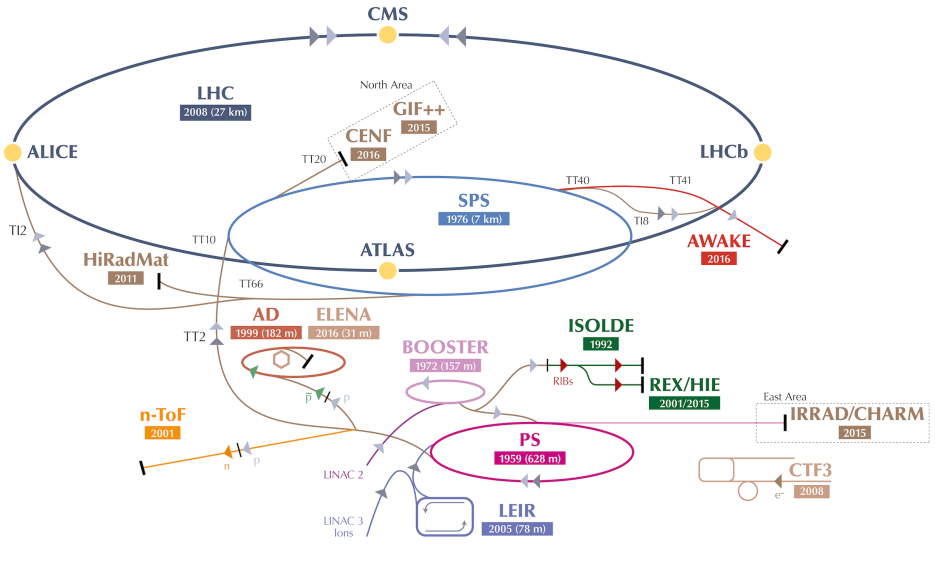


Figure 3.1 Schematic overview of the CERN accelerator complex. The LINAC2 accelerates protons from rest, after which they are injected into increasingly larger synchrotrons: the Proton Synchrotron Booster (PSB, or "Booster"), the Proton Synchrotron (PS), and the Super Proton Synchrotron (SPS). Finally, the protons are injected into the Large Hadron Collider (LHC) and are accelerated up to a design beam energy of 7 TeV. Figure adapted from Ref. [77].

three synchrotrons: the Proton Synchrotron Booster (PSB), with a circumference of 157 m, which accelerates the protons to 1.4 GeV, the Proton Synchrotron (PS), with a circumference of 628 m, which further increases the beam energy up to 25 GeV before the Super Proton Synchrotron (SPS), with circumference of 7 km, which accelerates the protons to energies up to 450 GeV prior to injection into the LHC. In each synchrotron, the protons are accelerated using RF cavities along the beam-pipe and steered using synchronously increasing dipole magnetic fields. Finally, the LHC will eventually accelerate the proton bunches up to a design energy of 7 TeV, for a pp centre-of-mass energy of $\sqrt{s} = 14$ TeV, compared to the 1.96 TeV achieved at the Fermilab Tevatron collider. In Run 1, between 2010 and 2012, the LHC operated at centre-of-mass energies of $\sqrt{s} = 7$ TeV and 8 TeV. During Run 2, in 2015–2018, the LHC has been running at $\sqrt{s} = 13$ TeV, close to its design energy.

Since it collides like-charged particles, the LHC is constructed as two separate beam-pipes containing counter-rotating proton bunches. The LHC beam-pipes intersect in four interaction regions along its perimeter, each housing one of the four large LHC experiments: ATLAS [80] and CMS [81] are so-called general-purpose experiments designed for high-energy pp collisions; the ALICE experiment [82] is designed

specifically for heavy-ion collisions; and LHCb [83] is a single-arm, forward experiment designed for flavour physics.

In each of the interaction regions, the two beams can be made to cross, thereby allowing the protons in each of the counter-circulating bunches to collide. The expected rate $\mathrm{d}N/\mathrm{d}t$ of collision events for a process with cross-section σ is given by

$$\frac{\mathrm{d}N}{\mathrm{d}t} = \mathcal{L}\sigma \quad \longrightarrow \quad N = \int \mathcal{L}\sigma \,\mathrm{d}t, \tag{3.1}$$

where $\mathcal{L}$ is the instantaneous luminosity. The total number of expected collision events N for any given process is proportional to the time-integrated luminosity, which is commonly referred to simply as $L = \int \mathcal{L}\,\mathrm{d}t$. The LHC is designed to operate with an instantaneous luminosity of $\mathcal{L} = 10^{34}\ \mathrm{cm}^{-2}\ \mathrm{s}^{-1}$ [78], or roughly two orders of magnitude larger than the peak instantaneous luminosity for the Tevatron. However, already since 2016, the second year of running at $\sqrt{s} = 13$ TeV, the LHC has managed to operate with up to twice its design luminosity. Each LHC bunch contains $\mathcal{O}(10^{11})$ protons, and the LHC is designed to store 2808 simultaneously circulating proton bunches in each beam, corresponding to a nominal bunch spacing of 25 ns [78]. This corresponds to a collision rate of approximately 40 MHz at each IP, including in the ATLAS experiment. This high luminosity means that the LHC has produced a peak number of more than 60 interactions per bunch-crossing — with average pile-up multiplicities during 2015-2016 data taking, and in Run 2 overall, of approx. 24 and 34, respectively [84] — making for highly complex final states.

Overview of the ATLAS Experiment

The ATLAS experiment, shown in Figure 3.2, has a cylindrical design with approximate rotational symmetry in the azimuthal plane and forward-backward symmetry along the beam axis. The nominal IP of the LHC beams is positioned at the centre of the detector, and serves as the origin of the standard coordinate system introduced in Section 1.2.

The ATLAS detector is designed to be almost hermetic in the azimuthal plane, and to provide large coverage in pseudo-rapidity. This is to minimise potential energy leakage in the transverse plane and to guarantee a large geometric acceptance of final state particles. Due to the presence of the beam-pipe, with an inner radius of 25 mm during Run 2 [86], it is not possible to construct a collider experiment with perfect 4π solid angle coverage.

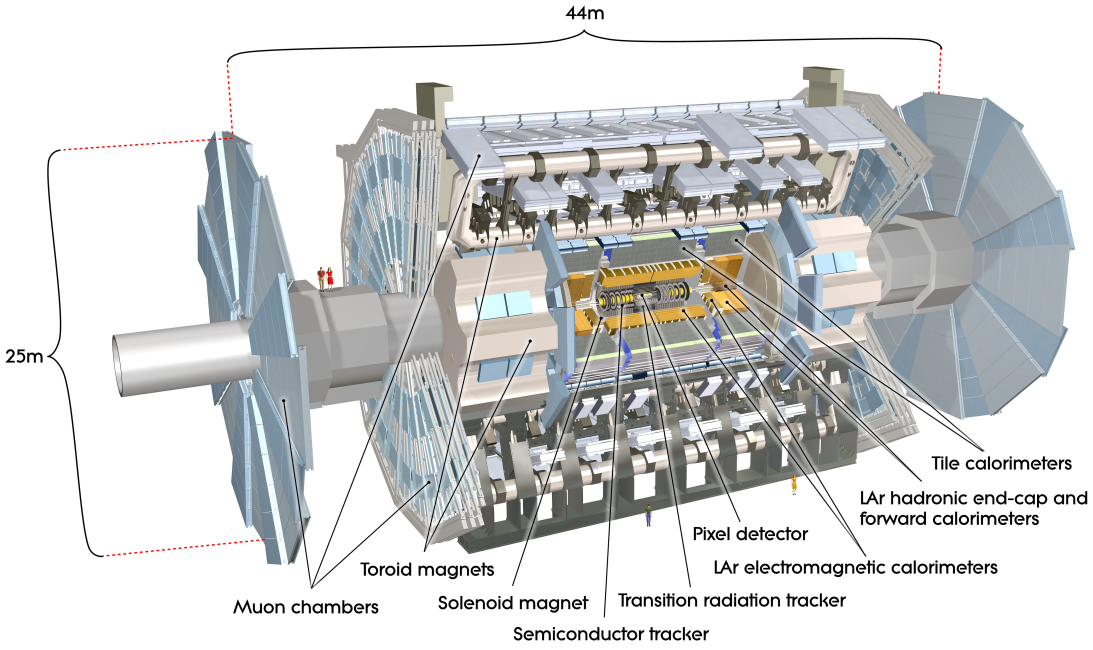


Figure 3.2 Computer-generated image of the ATLAS detector, highlighting each sub-detector. Figure from Ref. [85].

The ATLAS experiment is one of the two general-purpose detectors at the LHC that discovered the Higgs boson. To achieve this accomplishment, the ATLAS detectors was designed to be able to reconstruct *e.g.* photons, electrons, muons, taus, and hadrons — all particles that have been used in observations of the different Higgs boson decay modes [87]. To precisely reconstruct and identify this variety of particles, the ATLAS detector is constructed as a collection of nested sub-detectors at increasing radii from the IP, each specialised to measure different properties of the particles created in collision events. These sub-detectors are the

inner detector (ID), measuring the spatial location of points along the trajectory of electrically charged particles, with minimal impact on the trajectory itself. The ID comprises the pixel detector, the semi-conductor tracker (SCT), and the transition radiation tracker (TRT);

calorimeters, measuring the direction and total energy of most electromagnetically and hadronically interacting particles via full absorption. The calorimeter system consists of the electromagnetic calorimeter (ECAL) and hadronic calorimeter (HCAL); and

muon spectrometer (MS), measuring the direction and momentum of muons passing through the calorimeters, as well as providing specialised trigger capabilities. The

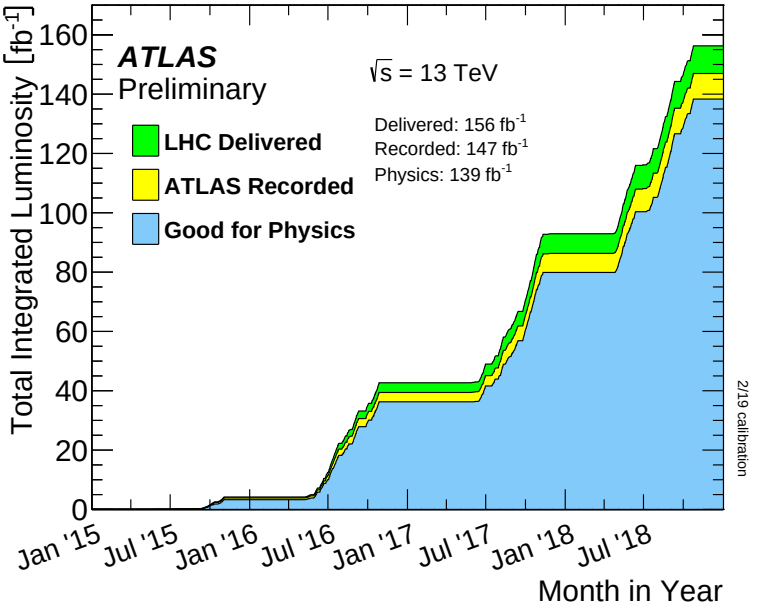


Figure 3.3 Cumulative integrated luminosity and data quality over the duration of the LHC Run 2. Figure from Ref. [84].

MS comprises the monitored drift tube chambers (MDTs), cathode strip chambers (CSCs), resistive-plate chambers (RPCs), and thin-gap chambers (TGCs).

Each sub-detector is highlighted in Figure 3.2. To achieve large coverage and granular measurements in the forward regions, *i.e.* at large $|\eta|$, the ATLAS sub-detectors are generally constructed as barrel and end-cap components: the central barrel component has active elements parallel to the beam-pipe while the forward end-cap components are placed perpendicular to the beam-pipe at each end of the barrel.

Finally, the 40 MHz nominal pp collision rate in the ATLAS experiment far exceeds the maximum rate at which the experiment is able to record collision events for reconstruction and later analysis, which is approx. 1 kHz [80]. The majority of collision events are not of primary interest, since processes involving *e.g.* Higgs boson production and hypothesised BSM processes, are exceedingly rare. Therefore, the ATLAS trigger system performs a fast online selection to achieve this massive reduction in event rate while retaining a high acceptance of interesting physics processes.

For the full duration of Run 2, the ATLAS experiment has been recording data with detector uptime and so-called "good data quality" efficiencies of $\geq 95\%$. This has allowed the ATLAS experiment to efficiently utilise the luminosity delivered by the LHC, recording 139 fb^{-1} of data which is good for physics analyses as shown in Figure 3.3.

3.2 Magnet system

ATLAS uses a combination of a solenoidal magnet and a three-part toroidal magnet, see Figure 3.2, to bend the charged particles through the Lorentz force, for charge identification and precise momentum measurement in the ID and MS.

The central solenoidal magnet envelops the ID with a radius of 1.2 m and an axial length of 5.3 m [80, 88]. It produces a 2 T axial magnetic fields, causing electrically charged particles to bend in the transverse plane. The solenoid is a thin superconducting magnet with a thickness of 4.5 cm (approx. 0.66 radiation lengths, X_0), designed to minimise the amount of material upstream of the calorimeters.

Downstream of the calorimeters, a set of one barrel and two end-cap toroidal magnet components are used to provide additional bending of muon trajectories for the MS [80]. These magnets produce a closed, circular $0.5 - 1$ T magnetic field in the transverse plane, almost perpendicular to the muon trajectories. These air-core toroid magnets extend from radius 4.7 m to 10 m with an axial length of 25.3 m. They create a strong magnetic field in a large volume with minimal material, providing the basis for high-precision measurements of muon momenta in the MS.

3.3 Inner detector

The task of the ATLAS ID, shown in Figure 3.4, is to perform high-resolution, non-destructive measurements of the trajectories of electrically charged particles. Using different detector techniques, the passage of a charged particle through a segment will result in a detectable, electrical signal. These so-called hits are used as spatial coordinates, or coordinate constraints, for the charged particle from which so-called tracks are constructed. For promptly produced charged particles, the track will originate from the IP and traverse the layers of the ID at increasing radii. Since the ID is embedded in the magnetic field of the central solenoid, the track produced by a charged particle will bend in the transverse plane with radius of curvature $r = p_{\mathrm{T}}/qB$ [11], where p_{T} is the transverse momentum of the particle, q is its electrical charge in units of the elementary charge, and B is the magnetic field strength. Using a minimum of three hits, the bending radius of a track can be measured and the charge and momentum of the charged particle can be determined. Additionally, charged particle tracking is crucial to identifying the so-called primary vertex (PV): the hard pp collision, possibly among

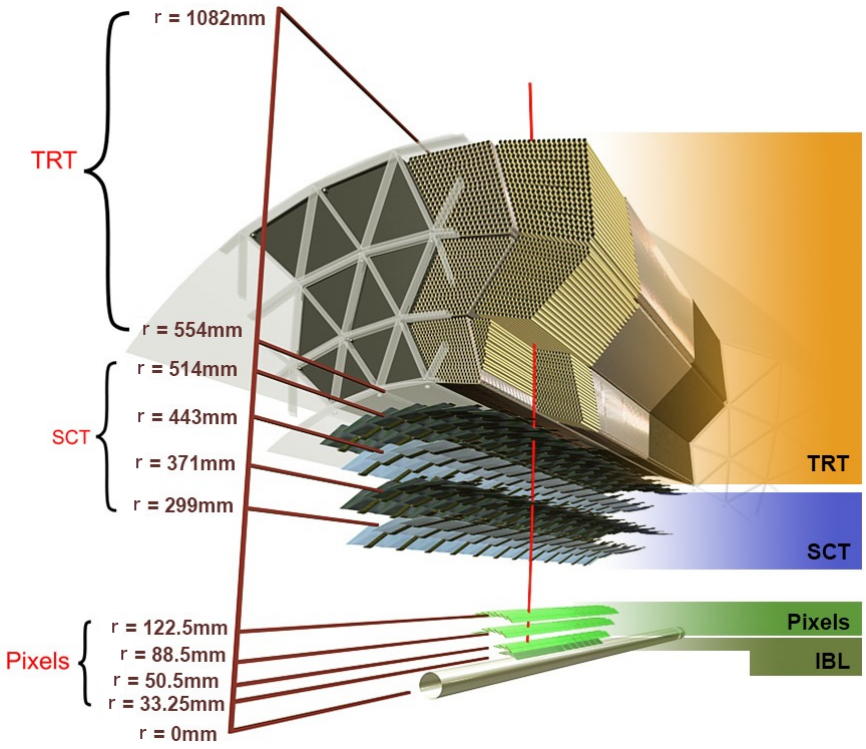


Figure 3.4 Computer-generated cut-out image of the ATLAS inner detector (ID), highlighting each sub-detector. The red line corresponds to the trajectory of a charged particle with $p_{\mathrm{T}} = 10$ GeV. Figure from Ref. [89].

many, from which the final state particles of interest originated. Collectively, the ID provides tracking capabilities for $|\eta| < 2.5$ for charged particles with p_{T} nominally down to 400 MeV [89].

Pixel detector

The two innermost sub-detectors in the ATLAS ID are the pixel detector and the SCT. Placed closest to the IP, these detectors must handle a large particle flux and provide measurements with high granularity to facilitate track and PV reconstruction with high resolution. Therefore, both are constructed as silicon detectors, as these can provide position resolution at the scale of tens of microns. The semiconductor detector sensors are operated in reverse-bias mode, creating a depletion region in the centre of the sensor which is devoid of free electrons [11]. Charged particles traversing the depletion region will excite a considerable number of electron-hole pairs which are then swept to opposite electrodes under the reverse bias electric field. This results in a signal current which is read out as a hit. In this way, semiconductor detectors act as solid-state ionisation chambers.

The ATLAS pixel detector [90] was initially constructed with three layers in the barrel region, between radii of 50.5 mm and 122.5 mm, as well as three end-cap disks providing coverage for $|\eta| < 2.5$ [80]. In 2014, the insertable B-layer (IBL) was installed in the barrel region, at a radius of 33 mm from the beam axis [86], providing improved resolution closest to the IP. All pixels are approx. 250 μm thick, so as to minimise energy loss in the ID. The nominal pitch in the barrel region is $50 \times 400\,\mu\text{m}^2$ in $\phi \times z$, with a similar pixel size in $\phi \times R$ in each of the end-caps. This corresponds to an intrinsic accuracy of $10 \times 115\,\mu\text{m}^2$ in $(R - \phi) \times z$ and $(\phi - z) \times R$, respectively. Since the ID is immersed in the axial magnetic field produced by the central solenoid, the spatial location of hits in the transverse plane are more important for the momentum determination than along the longitudinal axis. This guided the choice of $1 : 8$ sensor dimensions.

Semi-conductor tracker

Outside the ATLAS pixel detector, the SCT provides precision tracking using the same semiconductor technology [80]. The barrel SCT is comprised of 4 cylindrical layers, each with an axial length of 150 cm, located between radii of 29.9 cm and 55.4 cm. On each side of the barrel, the SCT end-cap consists of 9 disks between $|z| = 8.53$ m and 27.2 m, providing coverage for $|\eta| < 2.5$. Since the SCT has a considerably larger active area than the pixel detector (63 m^2 *cf.* 1.9 m^2), it is built using a cost-effective strip design. The sensors use silicon strips with 80 μm pitch, a thickness of approx. 285 μm, and a strip length of approx. 6 cm. The SCT strips are grouped in pair-wise back-to-back sensor modules, rotated at a relative stereo angle of 40 mrad. The axial strips in the barrel are oriented parallel to the beam-axis, and perpendicular to the beam-axis in the end-caps; the off-axis strips are rotated at a stereo angle, in the plane of the detector layers, with respect to these. Charged particles traversing an SCT module will normally hit two overlapping strips, allowing precise location determination even in the longitudinal direction. For this reason, SCT modules have an intrinsic accuracy of $17 \times 580\,\mu\text{m}^2$ in $(R - \phi) \times z$ and $(\phi - z) \times R$ in the barrel and end-caps, respectively. Similarly to the pixel detector, maximal resolution in the bending direction of the central solenoid is prioritised for the SCT.

Due to the remarkable spatial resolution of the ATLAS silicon detectors, hits in the pixel detector and SCT modules are referred to as “space-points” and provide a high-granularity basis for tracking and vertex-finding.

Transition radiation tracker

The TRT [80, 91] is constructed from drift tubes, also called straws, with a diameter of 4 mm, and is composed of a barrel and two end-cap components. The straws in the barrel region have a length of 144 cm and are aligned in parallel with the beam-axis out to $|z| = 712$ mm. The 37 cm long end-cap straws are positioned in wheels perpendicular to the beam between radii 64.4 cm $< r <$ 100.4 cm, providing a combined TRT acceptance of $|\eta| < 2.0$. As a drift tube detector, the TRT provides 2D position constraints in the plane perpendicular to the straws, in contrast to 3D space-points from silicon detectors.

At the centre of each straw is a 31 μm diameter gold-plated tungsten anode wire, which is connected to each end of the straw and is kept at ground potential. The walls of the drift tube, or the cathodes, have a thickness of 70 μm and are operated at a potential of −1530 V. Nominally, the straws were filled with a gas mixture of 70% Xe, 27% CO_2, and 3% O_2. However, due to leakages, parts of the TRT now uses as a gas mixture primarily made up of the cheaper argon, with similar performance [92].

Charged particles traversing a drift tube will ionise the gas, thereby creating electron-ion pairs. The negatively charged ionisation electrons drift towards the anode wire and are accelerated by the electrical field, initiating an ionisation cascade resulting in a gain of approx. 2.5×10^4. The positively charged ions similarly drift towards the cathode, albeit more slowly. The electrical pulse induced by the ionisation particles reaching the anode is read out as a signal. Based on the drift-time, *i.e.* the time between the 40 MHz LHC clock and the leading edge of the signal pulse, the impact parameter of the original charged particle with respect to the anode wire can be estimated. This provides an intrinsic drift-time accuracy of 130 μm in the plane transverse to the straw. Charged particles with $p_T > 500$ MeV and $|\eta| < 2.0$ will traverse at least 36 straws, except in the barrel–end-cap transition region at $0.8 < |\eta| < 1.0$, where particles will only traverse a minimum of 22 straws [80].

In addition to providing spatial measurements of charged particles, the TRT also has capacity for particle identification. The TRT straws are interleaved with polypropylene fibres in the barrel and foils in the end-caps. These cause highly relativistic, electrically charged particles — particularly electrons with large Lorentz factors γ — to emit transition radiation photons. Their photoelectric absorption in the straw gas yields significantly larger signals than lower-γ particles. By defining separate low and high thresholds for TRT hits, this can be used to identify electrons against the dominant charged pion background.

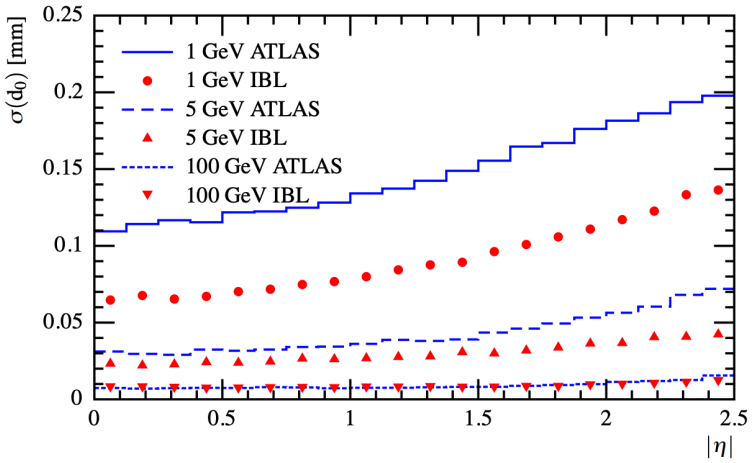


Figure 3.5 Expected resolution of the transverse impact parameter d_0 for single muons with energies of 1, 5, and 100 GeV as a function of the muon $|\eta|$. Resolution are shown for the original ATLAS inner detector (ID), as well as with the addition of the insertable B-layer (IBL). Figure from Ref. [86].

Performance

The main Phase-0 upgrade (*i.e.* prior to Run 2) of the ATLAS ID with respect to the original design is the addition of the IBL [86]. The IBL was installed to reduce the rate of fake tracks at high luminosities, mitigate deterioration over time of the original ATLAS pixel detector *e.g.* due to large radiation doses, and improve parameter resolution, particularly for low-p_{T} tracks, see Figure 3.5.

Apart from this, the ATLAS ID has performed on par with the design expectations in Ref. [80] during Run 2. For instance, the expected resolution of the transverse impact parameter d_0 for central tracks was expected to be $\sigma(d_0) \approx 20\,\mu\mathrm{m}$ at $p_{\mathrm{T}} = 10$ GeV, and a 2015 performance study in low–pile-up Run 2 data found the measured resolution to be in agreement with this value [89]. The ATLAS ID has had such a stable performance because the resolution of the transverse impact parameter, *e.g.*, is driven by the resolution of the track hit measurements in the pixel detector. This spatial resolution is determined by the detector elements themselves and therefore cannot be improved in subsequent offline reconstruction and calibration.

3.4 Calorimetry

After traversing the ID, particles will enter the ATLAS calorimeter system, see Figure 3.6. While the ID is designed for precise spatial measurements with minimal

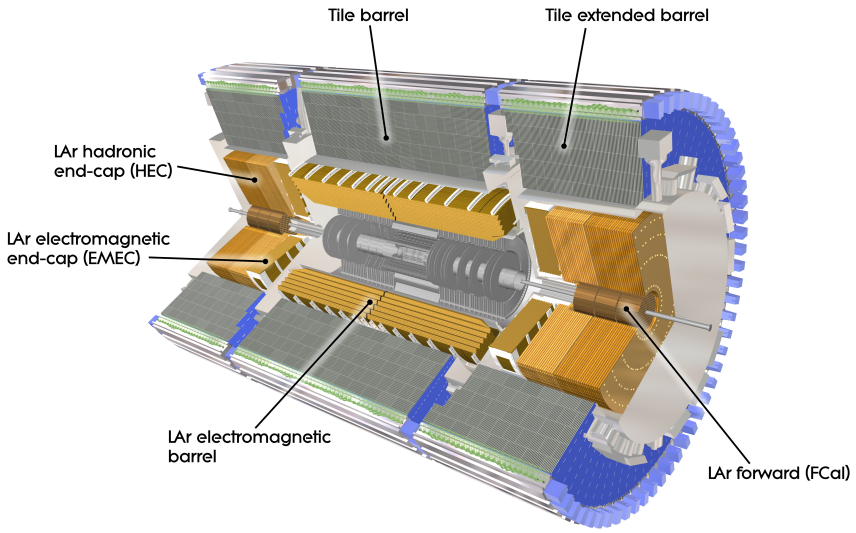


Figure 3.6 Computer-generated cut-out of the ATLAS calorimeter system, comprised of the electromagnetic calorimeter (ECAL) and hadronic calorimeter (HCAL). Figure from Ref. [93].

impact on the energy and trajectory of electrically charged particles, the calorimeters are designed to measure the total energy of electromagnetically and hadronically interacting particles. This is done by absorbing the particles, with sufficient granularity to also measure their position in $\eta - \phi$.

When encountering a dense material, electromagnetically interacting particles (*e.g.* electrons and photons) will typically initiate a cascade: *e.g.* an incident electron emits a bremsstrahlung photon, which subsequently decays to a pair of electrons, and so on. This results in a electromagnetic shower, characterised by the radiation length X_0, the distance over which a high-energy electron ($E_e \gg 1$ GeV) will have lost all but $1/e$ of its initial energy due to bremsstrahlung radiation [94]. The radiation length is also related to the mean free path for high-energy photons to decay to an electron pair, which is given by $9X_0/7$. Since the radiation length is material-specific, passive materials with short radiation lengths allow compact calorimeters to efficiently contain electromagnetic showers.

Similarly, energetic hadrons such as protons or neutrons encountering passive material will initiate hadronic showers. These are characterised by the nuclear interaction length λ_{int}, which is the average distance a high-energy hadron will travel in the material before interacting with the nuclei [94]. Generally, for the same material, the nuclear interaction length is considerably larger than the radiation length. This means that hadronic showers tend to extend much further in the longitudinal and lateral dimensions, necessitating

bigger calorimeters for efficient containment.

For both electromagnetic and hadronic showers, the longitudinal position of the point where the shower activity is at its maximum scales only with the logarithm of the energy of the incident particles [94]. This means that the same calorimeter system can be used for particles with energies spanning several orders of magnitude, from $\mathcal{O}$(GeV) to $\mathcal{O}$(TeV).

Separate calorimeter systems are typically designed for each process: ECALs are designed for almost full containment of electromagnetic showers, and are typically placed closest to the beam, while HCALs are designed to absorb hadronic showers, and are typically placed outside of the ECAL due to the large values of λ_{int} in common absorber materials. However, the distinction between electromagnetic and hadronic showers, and by implication between ECALs and HCALs, is not clear-cut: Hadronic showers may start already in the ECAL and will also have an electromagnetic component, *e.g.* from neutral pion decays to photons.

The different components making up the ATLAS calorimeter system provide coverage for $|\eta| < 4.9$. Similarly, the combined calorimeter has a depth corresponding to approx. 10 λ_{int} in both the barrel and end-cap regions. This minimises potential lateral leakage of hadronic showers, also called punch-through, thereby achieving near-hermetic containment. Given that the initial net transverse momentum in any collision event is effectively zero, this hermeticity allows for the determination of any net missing tranverse energy ($E_{\text{T}}^{\text{miss}}$) in an event. Large $E_{\text{T}}^{\text{miss}}$ values may be an indication of a final-state neutrino, which is the only SM particle that consistently does not interact with the detector.

Electromagnetic calorimeter

The ECAL comprises the liquid-argon (LAr) barrel and electromagnetic end-cap (EMEC) detectors, see Figure 3.6. Both of these are lead–liquid-argon (Pb-LAr) sampling calorimeters with an accordion geometry, see Figure 3.7 [80]. This unique layout avoids potential azimuthal gaps, such that no particles can escape the calorimeter undetected. The lead is used as absorber material with a radiation length of $X_0 = 5.6$ mm [94]. By contrast, the radiation length of LAr is $X_0 = 140$ mm [94]. The lead causes the incident particles to shower, resulting in a large number of secondary particles. These then ionise the LAr active material, resulting in a large number of low-energy electrons which drift to the electrodes and are detected as a signal. In sampling calorimeters such as the

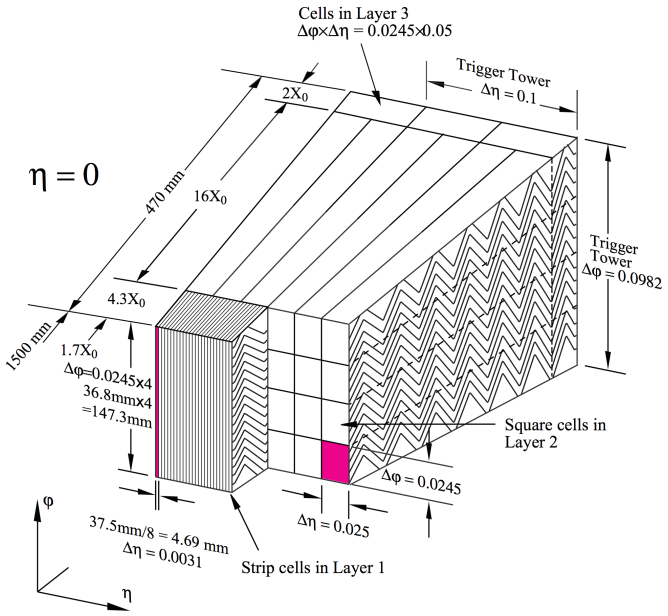


Figure 3.7 Schematic representation of a section of the ATLAS electromagnetic calorimeter (ECAL) at $\eta = 0$. Figure from Ref. [80].

ATLAS ECAL, only a fraction of the energy of the incident particle is converted into ionisation. This is because the remainder of the energy of the particle is deposited in the absorber material and is therefore not observed.

The LAr barrel covers $|\eta| < 1.475$ while the EMEC, constructed as two coaxial wheels partitioning the detector in R with the same extent in z (called the inner and outer wheels), extends coverage to $1.375 < |\eta| < 3.2$ [80]. In the range $|\eta| < 2.5$, comprising the LAr barrel and outer EMEC wheel (at larger radii R), the ECAL is divided into three depth layers with different granularity in $\eta - \phi$. This region overlaps with the ID coverage and is dedicated to precision physics. By contrast, the inner EMEC wheel (at smaller radii R), covering $2.5 < |\eta| < 3.2$, has two depth layers. The first, so-called strip layer is finely segmented in η, see Figure 3.7, allowing for precise separation of single-photon showers and the two overlapping showers from neutral pion decays [95]. Collectively, the three ECAL layers measure the total energy of electromagnetic showers and provide information about their lateral and longitudinal shape. Finally, a presampler detector, consisting of a thin active LAr layer, covers the range $|\eta| < 1.8$ and is used to estimate energy losses in passive material before the calorimeter. Combined together, the depth of the ECAL corresponds to $> 22\ X_0$ in the LAr barrel and $> 24\ X_0$ in the EMEC. Therefore, most electromagnetic particles are expected to be fully contained within the ECAL. However, in the transition region $1.37 < |\eta| < 1.52$ between the LAr barrel and the EMEC, also called the “crack” region, the amount of passive material before the ECAL is up to twice as large as the more central and forward regions. Therefore, ECAL

measurements in this region are typically not used for physics analyses, which will also be the case for the analysis described in Part II of this thesis. Finally, the ECAL depth only corresponds to approx. 2 λ_{int}, meaning that the ECAL alone has insufficient capacity to contain hadronic showers.

Hadronic calorimeter

The ATLAS HCAL is designed to absorb hadronic showers which start or extend beyond the ECAL. The HCAL comprises the scintillating-tile calorimeter (TileCal), the LAr hadronic end-cap (HEC), and the forward calorimeter (FCal), see Figure 3.6.

The TileCal is a three-layer sampling barrel calorimeter, covering $|\eta| < 1.7$ [80]. It uses a steel bulk with $\lambda_{\text{int}} = 16.8$ cm [11] as absorber, interleaved with scintillating tiles as the active medium. The scintillating tiles are made from polystyrene doped with approx. 1.5% fluorescent materials. The ionising secondary particles created in the steel absorber excite the fluorescent molecules which in turn emit ultraviolet scintillation light. Wavelength shifting fibres placed in contact with the tiles are used to read out the scintillation light as a signal using photomultiplier tubes (PMTs). The TileCal has a radial depth of approx. 7.4 λ_{int}.

The HEC is a copper–liquid-argon (Cu-LAr) sampling calorimeter with flat-plate design, in contrast to the accordion design of the ECAL. Placed immediately after the EMEC along the beam-axis, the HEC extends the coverage of the HCAL for $1.5 < |\eta| < 3.2$.

Finally, the FCal provides electromagnetic and hadronic calorimetry for $3.1 < |\eta| < 4.9$. The first FCal layer is optimised for electromagnetic showers and is constructed using copper rods embedded in a copper matrix in a hexagonal patterns. The rods are placed in parallel with the beam, and provide fast measurements in the high-flux forward region. The last two FCal layers are optimised for hadronic showers, using the same design as the first layer but with tungsten as the main absorber material around the copper rods.

Performance

The ATLAS ECAL is instrumental in the reconstruction of electromagnetically interacting particles, such as electrons (e) and photons (γ). These particles are reconstructed and calibrated using the same procedure, described in Chapter 7, which remained the same during Runs 1 and 2. The relative e/γ energy resolution was

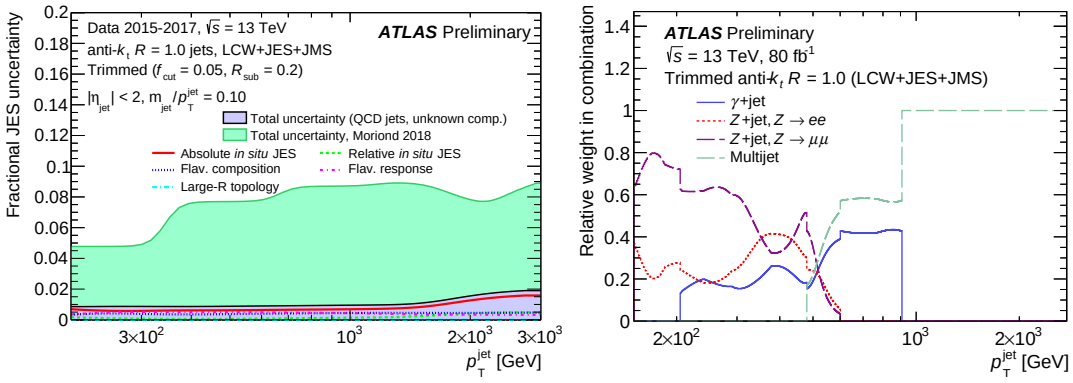
measured in Run 1 to be approx. 1% at $E_{\mathrm{T}} = 200$ GeV [96], which is in excellent agreement with the ATLAS design targets [80].

In combination, the ECAL and HCAL also enable the reconstruction of hadronic jets in the ATLAS detector. These are typically categorised as small-radius (small-R) jets, with radius parameter $R \approx 0.4$, and large-R jets, with radius parameter $R \approx 1$. ATLAS calorimeter resolution design targets are given for single charged pions, based on test-beam experiments, so no clear targets exist for hadronic jets.

The reconstruction and calibration procedure for standard small-R jets is generally unchanged since Run 1 [97]. The measured relative small-R jet energy resolution in Run 2 data at $p_{\mathrm{T}} = 500$ GeV is approx. 5% [98], consistent with measurements in early Run 1 data [99]. At high p_{T}, the constant terms dominates the jet energy resolution, whereas the stochastic terms dominates at low p_{T}, due to fluctuations in the amount of energy sampled from the hadronic shower. In this region, *e.g.* at a lower p_{T} of 20 GeV, the relative jet energy resolution is approx. 30%. However, this has been reduced to approx. 25% by the use of the ATLAS particle flow algorithm [100], which utilises the superior energy resolution of the ATLAS ID at low p_{T} compared to the calorimeter system. The particle flow jet definition will become the default for small-R jet in ATLAS in Run 3.

In addition to small-R jets, large-R jets are of particular relevance to the analysis in Part II. As with small-R jets, the large-R jet energy and mass resolutions are driven by stochastic effects, and have therefore been stable since Run 1 for the same jet input and grooming definitions [40]. For instance, the relative jet mass resolution at $p_{\mathrm{T}} \approx 500$ GeV for standard trimmed anti-k_t jets with a radius parameter of $R = 1.0$ is approx. 10% [33]. However, the calibration of large-R jets has recently been extended to include *in situ* calibration in addition to the previous MC-based procedures [101]. This has led to a reduction of the large-R jet energy scale uncertainty from approx. 8% at $p_{\mathrm{T}} = 1$ TeV to approx. 1 % [102], see Figure 3.8a.

The *in situ* calibration considers three different final states, where a large-R jet is balanced against a well-calibrated reference object or set of objects: large-R jet + γ, large-R jet + $Z \to ee$ and $\mu\mu$, and multijet events, *i.e.* large-R jet + multiple small-R jets. The relative weight assigned to each of these techniques in the combination of *in situ* measurements is shown in Figure 3.8b. This approach has been used for small-R jets since Run 1, but *in situ* large-R calibration was only introduced in 2019, and was therefore not available for use in the analysis in Part II. However, as the large-R jet calibration is found to be the second largest source of systematic uncertainty, this



(a) Fractional jet energy scale uncertianty **(b)** Weight in *in situ* combination

Figure 3.8 Novel *in situ* large-*R* jet calibration procedure in ATLAS. The filled teal curve in Figure **(a)** indicates the uncertainty in 2018, at the time of the analysis in Part II; the black curve indicates the reduced uncertainty from the *in situ* calibration, using the combination of final states shown in Figure **(b)**. Figures from Ref. [102].

improved calibration is likely to benefit future iterations of this analysis.

3.5 Muon spectrometer

Most particles that interact with the ATLAS detector are stopped by the calorimeter system. A notable exception are muons, with a lifetime of $c\tau \approx 660$ m, which typically will not deposit all of their energy in an electromagnetic shower. For particles with energies below some critical energy, energy loss by ionisation will be more important than radiative processes. This critical energy is approx. 40,000 times greater for muons than for electrons, meaning that muon interactions with matter will be overwhelmingly ionisation-dominated, even at large energies [94]. Four specialised sub-detectors located on both sides of, and embedded in, the toroidal magnet system are used to measure muons exiting the ATLAS calorimeters, see Figure 3.9.

The MDTs cover the region $|\eta| < 2.7$ using the same detector technology as the TRT. They are constructed as a number of separate chambers in a barrel component and two end-caps components. Muons will typically traverse three to four such chambers, each comprising between three and eight layers of drift tubes with diameters of 30 mm. Due to the field in the toroidal magnet, the drift tubes in both the barrel and end-caps are aligned along ϕ, *i.e.* parallel to circles around the beam-axis. Similarly to the TRT, the MDTs provide position measurements only in two dimensions ($R - z$).

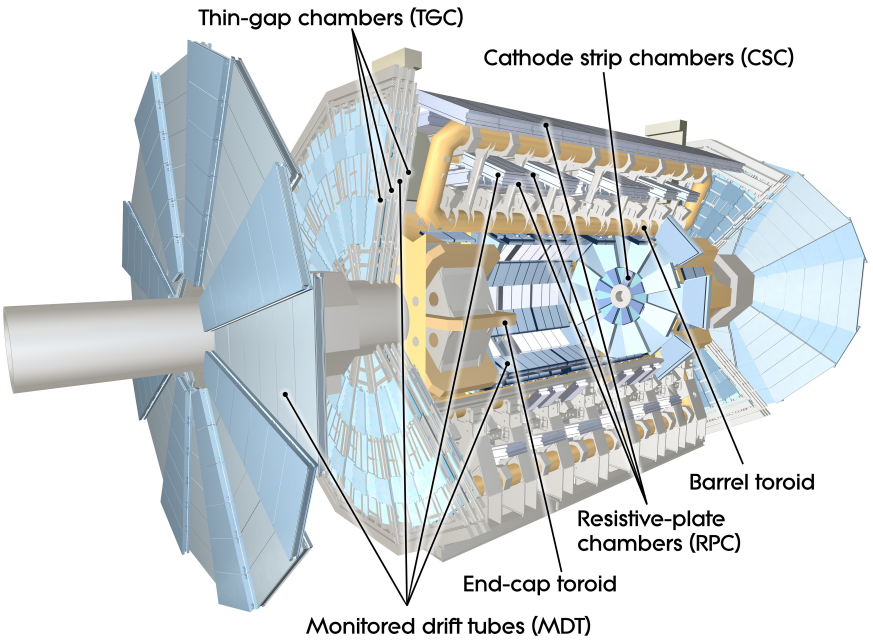


Figure 3.9 Computer-generated cut-out of the ATLAS muon spectrometer (MS), comprised of the monitored drift tube chambers (MDTs), cathode strip chambers (CSCs), resistive-plate chambers (RPCs), and thin-gap chambers (TGCs). Figure from Ref. [103].

Due to the large particle occupancy in the forward region, the first end-cap MDT layer is replaced with CSCs. The CSC in each end-cap comprises four layers of multiwire proportional chambers. The anode wires are aligned radially, while the cathode strips on each side are mutually orthogonal, with one set parallel to the anode wires. Electrical signals from the ions produced by a charged particle traversing the gas in a multiwire proportional chamber are read out from both cathode strip planes. Thereby, the CSCs provide measurements in both dimensions transverse to the beam. The CSCs cover the forward region $2.0 < |\eta| < 2.7$, and with a drift time of approx. 40 ns, compared to the 700 ns for the MDTs, they are capable of operating in large particle fluxes [80].

The MDT and CSC sub-detectors provide high-precision muon measurements over a large range in pseudo-rapidity, but both have drift-times beyond the nominal LHC bunch spacing of 25 ns. Therefore, they are complemented by the RPCs and the TGC, which provide fast trigger capability for $|\eta| < 2.4$. In the barrel region ($|\eta| < 1.05$), the RPCs are constructed as three concentric layers each with two active detector layers. The RPCs are constructed from parallel plate layers with a separation of 2 mm, for a time resolution of approx. 5 ns. Each layer is instrumented with parallel read-out strips, oriented such that the strips in each layer are orthogonal. Finally, the forward regions ($1.05 < |\eta| < 2.4$) are covered by the TGCs. These are multiwire proportional chambers

with an anode-to-cathode distance which is roughly half that of the CSC. The TGCs comprise a total of 9 layers, placed around the first and second MDT end-cap layers. The triggering detectors additionally provide measurements of the muon ϕ coordinate to complement the MDT measurements.

Performance

The performance requirement driving the design of the ATLAS MS was a relative p_{T} resolution of 10% for muon tracks with $p_{\mathrm{T}} = 1$ TeV, independently of the ID. The MS p_{T} reconstruction performance is limited by the track sagitta, which is approx. 500 µm at 1 TeV, corresponding to a requirement on MS chamber alignment precision to be better than 40 µm [104]. The muon momentum resolution in early Run 2 data has been measured to be less than 4% up to $p_{\mathrm{T}} = 200$ GeV, with limited statistic available above this point. Furthermore, studies in Run 2 MC simulation has indicated that by incorporating *in situ*–deterimined alignment effects in the muon track fitting procedure, it is possible to achieve an expected muon p_{T} resolution as low as 7-8% across η at $p_{\mathrm{T}} = 1$ TeV [104, 105].

3.6 Trigger system

During most of Run 2, the LHC has operated at the target proton bunch-spacing of 25 ns, resulting in a pp collision rate of 40 MHz. With an average size of 1.3 MB/event [80], recording all collision events in ATLAS is infeasible. The ATLAS experiment therefore uses a two-tier trigger system, consisting of the Level-1 (L1) trigger and the high-level trigger (HLT), to identify and record a small subset of interesting collision events [80, 106].

The L1 is a specialised, synchronous hardware-based trigger which uses coarse data from the calorimeter system and the MS, specifically the RPCs and the TGCs. With a latency of less than 2.5 µs, the L1 trigger identifies signs of high-p_{T} charged leptons, hadronic activity, or $E_{\mathrm{T}}^{\mathrm{miss}}$ along with their approximate orientation in $\eta - \phi$ as regions of interest (ROIs). The L1 trigger accepts interesting events up to a rate of 100 kHz, which are then passed on to the HLT.

The HLT is an asynchronous software-based trigger, which analyses the full event information based on the L1 ROI seeds. At this trigger stage, it is possible to perform

e.g. charged particle tracking, particle identification and isolation, jet building, *etc.*. Events which are selected by the HLT are recorded at a rate of approx. 1 kHz and used in offline physics analyses.

At both trigger stages, events can be selected based on a large number of final state hypotheses, implemented as so-called chains of trigger algorithms and selections. The menu of these trigger chains is constructed based on the needs of all ATLAS physics analyses, and is updated in response to different running conditions, special runs, *etc.*. Among other things, the ATLAS trigger system enables the selection and recording of events that could be attributable to DM particles or mediators produced in pp collision events.

3.7 Upgrades

Since the installation of the ATLAS detector in 2008 and first data-taking in 2010, the detector has undergone a number of hardware upgrades. The so-called Phase-0 upgrades took place during Long Shutdown 1 (LS1) between LHC Runs 1 and 2. The most prominent hardware upgrade during this time was the insertion of the IBL and the reduction in the diameter of the beam pipe near the ATLAS IP [86].

With the completion of Run 2 in 2018 the LHC is currently in LS2, where Phase-I upgrades are under way in preparation for Run 3, with planned commissioning in early 2021 [107]. These include the installation of the new small wheel (NSW) [108], which replaces the current so-called small end-cap wheels, comprising the CSCs and most of the TGCs in the first end-cap layer. It will provide improved tracking efficiency and resolution, as well as reduced L1 trigger rates in the forward region, $1.3 < |\eta| < 2.7$. Other crucial Phase-I developments are the upgrades of the ATLAS L1 trigger. These include an upgrade to the LAr calorimeter system, which replaces the trigger towers with more finely segmented cell information for each ECAL layer [109], as well as new topological trigger processor (L1Topo), combining information from the calorimeter and muon systems, allowing for the computation of angular separations and invariant masses of physics objects at the L1-level [110]. Finally, the Fast TracKer (FTK) is currently being developed to allow for real-time tracking in all events accepted at L1 for use by the HLT [111].

After Run 3, slated for completion in 2023, the Phase-II upgrade will be performed in anticipation of the high-luminosity LHC (HL-LHC), with planned commissioning at

the end of 2026 [107]. The Phase-II hardware upgrades include the Inner Tracker (ITk), which will replace the Phase-I ATLAS ID with an all-silicon pixel and strip detector covering $|\eta| < 4.0$ [112, 113] as well as the optional high-granularity timing detector (HGTD) which can provide improved pile-up mitigation and minimum bias triggering in the forward regions. Finally, the ATLAS trigger system will be upgrade to have a single-layer hardware trigger (L0) with a maximum accept rate of 1 MHz followed by a software-based event filter (EF) with a maximum output rate of 10 kHz, which will allow the ATLAS experiment to keep trigger thresholds low even at instantaneous luminosities of $\mathcal{L} = 5 \times 10^{34}$ cm^{-2} s^{-1}, as expected at the HL-LHC [114].

CHAPTER 4

Machine Learning

The HLT of the ATLAS experiment nominally accepts collision events at a rate of approx. 1 kHz as mentioned in Chapter 3. Each recorded event takes up approx. 1.3 MB of raw detector data [80], from which final state particle candidates need to be reconstructed, calibrated, and analysed to extract physics results. The complexity of challenges facing many high-energy physics (HEP) tasks — extracting information from vast amounts of high-dimensional data — is what makes them uniquely suited for machine learning (ML) [115]. Below, the basics of two common ML techniques, namely neural networks (NNs) and boosted decision trees (BDTs), are presented, with an emphasis on the former. This chapter presents a high-level overview, with additional technical details given in Appendix B.

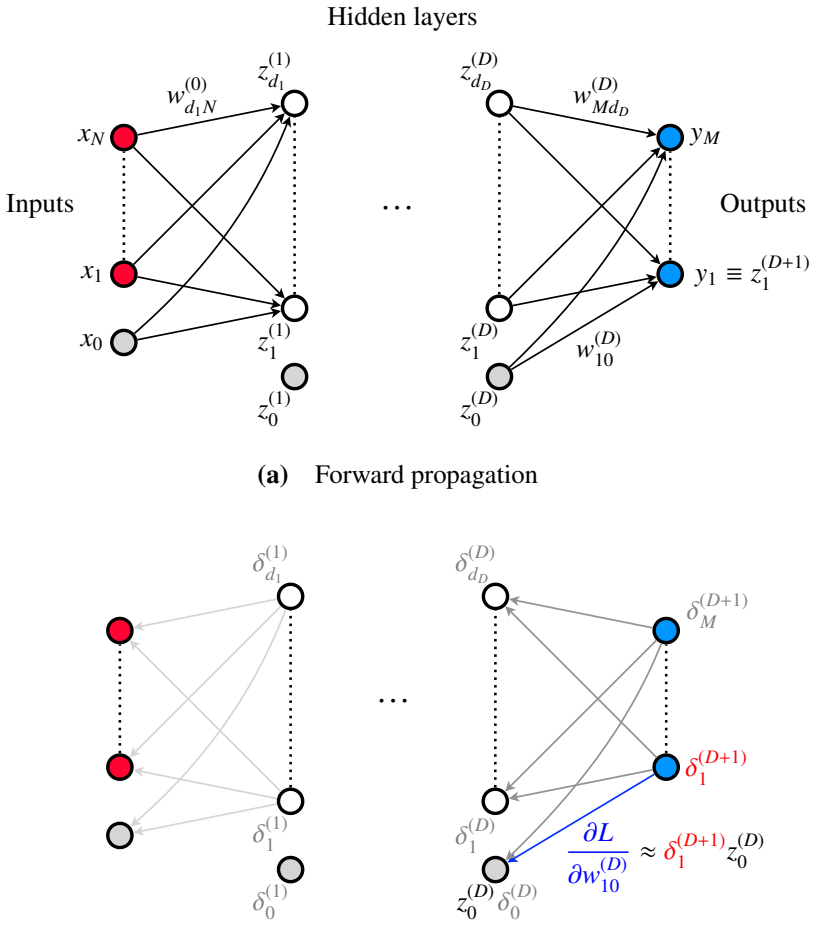
4.1 Neural networks

NNs are a general class of multivariate functions $f : \mathbb{R}^N \to \mathbb{R}^M$, mapping a vector of input features $\mathbf{x} = (x_1, x_2, \ldots, x_N)$ to some output $\mathbf{y} = (y_1, y_2, \ldots, y_M)$. Such mappings can generally be used either for classification or regression problems. This class of functions is constructed by connecting the so-called input and output layers, $\mathbf{x}$ and $\mathbf{y}$, by a number of hidden layers, as illustrated in Figure 4.1a.

The connection between adjacent layers is implemented as a linear transform of the inputs along with the application of a non-linear activation function $h : \mathbb{R} \to \mathbb{R}$

$$a_i^{(l+1)} = \sum_{j=1}^{d_l} w_{ij}^{(l)} z_j^{(l)} + b_i^{(l)} = \sum_{j=0}^{d_l} w_{ij}^{(l)} z_j^{(l)}, \qquad z_i^{(l)} = h(a_i^{(l)}). \tag{4.1}$$

Here, l is the layer index; $\mathbf{z}^{(l)} = \{z_i^{(l)}\}$ are the values, or activations, of each of the



(a) Forward propagation

(b) Backward propagation

Figure 4.1 Schematic diagram of a neural network (NN) with densely connected, feed-forward architecture. Each node represents an input, output, or activation value. Each line represents a network parameter or weight, the ones connected to 0-index variables corresponding to biases. Arrows indicate the direction of the data flow in the **(a)** forward and **(b)** backward propagation pass.

so-called nodes in the l^{th} layer; $d_l = \| \mathbf{z}^{(l)} \|$ is the number of nodes in the l^{th} layer; $\mathbf{W}^{(l)} = \{w_{ij}^{(l)}\}$ is the $d_{l+1} \times d_l$ weight matrix and $\mathbf{b}^{(l)} = \{b_i^{(l)}\}$ is the vector of biases with length $\| \mathbf{b}^{(l)} \| = d_{l+1}$, collectively parametrising the connection; and $\mathbf{a}^{(l)} = \{a_i^{(l)}\}$ is the associated linear combination, or output, to which the non-linear activation is applied. The full set of network parameters, including biases, is collectively labelled θ. Equation (4.1) has the inputs $z_i^{(0)} = x_i$ and outputs $z_i^{(D+1)} = y_i$ as special cases. The type and number of hidden layers, their mutual connections, and the number nodes per hidden layer is referred to as the architecture of the NN. The architecture and NN training parameters such as the learning rate (see below) are among the so-called hyperparameters of the NN, which will typically be optimised as part of a training phase. Each additional hidden layer, and each additional node in the hidden layers, increases the capacity of the NN, in the form of additional tunable network parameters θ. This allows the NN to approximate increasingly complex functional relations. Generally, NNs are so-called universal approximators, meaning that provided sufficient capacity, they are able to approximate any continuous function to arbitrary accuracy [116]. This property makes NNs well suited for complex computational tasks, provided sufficiently large datasets for tuning the network parameters.

The artificial neural networks used in this thesis are historically inspired by the simplistic modelling of biological neural networks [117]. In this simplified picture, the activity of a given node, or "neuron," is dependent on the weighted information feeding into it from the neurons in the preceding layer(s), and the activation function models the "firing" of the neurons. Examples of standard activation functions h are given in Appendix B.

Training

So far, the basics of NNs and the forward propagation of information from the input domain to the output domain, through Equation (4.1), have been described. However, a crucial concept is the training of the NN itself, *i.e.* the tuning of the network weights θ for a particular classification or regression task. This is the "learning" in machine learning. The network weights are not known *a priori*, and are therefore typically initialised by drawing at random from some distribution, *e.g.* a Gaussian. In standard functional χ^2-regression, the gradient of the objective with respect to each of the weights or fit parameters can be calculated exactly, and a gradient descent algorithm can be used to minimise the objective. For NNs, however, which may have millions of network weights, the dimensionality of the parameter space makes this approach infeasible. Therefore, an alternative approach to weight tuning is required.

In the discussion of NNs in thesis, X will denote the distribution of NN inputs in the feature space $\mathbb{R}^N$ introduced above, and $\mathbb{R}^N \ni \mathbf{x} \sim X$ will be the concrete values of the input features for a given (training or testing) example, drawn from X. Similarly, Y will denote the distribution of NN targets in $\mathbb{R}^M$, and $\mathbb{R}^M \ni \mathbf{y} \sim Y$ will be the concrete targets for a given example. When input features and targets are drawn coherently from the distributions X and Y for a given example, this is written as "$\mathbf{x} \sim X, \mathbf{y} \sim Y$" to indicate the relation between the two associated realisations $\mathbf{x}$ and $\mathbf{y}$. For vector-valued inputs and targets, the i^{th} element in *e.g.* $\mathbf{x}$ may be written as x_i, and similarly for $\mathbf{y}$. For scalar-valued inputs and targets, the vector notation is dropped such that *e.g.* $y \in \mathbb{R}$ denotes a scalar-valued target (*e.g.* a binary label or an energy).

Similarly to functional optimisation, NNs are be trained by minimising some objective, or loss function, L appropriate to the task. Given sets of inputs X and target values Y, a natural regression loss is the mean squared error (MSE)

$$L_{\text{MSE}}(\theta) = \mathbb{E}_{\mathbf{x}\sim X, \mathbf{y}\sim Y}\left[\frac{1}{M}\sum_{i=1}^{M}\left(y_i - p_i(\mathbf{x}\,|\,\theta)\right)^2\right], \tag{4.2}$$

where $\mathbb{E}$ denotes the average over a set of associated inputs and regression targets, θ is the set of network weights, and $p_i(\mathbf{x}\,|\,\theta)$ is the i^{th} NN output prediction for inputs $\mathbf{x}$ given θ. Here, M is the dimension of the target space, which in the case of $M = 1$ reduces $\mathbf{y} = \{y_i\}_{i=1\ldots M} \to y$, following the notation introduced above. This loss minimises the average distance between the NN output prediction and the regression targets in exact correspondence to χ^2-regression with equal uncertainty on the targets.

For classification of inputs $\mathbf{x}$ with associated scalar-valued labels $y \in \{0, 1\} \subset \mathbb{R}$, the sigmoid activation is typically used, see Figure B.1, restricting the NN output to $0 < p < 1$. In this case, the binary cross-entropy (BCE) loss can be used

$$L_{\text{BCE}}(\theta) = \mathbb{E}_{\mathbf{x}\sim X,\, y\sim Y}\left[-y \log p(\mathbf{x}\,|\,\theta) - (1-y)\log\left(1 - p(\mathbf{x}\,|\,\theta)\right)\right], \tag{4.3}$$

where $p(\mathbf{x}\,|\,\theta)$ is again the NN output prediction given network weights θ. The BCE loss, and the multi-class cross-entropy loss more generally, corresponds to the negative log-likelihood of the NN output, which can therefore be interpreted directly as the probability for label y given the input: $p(y\,|\,\mathbf{x}, \theta)$ [116].

To train the NN according to such losses L, so-called back-propagation with some variant of stochastic gradient descent is used. The basic procedure, described in Appendix B, allows for estimating the gradient $\partial L/\partial w_{ij}^{(l)}$ without varying each parameter in the network weight space individually. Given a dataset containing n pairs of inputs and

targets $\{(\mathbf{x}, \mathbf{y})_i\}$, the training can proceed by iterating through each pair, computing the activations by performing the forward pass in Equation (4.1); performing the error back-propagation in Equation (B.2); and updating the network weights in Equation (B.4). The weight gradients are scaled by a parameter, called the learning rate η, to control the speed of convergence and, conversely, to avoid divergence from too large gradient descent steps. One full iteration through the dataset is called an epoch, and in practice each weight update is typically performed by accumulating the gradients in Equation (B.1) over a batch of input-target pairs. Since the NN is trained through gradient descent, there is no guarantee that the loss minimisation will result in a global minimum. However, in practice all local minima typically yield roughly equal performance when evaluated on a dataset not used for training [118].

Techniques

The NNs weights are tuned on so-called training datasets, which are generally assumed to be drawn in a completely random manner from some underlying population. The assumption is that a NN optimised on the training data subset should generalise to the full population. However, considering their potentially vast number of free parameters, NNs are prone to over-fitting, *i.e*. learning features in the training dataset that are not representative of the underlying population. This leads the NN to generalise poorly to data not seen in the training phase. The converse case of under-fitting arises when the NN has insufficient capacity (number of layers, or nodes per layers) to efficiently capture features in the training data, leading to suboptimal performance. For this reason, it is customary to prepare separate, non-overlapping training and testing datasets, such that the performance of the network after training may be evaluated on an independent dataset. This provides an unbiased estimate of the ability of the NN to generalise to unseen data.

However, the testing dataset is ideally held out until final evaluation, and so cannot be used *e.g*. when optimising the NN architecture. In addition, the training dataset may therefore be partitioned into k smaller, non-overlapping datasets. For each of the k partitions, a new NN is randomly initialised and trained on $k - 1$ training partitions and evaluated on the remaining, held-out so-called validation partition. This is done k times, scanning all possible designations of the validation set, such that every example in the training dataset is included in a validation set exactly once. This so-called k-fold cross-validation provides k unbiased measurements of the NN performance, one from each of the k independent validation partitions. A variant of this is stratified k-fold

cross-validation, which can be used for classification tasks, in which each partition is constructed to contain equal proportions of training examples for each target label y.

Finally, a crucial concept for NNs is feature scaling. Before training, the network weights $w_{ij}^{(l)}$ are typically randomly initialised with mean of zero and variance $\lesssim \mathcal{O}(1)$. This means that initially all inputs and activations are given equal importance. This has the implication that if the input features x_i have markedly different scales (*e.g.* a particle energy in MeV compared to its electric charge in units of e), the NN will tend to ignore the features with the smaller scales. Similarly, features with characteristic scales $\gg \mathcal{O}(1)$ may lead to extremely large network weight gradients which in turn may complicate or prohibit a stable convergence of the training. To solve this problem, feature scaling is typically introduced as a data pre-processing step. That is, the mean μ_i and standard deviation σ_i of each input feature x_i are used to scale them as

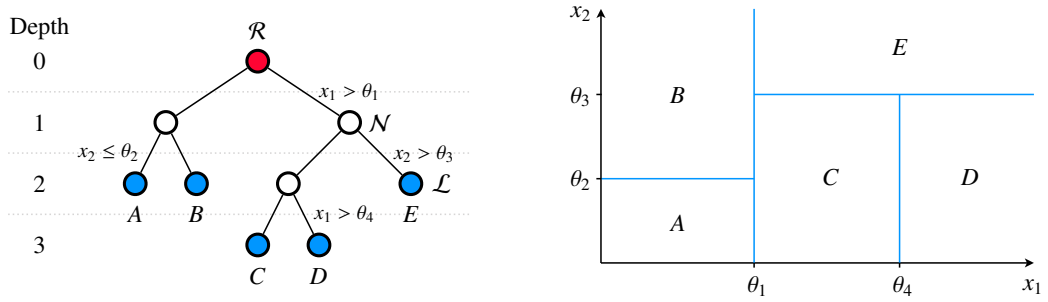
$$\tilde{x}_i = (x_i - \mu_i)/\sigma_i \,. \tag{4.4}$$

This may also be done dynamically for each layer using batch normalisation [119].

In Part III of this thesis, NNs will be used to classify hadronic jets according to their initiating process. To this end, a number of analytically calculated observables will be used as input features with the aim of performing a binary classification of the so-called signal and background processes. These input features are separately considered weak classifiers, and the NN will be tasked with extracting additional information from the collective set. Supervised training of a NN to perform this task requires a large labelled dataset, *i.e.* one for which the label y is known for each set of input features $\mathbf{x}$. Uniquely, this is possible in MC simulated pp collision events in the ATLAS experiment, where the generator-level information allows for unique assignment of a target label to each simulated event. This feature is what makes HEP a remarkably well-suited area for developing and applying ML techniques.

4.2 Boosted decision trees

An alternative ML algorithm is provided by decision trees (DTs) [116, 120]. Similarly to NNs, these map an N-dimensional set of input features $\mathbf{x} = (x_1, x_2, \ldots, x_N)$ to a class probability, in the case of classification, or a functional value, in the case of regression. For concreteness, binary classification will be used as an example, since it will be relevant for Part III of this thesis.



(a) Binary decision tree **(b)** Decision boundaries

Figure 4.2 Example of **(a)** a binary decision tree (DT) for two input features $\mathbf{x} = \{x_1, x_2\}$ with one root node ($\mathcal{R}$), three internal decision nodes ($\mathcal{N}$), and five leaf nodes ($\mathcal{L}$) and **(b)** the associated partitioning of the feature space into five disjoint decision regions A, B, C, D, and E. Figures reproduced from Ref. [116].

The standard classification and regression tree (CART) algorithm [120] constructs a binary DT by sequentially partitioning the input feature space using binary selections on individual features. The DT starts with a single so-called root node $\mathcal{R}$ comprising the entire training dataset. In the simplest case, the task is to perform a single split of the dataset such that the two resulting partitions optimally separate the two classes. Such a split results in two so-called child nodes $\mathcal{N}$ which have the root node as parent and which represent the two partitions of the dataset. To find the optimal split, each of the input features $x_i \in \mathbf{x}$ are scanned and potential splits are evaluated according to some metric for the given task. In the case of classification, a common metric is the Gini impurity, or Gini index [116]. The splitting procedure is detailed in Appendix B.

The CART algorithm is greedy, meaning that each split is performed to locally maximise the class purity after the split. After the first split on the root node, the procedure is recursively applied to each of the child nodes. This sequence of binary splits results in a binary tree where each internal node corresponds to a particular decision, hence the name "decision tree." At each step, the class purity will be improved. The branching proceeds either until every example in the training dataset is perfectly classified, or until some stopping condition is reached. For instance, a maximal depth of the DT may be imposed, or a minimal number of training examples may be required to perform a split. These properties are hyperparameters of the DT, similar to the NN architecture, and require optimising for each practical application. At the end of the procedure, the last nodes in the tree are called leaf nodes $\mathcal{L}$, and each path from the root node to a leaf node is called a branch. An example of a simple binary DT, and the resulting partitioning of the feature space, is show in Figure 4.2.

Using a single DT, the same output prediction is given for all examples within the

feature subspace represented by each leaf node. The predicted value is determined as the fraction of training examples of each class c on the leaf node in the case of classification and as the average value of the target variable in the case of regression. This has the downside of resulting in a discontinuous decision function, which is typically not ideal for HEP use cases where most relations are expected to be continuous: *e.g.* the classification of a physical process, observed in an experiment, is normally not expected to exhibit discontinuous "jumps" as a function of some kinematic variable. Similarly, since individual DTs may result in leaf nodes with 100% sample purity, these are prone to over-fitting.

Boosting

One way to mitigate this problem is through the use of boosting. This is the general concept for combining a set of weak learners to obtain a stronger predictor. In particular, the AdaBoost algorithm [121, 122] described in Appendix B is popular for constructing BDTs, but boosting as a general technique is applicable to other ML algorithms as well.

The AdaBoost algorithm trains a set of DTs in sequence, enumerated by the boosting step t, similar to the NN training epochs. At each boosting step, the weights assigned to training examples, which were misidentified in the previous step, are increased, thereby boosting their relative importance. Based on the weighted fraction of misclassified training examples at boosting step t, a DT weight a^t is calculated. The full set of boosted, weak learners is then combined as the weighted average

$$\mathrm{BDT}(\mathbf{x}) = \sum_t a^t \mathrm{DT}^t(\mathbf{x}). \tag{4.5}$$

Boosting DTs, in this way, yields a robust estimator which is constructed to provide good classification for all training examples; is less prone to over-fitting; and does not suffer from discontinuous decision functions to the same extent as the individual DTs.

The AdaBoost BDT and the standard, densely connected NN will be used as the basis for the ML algorithms studied in Part III of this thesis.

PART II

A search for low-mass leptophobic Dark Matter mediators

CHAPTER 5

Introduction and review

In Chapter 2, the experimental evidence for Dark Matter (DM) was presented along with the motivation for weakly interacting massive particles (WIMPs) as DM candidates. Simplified models were presented as a useful framework for characterising DM searches at hadronic collider experiments in a way which is compatible with direct detection experiments. This part presents a search for leptophobic DM mediators using the ATLAS experiment, described in Chapter 3. This analysis is part of a larger programme of searches for DM already performed in a variety of final states. An overview of previous ATLAS search results is shown in Figure 5.1.

Mono-*X*

The most direct topology is the production of pairs DM particles χ_{DM} through the decay of the mediator boson Z', produced in association with some Standard Model (SM) particle [71, 124–134]. Examples of pair-production of DM particles in association with a photon (γ) or a Z boson and a SM parton are shown in Figure 5.2.

In these processes, the DM mediator is produced in the annihilation of two quarks with coupling g_q and subsequently decays to a pair of DM particles with coupling g_{DM}, typically taken to be equal to 1. Since the DM particles are weakly interacting, they leave no visible signature in the ATLAS detector and therefore escape the experimental apparatus undetected; the only reconstructed physics object of interest in the final state is the initial-state radiation (ISR). The experimental signature for these "mono-X" processes is therefore an isolated, visible object (X) balanced by missing transverse energy ($E_{\mathrm{T}}^{\mathrm{miss}}$) in the opposite azimuthal hemisphere of the detector. The natural search discriminant in this type of final state is the magnitude of the reconstructed $E_{\mathrm{T}}^{\mathrm{miss}}$, since the mass of the DM particles cannot be reconstructed directly. Searches for direct

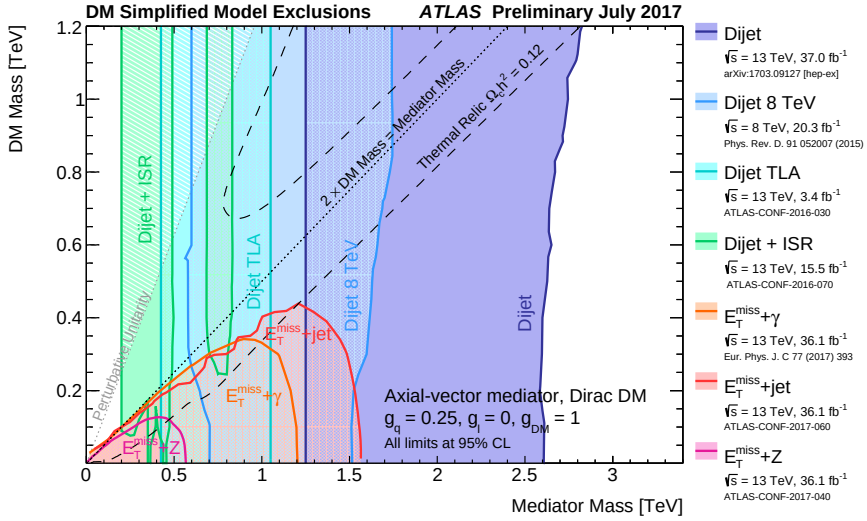
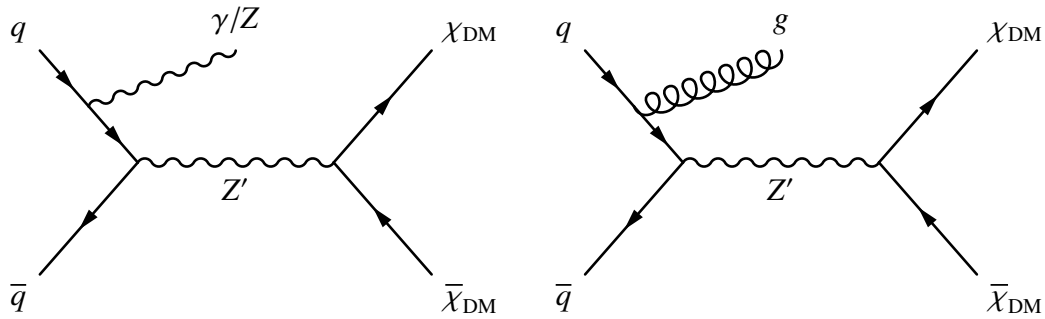


Figure 5.1 Summary plot of ATLAS searches for Dark Matter (DM) with leptophobic mediator particles coupling to Standard Model (SM) quarks with coupling constant $g_q = 0.25$. The coloured contours show 95% confidence level (CL) exclusion regions as a function of the mediator mass $m_{Z'}$ and DM particle mass m_{DM} resulting from analyses of the $E_{\mathrm{T}}^{\mathrm{miss}}$ + jet, $E_{\mathrm{T}}^{\mathrm{miss}}$ + γ/Z, dijet, and dijet + initial-state radiation (ISR) final states as of early 2017. Thermal relic and perturbative unitarity limits are discussed in Chapter 2. Mass configurations between the dashed thermal relic lines yield underproduction relative to the observed energy density $\Omega_{\mathrm{DM}} = 0.26$. Figure from Ref. [123].



(a) Mono-γ/Z signature **(b)** Mono-jet signature

Figure 5.2 Example Feynman diagrams for processes in which a DM mediator particle Z' is produced in association with initial-state radiation (ISR) in the form of **(a)** a photon or Z boson or **(b)** a parton (quark or gluon), giving rise to the so-called "mono-γ/Z" and "mono-jet" experimental signaltures, respectively. Diagrams made using Ref. [135].

production of DM particles require the presence of ISR to be identified by the ATLAS trigger system, since a $Z' \to \chi_{\mathrm{DM}}\overline{\chi}_{\mathrm{DM}}$ decay at rest would not result in any signature in the detector. The presence of ISR makes the invisible Z' decay "visible by inference." Additionally, mono-X searches are kinematically restricted to $2m_{\mathrm{DM}} < m_{Z'}$ for Z' decay to physical DM particles. The most sensitive search of this type, the ATLAS mono-jet search, has excluded axial-vector DM mediators Z' with masses $m_{Z'} < 1.55$ TeV for very light DM particles using an integrated luminosity of 36.1 fb^{-1} collected at a centre-of-mass energy of $\sqrt{s} = 13$ TeV, assuming a quark-flavour universal coupling $g_q = 0.25$ [124], see Figure 5.1. Due to the kinematic limitation of this search, no DM particles with $m_{\mathrm{DM}} > 440$ GeV are excluded. Near the $2m_{\mathrm{DM}} = m_{Z'}$ kinematic threshold, the cross-section for the Z' decay to DM particles is suppressed due to the available kinematic phase space, which is why the upper limit on m_{DM} is considerably lower than half of the upper limit on $m_{Z'}$. In principle, there is no lower limit on the DM particle and mediator masses which can be probed by mono-X searches; in practice, the ATLAS search studied DM particle masses down to 1 GeV and mediator masses down to 10 GeV. At large masses, searches of this type are limited by the energy available in the proton–proton (pp) collisions, as determined by the parton distribution functions (PDFs), to create increasingly massive particles.

Dijet

In simplified models of DM, an alternative to the mono-X signature is afforded by the possibility of searching for DM mediators directly, without the DM particle taking part in the process. In the context of simplified models, provided a non-zero DM coupling g_{DM} in Equation (2.5), the discovery of a mediator particle Z' would imply the existence of the DM particle χ_{DM}, which makes this type of search equally viable. In addition, various other models for physics beyond the Standard Model (BSM) give rise new resonances coupling to SM fermions, which means that DM mediator searches are not only useful within the context of simplified models for DM. In the event of a discovery of a Z'-like particle, measurements of the properties of this particle as well as searches for potentially associated DM particles are necessary to disentangle the theoretical models giving rise to similar experimental signatures.

Provided the Z' boson couples to SM leptons, with coupling g_l, it may be observed cleanly by its decay to e^+e^- or $\mu^+\mu^-$ pairs. However, due to the unambiguous nature and low rate of opposite-sign same-flavour dilepton final states, dilepton spectroscopy [136, 137] along with high-mass dilepton searches [138, 139] have effectively searched the

full range in dilepton invariant masses from below 1 GeV up to 5 TeV for signs of new physics. Searches for DM mediators coupling to leptons are therefore focussed on the rate-limited high mass range, with no inherent technical limitations similar to those faced by hadronic searches as discussed below.

While DM mediator production at pp colliders does not necessarily require a non-zero coupling to SM leptons g_l, it does require a non-zero coupling to SM quarks g_q. The UA1 and UA2 experiments at the CERN Super Proton–Antiproton Synchrotron (Sp$\bar{\text{p}}$S) [140, 141] and CDF and DØ experiments at the Fermilab Tevatron [142, 143] have performed searches for leptophobic resonances ($g_l = 0$). However, there are still unexplored regions of simplified DM model parameter space. For instance, UA2 and CDF have set limits on dijet resonance masses down to $m_{Z'} = 140$ GeV and 200 GeV, for couplings as low as $g_q \approx 0.3$ and 0.2, respectively [73, 144]. These results leave regions of simplified model parameter space, that are self-consistent and consistent with cosmological observations, uncovered, see Figure 5.1. Therefore, leptophobic DM mediators provide a promising model warranting further study.

A process in which the DM mediator is created from the annihilation of a pair of SM quarks, and subsequently decays back to quarks, is shown in Figure 5.3. Since the final state quarks manifest experimentally as hadronic jets, this process gives rise to the so-called "dijet" signature [145–149]. The natural search discriminant in this final state is the invariant mass of the dijet system, peaking at $m_{jj} \approx m_{Z'}$ for the DM process. An ATLAS dijet search in 37 fb^{-1} of $\sqrt{s} = 13$ TeV data set limits on DM mediator masses 1.35 TeV $< m_{Z'} <$ 2.6 TeV for $g_q = 0.25$ [145]. This exclusion range is largely independent of m_{DM} (specifically, for $m_{Z'} < 2m_{\text{DM}}$), illustrating how the dijet topology provides access to regions of the simplified model parameter space which are complementary to those available to mono-X–type searches. At large mediator masses, dijet searches are rate-limited, similarly to the mono-X searches. However, the main limitation at low values of $m_{Z'}$ are the p_{T} thresholds of the relevant jet triggers.

Due to the high rate of QCD processes in pp collisions, p_{T} thresholds of approx. 380 GeV are applied to jets at the trigger level to keep the rate of recorded events within the 1 kHz budget of the ATLAS high-level trigger (HLT) [150]. This means that for reconstructed jets with $p_{\text{T}} > 450$ GeV, standard dijet searches at the Large Hadron Collider (LHC) are limited to mediator masses above approx. 1 TeV. It is possible to lower the jet p_{T} thresholds by performing trigger-level analyses, also called data scouting, where the amount of information recorded for each event is reduced to approx. 0.5% of the size of standard event records [148, 149]. The reduced event size means that events can be stored at twice the standard HLT trigger rate, using only 1%

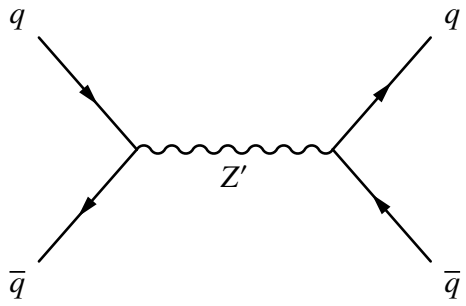
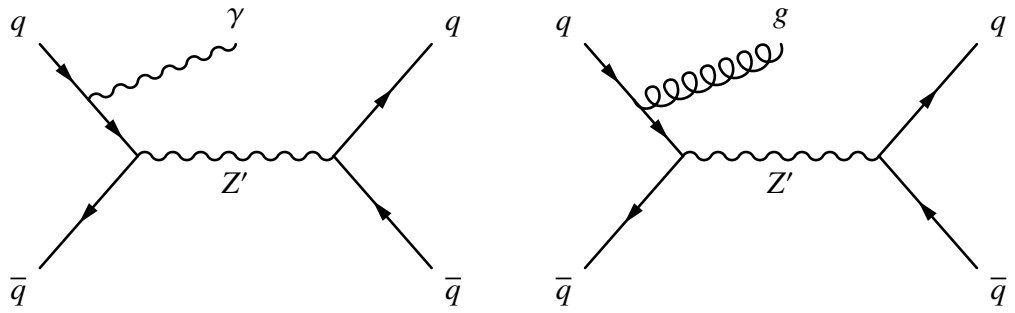


Figure 5.3 Example Feynman diagram for process in which a DM mediator particle Z' is produced and decays back to a pair of quarks in the so-called "dijet" experimental signalture. Diagram made using Ref. [135].

of the total bandwidth [148]. An ATLAS trigger-level dijet analysis used this procedure to lower the search range in $m_{Z'}$ to 450 GeV using 29.3 fb^{-1} of $\sqrt{s}$ = 13 TeV data.

Dijet + ISR

To reach even lower mediator masses, it has been proposed to search for hadronically decaying Z' particles produced in association with initial-state radiation (ISR) [151, 152]. Two such processes are illustrated in Figure 5.4. These are similar to the mono-X processes in Figure 5.2, but with the DM mediator decaying to SM quarks rather than to DM particles. In this final state, the ISR object may be used for triggering as in the mono-X final states, whereby mediator masses below two times the jet trigger p_{T} threshold may be explored. For masses 200 GeV $\lesssim m_{Z'} \lesssim$ 450 GeV, mediators can be reconstructed as two small-radius (small-R) jets ($R \approx 0.4$) recoiling of *e.g.* a photon or another small-R jet. For this reason, these processes are referred to as "dijet + ISR" [153], and the natural search discriminant is still the invariant mass of the dijet system m_{jj}. In these final states, as long as the ISR object has sufficiently large p_{T} to pass the corresponding trigger threshold(s), mediator masses can in be probed down to much lower masses than in the standard dijet topology. In the so-called "resolved" topology, where the two small-R jets from the Z' decay do not overlap in $\eta - \phi$ space, ATLAS has set limits for 200 GeV $< m_{Z'} <$ 950 GeV [154] using 15.5 fb^{-1} of $\sqrt{s}$ = 13 TeV data. However, the angular separation between the quarks in the Z' decay scales roughly linearly with the mediator mass, see Equation (1.5). This means that, for low mediator masses, the small-R jets begin to overlap and can no longer be reconstructed in a resolved dijet + ISR topology. This limits the resolved dijet + ISR topology of search from probing mediator masses below approx. 200 GeV assuming typical ATLAS Run 2 unprescaled trigger thresholds.



(a) "Dijet + ISR γ" signature **(b)** "Dijet + ISR jet" signature

Figure 5.4 Example Feynman diagrams for processes in which a DM mediator particle Z' is produced in association with initial-state radiation (ISR) in the form of **(a)** a photon or **(b)** a parton (quark or gluon). There are referred to as "dijet + ISR" signatures and categorised by the experimental signature of particle off which the mediator particle Z' recoils. Diagrams made using Ref. [135].

Regions of simplified DM model parameter space with leptophobic mediator masses below 200 GeV are therefore poorly explored. However, this region also covers parameter configurations which are consistent with cosmological constraints (yielding $\Omega_{\text{DM}} \leq 0.26$) as well as with perturbative unitarity, see Figure 5.1, as discussed in Chapter 2. This region of model parameter space therefore warrants further study, which is possible using the so-called "boosted dijet + ISR" final state. A natural extension of the resolved regime, the boosted topology reconstructs hadronically decaying mediators as a single large-radius ($R \approx 1.0$) jet. This final state allows for probing DM mediator masses down to $\mathcal{O}(10\text{ GeV})$. The boosted dijet + ISR final state is similarly characterised by the type of recoiling ISR object used for triggering. In this final state, the natural search discriminant is the invariant mass of the large-radius (large-R) jet, and the lower limit on the mediator masses which may be probed is determined mainly by the structure of the large-R jet mass spectrum and the rate of background SM processes. This final state had been explored by the CMS Collaboration [73, 155–157], though not by ATLAS prior to the work in this thesis.

The existing ATLAS search results for leptophobic DM mediators, for combinations of mediator masses $m_{Z'}$ and its coupling to SM quarks g_q, are shown in Figure 5.5 for high-mass DM. This represents the "state-of-the-art" prior to the work in this thesis, and illustrates the need for exploration of the $m_{Z'} < 200$ GeV region.

This second part describes the first ATLAS search for leptophobic DM mediators in the boosted dijet + ISR final state, the first ATLAS search to probe leptophobic DM mediators with masses $\lesssim 200$ GeV for high-mass DM particles. As the first of its kind in ATLAS, this analysis chooses to focus on mediator masses above $m_{W/Z}$. Specifically,

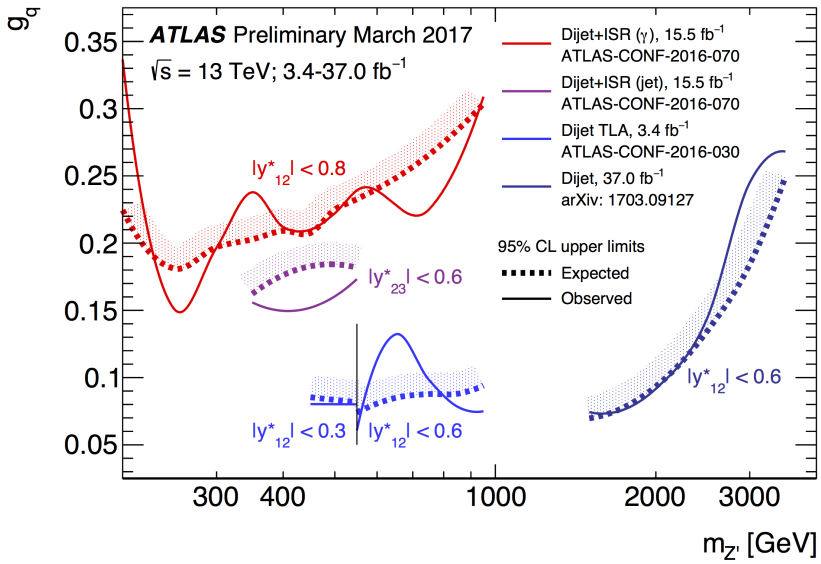


Figure 5.5 Summary plot of ATLAS searches for DM with leptophobic mediator particles coupling to Standard Model (SM) quarks. The coloured lines show 95% confidence level (CL) expected and observed exclusion limits on the coupling of DM mediator particles to SM quarks g_q, as a function of the mediator particle mass $m_{Z'}$, resulting from analyses of the dijet and dijet + ISR experimental signatures as of early 2017. The rapidity difference y^* between the leading jets is typically used to categorise dijet searches. The limits are computed for a DM particle mass of $m_\chi = 10$ TeV, restricting the mediator to decay to SM quarks. Figure from Ref. [123].

the search region is chosen to be 100 GeV $< m <$ 220 GeV to cover a large region of unexplored phase space while allowing for the use of the W/Z large-R jet mass peak as a known signal in data for *in situ* calibration and for validating the analysis strategy. The analysis focuses on two production mechanisms for the leptophobic Z', namely in association with ISR in the form of a photon or a quark or gluon, the latter manifesting as a jet in the detector. In order to target these two production mechanisms, the analysis is carried out in two separate channels, one for each type of ISR object, each with its own trigger and event selection strategy. These are called the ISR γ and ISR jet channels, respectively. This is the first time the ISR γ channel is explored in the boosted regime in either ATLAS or CMS. The ISR γ channel is analysed due to its clean experimental signature, low p_{T} threshold, and its expected sensitivity compared to alternative channels [152]. A similar choice was made in the previous ATLAS resolved dijet + ISR search, where the ISR γ channel was found to be more sensitive than the ISR jet channel [154]. This is due to the potential two-fold (in the boosted regime) ambiguity in ISR jet channel, since the ISR jet may also be reconstructed as the signal candidate large-R jet. However, the increased yield for the signal process in the ISR jet channel, due to the larger cross-section compared to the ISR γ channel, may offset the detrimental effect of this ambiguity. Results are presented with a focus on the ISR γ channel, with results from the ISR jet channel taken from Refs. [1, 158].

The recorded data and Monte Carlo (MC) simulated datasets used in the analysis are described in Chapter 6. Chapter 7 describes the reconstruction of physics objects used in the analysis. In Chapter 8, the event selection requirements used to define the final dataset used in each channel are described. The procedure for estimating the rate of the dominant background process in a data-driven fashion is detailed in Chapter 9. Finally, the search results obtained in this analysis are presented in Chapter 10.

CHAPTER 6

Datasets

Searches for leptophobic DM mediator bosons Z' in the boosted dijet + ISR final states were motivated in Chapter 5. This analysis performs a search for this hypothesised process (the "signal") in a dataset collected by the ATLAS experiment, see Chapter 3. In addition, simulated datasets are produced for the dominant SM processes contributing to the targeted experimental final state (the "backgrounds"), as well as for the benchmark signal process, using MC event generators. Using these simulated datasets, the analysis will assess the compatibility of the experimentally recorded data yields with the rates from known SM processes, as well as from potential BSM processes involving leptophobic Z' bosons. Below, the experimentally recorded, and MC simulated, datasets used in the analysis are described.

6.1 Experimental data

This analysis is performed with a sample of pp collision events collected by the ATLAS experiment in 2015 and 2016 at a centre-of-mass energy of $\sqrt{s} = 13$ TeV, corresponding to an integrated luminosity of 36.1 fb^{-1}. This data taking period was chosen since the ATLAS trigger menu, pile-up conditions, *etc*. were consistent across these two years, thereby simplifying the analysis.

The data sample is selected using a set of triggers suitable to the final states of the signal processes depicted in Figure 5.4. These triggers record events containing at least one isolated photon or at least one high-p_{T} small-R jet. In both channels, the p_{T} and E_{T} requirements for trigger-level objects is chosen to be the lowest value for which all triggered event are recorded for offline analysis within the data transfer constraints of the ATLAS trigger system for the entire 2015-2016 data taking period [106, 159]. The

HLT_g140_loose single-photon trigger chain, used in the ISR γ channel, is seeded at the L1 by an isolated energy deposit in the electromagnetic calorimeter (ECAL) with a transverse energy greater than $E_{\mathrm{T}} > 22$ GeV, constituting a region of interest (ROI) for further selection [150]. At the subsequent HLT stage, this ROI is further analysed to identify photons with $E_{\mathrm{T}} > 140$ GeV based on loose identification criteria using a multivariate likelihood technique [160]. The HLT_j380 single-jet trigger chain, used in the ISR jet channel, is seeded at the L1 by a jet element comprising 2×2 trigger towers with $\sum E_{\mathrm{T}} > 100$ GeV which is passed as an ROI to the HLT, where an anti-k_t jet with a radius parameter of $R = 0.4$ and $E_{\mathrm{T}} > 380$ GeV is required. The minimum p_{T} requirements on the reconstructed physics objects during offline analysis of the recorded data passing the above triggers are chosen to be 155 GeV for photons and 420 GeV for small-R jets, at which points the respective triggers are fully efficient. Only events satisfying beam, detector, and data quality requirements are retained for analysis [161]. The average number of simultaneous pp collisions per bunch crossing (pile-up) was $\langle\mu\rangle = 13.4$ in 2015 and 25.1 in 2016.

6.2 Simulated datasets

To estimate the rate of known background and hypothesised signal processes, MC generators are used to produce simulated datasets. These can be used to compare the recorded data distributions to the expectation from SM processes, and to test the compatibility of data with the hypothesised signal process.

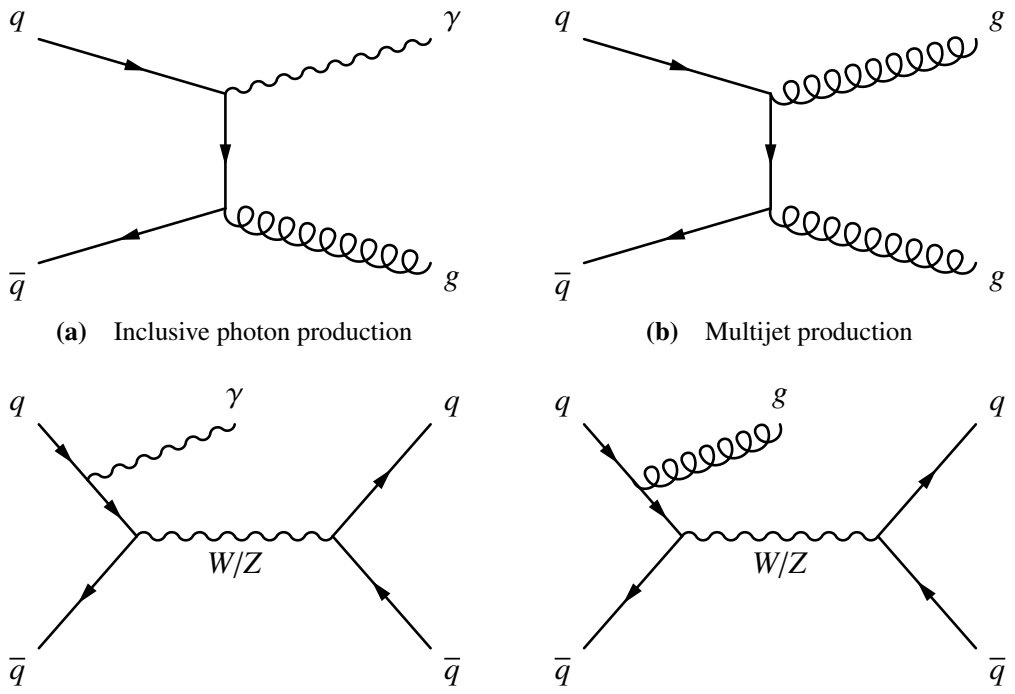
A simplified model of DM, see Chapter 2, is used to generate benchmark simulated datasets for the signal processes involving a leptophobic Z' particle [65, 144, 162], shown in Figure 5.4. In these benchmark models, the Z' has axial-vector, flavour-universal coupling to SM quarks, see Chapter 2, taken to be $g_q = 0.5$. Simulated datasets are generated for five different mass hypotheses, $m_{Z'} = 100, 130, 160, 190,$ and 220 GeV, to span the Z' mass range targeted by this analysis, see Chapter 5. In all cases, the mass of the DM particle is set to be $m_{\mathrm{DM}} = 10$ TeV to force a decay to SM particles. This large value is a common benchmark in ATLAS DM searches but is somewhat arbitrary in the context of the present analysis, where any value $m_{\mathrm{DM}} \gg 220\ \mathrm{GeV}/2$ would suffice, since these all kinematically prohibit the decay of the mediator to a pair of DM particles. Signal samples are generated using the next-to-leading order (NLO) MadGraph5_aMC@NLO generator [163] with the NNPDF2.3 PDF set [164]. Pythia 8.186 [165], with the ATLAS A14 set of tuned parameters [166], is used to perform the parton showering and to simulate multiple parton interactions.

Z' mass $m_{Z'}$ [GeV]	Kinematic filtering Min. $p_{\mathrm{T}}^{\mathrm{ISR}}$ [GeV]	Cross-section $\sigma(p_{\mathrm{T}}^{\mathrm{ISR}} > p_{\mathrm{T}}^{\mathrm{ISR,min}})$ [pb]
$Z'(\rightarrow q\bar{q}) + \gamma$		
100	150	6.74
130	195	5.86
160	240	5.04
190	285	4.32
220	330	3.70
$Z'(\rightarrow q\bar{q})$ + jet		
100		42.3
130		36.3
160	350	32.7
190		30.3
220		28.3

Table 6.1 Summary of the model parameters, generator-level kinematic filtering, and resulting cross-sections of the simulated signal processed used in the analysis, for the (*top*) ISR γ and (*bottom*) ISR jet channel, respectively.

Separate signal samples are generated for the ISR γ and ISR jet channels, with p_{T} requirements on the ISR object imposed at the generator level to efficiently populate the kinematic phase space relevant to the analysis. Specifically, the mass-dependent p_{T} filtering at the MC generator-level in the ISR γ channel ensures a sufficient number of MC simulated signal events passing the so-called boosted topology selection introduced in Chapter 8. In the ISR γ channel, an $m_{Z'}$-dependent filtering of the ISR photon $p_{\mathrm{T}} > 3m_{Z'}/2$ is used, whereas a fixed requirement on the recoil jet $p_{\mathrm{T}} > 350$ GeV is required in the ISR jet channel. The specific filtering in the ISR γ channel is chosen based on the fact that the boosted topology selection $p_{\mathrm{T}} > 2m$ is applied to signal candidate large-R jets in the analysis. Applying a similar but looser p_{T} filtering at the MC generator level ensures that sufficient MC signal events will pass the selection and populate the search region. The generator-level photon p_{T} filter is chosen such that it will always be sufficiently looser than the effective offline large-R jet p_{T} requirement, and will therefore not affect the final selection. In the ISR jet channel, where the effect of the boosted selection is minimal, due to the higher minimum large-R jet p_{T}, a similar filtering is not necessary. A summary of the simulated signal datasets is provided in Table 6.1.

The dominant SM processes producing similar final states to the benchmark signal model are inclusive photon production and continuum multijet production, with sub-dominant contributions from hadronically decaying W/Z boson produced in association



(a) Inclusive photon production **(b)** Multijet production

(c) $W/Z \to q\bar{q}$ decay in assoc. with photon **(d)** $W/Z \to q\bar{q}$ decay in assoc. with "jet"

Figure 6.1 Example Feynman diagrams for the **(a, b)** dominant and **(c, d)** sub-dominant background processes which may be reconstructed as a large-R jet recoiling off **(a, c)** a photon or **(b, d)** a parton reconstructed as a small-R jet, thereby resulting in the same final state signature as the targeted Z' signal processes. Diagrams made using Ref. [135].

with a photon or a jet. Tree-level diagrams illustrating examples of these processes are shown in Figure 6.1. Background processes involving the production of top quarks were considered but the contribution to the final event yield was found to be negligible [158]. The sub-dominant background W/Z processes are irreducible in cases where $m_{Z'} \approx m_{W/Z}$, as they lead to the same final state as the signal process, where the large-R jet reconstructs a hadronic two-body decay. By contrast, the dominant background processes are somewhat reducible, since they are characterised by a single hard parton emission to leading order (LO).

All background processes are simulated using the Sherpa 2.1.1 generator [167]. The inclusive photon and multijet samples all are generated at LO with up to three final state partons in the matrix element. The inclusive W/Z samples are generated at LO with three additional partons in the ISR γ channel and four additional partons in the ISR jet channel, and are corrected with an NLO k-factor to account for higher-order effects [158]. The CT10 PDF set is used [168] and parton showering is performed in Sherpa [169] using the ME+PS@LO prescription [170]. The dominant background

samples are generated in bins of the leading photon p_{T} in the ISR γ channel and of the leading jet p_{T} in the ISR jet channel to cover a broad kinematic range; the inclusive W/Z samples are generated in bins of the W/Z boson p_{T}. While the simulated inclusive W/Z samples are used in the final search, the MC simulated samples for the dominant backgrounds are only used for optimisation of the analysis selection criteria and for validation studies. A data-driven method is used instead to estimate the dominant background contribution in both channels, as detailed in Chapter 9.

Finally, all MC simulated samples are overlaid with additional pile-up events and the response of the ATLAS detector to the outgoing particles of the composite events is modelled using a full simulation of the detector [171] implemented in GEANT4 [172]. Pile-up events are simulated as minimum bias interactions in PYTHIA 8.186 [165] using the A2 tune [173] with the MSTW2008LO PDF set [174]. The simulated pile-up events are reweighted to yield the same number of expected of pp collisions per bunch crossing as in data.

CHAPTER 7

Reconstruction of physics objects

The recorded data studied in this analysis, and the MC simulated datasets used to estimate the expected contributions from SM processes as well as the signal DM process, were summarised in Chapter 6. To search for experimental evidence of the leptophobic Z' mediator in this data, the necessary physics objects have to be reconstructed from detector-level information. Physics objects are the reconstructed, experimental proxies for the particles produced in the pp collisions inside the ATLAS detector. In this analysis, the physics objects of interest are photons and small-R jets as the ISR objects, as well as large-R jets as candidate proxies for the hadronic decay of the Z' mediator. This chapter describes the reconstruction, calibration, and identification of these physics objects in the ATLAS experiment.

7.1 Photons

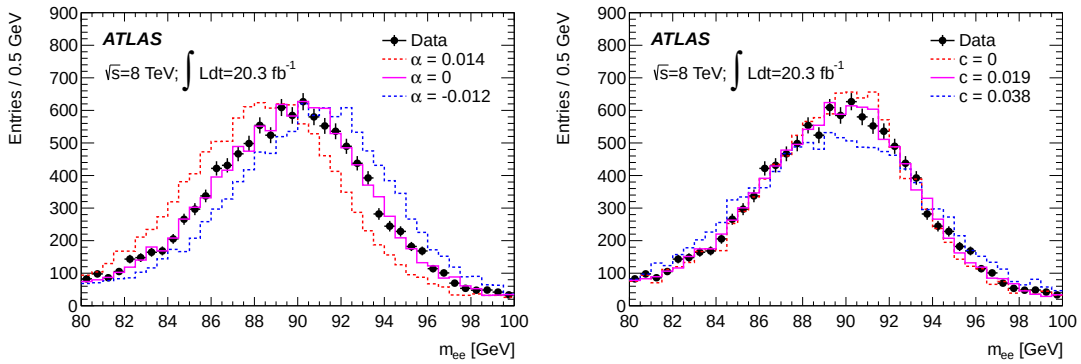
Photon candidates are reconstructed using clusters of energy deposited in the ATLAS ECAL and are required to have $|\eta| < 2.37$, excluding the transition region $1.37 < |\eta| < 1.52$ between the ECAL barrel and end-cap [95], see Figure 3.6. In the transition region, the ECAL is not sufficiently deep to contain the electromagnetic shower, the ECAL cell segmentation is not fine enough to allow for the separation of single photon showers from the double photon showers arising from the decay of neutral hadrons in the ECAL, and the amount of passive material before the ECAL degrades the energy resolution. Additionally, the lack of a tracking detector outside $|\eta| > 2.5$ means that it is not possible to use track information to distinguish between electron and photon candidates in this region. The ATLAS electron and photon reconstruction is based on clusters of energy deposits in the calorimeter cells, constructed with fixed size in the $\eta - \phi$ plane and spanning all layers of the ATLAS ECAL. The cluster used to

reconstruct photons in the ECAL barrel (end-cap) has a fixed size of 3×7 (5×5) cells in the middle layer, corresponding to an area in the pseudorapidity-azimuth plane of $\Delta\eta \times \Delta\phi = 0.075 \times 0.175$ (0.127×0.125) [160]. An ECAL cluster is classified as a photon candidate either if the cluster is not matched to any inner detector (ID) tracks in $\eta - \phi$ (so-called "unconverted photons"), or if it is associated with two ID tracks with opposite charges that are collinear at the production vertex and are both compatible with electron hypotheses in the transition radiation tracker (TRT) (so-called "converted photons," having pair-produced two opposite-charged electrons, e^- and e^+, in the ID volume).

The energy of the photon candidate is calibrated to account for lateral energy leakage into neighbouring cells outside the fixed cluster size; longitudinal energy leakage beyond the ECAL; and energy losses in the passive material upstream of the ECAL [96]. First, the photon candidate energy is corrected using calibration constants in bins of η and p_{T}, derived in MC simulated datasets using a boosted decision tree (BDT) algorithm taking a set of detector-level observables as input variables. These include the total calorimeter energy deposit and ratio of energy deposited in the first two layers of the ECAL. As this calibration relies on MC simulated datasets, it is highly dependent on the accurate modelling of the interaction of the photon with matter in the detector as well as on the detector geometry itself, including passive material. Second, additional data-driven corrections to the energy of the photon candidate and its relative resolution are applied using recorded $Z \to e^+e^-$ events [160]. Here, a so-called template method is used, similar to that shown in Figure 7.1. A χ^2-fit of MC simulation to data is performed in the distribution of the invariant mass of the two reconstructed electrons in the region around the mass of the Z boson. The residual energy mis-calibration and the difference in energy resolution between MC and data are parametrised as corrections to the electrons and photons in MC as

$$E_i^{\mathrm{data}} = E_i^{\mathrm{MC}}(1 + \alpha_i) \qquad \text{and} \qquad \left(\frac{\sigma(E)}{E}\right)_i^{\mathrm{data}} = \left(\frac{\sigma(E)}{E}\right)_i^{\mathrm{MC}} \oplus c_i\,, \tag{7.1}$$

respectively, where i enumerates bins in pseudorapidity, α_i and c_i are correction factors to be optimised, $\sigma(\,\cdot\,)$ is the absolute resolution, and '$\oplus$' denotes a sum in quadrature. By varying these correction factors, the MC-data agreement can be improved by minimising the χ^2 between the data and MC template distributions, as in Figure 7.1. These *in situ* corrections, which are at the level of 1%, are taken to be the combination of correction factors α_i and c_i in each pseudorapidity bin i which minimise the χ^2-fit to data. These residual data-MC differences are common to electrons and photons, which is why the correction factors derived in $Z \to e^+e^-$ are applied to photons as well. This assumption



(a) *In situ* e/γ energy scale corrections **(b)** *In situ* e/γ energy resolution corrections

Figure 7.1 *In situ* Monte Carlo (MC) template fit to the m_{ee} distribution in recorded data in the region around the Z mass. The MC templates are varied by parameters characterising changes in **(a)** the e/γ energy scale through the correction factor α and **(b)** the e/γ energy resolution through the correction factor c, respectively, to minimise the χ^2-fit to the data distribution. Figures from Ref. [96].

is validated in a separate photon-enriched $Z \to \ell^+\ell^-\gamma$ dataset [96]. The photon energy resolution in the kinematic range relevant to this analysis is approx. 1% [96] and the energy scale is accurate to $O(1‰)$ [160]

To identify prompt photons, *i.e.* ones produced in the hard scatter interaction, the photon candidates in this analysis are required to pass tight identification criteria in order to reject the dominant backgrounds of photons produced in hadron decays or hadrons being misidentified as photons [175]. The tight photon identification definition in ATLAS uses selections on a set of discriminating variables which characterise the shape of the calorimeter shower in the ECAL strip and middle layers as well as the amount of longitudinal leakage into the ATLAS hadronic calorimeter (HCAL). These tight identification criteria have a photon selection efficiency of approx. 95% in the kinematic range considered in this analysis.

The photon candidates in this analysis are also required to pass tight isolation requirements. This is intended to further reject the dominant backgrounds, with photons produced in hadronic decays, which are typically associated with substantial additional activity in the ECAL [175]. Therefore, the amount of energy in the ATLAS calorimeters in a cone of size $\Delta R = \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2} = 0.4$ around the barycenter of the energy in the photon candidate cluster which must be less than 2.45 GeV + $0.022 \times p_{\mathrm{T}}$, excluding the energy associated with the cluster itself, where p_{T} is the transverse momentum of the photon candidate [95]. This cone energy is corrected for leakage of the photon energy outside the cluster using correction coefficients derived from simulation, as

well as for contributions from pile-up activity determined on an event-by-event basis using an estimate of the average pile-up energy density [95, 176, 177]. On top of the tight identification, the tight isolation criteria has a photon selection efficiency which is greater than 90% in the kinematic region of interest to this analysis [178].

Finally, the photon candidate is rejected if it is deemed to arise from instrumental problems or non-collisions backgrounds, based on a set of quality criteria including liquid-argon (LAr) calorimeter noise bursts, masked calorimeter cells, and out-of-time calorimeter clusters [179].

7.2 Jets

The leptophobic Z' mediator in this analysis decays to a pair of SM quarks which, along with the ISR quark or gluon in Figure 5.4b, will hadronise due to colour-confinement, resulting in collimated hadronic jets as introduced in Section 1.3.

These hadronic jets manifest as showers of energy deposits in the cells of the ATLAS calorimeter system. To suppress the energy-equivalent noise from electronics ($\sigma_{\text{noise}}^{\text{electronic}}$) and from the average expected background from pile-up activity ($\sigma_{\text{noise}}^{\text{pile-up}}$), the calorimeter cells are grouped to form topological clusters using the iterative "4-2-0" procedure [180], the numbers referring to sequential noise thresholds. The topological clustering algorithm is seeded from calorimeter cells with an energy $|E_{\text{cell}}| > 4 \times \sigma_{\text{noise}}$, where $\sigma_{\text{noise}} = \sigma_{\text{noise}}^{\text{electronic}} \oplus \sigma_{\text{noise}}^{\text{pile-up}}$ is the total nominal energy-equivalent noise expected in the cell, found as the sum in quadrature of the two terms. Adjacent cells with an energy $|E_{\text{cell}}| > 2 \times \sigma_{\text{noise}}$ are iteratively included in the topological cluster. If an adjacent cell belongs to a different cluster, the two clusters are merged. Finally, all bounding cells with $|E_{\text{cell}}| > 0$ ($\times \sigma_{\text{noise}}$) are included in the cluster and the clustering algorithm stops iteration. Here, the zero-multiplication is customarily included to be consistent with the naming of the "4-2-0" procedure. Four-momenta are constructed from these topological clusters, with energy given by the cluster energy calibrated using the local hadronic cell-weighting (LCW) procedure; pointing in $\eta - \phi$ given by the cluster coordinates; and zero mass [180]. The LCW calibration is a multi-stage, MC-based procedure intended to provide cluster-by-cluster energy reconstruction, correcting for the non-compensating nature of the ATLAS calorimeters; out-of-cluster energy losses due to the noise-suppressing "4-2-0" topological clustering scheme; and energy lost in inactive material in and around the calorimeters.

Small-*R* jets

The ISR quark or gluon candidate in the ISR jet channel is reconstructed from topological clusters using the anti-k_t algorithm [30] as implemented in FastJet [181] with a radius parameter of $R = 0.4$, as introduced in Section 1.3. All such small-radius (small-R) jets with $|\eta| < 2.4$ and $p_{\mathrm{T}} > 20$ GeV are selected initially.

The small-R jets are constructed as the four-momentum sum of the topological cluster constituents found during the sequential recombination in the anti-k_t algorithm. These jets are then calibrated in a sequential, multi-step correction procedure [182], performing pointing correction of the jet to the primary vertex in the event, see Chapter 8; jet area-based and residual pile-up corrections; MC-based correction of the jet energy scale and η calibration to the particle level; a global sequential calibration (GSC) of five jet shape observables which are found to capture residual non-uniformities in the jet p_{T} response; and finally four residual *in-situ* calibration methods are applied to account for differences between data and MC simulation. The *in situ* calibrations rely on balancing the p_{T} of the jet against other well-calibrated physics objects, and are performed in dijet (η-intercalibration), Z + jets, γ + jets, and multijet events. The jet energy scale is accurate to $1 - 2\%$ [182], with a jet p_{T} resolution of approx. 5% [183].

In order to reject jets originating from pile-up interactions, small-R jets with $p_{\mathrm{T}} < 60$ GeV are required to originate from the primary vertex, as determined by a jet vertex tagger [23]. This tagger uses information about the ID tracks associated with a given small-R jet, since the precise tracking information can be used to accurately determine whether these are likely to have been produced at the primary vertex. For jets with $p_{\mathrm{T}} > 60$ GeV, no jet vertex tagging requirement is imposed, due to the lower probability for spurious jets to be produced with such large transverse momenta by other pp collisions in the same bunch crossing; so-called pile-up jets.

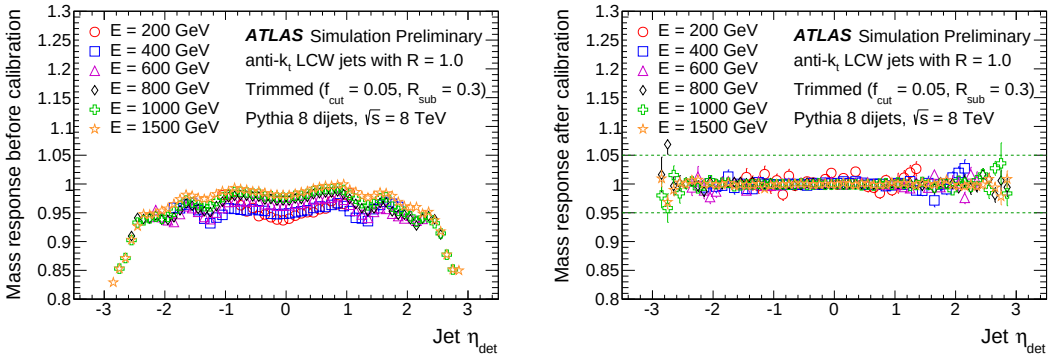
Finally, the small-R jet candidates are subjected to the "loose" set of quality requirements [184], intended to reject jets misreconstructed from calorimeter noise or non-collision backgrounds such as beam-induced background and cosmic ray muon showers. If any small-R jet with $p_{\mathrm{T}} > 20$ GeV fail any of these quality requirements, the event is discarded.

Large-*R* jets

The Z' mediator candidates are similarly reconstructed from topological clusters as single jets using the anti-k_t algorithm [30] as implemented in FastJet [181] with a distance parameter of $R = 1.0$. These are referred to as large-R jets, and are required to have $|\eta| < 2.0$.

The signal process results in the decay of the Z' mediator into a collimated pair of quarks the invariant mass of which, when reconstructed as a large-R jet, provides a way to infer the mass of the Z'. Additional soft radiation, either from the underlying event (UE) or from pile-up, will degrade the resolution of the reconstructed mass and bias the invariant mass of the large-R jet towards larger values. To mitigate this problem, the reconstructed large-R jets are groomed using the jet trimming algorithm [36], see Section 1.3. In this analysis, the constituents of the large-R jets are reclustered using the k_t algorithm with a radius parameter of $R_{\text{sub}} = 0.2$ to yield a collection of subjets. Subjets carrying less than $f_{\text{cut}} = 5\%$ of the total large-R jet momentum are considered to be due to soft radiation and are therefore discarded. The constituents of the subjets that are not discarded in the trimming procedure are taken to constitute the trimmed large-R jet. Jet observables, including the invariant mass, are then computed on this reduced set of topological clusters.

The large-R jets are calibrated in a two-step procedure that first corrects the jet energy scale and the jet pseudorapidity η, and then the jet mass scale [33, 182, 183]. This calibration is centrally provided for jets with $p_{\text{T}} > 200$ GeV, which is the region in which large-R jet performance is robust and well-understood [158]. The procedure is based on a comparison of isolated large-R jets in inclusive jet events found in MC simulated dataset, after the application of the trimming procedure. The jet energy scale and η calibrations are similar to those for small-R jets: First, calibration factors correcting the jet energy response $\mathcal{R}_E = E^{\text{reco}}/E^{\text{truth}}$ of the reconstructed jet energy E^{reco} to the particle level E^{truth} in simulated data are computed in bins of the particle-level large-R jet energy and reconstructed jet pseudorapidity η^{det} relative to the geometric center of the detector. Additionally, the large-R jet η is corrected for biases as a function of the η^{det}, due to differences in response in poorly instrumented regions of the detectors, primarily the transition regions [185]. Finally, the jet mass is susceptible to soft, wide-angle radiation which is not corrected for by the jet energy scale calibration. Therefore, a dedicated jet mass calibration is required. This procedure resembles the jet energy scale calibration, with the jet mass response $\mathcal{R}_m = m^{\text{reco}}/m^{\text{truth}}$ computed in bins of the particle-level jet p_{T} and mass as well as η^{det}, and used to correct the jet mass to



(a) Before jet mass calibration **(b)** After jet mass calibration

Figure 7.2 Large-R jet mass response as a function of the jet η^{det} **(a)** before and **(b)** after the application of the MC-based jet mass calibration. Figures from Ref. [183].

the particle level [33]. The result of the jet mass calibration is shown in Figure 7.2. In both cases, numerical inversion is used to express the calibration factors in terms of the reconstructed quantities before calibration so as to avoid any dependence on particle-level quantities [33]. The large-R jet mass resolution is approx. 10% in the kinematic region of interest to this analysis [33].

Large-R jet substructure

The ISR photons and jets are used exclusively for triggering, and the analysis therefore centres on the reconstruction of the Z' candidates as large-R jets. The dominant backgrounds in both search channels are characterised by the non-resonant emission of a single parton, in contrast to the signal process, which results in the hadronic two-body decay of a high-mass particle, see Figures 5.4, 6.1a, and 6.1b. Therefore, the invariant mass of the trimmed large-R jet, see Section 1.3, is the most characteristic variable distinguishing the signal from the dominant background processes, which is why it is used as the search discriminant. In addition, jet substructure observables that characterise the structure of the hadronic activity inside the jet may help reduce the dominant multijet and inclusive γ backgrounds, as explained in Section 1.3. In this analysis, the N-subjettiness ratio $\tau_{21} = \tau_2/\tau_1$ is used to distinguish jets from the hadronic decay of Z' to two quarks (so-called "two-prong" jets) from non-resonant jets characterised by a single hard parton emission (so-called "one-prong" jets) like those in the inclusive photon and multijet background processes. This variable was described in detail in Chapter 1. In practice, ATLAS employs a 'winner-takes-all' approach [186], where the direction of hardest constituent within each subjet is used in the calculation

of ΔR_{ji} in Equation (1.7) rather than the k_t subjet axis. The separation of large-R jets based on their initiating process will be used in this analysis to create a signal-enhanced search region.

CHAPTER 8

Event selection

The procedures for reconstructing and calibrating the physics objects relevant to this analysis were described in Chapter 7. A set of requirements *e.g.* on their kinematics are applied to define a signal-enhanced search region. The sequence of these requirements is referred to as the event selection. The event selection defining the analysis is presented below, and is applied equally to recorded and MC simulated datasets.

8.1 Basic selection

Events in the ISR γ and ISR jet channels are selected based on the single-photon and single-jet triggers, respectively, introduced in Chapter 6. These events are first required to contain a primary vertex, which is reconstructed from the set of ID tracks in the event. Vertices are reconstructed from at least two tracks with $p_{\mathrm{T}} > 400$ MeV [187]. The primary vertex is taken to be the vertex in the event with the largest $\sum p_{\mathrm{T}}^2$ over associated tracks.

The reconstructed photon candidates are required to have $p_{\mathrm{T}} > 155$ GeV, at which point the single-photon trigger is fully efficient. Similarly, the small-R jet candidates are required to have $p_{\mathrm{T}} > 420$ GeV. Each event is required to contain at least one ISR object candidate appropriate to the channel.

The signal candidate large-R jets are required to have a $p_{\mathrm{T}} > 200$ GeV in the ISR γ channel and $p_{\mathrm{T}} > 450$ GeV in the ISR jet channel. For the ISR jet channel, the chosen threshold guarantees full trigger efficiency. In the ISR γ channel, the p_{T} threshold is additionally restricted to the region for which ATLAS large-R calibrations described in Section 7.2 are valid [183].

In the ISR jet channel, the recoiling small-R jet may also be reconstructed as a large-R jet with sufficient p_{T} to be considered a signal candidate. This potential ambiguity is resolved by choosing the large-R, which is most consistent with a two-body decay hypothesis, as the signal candidate. This determination is made using the τ_{21}^{DDT} jet substructure observable, introduced below.

In both channels, the chosen signal candidate large-R jet is required to be in the opposite hemisphere of the detector relative to the ISR object, *i.e.* have a separation in the azimuthal angle of at least $\pi/2$. This is done to avoid overlap between the physics objects in the $\eta - \phi$ plane, and to primarily select events consistent with the back-to-back topology expected from the signal processes in Figure 5.4.

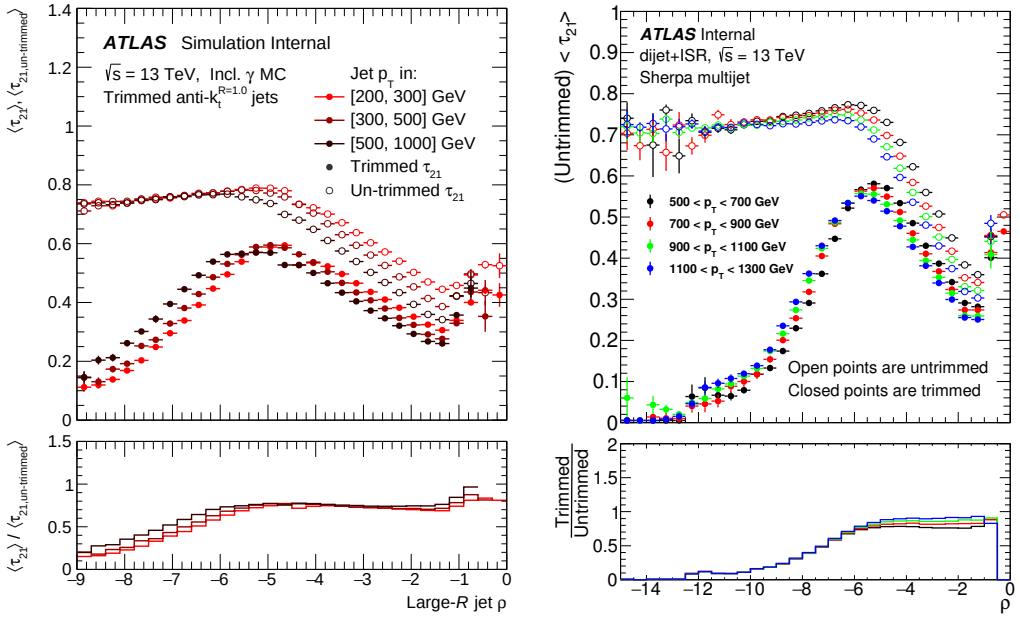
8.2 Substructure decorrelation

To increase sensitivity to new physics, this analysis uses the τ_{21} observable, introduced in Section 7.2, as the basis for selecting events consistent with hadronic Z' decays while reducing the rate of the background processes. However, it has been observed that jet substructure observables exhibit non-trivial correlations with the invariant mass of the large-R jets [188]. This means that a threshold selection on such an observable will introduce morphological changes in the large-R jet mass spectrum for the background processes. Since this analysis aims at performing a resonance search in the large-R jet mass spectrum, such sculpting effects complicate the determination of the leading background process contributions. This is because the background estimation procedure used in this analysis, described in Chapter 9, requires the average jet substructure observable to be independent of the jet mass and p_{T}.

To remove the mean dependence of τ_{21} on the jet mass and p_{T} in the kinematic region of interest, the designed decorrelated taggers (DDT) method [189] is used in this analysis. The dimensionless ρ variable [190], defined as

$$\rho = \log\left(\frac{m^2}{p_{\mathrm{T}}^2 R^2}\right) \xrightarrow[R=1.0]{} \log\left(\frac{m^2}{p_{\mathrm{T}}^2}\right), \tag{8.1}$$

provides a convenient means of simultaneously studying the correlation of τ_{21} with the jet mass m and transverse momentum p_{T}. Here, R is the radius parameter of the algorithm used to reconstruct the large-R jet, which is $R = 1.0$ in this analysis. Figure 8.1 shows the dependence of the mean value of τ_{21} on the large-R jet ρ after the above selection using simulated datasets for the leading background process in each channel.



(a) ISR γ channel **(b)** ISR jet channel

Figure 8.1 Mean value of τ_{21} as a function of the large-R jet ρ, in bins of large-R jet transverse momentum (p_{T}), in the **(a)** ISR γ and **(b)** ISR jet channel, with and without the application of jet trimming. Bottom panels show ratio of trimmed to untrimmed profiles. Figure **(b)** from Ref. [158].

Profiles are shown with τ_{21} calculated on the jet constituents with and without the application of jet trimming.

In both channels the τ_{21} profile is characterised by three distinct kinematic regions, separated by two "kinks" at $\rho \approx -5$ and -1.5. For a large-R p_{T} of 200 GeV, which is the minimum in ISR γ channel, the ρ values of these kinks correspond to large-R jet masses of approx. 15 GeV and 100 GeV, respectively, see Equation (8.1). This behaviour is consistent with Refs. [155, 189].

In the central region, $-5 \lesssim \rho \lesssim -1.5$, τ_{21} decreases linearly with ρ, *i.e.* becoming more signal-like, since jets that are consistent with a two-subjet hypothesis have lower values of τ_{21}, see Chapter 1. Assuming that, to first order, the masses of the background jets are dominated by a single energetic emission from the hard scatter parton, the approximation in Equation (1.5) implies that $\rho \sim \log R_{12}$, where R_{12} is the angular separation of the two leading partons. In this simplified picture, ρ can be considered a proxy for the angular separation of the two leading subjets as considered by the N-subjettiness variable. This picture is particular attractive for trimmed jets, which attempt to remove soft activity and leave the hard scatter components of the jet intact. As the subjet separation increases, the jet appears more two-prong–like, resulting in a lower

value of τ_{21} for increasing ρ; see also the discussion of N-subjettiness in Chapter 1. This effect results in a τ_{21} profile which happens to be linear with ρ, allowing for a simple linear correction.

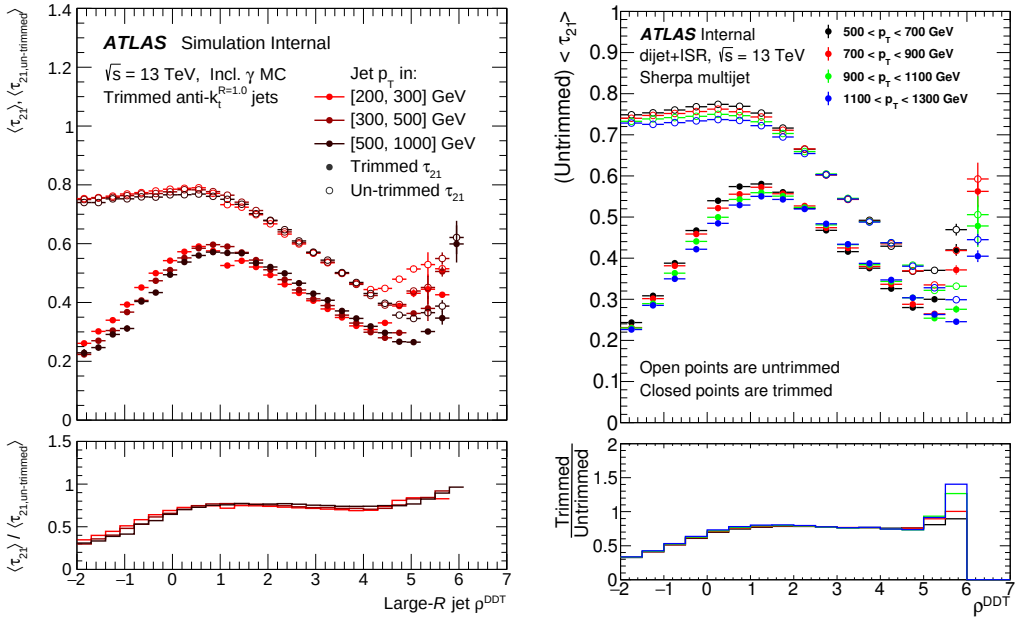
Jets in the low-mass region, $\rho \lesssim -5$, are not important for this analysis as they are far from the chosen search region $m_{Z'} \in [100, 220]$ GeV. Specifically, in the ISR γ channel, with a minimal large-R jet p_{T} of 200 GeV, the $\rho \lesssim -5$ region corresponds to jet masses $m \lesssim 15$ GeV. Here, the τ_{21} profile is also roughly linear, but the specific behaviour depends on whether jet trimming is applied. In this region, the emissions from the hard scatter parton in the dominant background processes are too soft/collinear to generate substantial mass, meaning that QCD is non-perturbative. The fact that soft QCD effects dominate in this region is substantiated by the large discrepancy in the τ_{21} profile with and without the application jet trimming, which exactly attempts to remove soft activity from the jet. The dominance of non-perturbative QCD in this region poses a challenge to the DDT method itself, but does not prevent future searches from probing even lower large-R jet masses. Theses searches may instead rely on different mass-decorrelation methods, *e.g.* those discussed in Part III of this thesis.

Another challenge in the low-ρ region is that of angular resolution. Equation (1.5) suggests that large-R jets at the ISR γ channel p_{T} threshold ($p_{\mathrm{T}} \approx 200$ GeV) which have $\rho \lesssim -5$ are characterised by angular emissions of $\Delta R_{12} \lesssim 0.15$. This is comparable with the size in $\eta - \phi$ of topological clusters in the ATLAS calorimeter [180]. This means that, in this kinematic region, the ATLAS calorimeter does not have sufficient granularity to resolve individual particles, which impacts the reconstruction of the jet mass as well as jet substructure observables. This might be mitigated *e.g.* through the use of particle flow algorithms [100, 191], which were not explored in this analysis.

Finally, the high-mass region, $\rho \gtrsim -1.5$, is characterised by an angular separation between leading subjets of $\Delta R \approx 2m/p_{\mathrm{T}} = 2\sqrt{\exp(\rho)} \gtrsim 1$. With a large-$R$ jet radius parameter of $R = 1.0$, jets in this region are characterised by angular separations on the scale of the jet radius parameter. The turn-over in the τ_{21} profile at high values of ρ is therefore understood as arising situations where the two hardest subjets in the large-R jet are not fully contained by the fixed-radius jet algorithm.

Although the τ_{21} profiles for jets with and without the application of trimming exhibit significant differences at low ρ, they are qualitatively similar in the kinematic regions of interest to this analysis. To mitigate the effects of pile-up, the τ_{21} calculated from the trimmed large-R jet constituents are used in this analysis.

Figure 8.1 shows a dependence of the τ_{21} profiles on the large-R jet p_{T} in the form



(a) ISR γ channel **(b)** ISR jet channel

Figure 8.2 Mean value of τ_{21} as a function of the large-R jet ρ^{DDT}, in bins of large-R jet transverse momentum (p_{T}), in the **(a)** ISR γ and **(b)** ISR jet channel, with and without the application of jet trimming. Bottom panels show ratio of trimmed to untrimmed profiles. As a function of ρ^{DDT}, the residual p_{T}-dependence in Figure 8.1 is considerably mitigated. Figure **(b)** from Ref. [158].

of a roughly constant offset between p_{T}-slices. With the aim of removing the mean dependence of τ_{21} on both the jet mass and p_{T} with a simple linear transform, a modified variant of ρ was proposed in Ref. [189] as part of the DDT method

$$\rho^{\text{DDT}} = \log\left(\frac{m^2}{p_{\text{T}} \times \mu}\right) \tag{8.2}$$

on the basis that it is empirically found to reduce the residual dependence on p_{T}. Here, m is the mass of the large-R jet, p_{T} is the transverse momentum, and μ is an energy constant to balance units. In this analysis, a value of $\mu = 1$ GeV is used similar to Ref. [189], which found that this value leads to minimal residual p_{T}-dependence. Profiles of τ_{21} as a function of ρ^{DDT} are shown in Figure 8.2.

Indeed, as a function of ρ^{DDT}, the residual p_{T}-dependence in Figure 8.1 is considerably mitigated such that all p_{T}-slices now overlap. Additionally, the qualitative behaviour of the τ_{21} profile is unchanged and so still admits a linear fit in the central region. This is true for both channels considered in this analysis.

The low-mass region $\rho^{\text{DDT}} < 1.5$ is not used in the rest of the analysis, without any effect

on the sensitivity to the signal particle mass hypotheses considered. Conversely, the high-mass region ($\rho^{\mathrm{DDT}} \gtrsim 5$) is an essential part of the search region, but the non-linear behaviour in this region will prevent a robust background estimation. However, due to the large range in p_{T} probed across the two channels, the value of ρ^{DDT} at which the non-linearity occurs is not fixed. This means that a simple maximal ρ^{DDT} selection, common to both channels, is suboptimal. Instead, since the kink in the τ_{21} profile at high $\rho^{(\mathrm{DDT})}$ is due to fixed-radius effects, a so-called boosted topology selection is used, requiring $p_{\mathrm{T}} > 2m$. That is, the jet is required to have sufficient transverse momentum ("boost") to be fully contained within a jet with radius parameter of $R = 1.0$ according to the approximation in Equation (1.5). This selection is intended to remove the edge at large ρ^{DDT} in Figure 8.2 while, at the same time, not flatly discard all jets with ρ^{DDT} above some fixed threshold value.

The resulting profiles of the mean value of τ_{21} as a function of ρ^{DDT} in the ISR γ and ISR jet channels are shown in Figures 8.3a and 8.3b for simulated datasets of both the signal process and the dominant background process in each channel.

The requirement that $\rho^{\mathrm{DDT}} > 1.5$, along with the boosted topology selection, results in τ_{21} profiles which, in both channels, are roughly linear as a function of ρ^{DDT} and which have substantially reduced residual p_{T}-dependence. The signal processes generally result in jets with smaller values of τ_{21}, consistent with the two-prong hypothesis, see also the discussion in Chapter 1.

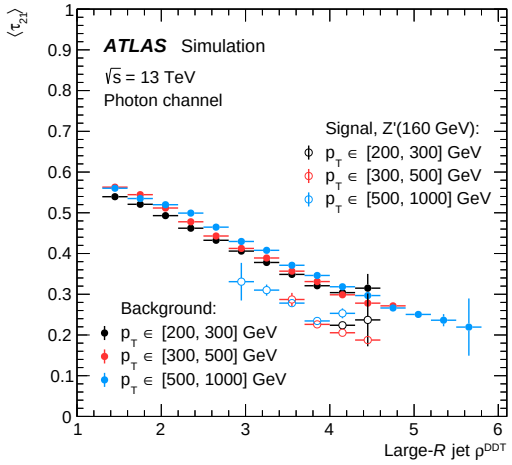
An alternative approach to remove the turn-over in the τ_{21} profile at large values of ρ^{DDT}, which was not explored in this analysis, would be to employ variable radius jets [192]. These jets are reconstructed using an effective radius parameter, $R_{\mathrm{eff}}(p_{\mathrm{T}}) \propto p_{\mathrm{T}}^{-1}$, which, for equal large-R jet masses, would lead to wider jets for larger values of $\rho^{(\mathrm{DDT})}$. This could potentially mitigate the above issue related to fixed-radius jet reconstruction.

To remove the dependence of τ_{21} on ρ^{DDT}, the leading background profiles in each channels are fitted separately with linear functions with slopes a. In both channels, the fits yield slopes consistent with a value of $a = -0.094$, with fit uncertainties at the sub-percent level. These are then used to perform the linear transform

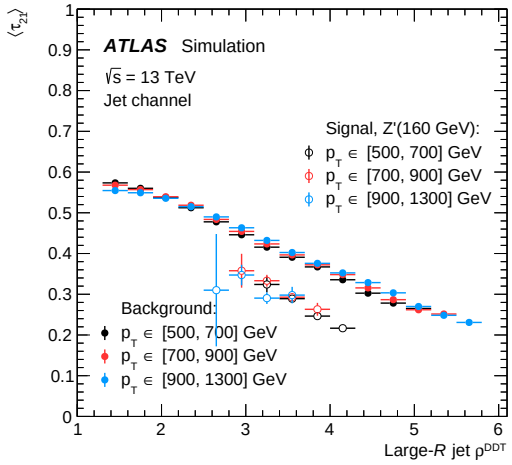
$$\tau_{21}^{\mathrm{DDT}} = \tau_{21} - a \times (\rho^{\mathrm{DDT}} - 1.5). \tag{8.3}$$

This results in a modified N-subjettiness observable, τ_{21}^{DDT}, which distinguishes one- and two-prong jets similarly to τ_{21}, but which is decorrelated from the jet mass and p_{T}.

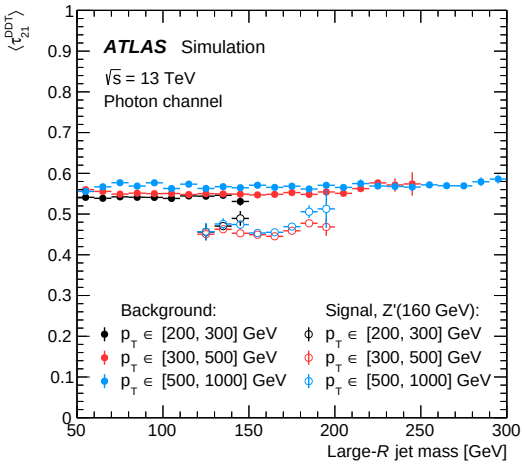
To study the effect of this transform, profiles of τ_{21}^{DDT} as a function of the large-R jet mass



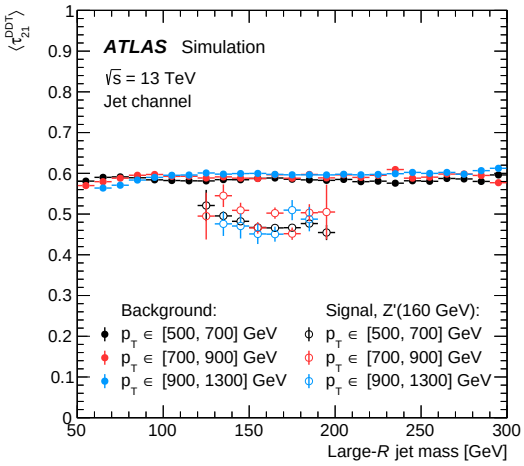
(a) τ_{21} profile for ISR γ channel



(b) τ_{21} profile for ISR jet channel



(c) τ_{21}^{DDT} profile for ISR γ channel



(d) τ_{21}^{DDT} profile for ISR jet channel

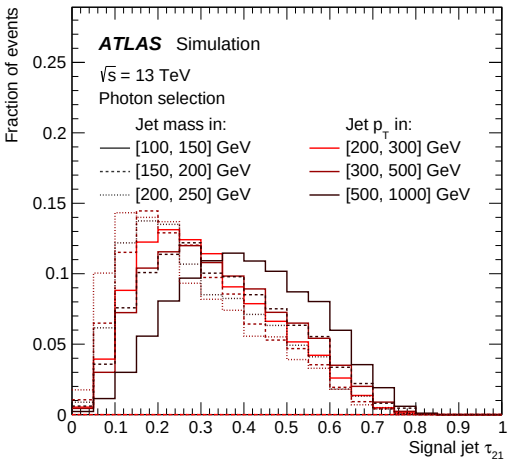
Figure 8.3 Mean value of **(a, b)** τ_{21} and **(c, d)** τ_{21}^{DDT} as a function of the large-R jet mass, in bins of large-R jet transverse momentum (p_{T}), in the **(a, c)** ISR γ channel and **(b, d)** ISR jet channel. The τ_{21} profiles are roughly linear as a function of ρ^{DDT} and have minimal residual p_{T}-dependence. Similarly, the τ_{21}^{DDT} profiles are roughly constant as a function of the jet mass, across p_{T}-bins, for the leading background processes. Figures **(b)** and **(d)** from Ref. [1].

are shown in Figures 8.3c and 8.3d. The average value of τ_{21}^{DDT} is roughly constant as a function of the jet mass for the leading background processes, with similar behaviour observed across p_{T}-bins. The DDT procedure has provided a mass- and p_{T}-decorrelated jet substructure observable which will be used in the analysis to reduce the dominant background as well as for the background determination itself. The DDT method was also explored for other jet substructure observables, but the linear relationship observed in Figures 8.1 and 8.2, required by the DDT method, is unique to τ_{21}.

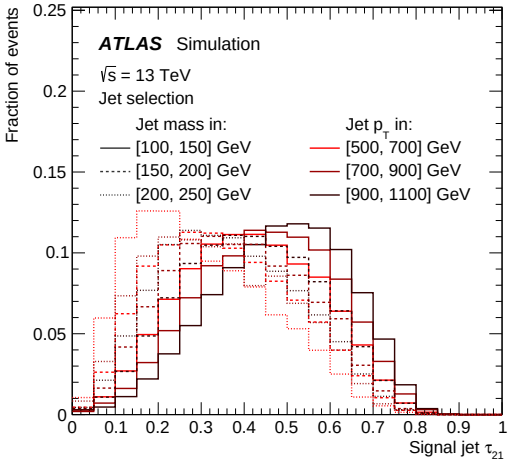
Finally, the stability of the $\tau_{21}^{(\text{DDT})}$ distributions themselves in bins of the jet mass and p_{T} are shown in Figure 8.4.

Whereas the τ_{21} distributions exhibit substantial variations across the jet mass and p_{T} bins in both channels, the τ_{21}^{DDT} distributions are more robust against changes of these variables. However, while the DDT method is able to remove the mean bias of τ_{21}, it does not affect the shape of the distributions. Therefore, while the centres of the τ_{21}^{DDT} distributions are stable across the mass and p_{T} bins in Figure 8.4, there are still noticeable shape difference between the high- and low-ρ^{DDT} distributions. This can be seen *e.g.* in Figure 8.4c by comparing the lowest-ρ^{DDT} bin (*i.e.* lowest mass, highest p_{T}; see Equation (8.2)) and highest-ρ^{DDT} bin (*i.e.* highest mass, lowest p_{T}); that is, the bins with $m \in [100, 150]$ GeV $\wedge$ $p_{\text{T}} \in [500, 1000]$ GeV and $m \in [200, 250]$ GeV $\wedge$ $p_{\text{T}} \in [200, 300]$ GeV, respectively. The former distribution is seen to be more symmetrical around $\tau_{21}^{\text{DDT}} \approx 0.55$ whereas the latter is more asymmetrical and peaking at $\tau_{21}^{\text{DDT}} \approx 0.4$. This limitation of the DDT method is discussed further in Part III.

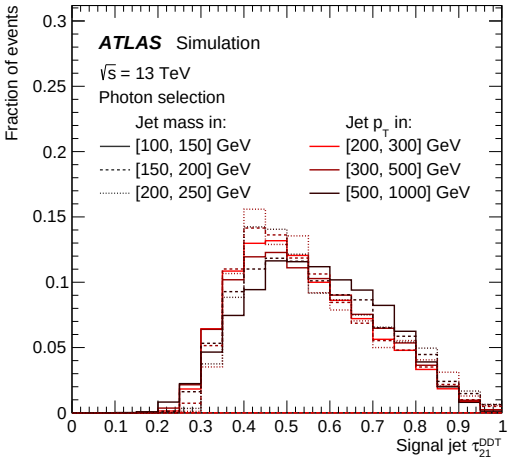
The baseline large-R jet selection now consists of a channel-dependent preliminary p_{T} selection of at least 200 GeV or 450 GeV in the ISR γ and ISR jet channel, respectively; a common selection of $\rho^{\text{DDT}} > 1.5$, discarding low-mass jets that are characterised by soft radiation, see the discussion above; and a common boosted topology selection of $p_{\text{T}} > 2m$, intended to ensure full collimation of the Z' decay products inside a jet with a radius parameter of $R = 1.0$ and to remove the high-mass kink in *e.g.* Figure 8.1. This selection defines a sample of large-R jets which are well-understood and for which a simple mass-decorrelated jet substructure observable, τ_{21}^{DDT}, can be defined. However, none of these selections have been chosen to specifically increase the purity of the signal process in the final search sample. Such a selection is performed using the τ_{21}^{DDT} observable.



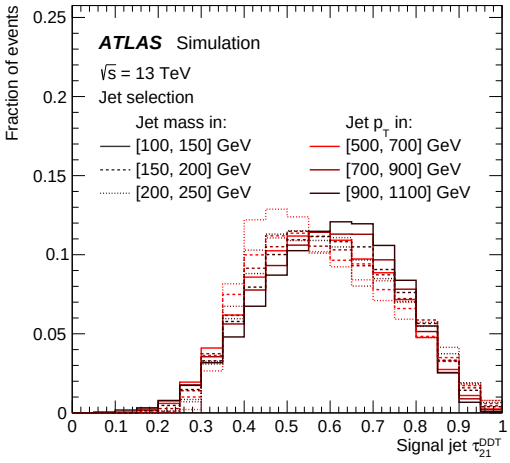
(a) τ_{21} distribution for ISR γ channel



(b) τ_{21} distribution for ISR jet channel



(c) τ_{21}^{DDT} distribution for ISR γ channel



(d) τ_{21}^{DDT} distribution for ISR jet channel

Figure 8.4 Distribution of the signal candidate large-R jet **(a, b)** τ_{21} and **(c, d)** τ_{21}^{DDT} for the leading background process in bins of the large-R jet mass and jet transverse momentum (p_{T}), in the **(a, c)** ISR γ channel and **(b, d)** ISR jet channel. Figures **(b)** and **(d)** from Ref. [1].

8.3 Substructure optimisation

Equipped the τ_{21}^{DDT} observable, a suitable value for a threshold selection must be chosen to define the search region of the analysis. The τ_{21}^{DDT} distribution for the leading background in each channel, as well as an example signal mass hypothesis, is shown in Figure 8.5. Since the choice of this selection value is not given from first principles or external constraints, the selection on τ_{21}^{DDT} is optimised with respect to the expected sensitivity of the hypothesised Z' signal over the leading background in each channel.

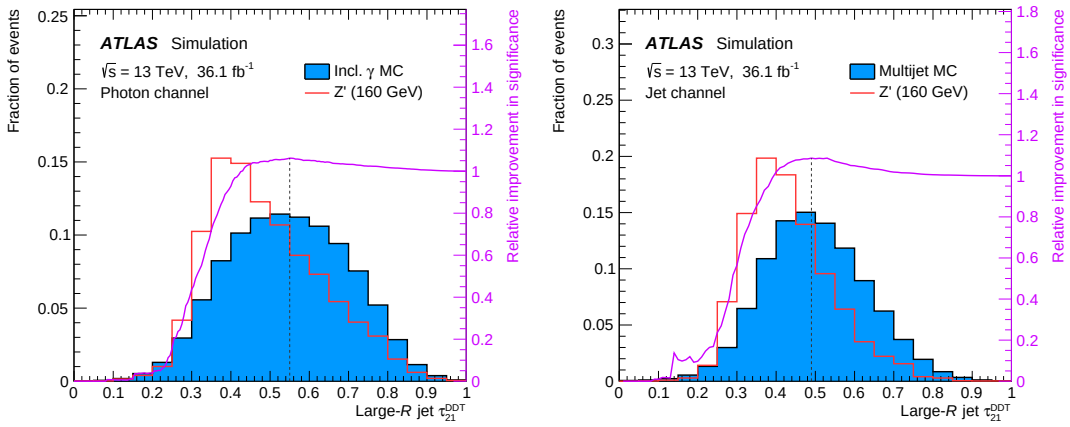
To do this, selection thresholds on τ_{21}^{DDT} are scanned and, for each value, the jets in the MC simulated datasets passing the selection — *i.e.* the ones with a value of τ_{21}^{DDT} below the selection threshold — are retained. The distributions of large-R jet masses for the retained signal and leading background MC large-R jets are compared, and the expected sensitivity in the jet mass spectrum, which is the search discriminant, is estimated as [193, 194]

$$\sigma_i = \sqrt{2\left[(s_i + b_i)\log\left(1 + \frac{s_i}{b_i}\right) - s_i\right]}, \tag{8.4}$$

where s_i and b_i are the bin contents of the signal and background large-R jet mass distributions, respectively, scaled to an integrated luminosity of 36.1 fb^{-1}. Equation (8.4) is the approximate Poisson significance for a single bin assuming observed data $o_i = s_i + b_i$ and a background-only expectation. In the limit of $s_i \ll b_i$, it reduces to the typical form $\sigma_i = s_i / \sqrt{b_i}$. The terms in Equation (8.4) are then summed in quadrature for all bins in the jet mass distribution to yield an estimate for the total expected significance for a given signal mass hypothesis. This procedure is performed for a range of potential τ_{21}^{DDT} selection values. The improvement in expected sensitivity, relative to the inclusive sample — *i.e.* without any selection on τ_{21}^{DDT}— is shown in Figure 8.5, along with an example τ_{21}^{DDT} distribution for $m_{Z'} = 160$ GeV. This is a simplified approach, only employed in the optimisation of the τ_{21}^{DDT} selection; the statistical approach used for extracting the search results is described in Chapter 10.

A selection threshold of $\tau_{21}^{\text{DDT}} < 0.5$ is found to be close to optimal in both channels, across all signal mass hypotheses, leading to a modest improvement in the expected sensitivity of the Z' signal. This selection is used to define the substructure "pass" and "fail" regions of the analysis. The search itself is performed in the pass region, which is enriched in events from the signal process following this jet substructure selection.

Finally, the τ_{21}^{DDT} variable is used to resolved the possible signal candidate ambiguity in



(a) ISR γ channel **(b)** ISR jet channel

Figure 8.5 Normalised distributions of τ_{21}^{DDT} for simulated signal samples with a mass of $m_{Z'} = 160$ GeV and **(a)** inclusive photon and **(b)** multijet samples in the ISR γ and ISR jet channel, respectively. The violet curve shows the relative improvement in the expected significance of the signal process relative to the leading background as evaluated in the large-R jet mass spectrum after a threshold selection on τ_{21}^{DDT} is performed for a cut corresponding to the value on the x-axis. See text for details. Figure **(b)** from Ref. [1].

the ISR jet channel: If multiple large-R jets are reconstructed and passing the outlined object selection, the jet with the lower value of τ_{21}^{DDT} is retained as the signal candidate.

The event selection defining the search regions in each of the two channels studied in this analysis is summarised in Table 8.1.

	Channel	
Selection	ISR γ	ISR jet
ISR object		
Type	Photon (γ)	Small-R jet (j)
p_{T} [GeV]	> 155	> 420
Large-R jet (J)		
p_{T} [GeV]	> 200	> 450
ρ^{DDT}	> 1.5	
Boosted topology	$p_{\mathrm{T}}^{J} > 2m^{J}$	
Angular separation	$\lvert\Delta\phi(J,\gamma)\rvert > \pi/2$	$\lvert\Delta\phi(J,j)\rvert > \pi/2$
Signal jet candidate selection		
Ambiguity resolution	—	Lower τ_{21}^{DDT}
τ_{21}^{DDT}	< 0.5	

Table 8.1 Overview of the physics object and event selection criteria applied in the definition of the search region for the analysis.

CHAPTER 9

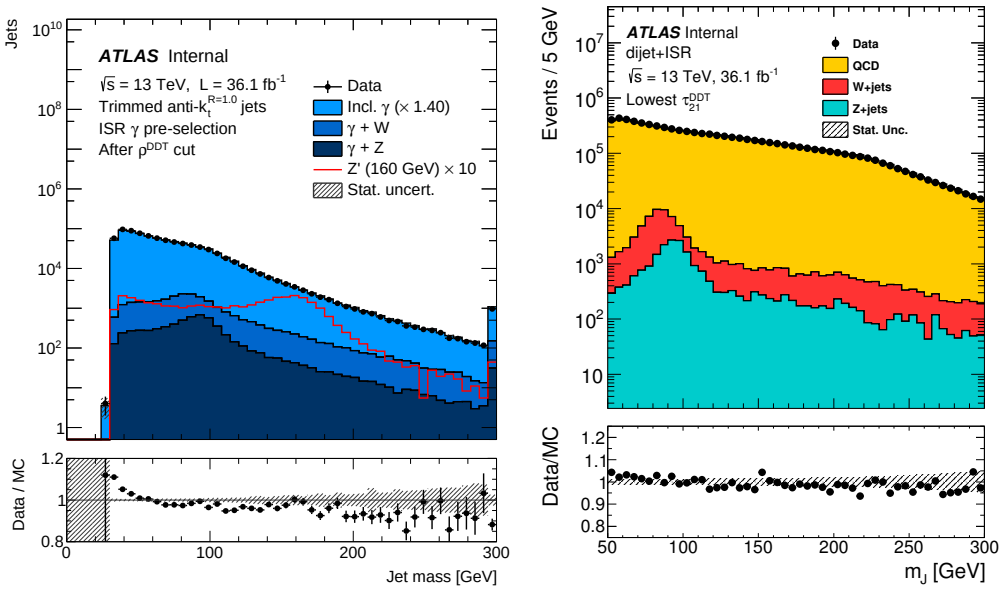
Background estimation

The selection criteria detailed in Chapter 8 define the search region in each of the two channels in this analysis. These criteria are applied equally to recorded and MC simulated datasets. This provides a basis for comparing observations in the recorded data to the expectation from known SM processes. Figure 9.1 shows the large-R jet mass distribution in the ISR γ and ISR jet channels after all selections with the exception of the jet substructure selection. It compares the recorded data yield to the expected yield from the dominant and sub-dominant background processes in each channel using MC simulated datasets.

The inclusive photon and multijet processes make up approx. 99% of the expected background yield in the ISR γ and ISR jet channel, respectively. The remaining, sub-dominant component is comprised by events with hadronically decaying electroweak bosons, W and Z. The data follows smooth distributions in both channels, with kinks around a large-R jet masses of $m = 100$ GeV and 225 GeV in the ISR γ and ISR jet channels respectively. These kinks arise from the onset of the boosted topology selection, as mentioned in Chapter 8: The minimum large-R jet p_{T} in the ISR γ and ISR jet channels are $p_{\mathrm{T,min}} = 200$ GeV and 450 GeV, respectively, which correspond to boosted topology selections of $m < m_{\mathrm{max}} = p_{\mathrm{T,min}}/2 = 100$ GeV and 225 GeV, respectively.

The inclusive W/Z backgrounds peak at large-R jet masses around the corresponding boson masses $m_{W/Z}$, see Figure 9.1a. Similarly, the hypothesised signals peaks around corresponding signal mass $m_{Z'}$.

However, in both channels the dominant background estimate from MC simulation deviates significantly from the distribution in data. For instance, the total data yield is 40% larger than the dominant MC background estimate in the ISR γ channel, see Figure 9.1a. A similar discrepancy in yield is found in the ISR jet channel [158, 195].



(a) ISR γ channel **(b)** ISR jet channel

Figure 9.1 Distribution of large-R jet mass in the **(a)** ISR γ and **(b)** ISR jet channel after application of all event selection criteria except the τ_{21}^{DDT} jet substructure selection. In both channels, all background processes are estimated using simulated Monte Carlo (MC) datasets and the dominant background is scaled up by approx. 40% to match the data yield [195]. Figure **(b)** from Ref. [158].

Finally, the shape of the dominant background MC estimate is observed to be in disagreement with the observed data in both channels, as evident in the residuals in Figure 9.1. Such discrepancies are typical for hadronic final states, and a similar normalisation difference of 40% is found in another ATLAS result in the same final state [196]. This is because Sherpa — used for generating the MC simulated datasets for the dominant backgrounds in both channels, see Chapter 6 — uses an LO tree-level matrix element, resulting in a large uncertainty on the overall cross-section due to missing higher-order terms [196]. As a consequence, searches in final states similar to the ones considered in this thesis overwhelmingly favour data-driven background estimates, see *e.g.* Refs. [145, 154, 197]. These shortcomings in the available MC simulated datasets for the dominant backgrounds mean that these datasets will not be sufficiently accurate for the search itself.

However, apart from the merging scheme [170], the Sherpa matrix element calculation factorises from the parton showering and hadronisation [167], meaning that it does not affect the modelling of the individual large-R jets, and in particular their substructure. This has been studied explicitly in other ATLAS publications focusing on the detailed measurement of large-R jet substructure. For instance, Ref. [39] compared several jet

substructure observable distributions in data to those found in Sherpa MC simulation, as well as Pythia8 [165] and Herwig++ [198]. In all cases, the Sherpa MC in the γ + jet final states was found to provide an excellent description of the substructure observable distributions in data, even though similar disagreements in overall normalisations were observed. At most, any non-negligible mismodelling of the jet substructure would shift the τ_{21}^{DDT} distributions in Figure 8.5. This could lead to suboptimal performance of the chosen τ_{21}^{DDT} selection, but any other disagreement would be accounted for through the data-driven background estimation procedure described below. Therefore, the MC simulated dominant background samples are considered adequate for the jet substructure decorrelation studies and event selection optimisation in Chapter 8.

Based on the discussion above, this analysis chooses to use a so-called data-driven estimate of the dominant background component, derived without reliance on MC simulation of the processes in question. Considering the marginal impact of the inclusive W/Z backgrounds and the application of a dedicated k-factor correcting the normalisation to NLO, see Chapter 6, the available MC samples are deemed sufficient for these.

9.1 Transfer factor method

A so-called transfer factor (TF) method is used to perform the data-driven estimate of the dominant backgrounds. This method uses the orthogonal "pass" and "fail" regions defined by the τ_{21}^{DDT} selection, see Chapter 8, to compute the TF, defined as the ratio of the number of events passing the selection over the number of events failing the selection. For this reason, the TF is also called the "pass/fail ratio," and is computed as the ratio of two distributions

$$\mathrm{TF}_{\mathrm{meas},i} = \frac{N_{\mathrm{pass},i}}{N_{\mathrm{fail},i}}, \tag{9.1}$$

where i enumerates the bins of the distributions. Using this TF profile, the number of events in the signal-depleted ("one-subjet"–like) fail region can be used to estimate the number of background events in the signal-enriched ("two-subjet"–like) pass region using the procedure described below. To estimate only the dominant background component, the expected inclusive W/Z components in MC simulation are subtracted from both the pass- and fail histogram before calculating the TF.

For each signal mass hypothesis $m_{Z'}$, the aim is to obtain an unbiased estimate of

the dominant background under the peak of a possible enhancement in the large-R jet mass distribution due to a hypothesised signal process. Therefore, for each $m_{Z'}$ hypothesis, all jets with a large-R jet mass m within ±20% of the Z' mass are excluded from the calculation of the TF. This window size corresponds to roughly two times the jet mass resolution in the relevant kinematics range [33], see also Section 7.2. This ±20% exclusion window is referred to as the signal region (SR) window, and each SR is expected to contain the approx. 95% of events from the corresponding hypothesised signal process. The TF then needs to be interpolated into this SR window, from the sidebands in the large-R jet mass. This is done to avoid biasing the dominant background estimate in the pass region SR window by the potential contamination of signal events in the fail region. The fact that the dominant background is estimated separately for each signal mass hypothesis $m_{Z'}$ means that the statistical analysis performed for each $m_{Z'}$ will be conducted using unique background estimates. These will only be moderately correlated between signal mass hypotheses, and only for adjacent values of $m_{Z'}$.

Since τ_{21}^{DDT} is constructed to be independent of the large-R jet p_{T} and ρ^{DDT}, see Chapter 8, the TF profile is parametrised in terms of these variables. Through the definition of ρ^{DDT} in Equation (8.2), this was used to indirectly decorrelate τ_{21}^{DDT} from the large-R jet mass. In practice, the dimensionless quantity $\log(p_{\text{T}}/\mu)$ is used, with $\mu = 1$ GeV, to bring the two variables to similar numerical scales.

For concreteness, Figure 9.2 shows the measured TF profile in the ISR γ channel for the full event selection applied to the MC simulated inclusive γ dataset, with the exclusion of a ±20% SR window around a large-R jet mass of $m = 160$ GeV. The SR window is seen as a strip of empty bins across the TF profile. The background estimation method is illustrated using MC simulated datasets to validate the procedure. However, for the search itself, the TF method is applied directly to recorded data to estimate the dominant background. The histogram binning is chosen to ensure sufficient events in each bin to guarantee stability of the method [158]. Jets exceeding the upper edges of the binning along each axis are included in the last bin along the axis in question.

Interpolation

The dominant background estimate in the pass region is then estimated as

$$N_{\text{pass}}\left(\rho^{\text{DDT}}, \log(p_{\text{T}}/\mu)\right) = \text{TF}_{\text{pred}}\left(\rho^{\text{DDT}}, \log(p_{\text{T}}/\mu)\right) \times N_{\text{fail}}\left(\rho^{\text{DDT}}, \log(p_{\text{T}}/\mu)\right), \quad (9.2)$$

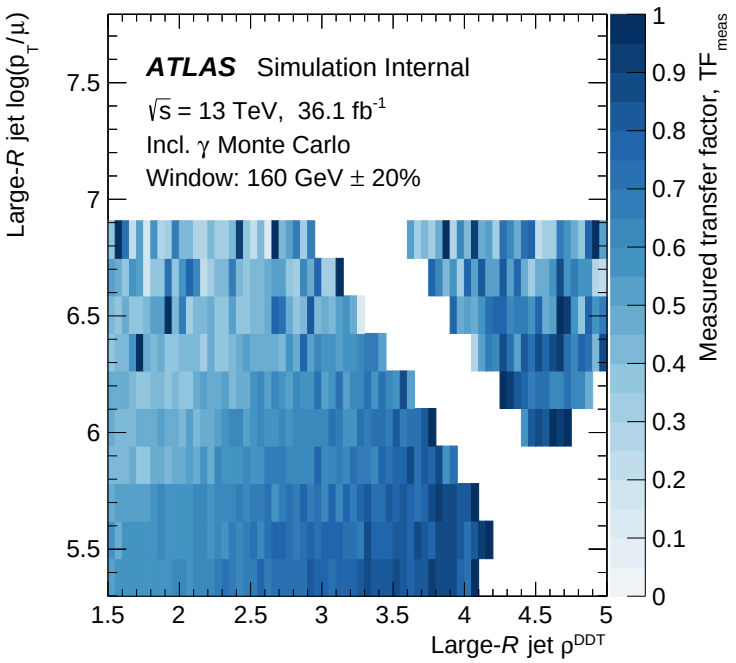


Figure 9.2 Profiles of the average value $\mathrm{TF}_{\mathrm{meas}}$ of the transfer factor (TF) for the $\tau_{21}^{\mathrm{DDT}} < 0.5$ selection, measured in a Monte Carlo (MC) simulated inclusive γ dataset with a signal region window of 20% around a large-R jet mass $m = 160$ GeV, seen as a white strip across the TF profile.

where $\mathrm{TF}_{\mathrm{pred}}$ is the predicted TF value obtained through a fit to the measured TF histogram defined in Equation (9.1). This is used to fit the TF profile in the large-R jet mass sidebands, *i.e.* the histogram bins outside the ±20% SR window, and interpolate into the excluded SR window.

The fact that the DDT method is constructed to mitigate the dependence of the pass/fail ratio on p_{T} and ρ^{DDT} is exactly intended to simplify the task in Equation (9.2). The ideal mass-decorrelation procedure would make the TF profile completely uniform, reducing Equation (9.2) to a simple constant scaling. However, as mentioned in Chapter 8, DDT only corrects first-order biases by removing the dependence of the mean value of τ_{21}^{DDT} on the large-R jet kinematics, and even this correction is limited by the validity of the linear approximation in Figure 8.3. Furthermore, since the τ_{21}^{DDT} selection threshold is chosen through an optimisation of the expected search significance, see Figure 8.5, and not as the mean value of τ_{21}^{DDT} for the background processes, the decorrelation will be further degraded. Finally, mismodelling of recorded data in the MC simulated datasets used to derive the DDT transform will lead to residual deviations from uniformity of the TF profile when measured in recorded data. The non-uniformity of the TF profile in Figure 9.2 is an illustration of the two first points. These inherent limitations of the

DDT method are discussed further in Part III.

To perform the interpolation into the SR, Gaussian process (GP) regression is used [199], as it allows for non-parametric regression to a set of data points; in this case the centres of the $\left(\rho^{\mathrm{DDT}}, \log(p_{\mathrm{T}}/\mu)\right)$-bins of the TF histogram and the associated values of $\mathrm{TF}_{\mathrm{meas}}$. GP regression provides a mean function $\mu(x)$ and a variance function $\mathrm{var}(x)$, representing the best-fit estimate of the underlying function and the associated variance at the point x. These two functions are in turn determined by a so-called kernel function, $K(x_1, x_2) = \mathrm{cov}(y_1, y_2)$, which expresses the covariance of function values y at different measurement sites x in terms of a relation between the latter. A common choice for the kernel function, which is also employed in this analysis, is the so-called Gaussian, or squared exponential

$$K(x_1, x_2) = \exp\left(-\frac{|x_1 - x_2|^2}{2\ell^2}\right), \tag{9.3}$$

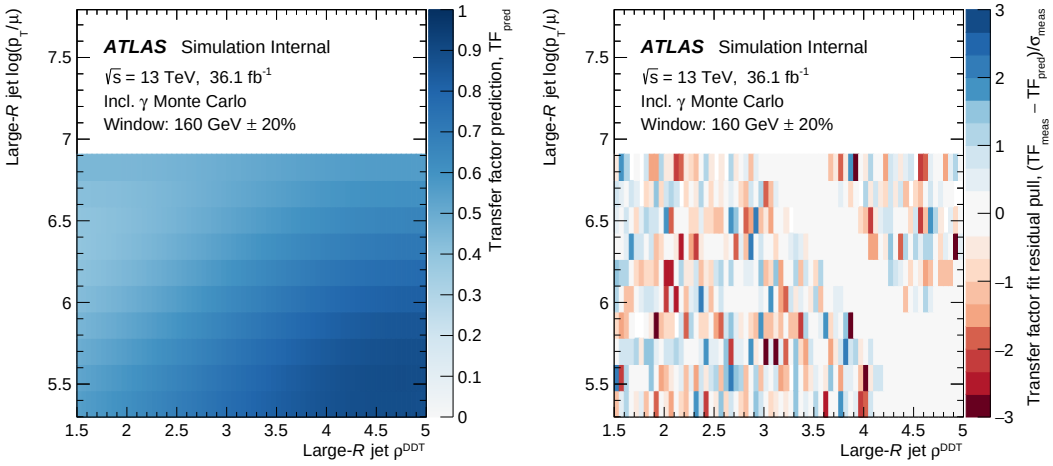
where ℓ is referred to as the length scale of the kernel. In situations with more than one input, such as this analysis, each input dimension d has an associated length scale ℓ_d. These length scales are the only free parameters in the GP regression, and are not given *a priori*, but are instead determined by maximising the log-likelihood [200, 201]

$$\log L(\mathbf{y} \,|\, X, \{\ell_d\}) = -\frac{n}{2}\log\left[\mathbf{y}^{\top}\left(K(X,X) + \sigma_n^2\mathbb{I}\right)^{-1}\,\mathbf{y}\right] - \frac{1}{2}\log\left|K(X,X) + \sigma_n^2\mathbb{I}\right|, \tag{9.4}$$

where n is the number of measurements $\{(\mathbf{x}, y)\}$ in the dataset $(X, \mathbf{y})$, K is the kernel function, σ_n is the uncertainty on each of the n measurements, $\mathbb{I}$ is the identity matrix, and $|\cdot|$ denotes the matrix determinant. The first term in Equation (9.4) quantifies the quality of the regression to the measurement data, and the second term penalises model complexity. Additional details are given in Appendix C.

This analysis uses the `GaussianProcess` class as implemented in the SCIKIT-LEARN (v0.17.1) library [201]. The GP regression to the TF profile, shown in Figure 9.3a, therefore provides a non-parametric means for interpolating into the SR window.

For this example, in the ISR γ channel, considering only the inclusive γ process in MC simulation, the optimal GP length scales are found to be $(\ell^*_{\rho^{\mathrm{DDT}}}, \ell^*_{\log(p_{\mathrm{T}}/\mu)}) = (3.63, 2.95)$. Figure 9.3b shows the residuals of the GP regression with respect to the measured profile, $(\mathrm{TF}_{\mathrm{meas}} - \mathrm{TF}_{\mathrm{pred}})/\sigma_{\mathrm{meas}}$, exhibiting no discernible structure. These residuals have a mean value consistent with zero and a standard deviation consistent with unity, indicating a robust regression.



(a) Transfer factor fit

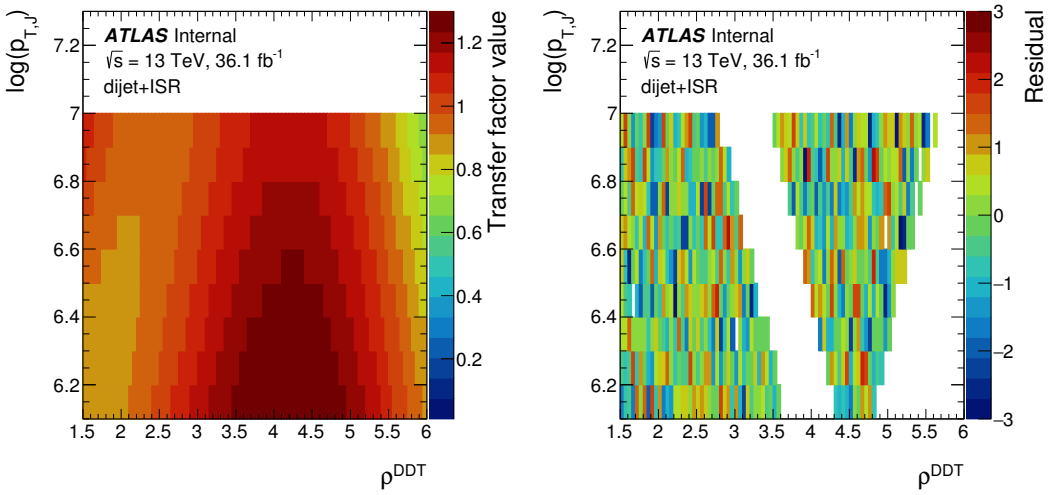
(b) Transfer factor fit residual pulls

Figure 9.3 Profiles of **(a)** the average value $\mathrm{TF}_{\mathrm{pred}}$ of the transfer factor (TF) as predicted by the Gaussian process (GP) regression to the measured profile in a Monte Carlo (MC) simulated inclusive γ dataset with a signal region window of $\pm 20\%$ around a large-R jet mass $m = 160$ GeV and **(b)** the associated residual pulls $(\mathrm{TF}_{\mathrm{meas}} - \mathrm{TF}_{\mathrm{pred}})/\sigma_{\mathrm{meas}}$ where $\mathrm{TF}_{\mathrm{meas}}$ and σ_{meas} are the average and standard deviation, respectively, of the TF value in each bin as measured in the inclusive γ dataset. The excluded $\pm 20\%$ signal region (SR) window is seen in **(b)** as a white strip across the TF profile.

For completeness, Figure 9.4 shows the fitted TF profile and the associated residuals with respect to the measured profile both in the ISR jet channel, similarly with the exclusion of a 20% signal window around $m = 160$ GeV.

In the ISR jet channel, the predicted TF value varies as a function of ρ^{DDT} and $\log(p_{\mathrm{T}}/\mu)$ with optimal length scales which are typically a factor of 2-3 shorter than for the ISR γ channel [195], see Figure 9.4a, Finally, from Figure 9.4b the residual pull values do not exhibit any structure that would indicate a lack of capacity of the GP regression model to represent the simulated multijet dataset. In this channel as well, the distribution of residual pulls are consistent with a Gaussian distribution with zero mean and unit width. This confirms that the GP length scales obtained by maximising the likelihood in Equation (9.4) provide a good estimate of the underlying data.

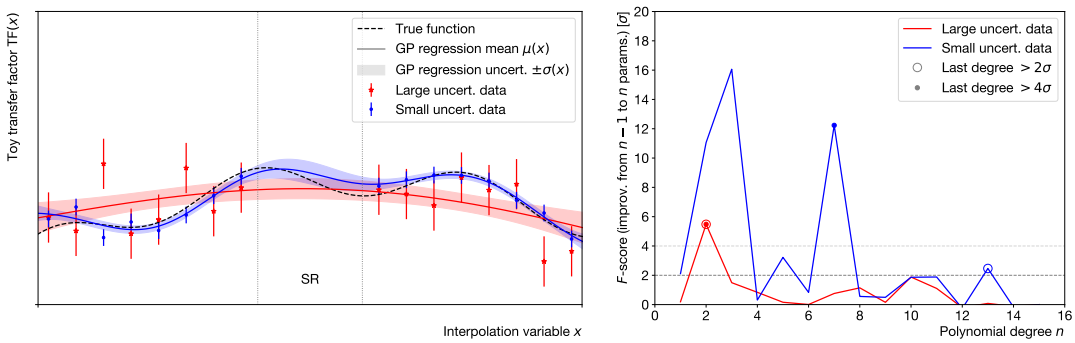
The reason that the GP regression in the ISR γ channel has length scales which are considerably longer than in the ISR jet channel is somewhat subtle. The situation is illustrated by an example in Figure 9.5. Figure 9.5a shows two datasets, one (*red*) with considerably larger uncertainties than the other (*blue*), similar to the ISR γ and ISR jet channel, respectively. Both datasets are generated from the same underlying



(a) Transfer factor fit **(b)** Transfer factor fit residual pulls

Figure 9.4 Profiles of **(a)** the average value $\mathrm{TF}_{\mathrm{pred}}$ of the transfer factor (TF) as predicted by the Gaussian process (GP) regression to the measured profile in a Monte Carlo (MC) simulated multijet dataset with a signal region window of ±20% around a large-R jet mass $m = 160$ GeV and **(b)** the associated residual pulls $(\mathrm{TF}_{\mathrm{meas}} - \mathrm{TF}_{\mathrm{pred}})/\sigma_{\mathrm{meas}}$ where $\mathrm{TF}_{\mathrm{meas}}$ and σ_{meas} are the average and standard deviation, respectively, of the TF value in each bin as measured in the multijet dataset. The excluded ±20% signal region (SR) window is seen in **(b)** as a white strip across the TF profile. Figure from Ref. [158].

function. The full lines show the best-fit GP regression to the respective datasets, and the shaded band shows the uncertainty on the regression. Since the underlying function exhibits variations over the range considered here (as the underlying TF profile might), the GP regression to the dataset with smaller uncertainties requires a shorter characteristic length scale to capture this behaviour. By contrast, for the dataset with larger uncertainties, these variations are not discernible, meaning that the GP regression can allow a more rigid fit with a larger characteristic length scale, which is favoured by the second term in Equation (9.4). This corresponds to the behaviour observed in Figures 9.3a and 9.4a. However, this behaviour would also be observed for methods of interpolation other than GP regression. For instance, a parametric, polynomial interpolation might have been chosen instead. Figure 9.5b shows the result of an F-test for iteratively increasing the number of polynomial terms included in a fit of the two datasets in Figure 9.5a. It shows the significance of the improvement in χ^2 per degree of freedom achieved by the addition of an n^{th} polynomial degree to a function which is otherwise of degree $n - 1$. Regardless of the threshold for considering the F-test improvement at degree n significant, the higher-uncertainty dataset is adequately described by a 2^{nd}-degree polynomial. By contrast, the



(a) GP regression example **(b)** Parametric fit F-test example

Figure 9.5 Toy example illustrating the impact of uncertainties on optimal length scales. Figure **(a)** show two datasets with different statistics, the true function generating both datasets (*dashed black*), and the Gaussian process (GP) regression to each (*full line*) along with its uncertainty (*shaded band*), interpolating into a signal region (SR). Figure **(b)** shows the F-score for a polynomial fit to the same datasets.

lower-uncertainty dataset requires at least 7, and possibly 13, terms to be completely described within uncertainties. Therefore, for both methods of interpolation, the dataset with smaller uncertainties requires shorter GP lengthscales or, correspondingly, more polynomial terms for an adequate description than the dataset with larger uncertainties.

As this examples shows, the observed behaviour is not an artefact of the chosen method for interpolation, but rather an expression of the fact that two very different final states, with very different statistics, are being compared. This analysis chooses to use the GP regression since it provides a natural notion of uncertainty through the variance function var(x), see also Appendix C.

Validation region

Finally, to account for possible imperfections in the GP interpolation, as well as to safeguard against the possibility of signal contamination in the large-R jet mass sidebands, a validation region (VR) is used. This region covers the range in large-R jet masses within ±30% of any given signal mass hypothesis, excluding the ±20% SR., *i.e.* the region in m corresponding to $[-30\%, -20\%] \cup [+20\%, +30\%]$ of $m_{Z'}$. In this analysis, a GP fit is first performed to the TF profile excluding the SR. In this fit, the GP length scales are left free to vary and determined by maximising the log-likelihood in Equation (9.4), yielding a set of optimal length scales $\{\ell_d^*\}$. Another GP fit is then performed, with $\{\ell_d^*\}$ fixed, to the TF profile excluding both the SR and the VR, *i.e.* to the large-R jet mass

sidebands outside the ±30% region around the signal mass hypothesis in question. An estimate of the dominant background contribution in the VR of the large-R jet mass spectrum for this validation GP fit is then found using Equation (9.2). The data in the VR is compared to the dominant background estimate from the TF method according to

$$\delta_{\mathrm{TF}} = \frac{1}{n_{\mathrm{VR}}} \sum_{i \in \mathrm{VR}} \frac{|N_i^{\mathrm{data}} - N_i^{\mathrm{est}}|}{\sqrt{(\sigma_i^{\mathrm{data}})^2 + (\sigma_i^{\mathrm{est}})^2}} \,. \tag{9.5}$$

Here, i enumerates the n_{VR} bins comprising the VR in the large-R jet mass spectrum; N_i^{data} and N_i^{est} are the number of observed and expected events, respectively, in the i^{th} VR bin; σ_i^{data} is the statistical uncertainty on the number of observed data events in the i^{th} VR bin; and σ_i^{est} is the uncertainty on the background expectation in the i^{th} VR bin, based on the GP uncertainty band. If $\delta_{\mathrm{TF}} \leq 1$, the TF background estimate is considered consistent with the data in the VR within uncertainties and the original fit to the sidebands of the SR is used with no modifications. However, if $\delta_{\mathrm{TF}} > 1$ some residual disagreement may be present, and the systematic uncertainty associated with the original fit to the sidebands of the SR is inflated by a multiplication by δ_{TF}.

In the search in the ISR γ channel, presented in Chapter 10, the average value of δ_{TF} across $m_{Z'}$ is approx. 1.5. In the ISR jet channel, δ_{TF} is less than one for all signal mass hypotheses. This may be understood as an effect of the difference in optimal length scales for the ISR γ and ISR jet channel, as discussed above. This difference in length scales arises because the number of events used to populate the TF profile in the ISR jet channel is an order of magnitude larger than in the ISR γ channel. This leads to smaller statistical uncertainties on the value of TF in each bin in the ISR jet channel, see Equation (9.1). This, in turn, necessitates shorter length scales according to the first term in Equation (9.4). Conversely, the larger statistical uncertainties on the TF profile in the ISR γ channel means that a more rigid GP fit is favoured, due to the inability to discern variations in the data which are of the same order as, or smaller than, the statistical uncertainty; this results in larger length scales through the second term in Equation (9.4). This difference in length scales is evident in the comparison of Figures 9.3a and 9.4a. However, as discussed above, this effect is not specific to the GP regression, since parametric methods for interpolation such as a polynomial fit exhibit the same characteristics. Finally, it is noted that while the relative uncertainty on the SR interpolation in the ISR jet channel is larger than in the ISR γ channel, due to the shorter characteristic length scales, the absolute uncertainty associated with the TF method is still smaller in the ISR jet channel due to the increased statistics, see Chapter 10.

9.2 Validation of the transfer factor method

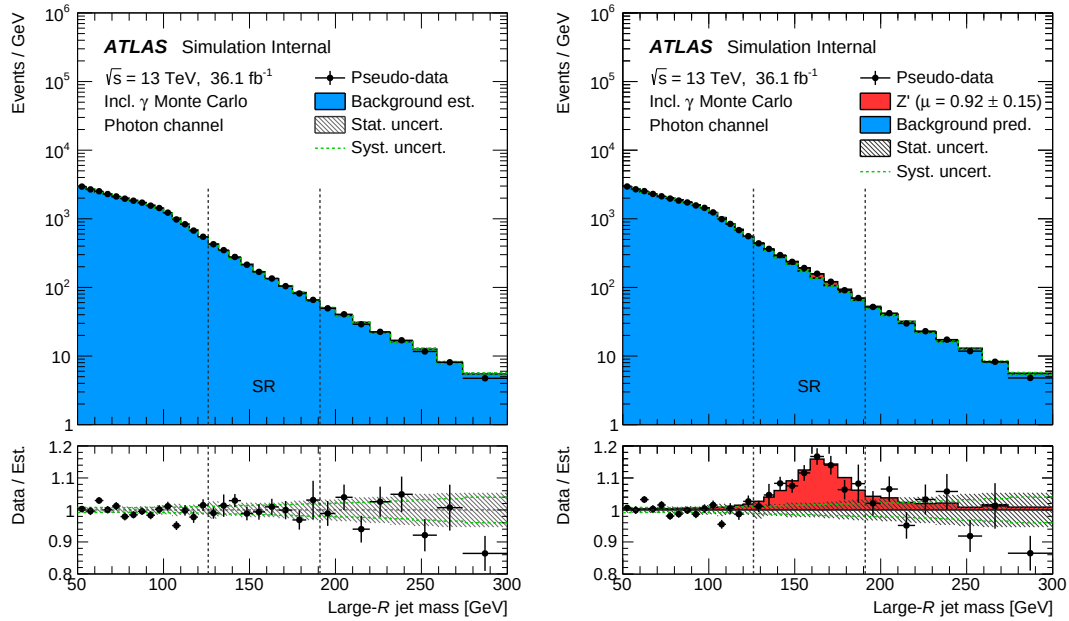
With the data-driven method of estimating the dominant background in the large-R jet mass spectrum presented above, the method is validated in MC simulated datasets before being applied to real data. Two closely related validation studies are performed in MC simulation: so-called closure and signal injection tests.

In the closure test, the full selection and background estimation procedure is applied to the MC simulated inclusive γ dataset, treated as pseudo-data. For each signal mass hypothesis, the data in the SR window of the large-R jet mass distribution is compared to the TF background estimate. Since the composition of the inclusive γ dataset used for these studies is completely known, the two samples should agree within statistical and systematic uncertainties. Any statistically significant deviation would be an indication of an inadequacy of the background estimation method itself. Figure 9.6a shows the result of a closure test for the $m_{Z'} = 160$ GeV mass point. Across all mass points, the TF background estimate is consistent with the pseudo-data in the signal region within uncertainties, indicating that the method itself is able to reliably reconstruct the dominant background in the SR in the absence of signal process contributions.

For the signal injection test, MC simulated signal samples with various mass hypotheses are injected into the pseudo-data to emulate the effect of a signal in the actual, recorded dataset. In this case as well, the full background estimation procedure is applied to the pseudo-data with the injected signal, emulating an attempt to reliably reconstruct a small excess in the recorded data. An example is shown in Figure 9.6b for the signal mass hypothesis $m_{Z'} = 160$ GeV. In this example, the validation region inflation scale factor in Equation (9.5) is found to be $\delta_{\mathrm{TF}} = 0.97$ and thus the systematic uncertainty associated with the GP regression to the TF profile is not inflated.

The aim of the signal injection test is then to extract the signal normalisation scale factor μ from a fit of the signal plus background MC component to the pseudo-data with signal injected. A signal strength of $\mu = 1$ corresponds to a signal yield exactly as predicted by the MC simulation of the signal model; a signal strength of $\mu = 0$ corresponds to a complete absence of signal. The extracted value of μ should be consistent with unity within uncertainties, otherwise this would be an indication that the TF method and the subsequent fitting strategy could systematically under- or overestimate a signal in data.

Although the τ_{21}^{DDT} selection is intended to create a signal-enhanced pass region and a signal-depleted fail region, Figure 8.5 shows that some signal contribution will



(a) Closure test

(b) Signal injection test

Figure 9.6 Validation of the transfer factor (TF) method for estimating the dominant background in the ISR γ channel in the form of **(a)** a closure test and **(b)** a signal injection test using Monte Carlo (MC) simulated datasets. In both figures, a ±20% signal region (SR) window around a large-R jet mass m = 160 GeV is excluded in the TF fit.

unavoidably contaminate the fail region, by failing the $\tau_{21}^{\text{DDT}} < 0.5$ selection. This potential signal contamination will be propagated into the pass region as part of the dominant background estimate, through Equation (9.2), effectively diluting the observed signal in the pass region. To mitigate this problem, the signal contribution in the fail region is estimated using the MC signal sample, propagated into the pass region using Equation (9.2), and subtracted from the dominant background estimate. This is done separately for each signal mass hypothesis $m_{Z'}$.

The signal component in the large-R jet mass distribution in Figure 9.6b is extracted from a binned χ^2-fit, as implemented in the RooFit toolkit [202]. The fit to the signal-injected pseudo-data is performed using the dominant background distribution and the signal distribution as templates. The total number of expected events is normalised to the number of events in the pseudo-dataset, with the variable to be extracted — the relative normalisation of the signal component, μ — left free to vary in the fit. The fit is performed for all bins in the range $m \in [50, 300]$ GeV. The χ^2 for the simplified fit in

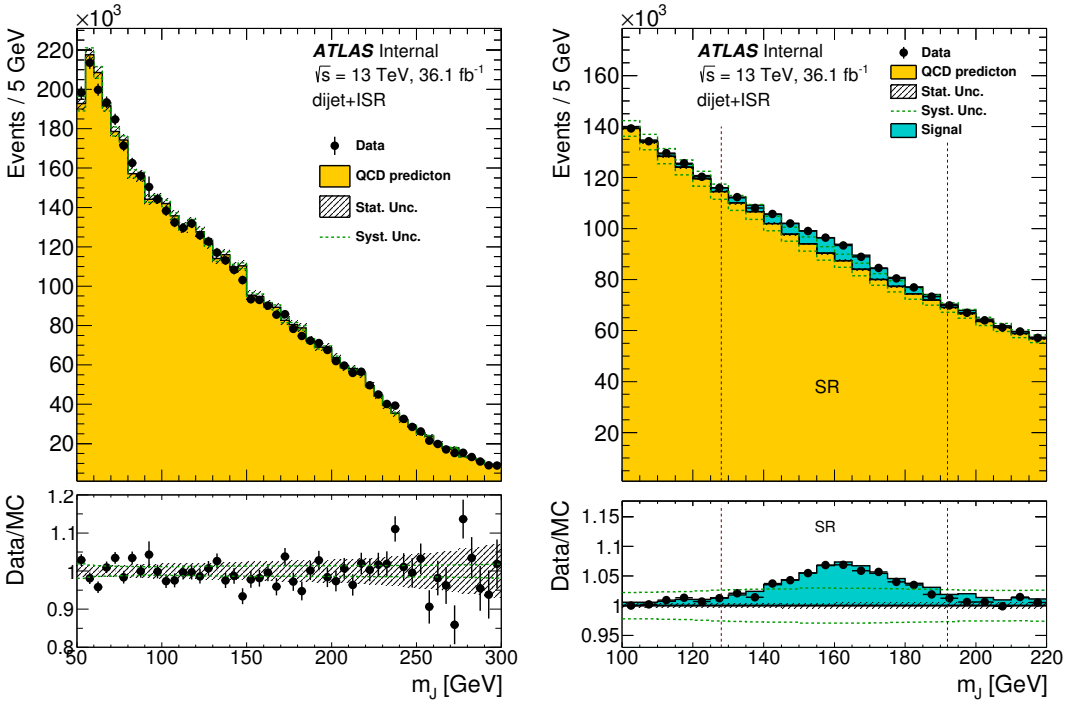
this validation study is given by [202]

$$\chi^2 = \sum_{i \in \text{bins}} \frac{|N_{\text{data},i} - N_{\text{exp},i}|^2}{\sigma^2_{\text{data},i}}, \qquad N_{\text{exp},i} = N_{\text{bkg},i} + \mu \times (N_{\text{sig},i} - N^{\text{fail}}_{\text{sig},i}). \tag{9.6}$$

Here, i enumerates all bins in the large-R jet mass spectrum, $N_{\text{data},i}$ is the number of observed pseudo-data events in the i^{th} bin, $\sigma_{\text{data},i}$ is the associated statistical uncertainty, $N_{\text{exp},i}$ is the total expected number of background events $N_{\text{bkg},i}$ and signal events $N_{\text{sig},i}$ given signal strength μ, accounting also for signal contamination in the fail region propagated into the pass region $N^{\text{fail}}_{\text{sig},i}$. The best-fit value for the signal strength $\hat{\mu}$ of the injected signal is found by minimising the χ^2 defined in Equation (9.6). This is equivalent to fitting only for $N_{\text{sig},i}$ with an effective signal strength $\tilde{\mu} = \mu \times f$, multiplied by a scale factor $f = (1 - \sum_i N^{\text{fail}}_{\text{sig},i} / \sum_i N_{\text{sig},i})$ determined directly using MC simulation, to account for this contamination. The statistical uncertainty $\Delta\hat{\mu}_{\text{stat}}$ is given by the values of μ for which $\Delta\chi^2 = 1$ with respect to the minimum.

To account for the systematic uncertainty associated with the TF background estimate, the fit is performed two more times with the TF prediction shifted up and down by the GP uncertainty band. This results in two additional best-fit signal strengths $\hat{\mu}^{\pm}$. The systematic uncertainty on $\hat{\mu}$ due to the GP regressions is then estimated as $\Delta\hat{\mu}_{\text{syst}} = |\hat{\mu}^{+} - \hat{\mu}^{-}|/2$ since the GP uncertainty is symmetric, see Appendix C. The combined uncertainty on the best-fit signal strength is then defined as $\Delta\hat{\mu} = \sqrt{\Delta\hat{\mu}^2_{\text{stat}} + \Delta\hat{\mu}^2_{\text{syst}}}$. An example of the post-fit large-R jet mass distribution for a signal injection test for the $m_{Z'} = 160$ GeV hypothesis is shown in Figure 9.6b. For this signal mass hypothesis, the extracted best-fit signal strength is $\hat{\mu} \pm \Delta\hat{\mu} = 0.92 \pm 0.15$. Similar studies in the ISR γ channel for other signal mass hypotheses also result in best-fit signal strengths consistent with $\hat{\mu} = 1$. Therefore it is concluded that the TF method provides a robust way of estimating the dominant background, even in the presence of signal, and allows for the reliable extraction of potential contributions from signal processes.

In the signal injection validation studies, the fit is performed for the same signal mass hypothesis $m_{Z'}$ at which a signal is injected. A related, relevant scenario is that in which a signal with mass $m_{Z'}$ is present in data but a fit is performed with a different mass hypothesis $m'_{Z'}$. In this scenario, the signal would be indistinguishable from the dominant background and would therefore be included in the TF background estimate, effectively erasing any signal peak at masses outside the SR window around $m'_{Z'}$. This is exactly why the analysis decided to define a SR, to provide an estimate of the dominant background contribution, in a region of the large-R jet mass distribution, which is not biased by the possible presence of a signal. Additionally, the analysis will scan signal



(a) Closure test **(b)** Signal injection test

Figure 9.7 Validation of the transfer factor (TF) method for estimating the dominant background in the ISR jet channel in the form of **(a)** a closure test and **(b)** a signal injection test using Monte Carlo (MC) simulated datasets. In both figures, a ±20% signal region (SR) window around a large-R jet mass $m = 160$ GeV is excluded in the TF fit. Figures from Ref. [158].

mass hypotheses in the range in $m \in [100, 220]$ GeV in steps of 10 GeV, see Chapter 10, which is less than the large-R jet mass resolution, see Chapter 7. This means that any potential signal process with a mass $m_{Z'}$ in the targeted range would be included in the SR of at least one analysed mass hypothesis; *i.e.* the analysis is sensitive to all signal masses in this range, but not for all analysed mass hypotheses. The fact that resonant processes outside of the SR are included in the dominant background estimate is also the case for the sub-dominant W/Z background processes, which are explicitly subtracted in the analysis, as mentioned above.

For completeness, examples of closure and signal injection tests in the ISR jet channel are show in Figure 9.7. In this channel, closure is also observed across all signal mass hypotheses, and the extracted best-fit signal strengths $\hat{\mu}$ are similarly consistent with one.

CHAPTER 10

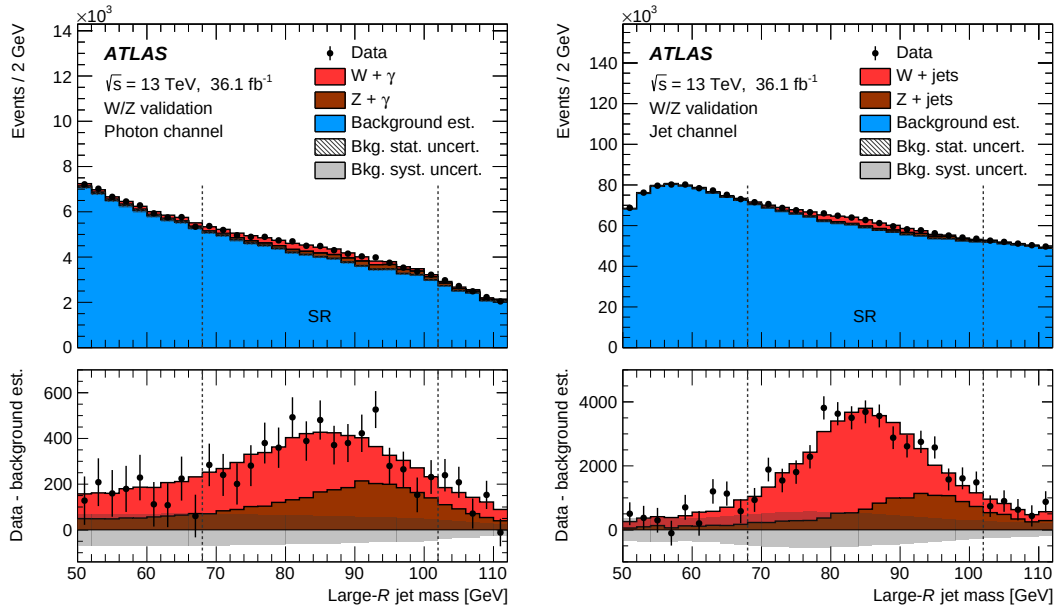
Statistical analysis and search results

The TF method for performing an estimate of the dominant background in each of the two search channels was described in Chapter 9, and validated using MC simulated datasets. Equipped with an event selection defining a signal-enhanced search region and a method for estimating the dominant background, the recorded data can be analysed to search for signs of leptophobic DM mediator particles Z'.

10.1 W/Z validation study

As an *in situ* validation of the analysis technique, the W/Z peak in the large-R jet mass spectrum is treated as a "known signal" in data and a search for this peak is performed. Since the W/Z peak resembles the targeted signal model for a mass hypothesis of $m_{Z'} \approx m_{W/Z}$, this validation study tests the ability of the analysis to recover a known excess in data.

The full event selection described in Chapter 8 is applied to the recorded dataset selected using the triggers described in Chapter 6. The TF method is applied to data with a signal mass hypothesis of $m_{Z'} = 85$ GeV, with a ±20% signal region (SR) covering large-R jet masses in the range $m \in [68, 102]$ GeV. The inclusive W/Z process is treated as a joint signal, and this component is therefore not subtracted from the pass and fail histograms during the TF procedure. Due to the near-degeneracy of the W and Z peaks, the two are fitted together with a single signal strength parameter μ, with relative normalisations determined with MC simulation. The agreement between data and the dominant background estimate in the ±30% VR, *i.e.* for large-R jet masses



(a) ISR γ channel **(b)** ISR jet channel

Figure 10.1 Distribution of large-R jet mass in the search region around the W/Z mass peak in the **(a)** ISR γ and **(b)** ISR jet channel. The dominant background contribution in each channel is estimated using the transfer factor (TF) method described in Chapter 9. The signal regions around the W/Z mass, excluded in the TF fit, are indicated by dashed vertical lines. The inclusive W/Z process contributions are estimated using Monte Carlo (MC) simulated datasets and have been scaled by their best-fit normalisation values $\hat{\mu}_{W/Z}$. The bottom panels show the background-subtracted data distributions, which are nicely described by the MC simulated W/Z contributions.

in the ranges $m \in [59.5, 68]$ GeV $\cup$ $[102, 110.5]$ GeV, is found to be $\delta_{\text{TF}} = 0.62$ in the ISR γ channel, see Section 9.1. Since this value is less than unity, there is no indication of a disagreement between the data and dominant background estimate in the VR. Therefore, the nominal estimation of the dominant background under the W/Z peak is deemed to be robust. A fitting procedure similar to that used in the signal injection test in Chapter 9.2 is employed, yielding a best-fit signal strength $\hat{\mu}_{W/Z}$ of the combined inclusive W/Z production relative to the SM prediction. In addition, a statistical uncertainty from the data in the pass region and a systematic uncertainty from the TF background estimate are obtained in the fit. The best-fit signal strength is found to be $\hat{\mu}_{W/Z} = 1.07 \pm 0.13\,(\text{stat.}) \pm 0.35\,(\text{syst.})$ in the ISR γ channel and $\hat{\mu}_{W/Z} = 0.93 \pm 0.03\,(\text{stat.}) \pm 0.24\,(\text{syst.})$ in the ISR jet channel. The post-fit large-R jet mass distributions around the W/Z peak in the two channels are shown in Figure 10.1. In both cases, the TF systematic uncertainty is seen to be dominant; this is discussed further in Section 10.3.

The difference in the resolution of the W/Z peak in the two channels is due to the difference in large-R jet p_{T} threshold, see Chapter 8. The fact that these signal strengths are consistent with unity within the quoted uncertainties means that the results are consistent with the SM prediction. This is a sanity check that demonstrates that the method is able to extract search results, and suggests that the analysis is able to accurately recover known resonances in the large-R jet mass spectrum. In both channels, the TF systematic uncertainty is significantly larger than the statistical uncertainty. This is an indication that, while the analysis yields accurate results, the precision could be improved by better constraining the dominant background systematic uncertainty. The TF uncertainty is driven by the challenge of fitting and interpolating the TF profile. A more comprehensive method for large-R jet mass-decorrelation than the DDT method, employed in this analysis, would simplify this task and drive down the dominant systematic uncertainty. This is discussed further in Section 10.3, and potential techniques for improved mass-decorrelation are detailed in Part III.

10.2 Systematic uncertainties

The main systematic uncertainty affecting this analysis is the one associated with the TF background estimation procedure, see Sections 9.1 and 10.1. The dominant background estimate found using the mean GP regression function $\mu(x)$ is referred to as the nominal estimate, see Appendix C. The associated uncertainty is found by varying the value of the GP fit to the TF profile by $\pm\sqrt{\mathrm{var}(x)}$ around the mean function $\mu(x)$; these are referred to as the up and down variations, respectively. These may be inflated based on the agreement between data and the background estimate in the VR in the large-R jet mass spectrum, as quantified in Equation (9.5). To allow for more flexibility in the dominant background modelling, the TF uncertainty is decomposed into a normalisation uncertainty and a shape uncertainty in the search, as illustrated in Figure 10.2.

The up and down variations of the normalisation uncertainty are found by scaling the nominal TF background estimate in the large-R jet mass spectrum to have the same integral as the corresponding (up, down) variation of the standard GP uncertainty band. This keeps the shape of the dominant background, but changes the yield. Conversely, the up and down variations of the shape uncertainty are found by scaling the corresponding (up, down) variation of the standard GP uncertainty band to have the same integral as the nominal TF background estimate. This leaves the total yield constant, but changes the shape of the dominant background spectrum.

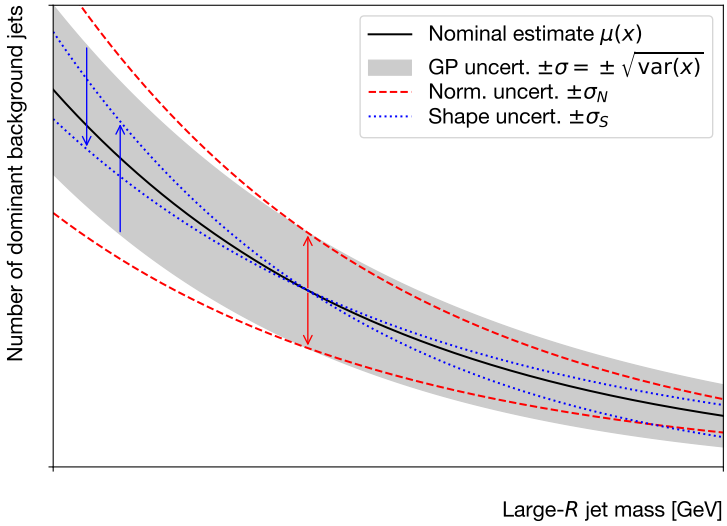


Figure 10.2 Schematic illustration of the decomposition of the Gaussian process (GP) uncertainty band $\pm\sigma = \pm\sqrt{\mathrm{var}(x)}$ around the nominal estimate $\mu(x)$ into a normalisation uncertainty $\pm\sigma_N$ and a shape uncertainty $\pm\sigma_S$. See text for details.

Since MC simulated datasets are used for modelling the otherwise unconstrained hypothesised signal processes, parametrised systematic uncertainties provided centrally by ATLAS are used for these. The dominant uncertainties on the signal process relate to the absolute scale and resolution of the signal candidate large-R jet energy, mass, and τ_{21} observables. These uncertainties affect both the normalisation of the hypothesised signal yield, with effects of roughly 10% [158], as well as the shape of the signal large-R jet mass distribution. Smaller effects arise from modelling uncertainties related to the ISR objects in each channel, namely the photon and small-R jet energy scale and resolution in the ISR γ and ISR jet channel, respectively. These uncertainties correspond to an overall uncertainty on the signal yield of approx. 2% in both channels [158]. An additional source of uncertainty arises from the choice of PDF used for generating the signal process in MC simulation, which also affects the yield of the signal process. This analysis uses signal processes generated using the NNPDF2.3 set [164], which comprises an ensemble of 101 PDFs. A baseline PDF is compared to the remaining 100 variations, and the standard deviation on the resulting signal process yield is assigned as an overall normalisation uncertainty on the signal models [158]. This PDF uncertainty is combined with a theory unceratinty on the magnitiude of the QCD coupling strength α_{QCD}, resulting in a total theory unceratinty on the signal yield normalisation of 3% in the ISR γ and 4% in the ISR jet channel [158].

The inclusive W/Z background is taken directly from MC simulated samples, scaled

Source	Uncertainty on signal yield	Process		
		Dominant bkg.	$W/Z + \gamma$, jet	Signal
TF norm.	—	●		
TF shape	—	●		
W/Z norm.	—		●	
Large-R jets (4)	$\approx 10\%$		⊕	●
ISR objects (5)	$\approx 2\%$			⊕
Theory (2)	$\approx 3\%$			⊕
Luminosity	2.2%			●

Table 10.1 Summary of systmatic uncertainties assigned in this analysis. The number in parenthesis (if any) indicates the number of separate uncertainties grouped for the given source. An "●" indicates that the uncertainty or uncertainties are each applied to the process in question; an "⊕" indicates that the uncertainties are summed in quadrature before application to the source in question. Table reproduced from Ref. [158].

by the best-fit signal strengths found in the validation study in Section 10.1. The uncertainty on the normalisation of the W/Z background is similarly taken to be the combined uncertainty found in each channel of the validation study. In addition, the large-R jet uncertainties are also applied to the inclusive W/Z background processes.

Finally, the uncertainty on the combined integrated luminosity for the ATLAS 2015-2016 data taking period, used in this analysis, is 2.2% [158, 203, 204].

The complete set of systematic uncertainties considered in this analysis, the processes to which they are assigned, and their impact on the signal yield (if relevant), as described above, are summarised in Table 10.1.

10.3 Statistical tests and results

Statistical tests are performed using the datasets selected in each channel, with the aim of calculating upper exclusion limits on the signal strengths μ for the considered signal models that are consistent with the observed data. This amounts to a statistical hypothesis testing, comparing the null (or background-only) hypothesis with alternative (or signal-plus-background) hypotheses, parametrised by the signal strength μ. Limits on the signal strength can then be translated to limits on the product of the production cross-section of the signal process with the acceptance of the analysis event selection, $\sigma \times A$, as well as on the coupling of the Z' particle to Standard Model (SM) quarks,

g_q. All search results are extracted from binned likelihood fits to the large-R jet mass spectrum in the pass region following the procedure described below.

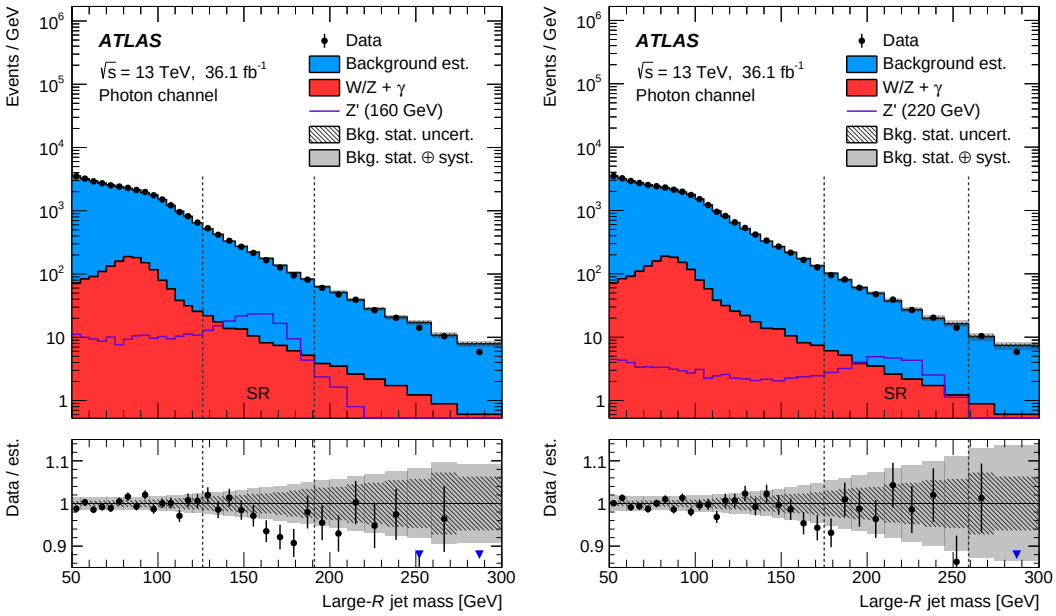
Search distributions

The MC simulated signal samples are generated with a mass spacing of 30 GeV, see Chapter 6. With a large-R jet mass resolution of approx. 10%, see Chapter 7, a finer grid of signal mass hypotheses is needed for the statistical tests to discover or exclude potential new physics processes. An interpolation of the shape of the signal large-R jet mass distribution, for different signal mass hypotheses $m_{Z'}$, is performed in steps of 10 GeV using a morphing method for interpolating probability density functions (p.d.f.s) [205] as implemented in RooFit's `RooMomentMorph` class [202].

Examples of the search distributions used in this analysis are shown in Figure 10.3 for two representative signal mass hypotheses in both the ISR γ and ISR jet channels.

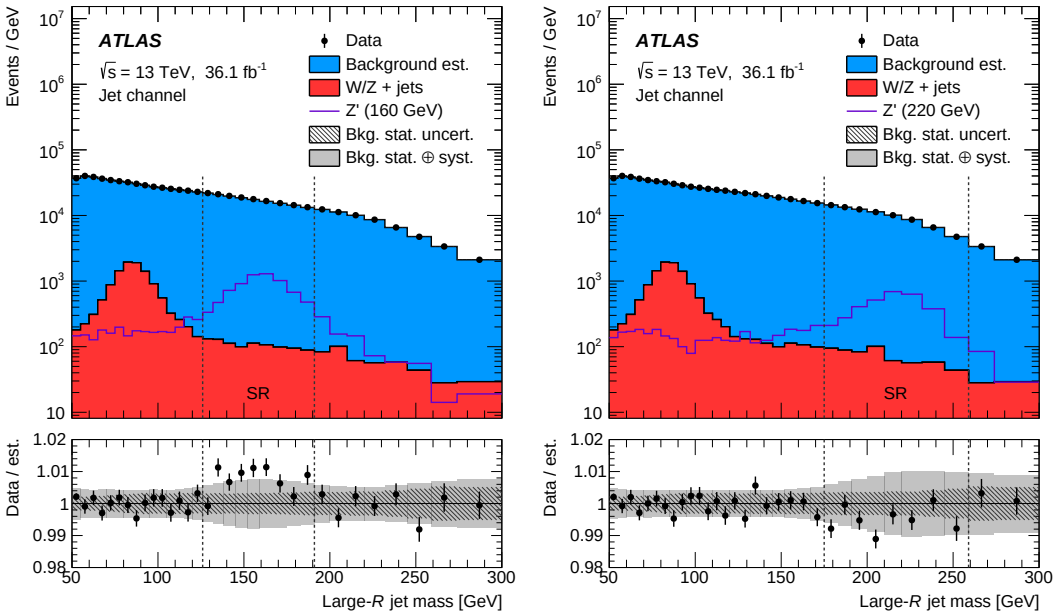
The full analysis selection is applied to both MC simulated and recorded datasets, and the search distributions in the substructure pass region are shown. The binning of the large-R jet mass is chosen to yield bin widths roughly corresponding to half the jet mass resolution of 10%. The dominant background component in each channel is estimated using the full TF procedure described in Section 9.1. The inclusive W/Z processes are combined in the fit, normalised according to the best-fit signal strengths $\hat{\mu}_{W/Z}$ found in Section 10.1. Each search distribution is overlaid with the large-R jet mass distribution expected from MC simulation for the relevant signal process, to visualise its potential contribution in the same figure. The limits of the ±20% TF SR are also indicated on the figure for each signal mass hypothesis $m_{Z'}$.

In both channels, the onset of the boosted topology selection manifests as a kink in the large-R jet mass spectrum around 100 GeV and 225 GeV in the ISR γ and ISR jet channel, respectively, which is most pronounced in the ISR γ channel. The behaviour of the TF uncertainty band, seen in the bottom panels of Figure 10.3 reflects the discussion in Section 9.1: the GP regression length scales in the ISR jet channel are shorter than in the ISR γ channel, due to the larger event count in the former, leading to a more uncertain interpolation into the SR, manifesting as a larger relative TF uncertainty band in the ISR jet channel SR.



(a) ISR γ channel, $m_{Z'} = 160$ GeV

(b) ISR γ channel, $m_{Z'} = 220$ GeV



(c) ISR jet channel, $m_{Z'} = 160$ GeV

(d) ISR jet channel, $m_{Z'} = 220$ GeV

Figure 10.3 Distribution of large-R jet mass in the search region of the **(a, b)** ISR γ channel and **(c, d)** ISR jet channel for signal mass hypotheses of **(a, c)** $m_{Z'} = 160$ GeV and **(b, d)** $m_{Z'} = 220$ GeV. The signal regions (SRs) around the Z' mass, excluded in the transfer factor (TF) fit for the dominant background estimate, are indicated by dashed vertical lines. The bottom panels show the ratio of the data to the combined background estimate. The dominant background estimates are different for each Z' mass hypothesis; see text for details.

Likelihood

To quantify the compatibility of observed data with the expectation from the predicted background and a potential signal contribution, a likelihood function is built. The number of observed data events n_i in each bin i in the large-R jet mass spectrum will follow a Poisson distribution with a mean equal to the expected value of $b_i + \mu s_i$, where b_i is the total expected number of background events from all considered processes, and μs_i is the expected number of signal events. Here, μ is the effective signal strength, see Section 9.2, controlling the normalisation of the signal contribution: a signal strength of $\mu = 0$ corresponds to a background-only hypothesis, and a signal strength of $\mu = 1$ corresponds to a signal-plus-background hypothesis exactly as predicted using the MC simulated datasets. The purely statistical likelihood of observing a particular dataset $\mathcal{D} = \{n_i\}$ given signal and background models s and b is then given by [194]

$$\mathcal{L}_{\text{stat.}}(\mathcal{D}\,|\,\mu, \boldsymbol{\theta}) = \prod_{i \in \text{bins}} \frac{(b_i + \mu s_i)^{n_i}}{n_i!}\, e^{-(b_i+\mu s_i)}, \tag{10.1}$$

where $\boldsymbol{\theta}$ is the set of so-called nuisance parameters (NPs). These provide a parametrisation of the systematic uncertainties on each background component, see *e.g.* Table 10.1. The NPs are not known *a priori*, but must be fitted to the data as part of the likelihood maximisation under Gaussian constraints. An example of this is the W/Z normalisation uncertainty, derived in Section 10.1, with associated NP $\theta_{W/Z}$: a value of $\theta_{W/Z} = 0$ yields the nominal W/Z contribution, whereas values $\theta_{W/Z} = \pm k$ yield W/Z peaks with normalisation shifted up and down, respectively, by k times the combined uncertainty on the W/Z normalisation. With an assumption of Gaussian uncertainties, these NPs contribute the following term to the likelihood

$$\mathcal{L}_{\text{syst.}}(\boldsymbol{\theta}) = \prod_{\theta \in \boldsymbol{\theta}} \frac{1}{\sqrt{2\pi}} \exp(-\theta^2/2) \tag{10.2}$$

penalising disagreement of each systematic variation with the associated prior. Notice that this NP constraint term does not depend on the dataset $\mathcal{D}$. The combined likelihood to be maximised in the fitting is therefore given by

$$\mathcal{L}(\mathcal{D}\,|\,\mu, \boldsymbol{\theta}) = \mathcal{L}_{\text{stat.}}(\mathcal{D}\,|\,\mu, \boldsymbol{\theta}) \times \mathcal{L}_{\text{syst.}}(\boldsymbol{\theta}). \tag{10.3}$$

For performing the statistical tests to extract upper exclusion limits, the profile likelihood ratio is used [194]

$$\lambda(\mu) = \frac{\mathcal{L}(\mathcal{D}\,|\,\mu, \hat{\hat{\boldsymbol{\theta}}})}{\mathcal{L}(\mathcal{D}\,|\,\hat{\mu}, \hat{\boldsymbol{\theta}})}\,. \tag{10.4}$$

Here, $\hat{\hat{\boldsymbol{\theta}}}$ is the conditional maximum likelihood estimator of $\boldsymbol{\theta}$ given μ, while $\hat{\mu}$ and $\hat{\boldsymbol{\theta}}$ are the unconditional maximum likelihood estimators. Equation (10.4) is referred to in this way, since the likelihood in the numerator is profiled over $\boldsymbol{\theta}$, making it a function only of the parameter of interest, μ. Following the Neyman-Pearson lemma [206], this is the most powerful statistical test comparing hypotheses parametrised by μ. The added flexibility afforded by the NPs will tend to lead to a broader profile of $\lambda(\mu)$, as tuning these can improve agreement with data leading to larger likelihoods for a broader span of signal strengths. This illustrates the need to mitigate and constrain systematic uncertainties to retain the maximal sensitivity to anomalous observations.

Limit setting

For setting upper exclusion limits on cross-sections for a signal-plus-background hypothesis, the aim is to identify the largest value of μ for which the signal-plus-background model is consistent with the observed data at some pre-determined confidence level (CL). To this end, the test statistic [194]

$$q_\mu = \begin{cases} -2\log\lambda(\mu) & \hat{\mu} \leq \mu \\ 0 & \hat{\mu} > \mu \end{cases} \tag{10.5}$$

is used for minimisation. The negative log-likelihood of the profile likelihood ratio in Equation (10.4) is used for convenience, with identical results since the logarithm is monotonic. The $\hat{\mu}$-dependent clause means that, for a given μ, a downwards fluctuation of the data ($\hat{\mu} \leq \mu$) reduces the likelihood of the signal-plus-background hypothesis, whereas an upwards fluctuation ($\hat{\mu} > \mu$) is not taken to represent a lesser degree of compatibility with data, since the test is only concerned with setting an upper limit on the values of μ that are compatible with the observed data.

For quantifying the agreement between expectation and the observed data, the p-value is typically used: for a certain observed test statistics q_μ, the p-value is the probability that an identically repeated experiment would yield as extreme, or greater, disagreement, *i.e.* $q'_\mu > q_\mu$. To compute such p-values, Wilks' theorem [207] can be used, as it states

that for nested hypotheses, the log-likelihood ratio test statistic in Equation (10.5) will, asymptotically for large samples sizes, follow a χ^2-distribution with one degree of freedom, corresponding here to the signal strength parameter μ. Furthermore, the Wald approximation [194, 208] states that the test statistic in Equation (10.5) is approximately quadratic around $\hat{\mu}$. With these results, it can be shown that using the q_μ test statistic, the p-value for a given signal strength μ is given by [194]

$$p_\mu = P(q'_\mu \geq q_\mu \,|\, \mu) = \int_{q_\mu}^{\infty} f(q'_\mu \,|\, \mu)\, \mathrm{d}q'_\mu = 1 - \Phi(\sqrt{q_\mu}), \tag{10.6}$$

where Φ is the cumulative density function (c.d.f.) for a unit Gaussian. This probability has the associated Gaussian significance

$$z_\mu = \Phi^{-1}(1 - p_\mu) = \sqrt{q_\mu}\,. \tag{10.7}$$

That is, z_μ is the number of standard deviations above the mean of a Gaussian distributed variable, that yields an upper-tail probability of p_μ. Similarly, the p-value under the background-only hypothesis is conventionally defined as [209]

$$p_0 = P(q'_\mu < q_\mu \,|\, \mu = 0) = \int_{-\infty}^{q_\mu} f(q'_\mu \,|\, \mu = 0)\, \mathrm{d}q'_\mu\,. \tag{10.8}$$

The p-value in Equation (10.6), may lead to non-physical exclusion limits for small or negative values of $\hat{\mu}$. This is because a downward fluctuation of the data with respect to the expected background might lead to artificially stringent limits on μ, through Equation (10.5), in cases with $s_i \ll b_i$ *i.e.* cases with very limited sensitivity. The modified frequentist CL_s statistical method [209] is used to avoid this problem by modifying p_μ in cases with a very small expected signal yield or, equivalently, cases where the background-only hypothesis provides a good representation of data. The CL_s p-value is given by [209]

$$\mathrm{CL}_s(\mu) = \frac{\mathrm{CL}_{s+b}(\mu)}{\mathrm{CL}_b} \equiv \frac{p_\mu}{1 - p_0}\,, \tag{10.9}$$

where the signal-plus-background and background-only p-values, p_μ and p_0, are given by Equations (10.6) and (10.8), respectively. In situations with well-separated hypotheses, where data favours the signal-plus-background hypothesis (*i.e.* $p_0 \to 0$), Equation (10.9) reduces to the standard frequentist p-value $\mathrm{CL}_s(\mu) \to \mathrm{CL}_{s+b}(\mu) = p_\mu$. Conversely, in cases with poorly separated hypotheses (*e.g.* $s_i \ll b_i$), the CL_s conservatively reduces the standard frequentist p-value according to the agreement of

the data with the background-only hypothesis (*i.e.* p_0 large in this case).

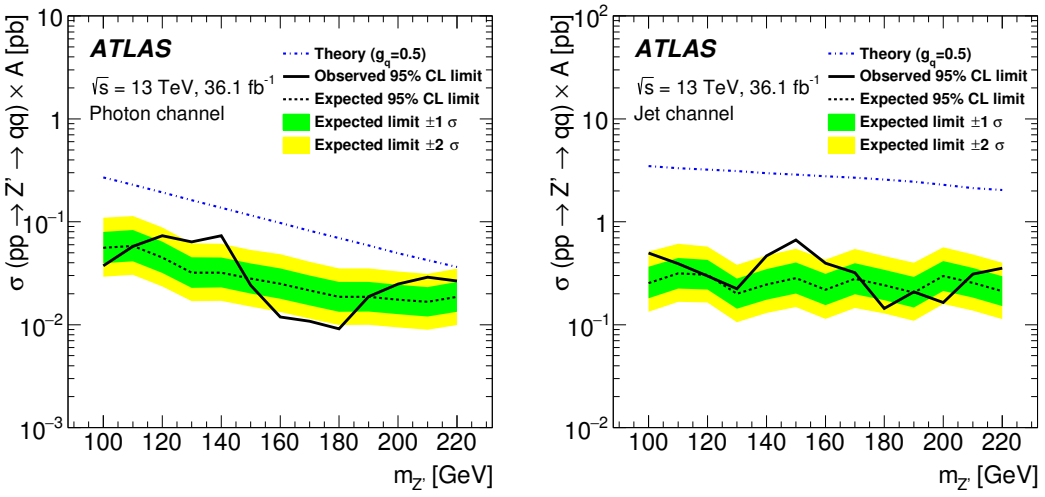
By scanning values of μ and testing the resulting compatibility of the signal-plus-background model with data through Equation (10.9), upper exclusion limits on μ at some CL α can be found as those for which $\mathrm{CL}_s(\mu) \leq 1 - \alpha$. In high-energy physics (HEP) the CL used for model exclusion is typically taken to be $\alpha = 95\%$, corresponding to a Gaussian significance of $z_\mu \approx 2\sigma$. Here, the σ-notation emphasises the nature of z_μ as a number of Gaussian standard deviations. For a discovery of a new physical process under a background-only hypothesis, a much more stringent requirement on the search significance is imposed: the threshold for discovery in HEP is customarily set at $z_0 \geq 5\sigma$, or $p_0 \leq 2.87 \times 10^{-7}$.

Results

Using the procedure outlined above, the analysis allows for setting 95% CL upper exclusion limits on the signal strength μ of the hypothesised Z' particle for different mass hypotheses $m_{Z'}$. The statistical tests are performed using the HistFitter package [202, 210–212], and in practice the likelihood term in Equation (10.1) is computed only for bins within the ±20% SR for each mass hypothesis $m_{Z'}$.

Initially, the compatibility of the observed data for each $m_{Z'}$ with the background-only hypothesis is tested. The largest upward fluctuation in each channel results in a local Gaussian significance of $z_0 = 2.2\sigma$ for signal mass hypothesis $m_{Z'} = 140$ GeV in the ISR γ channel and a local significance of $z_0 = 2.5\sigma$ for $m_{Z'} = 150$ GeV in the ISR jet channel. However, these significances do not account for the fact that 13 different signal mass hypotheses are tested independently in each channel. The larger the sample of independent trials, the more likely it is for the sample to contain a result which is discrepant according to some fixed criterion. To mitigate this so-called "look elsewhere" effect, a trials factor [213] is used to inflate the local probability p_0 to account for the multiple signal mass hypotheses tested. This results in a so-called global probability, which be approximated by $p_{\mathrm{global}} \approx p_0 \times k$ [213], where k is approximately equal to the number of distinct mass hypotheses tested, *i.e.* 13 in this analysis. Using this approach, the maximum global probabilities yield the so-called global significances of 0.8σ and 1.1σ in the ISR γ and ISR jet channel, respectively, which are considerably smaller than the local significances above.

These excesses are not significant according to the established discovery criterion. Upper exclusion limits on the Z' signal strength μ are set. To do so, it is noted that the expected



(a) ISR γ channel **(b)** ISR jet channel

Figure 10.4 Observed and expected 95% confidence level (CL) exclusion limits on the production cross section (σ) times kinematic acceptance (A) of the leptophobic Z' particle in **(a)** the ISR γ channel and **(b)** the ISR jet channel. The production cross section times kinematic acceptance for a value of the Z' coupling to SM quarks of $g_q = 0.5$ is shown as a guide. Figures from Ref. [1].

signal yield, parametrised by μ, scales linearly with the signal process cross-section. This means that upper exclusion limits on μ can be used directly to compute similar 95% CL limits on the signal cross-section times kinematic acceptance $\sigma \times A$ using the nominal signal normalisation ($\mu = 1$) as reference point. These limits are shown in Figure 10.4.

This figure shows the expected $\sigma \times A$ for a signal model with a coupling to SM quarks of $g_q = 0.5$, corresponding to $\mu = 1$ in the MC simulated signal samples; the expected 95% CL limit for each channel and the expected $\pm 1/2\sigma$ uncertainty on this limit; as well as the actual upper exclusion limit as observed in data along with the expected theory limit given a Z' coupling to SM quarks of $g_q = 0.5$. The expected limits are found by choosing the value of q_μ to be the median of the approximate distribution under the background-only hypothesis [194], and using the CL_s method to find the corresponding upper exclusion limit on the signal strength. Similarly, by repeating the process for the quantiles corresponding to $\pm N$ Gaussian standard deviations, the $\pm N\sigma$ uncertainty bands on the expected limit can be determined, illustrating the expected variability of the result.

In both channels, the observed 95% CL upper exclusion limit is seen to agree with the expectation within the shown uncertainty bands, with no significant deviations. In both

channels, the observed but also the expected limits are not entirely smooth as a function of the hypothesised signal mass $m_{Z'}$. This is because the fit is only performed within the ±20% SR window, in which the dominant background estimate is unique for each $m_{Z'}$. This means that the background estimate for adjacent signal mass hypotheses are only somewhat correlated, allowing for the potential for abrupt changes in the expected limits between neighbouring mass hypotheses $m_{Z'}$. This is corroborated by that the fact that the "bumpiness" in the expected limits is most pronounced in the ISR jet channel, which is characterised by shorter GP length scales in the regression to the TF profile, leading to a less robust interpolation and thus less correlation between the expected limits for neighbouring signal mass hypothesis. See also Section 9.1 for a discussion of this behaviour.

In addition to the cross-section limits, which allow for re-interpretation of the results in the context of different signal models, upper exclusion limits on the model-dependent coupling of the leptophobic Z' particle to SM quarks, g_q, are also calculated. These limits are readily obtained from the cross-section limits, as the cross-section scales with the square of g_q. Due to the model-dependent nature of these limits, the two channels are analysed in combination to fully utilise the sensitivity of the analysis. The systematic uncertainties related to the large-R jets and the luminosity are treated as correlated across channels, due to their common provenance, whereas the remaining uncertainties are treated as uncorrelated. The resulting upper exclusion limits on g_q for the combined analysis is shown in Figure 10.5.

In the combination, the observed limits are also in agreement with expected limits and no deviations with a significance at the level of discovery, or even evidence (customarily 3σ), are observed. The largest local deviation is found for $m_{Z'} = 140$ GeV with a local significance of 2.4σ, and a global significance of 1.2σ. The analysis is able to exclude coupling values much below the benchmark value of $g_q = 0.5$.

To illustrate the effect of the combination, Figure 10.6 shows the limits on q_g split by analysis channel. This result shows that the increased data statistics in the ISR jet channel, due to the larger cross-sections of the hypothesised signal processes as well as the later onset of the boosted topology selection, leads this channel to dominate the search sensitivity. The combined results, however, are generally improved by the inclusion of the ISR γ channel, even if marginally so.

Finally, the impact of each source of uncertainty is studied. First, the full likelihood is minimized, resulting in a best-fit signal strength $\hat{\mu}$ with an associated uncertainty $\Delta\hat{\mu}_{\text{tot}}$ found through the CL_s method using the profile likelihood ratio in Equation (10.4).

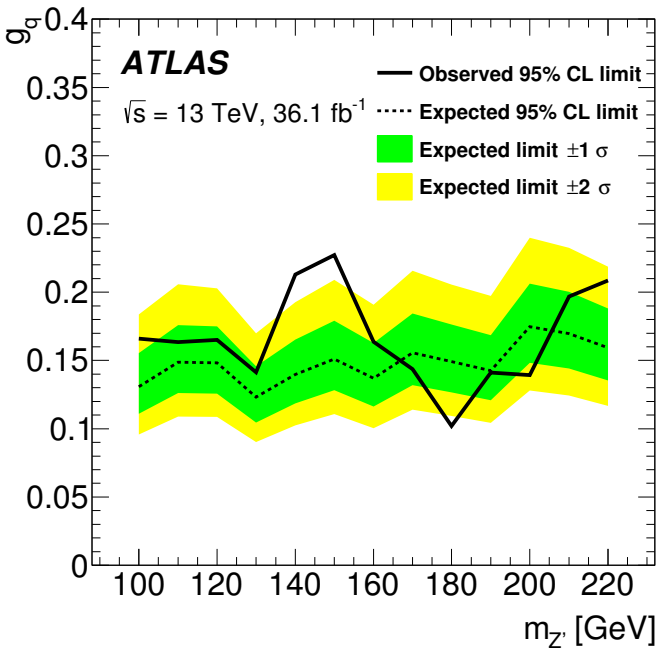


Figure 10.5 Observed and expected 95% confidence level (CL) exclusion limits on the coupling (g_q) of the leptophobic Z' particle to Standard Model (SM) quarks for the combination of the ISR γ and ISR jet channels. Figure from Ref. [1].

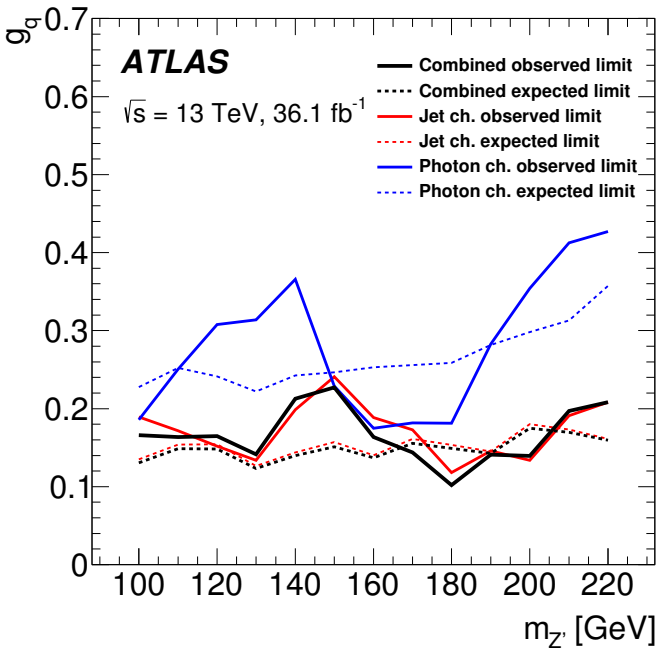


Figure 10.6 Observed and expected 95% confidence level (CL) exclusion limits on the coupling (g_q) of the leptophobic Z' particle to Standard Model (SM) quarks for the ISR γ and ISR jet channels separately. Figure from Ref. [1].

	$\Delta\hat{\mu}/\hat{\mu}$[%]		
Uncertainty source	$m_{Z'}$ = 100 GeV	$m_{Z'}$ = 160 GeV	$m_{Z'}$ = 220 GeV
Transfer factor	86	90	88
Large-R jet calib. and modelling	19	25	17
W/Z normalisation	43	$\ll 1$	$\ll 1$
Signal PDF	$\ll 1$	$\ll 1$	1
Luminosity	2	$\ll 1$	$\ll 1$
Total systematic uncertainty	91	93	91
Statistical uncertainty	9	10	11

Table 10.2 Overview of the impact of each of the largest uncertainties on the expected signal in the combined analysis, quantified as the uncertainty on the best-fit signal strength $\Delta\hat{\mu}$ over the best-fit signal strength $\hat{\mu}$ for signal mass hypotheses of $m_{Z'}$ = 100, 160, and 220 GeV. Reproduced from Ref. [1].

Then, each set of NPs are, in turn, fixed to their maximum likelihood estimator values and the uncertainty on the best-fit signal strength is re-evaluated with this reduced set of NPs. Following the discussion above, a reduced set of NPs will lead to a narrower profile likelihood ratio as a function of μ, due to a reduced flexibility in the fitting, resulting in a smaller uncertainty $\Delta\mu$. The difference in quadrature between the total uncertainty $\Delta\mu_{\text{tot}}$ and this reduced uncertainty $\Delta\mu$ constitutes the absolute impact on the signal strength due to the set of NPs in question. The relative impacts of each set of uncertainties affecting the combined analysis are shown in Table 10.2.

Generally, the TF uncertainties are dominant, contributing approx. 90% of the total uncertainty in the combined search. For the lowest-mass hypothesis, the inclusive W/Z normalisation starts contributing, due to the overlap between the W/Z peak in the large-R jet mass spectrum, and the signal peak for $m_{Z'} \approx 100$ GeV. The large-R jet uncertainties affecting the signal process yield contribute approx. 20% across signal mass hypotheses, whereas the uncertainties assigned to the signal PDF and the luminosity measurement are both minor. Finally, the systematic uncertainties dominate the statistical uncertainties, which have an impact of approx. 10%. As noted above, the TF uncertainty is driven by the challenge of fitting and interpolating the TF profile. Reducing the correlation of the chosen jet substructure observable with the jet mass and p_{T} would simplify this task, thereby reducing the dominant systematic uncertainty. Possible methods for improving large-R jet mass-decorrelation are studied in Part III.

CHAPTER 11

Conclusion and outlook

This part has presented a search for leptophobic DM mediators Z' using 36 fb^{-1} of data collected by the ATLAS experiment at a centre-of-mass energy of $\sqrt{s} = 13$ TeV during 2015 and 2016. The analysis targets two signal processes, in which the Z' boson is produced in association with either a photon or a small-radius jet. Dedicated event selections are devised to target each of these production mechanisms, referred to as the ISR γ and ISR jet channels. The search is conducted in the so-called boosted regime, where the particles produced in the decay of the Z' to SM quarks are reconstructed as a single large-R jet. The large-R jet substructure observable τ_{21} is decorrelated from the large-R jet mass using the DDT technique. This provides a way of creating a signal-enhanced search region with minimal effect on the shape of the large-R jet mass distribution. A TF method is used to estimate the leading background processes in both search channels, using the jets from the signal-depleted substructure fail region. The analysis is validated *in situ* by performing a search for the W/Z peak in the large-R jet mass spectrum, as a "known signal" in data. The best-fit signal strength for the inclusive W/Z process is consistent with the SM value ($\mu = 1$) in both channels, and is used to constrain the W/Z background normalisation in the search. Signal mass hypotheses between $m_{Z'} = 100$ GeV and 220 GeV are tested, and 95% CL upper exclusion limits are set on the product of the cross-section of the benchmark signal model with its acceptance under the kinematic selection in each channel, as well as on the coupling of the Z' boson to SM quarks, g_q. No significant excesses are observed in either channel. This is the first "boosted dijet + ISR" search performed in ATLAS. In addition, this is the first ever search in the boosted ISR γ channel. The resulting limits on g_q are shown in Figure 11.1 along with results from complementary ATLAS searches. This analysis has extended the sensitivity of the ATLAS experiment to leptophobic DM mediator particles down to the W/Z mass.

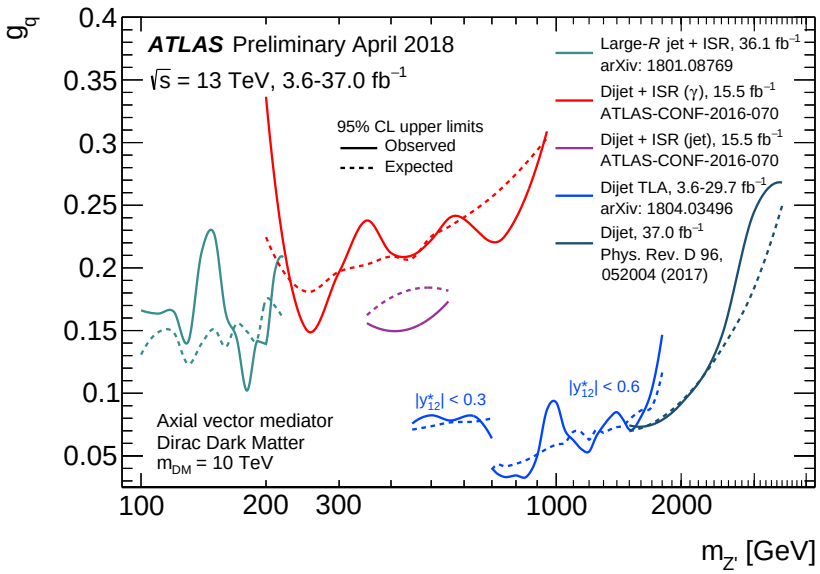


Figure 11.1 Summary plot of ATLAS searches for Dark Matter (DM) with mediator particles coupling to Standard Model (SM) quarks. The coloured lines show 95% confidence level (CL) expected and observed exclusion limits on the coupling of DM mediator particles to SM quarks g_q, as a function of the mediator particle mass $m_{Z'}$, resulting from analyses of the dijet and dijet + ISR experimental signatures as of early 2018. The limits are computed for a DM particle mass of $m_\chi = 10$ TeV. The results from this analysis are indicated in the left region ("large-R jet + ISR") and provide limits on the lowest DM mediator particle masses with high-mass DM particles ever probed in ATLAS. Figure from Ref. [123].

The main limitation of the analysis is the systematic uncertainty associated with the data-driven background estimate, derived from the variance function of the GP regression to the TF profile. This is mainly due to limitations of the DDT technique, which leads the TF "pass/fail" profile to have significant deviations from constancy even after decorrelation. This introduces a residual mass and p_{T}-dependence which necessitates short GP length scales, particularly in the ISR jet channel. This problem can be mitigated by employing a more sophisticated mass-decorrelation procedure. As an added benefit, this would also allow the analysis to use more powerful jet taggers than the τ_{21} observable, which was otherwise necessitated by the DDT method. Part III details a study of various approaches to the development of mass-decorrelated jet taggers which may benefit future iterations of this analysis.

Analyses from the CMS Collaboration suggest that using a k-nearest neighbours (k-NN) based mass-decorrelation method, see Part III, allows for a robust TF regression with only third order polynomials which may bring the uncertainty on the dominant backgrounds to the same level as *e.g.* large-R jet uncertainties [73, 214]. Other

analyses from the ATLAS Collaboration suggest that alternative background estimation procedures, such as directly fitting the dominant background process contribution in the large-R jet mass spectrum, may have similar potential for improvement [215]. Such a reduction could lead to an improvement in the limits on g_q of approx. 30% In addition, since the completion of this analysis, the ATLAS Collaboration has provided an *in situ* large-R jet calibration which reduces *e.g.* the jet energy scale uncertainty from approx. 8% to 1% [101, 102]. These are all promising developments which may benefit future iterations of this analysis.

PART III

Mass-decorrelated jet substructure taggers

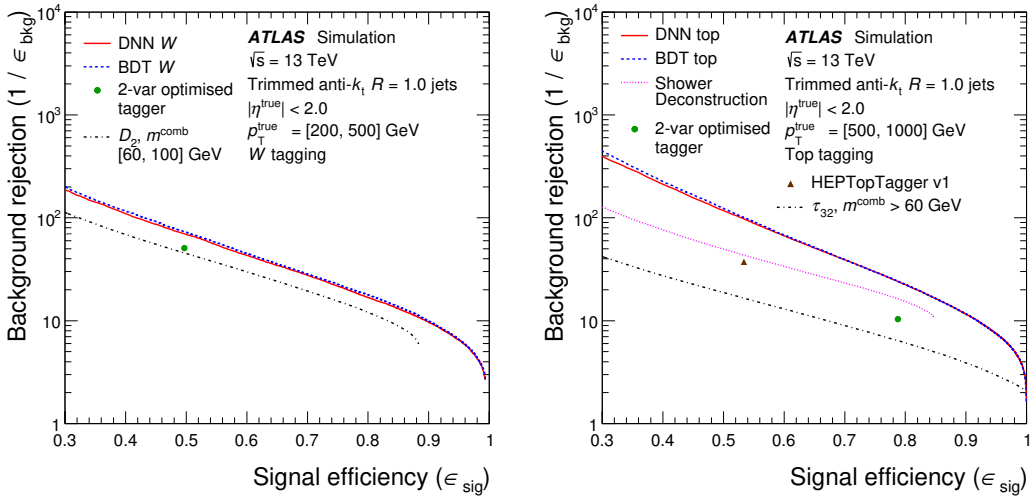
CHAPTER 12

Introduction and review

Part II described a search for Dark Matter (DM) mediator particles in proton–proton (pp) collisions in the ATLAS experiment in final states with large-radius (large-R) jets. This analysis relied on jet substructure for enhancing the purity of the hypothesised signal process and for estimating the leading background contribution in the search region. The chosen jet substructure observable (τ_{21}) was decorrelated from the large-R jet mass using the designed decorrelated taggers (DDT) method to simplify the background estimation. However, this method was found to be sub-optimal, leading to a relatively large systematic uncertainty associated with the leading background estimate.

This analysis is just one of many searches for physics beyond the Standard Model (BSM) which rely on the identification of hadronically decaying resonances reconstructed as jets in the calorimeters. These are characterised by vast backgrounds of non-resonant jet production, which makes the task of identifying jets from resonant hadronic decays crucial to searches for new physics [1, 73, 155–157, 216]. This identification relies on jet substructure observables, introduced in Section 1.3, which quantify the angular correlations between jet constituents. These correlations reflect differences in the radiation patterns for jets produced through resonant and non-resonant decays, and can therefore be used to distinguish the two. A substantial number of jet substructure variables have been proposed based on theoretical considerations, including τ_{21}, many of which have been used for jet classification in ATLAS [217–221], see Appendix A. Furthermore, it has been shown that improvements in the jet identification can be achieved through a combination of several jet substructure variables using multivariate analysis (MVA) approaches, in particular boosted decision trees (BDTs) and neural networks (NNs), which were introduced in Chapter 4.

In Refs. [39, 222], BDT- and deep neural network (DNN)-based W and top jet classifiers were compared to common analytically computed ('analytical') jet substructure

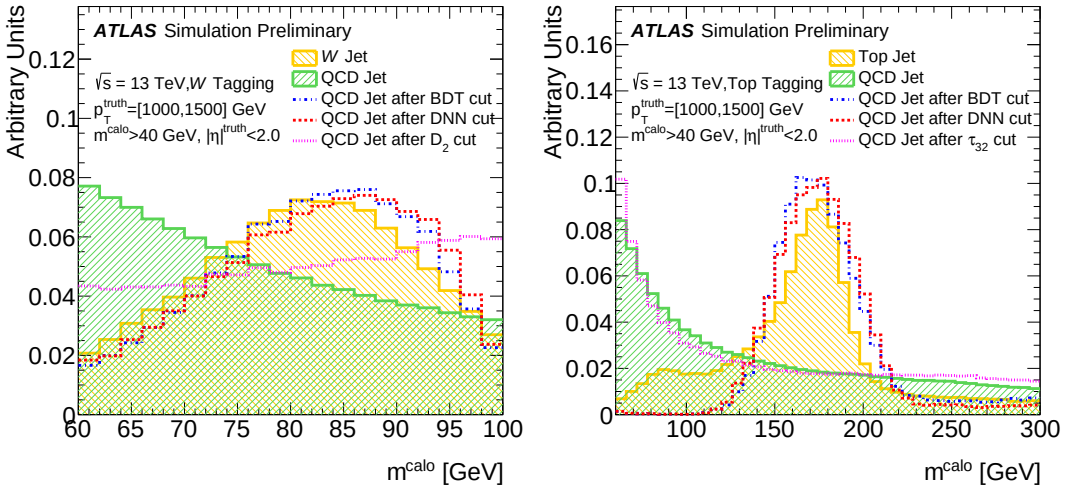


(a) W tagging **(b)** Top tagging

Figure 12.1 Classification performance of multivariate analysis (MVA)-based large-R jet taggers compared to analytical taggers for **(a)** W tagging and **(b)** top tagging. Figures from Ref. [39].

observables, such as the N-subjettiness ratio τ_{21} described in Section 1.3 and the ones described in Appendix A. The MVA taggers were tasked with performing binary classification of hadronically decaying W bosons or top quarks from non-resonant multijets based on multiple jet substructure observables. Both classes of MVA taggers were trained using a sample of jets with $m > 40$ GeV, requiring 200 GeV $< p_{\mathrm{T}} <$ 2000 GeV for W tagging and 350 GeV $< p_{\mathrm{T}} <$ 2000 GeV for top tagging, reweighted to uniformity in p_{T}. The hyperparameters of the BDT and NN classifiers, see Chapter 4, were optimised for each signal process to maximise the multijet rejection at a fixed signal efficiency of 50% for W tagging and 80% for top tagging. The MVA taggers were found to consistently improve classification compared to analytical approaches as shown in Figure 12.1. The BDT- and NN-based versions perform similarly well for all signal efficiencies studied, indicating that the use of information is comparable for these machine learning (ML) approaches.

However, it turns out that MVA-based large-R jet classifiers learn to exploit the fact that jet mass is a powerful feature for discriminating against the non-resonant background. Therefore, MVA classifiers trained using jet substructure variables to identify hadronically decaying resonances exhibit non-linear correlations with the reconstructed large-R jet mass. This means that a simple selection based on such MVA classifier observables tends to distort the background large-R jet mass distribution, making it resemble the signal jet mass distribution. This effect is shown in Figure 12.2



(a) *W* tagging **(b)** Top tagging

Figure 12.2 Large-R jet mass distribution for **(a)** W jets and **(b)** top jets, as well as multijets, before and after selection on either the deep neural network (DNN) or boosted decision tree (BDT)-based tagger, or an analytical jet substructure observable (D_2 or τ_{32}). Figures from Ref. [222].

for W and top tagging, which compares the ATLAS MVA taggers to a relevant analytical substructure observable in each case. For both signal processes, a threshold selection on the single-variable substructure observable moderately distorts the original multijet spectrum. This was what necessitated the mass-decorrelation procedure in the analysis described in Part II. In contrast, the MVA taggers — exploiting mass information to improve classification — result in multijet mass distributions which are close to degenerate with the signal distribution. This so-called sculpting effect also depopulates the jet mass side-band regions around the W boson or top quark mass. This complicates the background estimation, as the side-band regions are often used to fit a function to determine the expected background contribution in the search region, see Chapter 9. These sculpting effects have limited impact for identification of Standard Model (SM) resonances like hadronically decaying W/Z bosons or top quarks, where the resonance mass is known. However, it renders standard MVA jet classifiers less useful to searches for new hadronically decaying particles reconstructed as jets, where the masses of hypothesised resonances are not known *a priori*.

This third and final part will study and assess a number of techniques for decorrelating jet substructure classifiers from the large-R jet mass. Such mass-decorrelation will mitigate sculpting and thereby provide better sensitivity to searches for new resonances, such as future iterations of the analysis described in Part II. Methods for mass-decorrelation are considered for both analytical single-variable and MVA-based taggers.

In particular, DDT [189], fixed-efficiency k-nearest neighbours (k-NN) regression [223], and convolved substructure (CSS) [224] provide analytical methods for decorrelating a single jet substructure observable from the jet mass. The use of adversarial training of neural networks (ANN) [225] and adaptive boosting for uniform efficiency (uBoost) [226] have been proposed as ways to leverage the classification power of MVA taggers while reducing their correlation with the jet mass through specialised training methods. For concreteness, the identification of jets from the hadronic decay of W bosons is studied and compared to the dominant multijet background. Here, the W boson is used as a typical example of a resonance decaying to two quarks. However, the results should hold for other resonance masses as a direct result of the mass-decorrelation. Alternatively, the studied mass-decorrelation procedures could be applied for other resonance mass hypotheses. This should make the results in this part applicable to Z' bosons as well.

The Monte Carlo (MC) datasets used for the study are described in Chapter 13, along with the event selection, sample weights, and MVA tagger input features. In Chapter 14, the metrics chosen to quantify tagger classification power and mass-decorrelation are detailed. The five mass-decorrelation techniques considered in this study are described in Chapter 15. Finally, results for the various techniques are presented in Chapter 16.

CHAPTER 13

Datasets

In Chapter 12, the problem of correlations between jet taggers and the large-R jet mass was presented. The study of different techniques for constructing mass-decorrelated jet substructure taggers is performed using two MC samples, described in this chapter. Furthermore, the event selection used in this study, the event weighting schemes, and the choice of jet substructure observables for the ML-based jet taggers is detailed.

13.1 Simulated samples

The background process is taken to be QCD multijet production, which simulates the non-resonant production of jets predominantly originating from gluons and light-flavour quarks. This sample is simulated using the Pythia8 (v8.186) [165] generator with the NNPDF2.3LO [227] parton distribution function (PDF) set and the A14 tune [166], which is an ATLAS-specific optimisation of the Pythia8 parton shower and hadronisation models aimed *e.g.* at underlying event (UE) and jet substructure. The signal process is taken to be the decay of a high-mass W' resonance to vector bosons, in turn decaying hadronically: $W' \to WZ \to qqqq$. This allows for generating a large sample of hadronically decaying W bosons with sufficient p_{T} to let the decay products be reconstructed as large-R jets. Signal events are similarly generated using Pythia8 with the NNPDF2.3LO PDF set and the A14 tune, for W' mass values ranging from 400 GeV to 5 TeV, to populate of the region between 200 GeV and 2 TeV in W jet p_{T}. The simulated signal sample is reweighted according to the generator-level jet p_{T} to a distribution similar to that of the background sample. This removes the dependence on the choice of signal model used to produce the hadronic W jets.

The detector response is modelled with a detailed simulation of the ATLAS de-

tector [171] based on GEANT4 [172]. The MC samples are overlaid with additional pile-up events, generated using PYTHIA8 with the A2 tune [173] and MSTW2008LO PDF set [174] to roughly match the 2015 and 2016 data pile-up conditions with a mean number of 24 collisions per bunch crossing. The simulated events are processed using the same reconstruction algorithms, calibrations, *etc*. as for recorded data.

13.2 Reconstruction and event selection

Reconstruction and calibration of hadronic jets in the ATLAS experiment is described in Section 7.2. In this study, the jets are reconstructed from topological clusters using the anti-k_t algorithm [30] as implemented in FASTJET [181] with a distance parameter of $R = 1.0$. To remove the contribution of soft radiation, either from the UE or from pile-up, and to improve the jet mass resolution, the reconstructed jets are groomed using the jet trimming algorithm [36] with parameters $R_{\text{sub}} = 0.2$ and $f_{\text{cut}} = 5\%$, see Section 1.3. The trimming procedure yields a subset of constituent topological clusters from which substructure observables, including the calorimeter jet mass, are computed. In addition to energy deposits in the calorimeter, the charged jet constituents leave tracks in the inner detector (ID). The so-called track-assisted jet mass m^{TA} [33] can be computed as $m^{\text{TA}} = (p_{\text{T}}^{\text{calo}}/p_{\text{T}}^{\text{track}}) \times m^{\text{track}}$, where $p_{\text{T}}^{\text{calo}}$ is the transverse momentum of the trimmed calorimeter jet and m^{track} and $p_{\text{T}}^{\text{track}}$ are the invariant mass and p_{T}, respectively, of the four-momentum sum of tracks associated with the jet. Calorimeter jet resolution mass degrades with p_{T}, and for W jets with $p_{\text{T}} \gtrsim 1$ TeV the track-assisted jet mass has better resolution [33], see Figure 13.1. From the two, a combined jet mass (m) is defined as the average weighted by the relative resolutions as a function of p_{T}, *i.e*.

$$m \equiv m^{\text{comb}} = \frac{w^{\text{calo}}\, m^{\text{calo}} + w^{\text{TA}}\, m^{\text{TA}}}{w^{\text{calo}} + w^{\text{TA}}} \qquad \text{with} \quad w^{\text{calo}} = \sigma_{\text{calo}}^{-2},\ w^{\text{TA}} = \sigma_{\text{TA}}^{-2}\,, \tag{13.1}$$

where σ_{calo} and σ_{TA} are the jet mass resolutions of the calorimeter-based and track-assisted jet mass definitions, respectively, as a function of the jet p_{T}. The combined mass improves the jet mass resolution across a large range of jet p_{T}, and is therefore used as the default jet mass variable throughout this study.

From the simulated MC samples, a baseline selection is imposed on all jets to ensure that they are well-reconstructed and representative, either of the hadronic decay of a boosted W boson (signal) or of non-resonant multijet production (background). In order to have a correct jet-by-jet labelling, a separate set of jets are reconstructed from the stable simulated particles with a lifetime $c\tau > 10$ mm, excluding muons, neutrinos, and

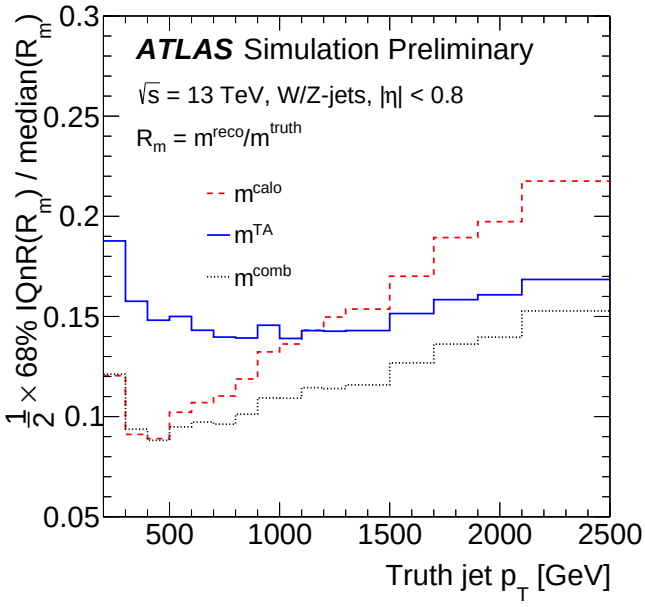


Figure 13.1 Mass resolution for jets produced in the hadronic decay of W bosons, as a function of the truth-level p_{T}. The resolution is shown for the calorimeter-based jet mass (m^{calo}), the track-assisted mass (m^{TA}), and the combined mass (m^{comb}). Figure from Ref. [33].

pile-up activity. These so-called truth jets are reconstructed using the anti-k_t algorithm with $R = 1.0$ similar to the calorimeter jets, but without the application of any jet grooming algorithms. Calorimeter jets are paired with truth jets by a matching in $\eta - \phi$. To define a clean ensemble of reconstructed W jets, a three-step matching procedure identical to that in Refs. [222, 228] is performed. This procedure requires the W and both quarks from the hadronic decay process $q_{1,2}$ to be within $\Delta R < 0.75$ of the truth jet J^{truth} corresponding to a particular reconstructed jet for it to be labelled as a signal jet. In each event, the two highest-p_{T} jets are selected, provided they satisfy $p_{\mathrm{T}}^{\mathrm{truth}} > 200$ GeV and $|\eta^{\mathrm{truth}}| < 2$, in order to obtain a realistic kinematic regime and containment within the central detector, respectively. In signal events, only jets satisfying the truth-jet–matching are selected

From the jets satisfying the truth-level selection, the initial reconstruction-level selection retains only events with at least one reconstructed primary vertex (PV) formed from at least two reconstructed tracks in the inner detector. From these, a subset of events is selected to focus on the kinematic regime relevant to the searches for new physics in the large-R jet mass spectrum: reconstructed jets are required to have a combined mass in the range $m \in [50, 300]$ GeV and a reconstructed transverse momentum in the range $p_{\mathrm{T}} \in [200, 2000]$ GeV. These kinematic bounds are chosen to simplify and concretise the mass-decorrelation task, in a way which is consistent with the analysis

	Multijets	W jets
Reconstruction-level event selection		
N_{PV}	≥ 1	
$N_{\text{track}\,\vert\,\text{PV}}$	≥ 2	
Truth-level jet selection		
$\Delta R(J^{\text{truth}}, x)$ $x = W, q_1, q_2$	—	< 0.75
$\vert\eta^{\text{truth}}\vert$	< 2	
$p_{\text{T}}^{\text{truth}}$ [GeV]	> 200	
Reconstruction-level jet selection		
N_{const}	> 2	
p_{T} [GeV]	$[200, 2000]$	
m [GeV]	$[50, 300]$	

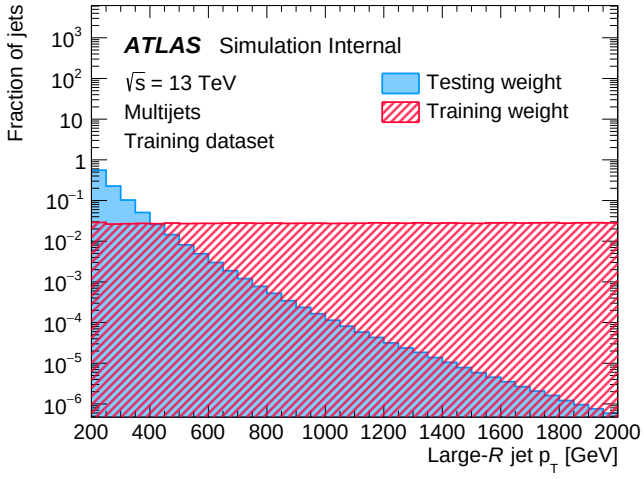
Table 13.1 Summary of the baseline selection applied to the multijet and W jet samples. See text for details.

presented in Part II of this thesis. Finally, all jets are required to have more than two constituent clusters ($N_{\text{const}} > 2$) such that all substructure observables are well-defined. Across p_{T}, the number of multijets failing this requirement is less than 1%. Among W jets, less than 1% of jets fail this requirement at $p_{\text{T}} = 200$ GeV and approximately 3% at $p_{\text{T}} = 2$ TeV. This selection, summarised in Table 13.1, defines the samples of jets based on which jet tagger performance is evaluated throughout this study.

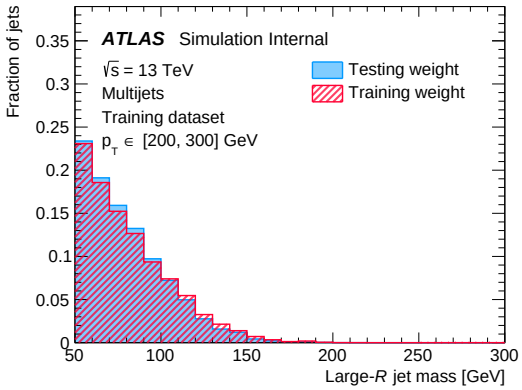
After the baseline selection, a class-balanced sample of one million (1M) signal and 1M background jets are retained for training. In addition, a separate set of 1M signal and 10M background jets are used for final performance evaluation, or "testing," to have sufficient statistics for differential studies in Chapter 16. Throughout Chapter 15 and Appendices D and E, the training dataset is used unless explicitly noted otherwise. Conversely, in Chapter 16, only the testing dataset is used, such that all results in the study are reported on data unseen during the optimisation and training of each jet tagger.

13.3 Sample weights

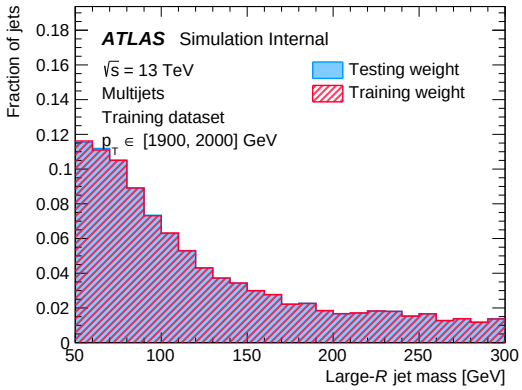
For performance evaluation throughout the study, the multijets are weighted by cross-section, resulting in a physical, smoothly falling p_{T}-spectrum. As mentioned above, the p_{T} distribution for the W jets is reweighted to resemble that of the multijets. When training the MVA classifiers, however, the signal and background jets are separately



(a) Jet transverse momentum.



(b) Jet mass, $p_{\mathrm{T}} \in [200, 300]$ GeV



(c) Jet mass, $p_{\mathrm{T}} \in [1900, 2000]$ GeV

Figure 13.2 Normalised distributions of the large-R jet **(a)** transverse momentum (p_{T}) and **(b, c)** mass for multijet events in the training dataset, weighted either by the training or the testing event weights.

reweighted to a uniform p_{T} distribution. This is done using the `BinsReweighter` tool from the HEP_ML (v0.5.0) library [229]. Reweighting to uniform p_{T} provides equal training attention across the entire p_{T} range of $[200, 2000]$ GeV. The reweighting is performed starting from the cross-section–weighted distributions. This guarantees that physical distributions of all other variables are retained as a function of p_{T}, while only the p_{T} distribution is affected. Additionally, the training weights are normalised to have a mean of one for the W jet and multijets samples separately, in order to achieve class-balanced training of the MVA jet taggers. Figure 13.2 shows large-R jet p_{T} and mass distributions for the multijet sample, weighted according to the training and testing weights.

Figure 13.2a shows the effect of reweighting the large-R jet sample to achieve a uniform

Variable	Type	Reference(s)
τ_{21}	N-subjettiness	[34]
C_2, D_2	Energy correlation function ratios	[230]
a_3	Angularity	[231]
A	Aplanarity	[232]
$\mathcal{P}$	Planar flow	[231]
R_2^{FW}	Fox–Wolfram moment	[233]
$KtDR$	k_t-subjet ΔR	[28]
$\sqrt{d_{12}}$, z_{12}	Splitting scales	[234, 235]

Table 13.2 Substructure variables used for the neural network (NN) and boosted decision tree (BDT)-based jet classifiers in this study. Choice of features is based on the 'DNN' selection in Ref. [39].

p_{T} distribution. Figures 13.2b and 13.2c show the large-R jet mass distribution for multijets in two bins of p_{T} around the lower and upper selection bounds, see Table 13.1. Using this flat-p_{T} reweighting, the large-R jet mass distribution with training weights is identical to that with testing weights, in sufficiently small bins of p_{T}. The same is true for all other kinematic and substructure variables. For distributions inclusive in p_{T}, this is not the case. Similar results hold for the sample of W jets.

13.4 Choice of features

For the two ML-based jet tagger algorithms presented in Chapter 15, a suitable set of input features must be selected. This need not be the largest possible set of features, as the added complexity of an excessive set of features may complicate the task of the ML model and not lead to improvements in performance. Therefore, feature selection may be treated similarly to hyperparameter optimisation, see Appendix E.

This study focuses on jet substructure variables as highly engineered, physics-motivated input features. The results of the 'DNN' feature selection in Ref. [39] are used in lieu of a full input feature optimisation due to the similarity of the problem and the datasets used. Additionally, this provides as direct a comparison to other ATLAS results as possible. The 10 jet substructure variables used in the training of the NN and BDT-based jet classifiers in this thesis are listed in Table 13.2, and their distributions for signal and background jets are shown in Figure 13.3. The N-subjettiness ratio τ_{21} was described in Chapter 1.3 and employed in the analysis in Part II of this thesis. The remaining substructure observables used in this study are described in detail in Appendix A.

Figure 13.3 show how, to a greater (*e.g.* D_2) and lesser extent (*e.g.* a_3), each of these variables provide some separation of the two studied processes. The 10 jet substructure observables used in this study all probe differences in radiation patterns inside jets originating from different processes, and the physical motivation for the observed behaviours of these variables are discussed in Appendix A. Each of these variables may serve as a W jet tagger on their own, but their mutual correlations may provide additional information which may be exploited by an MVA tagger.

The linear correlation coefficients for each possible pair of the jet substructure variables listed in Table 13.2 are shown in Figure 13.4 for both classes of jets. Groups of variables which are conceptually similar are generally also correlated (*e.g.* energy correlation functions variables such as C_2 and D_2, but also τ_{21}; and the splitting scales $\sqrt{d_{12}}$ and z_{12}), while other pairs of variables are not, despite addressing the same issue of identifying hadronic two-body decays. Therefore, despite their similarities, these variables are complementary in the sense that using the full set of 10 jet substructure observables enables better jet classification than any subset [222]. This indicates that they elucidate slightly different aspects of the substructure of the jets in this study, which MVA-based taggers are able to exploit.

In contrast to Ref. [39], the large-R jet mass and p_{T} are not used as input features. Using only substructure observables as input features to the MVA-based jet taggers is intended to provide a more direct comparison to the analytical mass-decorrelated taggers described in Section 15, all of which are based on single substructure observables. Although the training and testing datasets are reweighted to have identical p_{T} distributions for signal and background, including the jet p_{T} as an input feature could allow MVA taggers to better and more directly utilise p_{T}-dependent information about the different substructure observables.

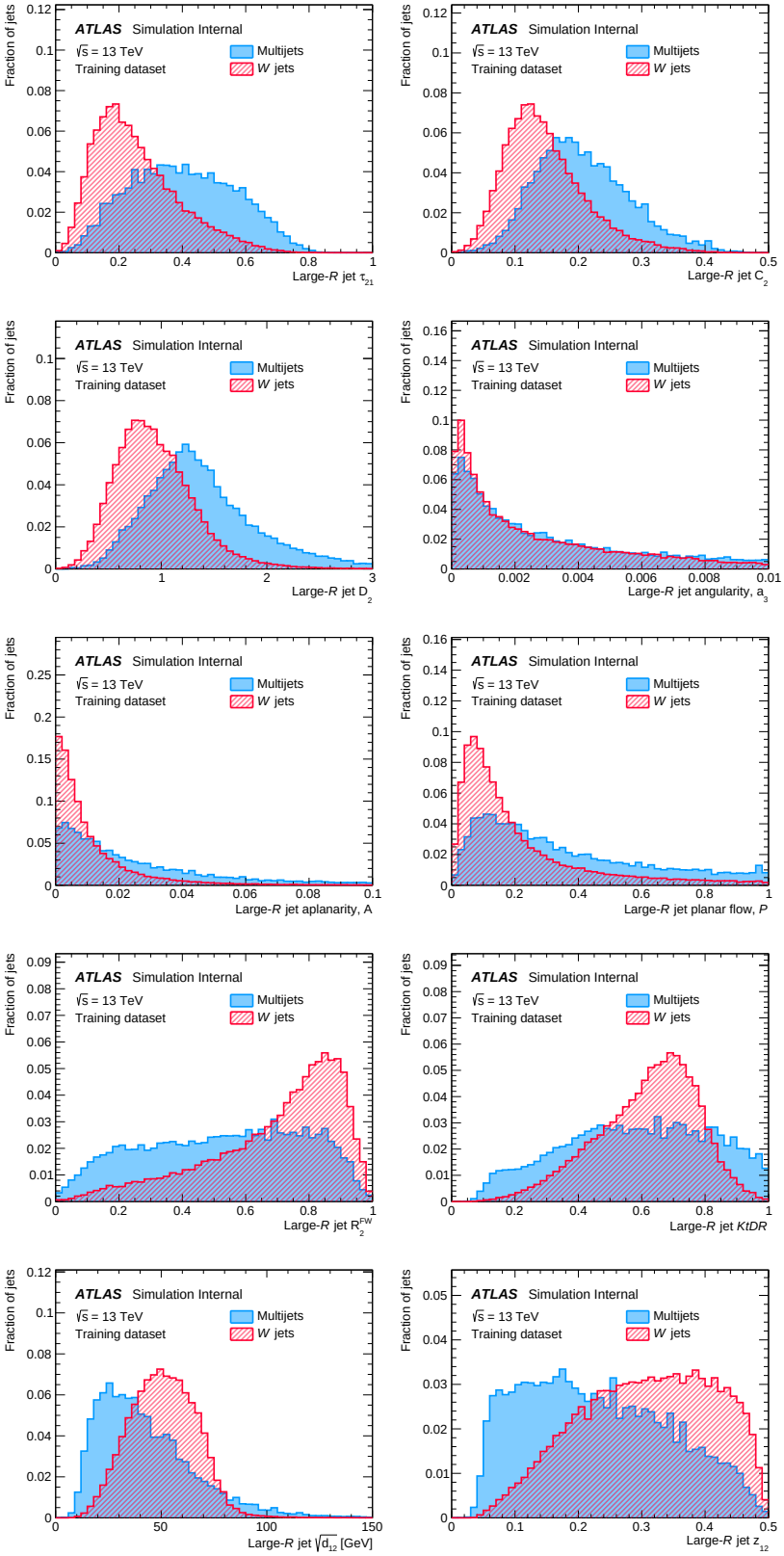
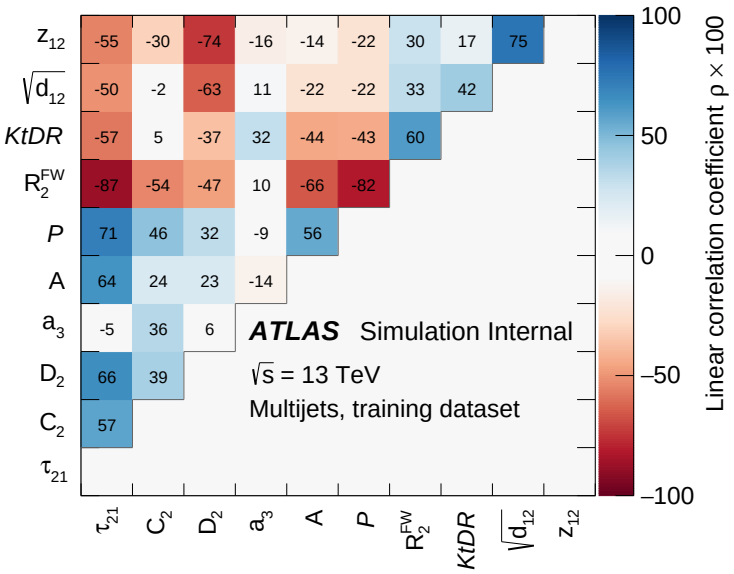
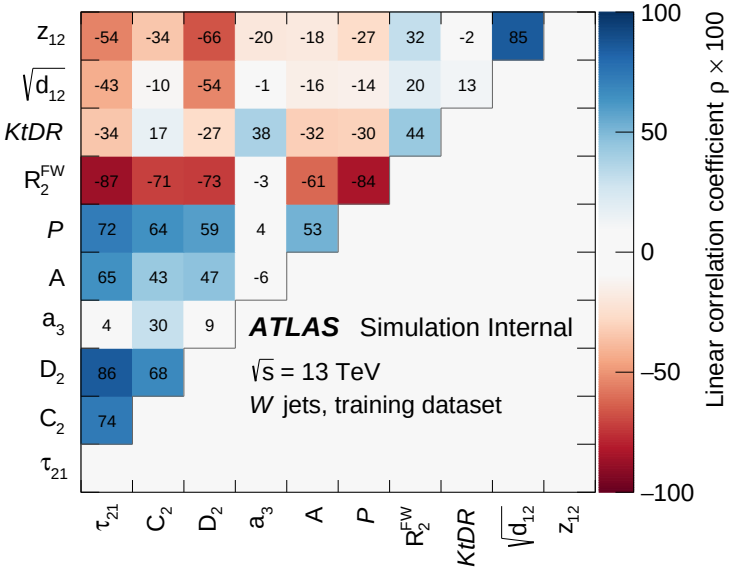


Figure 13.3 Distributions of the jet substructure variables used for the neural network (NN) and boosted decision tree (BDT)-based jet classifiers in this study, for multijets and W jets, see also Table 13.2.



(a) Multijets



(b) *W* jets

Figure 13.4 Linear correlation coefficient matrices of all pairs of the jet substructure variables listed in Table 13.2, for **(a)** multijets and **(b)** *W* jets.

CHAPTER 14

Evaluation metrics

The standard objective when developing jet substructure taggers is to increase their classification power, which has been done in previous jet tagging studies in ATLAS [39, 222]. However, in the development and study of mass-decorrelated jet taggers, the evaluation of classification performance must be complemented by a metric for quantifying the degree of mass-decorrelation. As classification and mass-decorrelation are generally opposing objectives, any combined figure of merit will be ambiguous and subject to a use case–dependent weighting of the two tasks. Below, the chosen metrics for classification and mass-decorrelation performance are described. Alternative metrics for mass-decorrelation, such as the Kolmogorov–Smirnov test [236, 237] (measuring the largest difference in cumulative density functions (c.d.f.s) between two distributions) or the Wasserstein or "earth-movers" distance (measuring the minimal amount of information transport required to convert one probability density function (p.d.f.) into another), could be used as well. However, the chosen mass-decorrelation metric, the Jensen-Shannon divergence (JSD), is motivated by the approach chosen for the adversarial neural network (ANN) training described in Section 15.5.

14.1 Classification

Classification performance is measured by the selection efficiency of signal jets $\varepsilon_{\text{sig}}^{\text{rel}}$ and the associated background rejection factor $1/\varepsilon_{\text{bkg}}^{\text{rel}}$ for a particular threshold selection on the jet tagger in question, relative to the inclusive sample after the baseline selection in

Chapter 13. These measures are defined as

$$\varepsilon_{\text{sig}}^{\text{rel}} = \frac{N_{\text{sig}}^{\text{tagged}}}{N_{\text{sig}}^{\text{total}}} \quad \text{and} \quad \varepsilon_{\text{bkg}}^{\text{rel}} = \frac{N_{\text{bkg}}^{\text{tagged}}}{N_{\text{bkg}}^{\text{total}}}, \tag{14.1}$$

where N^{total} refers to the total number of events of a given class passing the baseline selection, and N^{tagged} refers to the number of events passing both the baseline selection and the tagging selection. The background rejection at $\varepsilon_{\text{sig}}^{\text{rel}} = 50\%$ is used as a summary metric to evaluate the classification performance. The chosen 50% working point is typical for jet tagging studies [39, 222], and is indicative of the selection efficiency aimed for by searches in this kinematic regime [1, 156].

14.2 Mass-decorrelation

The linear correlations between the jet mass, substructure variables, and MVA jet classifier outputs were studied in Ref. [222]. However, a linear correlation coefficient has insufficient capacity to express the highly non-linear correlations observed between the jet mass and existing MVA jet taggers [39, 222], see also Figure 12.2. Therefore, it is not an ideal metric for the correlation with the jet mass, which is the main subject of this study. Instead, a figure of merit which directly quantifies the sculpting of the background jet mass distribution, caused by a threshold selection on a jet tagger observable, is proposed. Concretely, distributions of jet masses between 50 GeV and 300 GeV, with a bin width of 5 GeV, will be used.

Kullback-Leibler divergence

The Kullback-Leibler (KL) divergence [238] for discrete probability distributions P, F is defined as

$$\text{KL}(P \parallel F) = -\sum_i P_i \log_n\left(\frac{F_i}{P_i}\right), \tag{14.2}$$

where i enumerates the discrete bins of the distributions and n is the base of the logarithm. In this study, these probability distributions are taken to correspond to normalised jet mass distributions. The KL divergence measures the relative entropy of P with respect

to F, since

$$\mathrm{KL}(P \parallel F) = -\sum_i P_i \log_n F_i + \sum_i P_i \log_n P_i = H(P, F) - H(P). \tag{14.3}$$

Here, $H(P)$ is the entropy of the discrete probability distribution P, and $H(P, F)$ is the cross-entropy of distributions P and F. For identical distributions P and F, each summand in Equation (14.2) will be zero, the cross-entropy of P and F will be equal to the entropy of P, and therefore the KL divergence will be zero. The larger the difference between P and F, the larger the cross-entropy $H(P, F)$, and the larger $\mathrm{KL}(P \parallel F)$ will be. The KL divergence can therefore be used to measure the similarity of discrete distributions. However, the KL divergence is prone to numerical instabilities: for any bin i where $P_i > 0$ and $F_i = 0$, the cross-entropy $H(P, F)$ will go to infinity. Similarly, the KL divergence is asymmetric with respect to its arguments. However, the metric for mass-decorrelation is intended to measure the differences between jet mass distributions for jets passing and failing a certain selection on a jet tagger observable. The same two jet mass distributions ought to yield the same metric value, regardless of which is designated as 'pass' or 'fail.' Therefore, the chosen metric should ideally be symmetric with respect to its arguments.

Jensen-Shannon divergence

The Jensen-Shannon divergence (JSD) [239] is a generalisation of KL which avoids the instabilities mentioned above and symmetrises the metric with respect to P and F:

$$\mathrm{JSD}(P \parallel F) = \frac{1}{2}\big(\mathrm{KL}(P \parallel M) + \mathrm{KL}(F \parallel M)\big), \qquad \text{with} \quad M = \frac{P + F}{2} \tag{14.4a}$$

$$= H(M) - \frac{1}{2}\left[H(P) + H(F)\right]. \tag{14.4b}$$

In the case of identical distributions P and F, $H(M) = H(P) = H(F)$ and $\mathrm{JSD}(P \parallel F) = 0$. The converse case of maximal sculpting arises when the distributions P and F have no common support, *i.e.* no overlapping non-zero bins. In this case, $H(M) = (H(P) + H(F))/2 + \log_n 2$ meaning that $\mathrm{JSD}(P \parallel F) = \log_n 2$. Therefore, by using the logarithm base $n = 2$ in Equation (14.2), JSD will be in the range $[0, 1]$ with smaller values indicating less sculpting of the background jet mass distributions; and *vice versa*.

In this study, the JSD is used to measure the difference between the normalised mass distributions of the background jets passing and failing a given threshold selection on a

jet tagging variable

$$\mathrm{JSD} \equiv \mathrm{JSD}(P \parallel F) = \mathrm{JSD}\left(N^{\mathrm{pass}}_{\mathrm{bkg},i} \,/\, \textstyle\sum_j N^{\mathrm{pass}}_{\mathrm{bkg},j} \,\middle\|\, N^{\mathrm{fail}}_{\mathrm{bkg},i} \,/\, \textstyle\sum_j N^{\mathrm{fail}}_{\mathrm{bkg},j}\right), \tag{14.5}$$

where i and j enumerate the bins in the large-R jet mass spectrum. This metric will hereafter be referred to simply as JSD for any given signal or background selection efficiency. For summary performance evaluation, JSD or $1/\mathrm{JSD}$ at $\varepsilon^{\mathrm{rel}}_{\mathrm{sig}} = 50\%$ will be used to quantify the mass-decorrelation of a given jet classifier.

To illustrate the two divergences discussed above, Figure 14.1 shows the calculation of JSD for the τ_{21} jet substructure observable, see Chapter 1. The top panel shows the normalised jet mass distributions for jets with $p_{\mathrm{T}} \in [500, 1000]$ GeV passing (P) and failing (F) a threshold selection on τ_{21}, with a threshold value chosen such that $\varepsilon^{\mathrm{rel}}_{\mathrm{sig}} = 50\%$. Also shown is the average of the two distributions (M), used in Equation 14.4a. The two middle panels show the bin-wise KL divergence summands, computed for the the pass and fail distributions P and F relative to the mean distribution M, see also Equation (14.2). The bottom panel shows the corresponding JSD summands found by taking the bin-wise average of the two KL divergences above, see Equation (14.4a). Finally, the cumulative sum of the JSD summands is shown in violet in the bottom panel, such that the total value of $\mathrm{JSD}(P \parallel F)$ can be read off at the far right end of the curve. Characteristically, the JSD summands are non-negative, they are largest in regions where the normalised P and F distributions are the most discrepant, and zero in the regions where they cross or overlap. This shows how the JSD is a suitable metric for quantifying localised morphological differences between normalised, binned jet mass distributions.

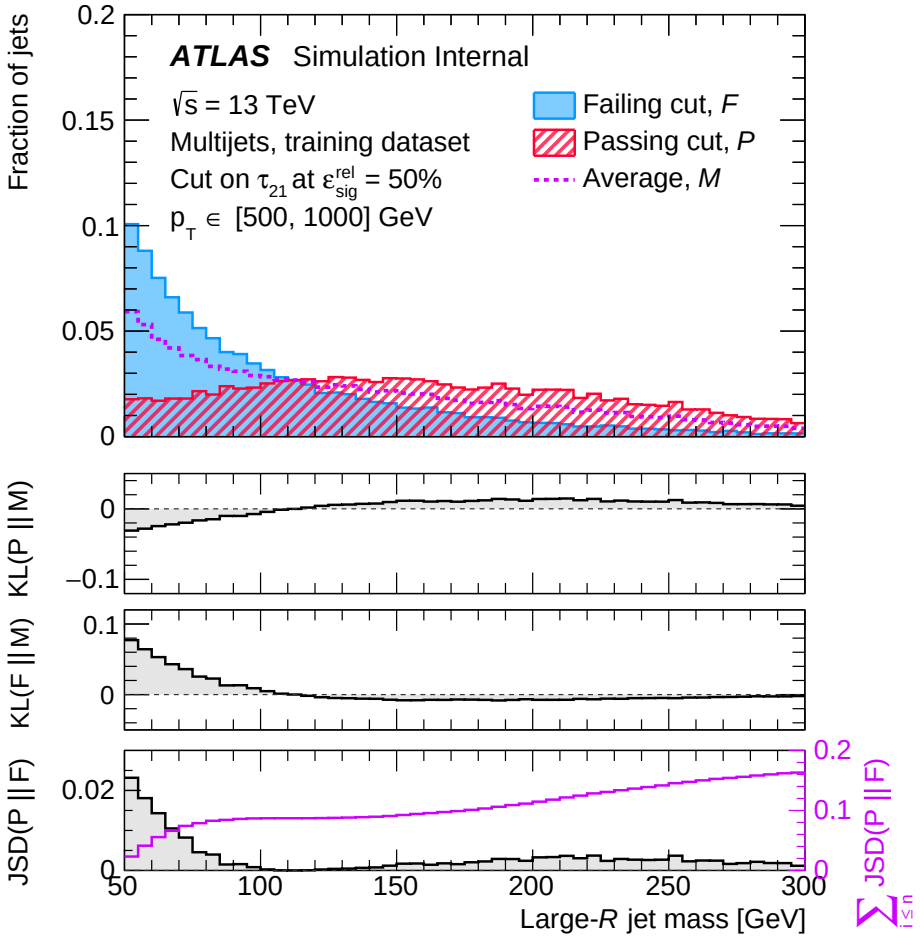


Figure 14.1 Distribution of the large-R jet mass for multijet events with $p_{\mathrm{T}} \in [500, 1000]$ GeV in the training dataset, either passing or failing a selection on τ_{21} chosen to give a signal efficiency of $\varepsilon_{\mathrm{sig}}^{\mathrm{rel}} = 50\%$, along with the associated Kullback-Leibler (KL) divergences and Jensen-Shannon divergence (JSD). See the text for details.

CHAPTER 15

Mass-decorrelation techniques

This chapter introduces the five mass-decorrelation techniques studied in this thesis. Particular attention is paid to ANNs, which is a promising algorithm for exploiting the classification power of ML while minimising jet mass sculpting. The first four ("alternative") mass-decorrelation techniques are described in more depth in Appendix D. Finally, additional details on the ANN training characteristics and hyperparameter optimisation are provided in Appendix E.

15.1 Designed decorrelated taggers

The simplest mass-decorrelation method studied in this thesis is DDT [189], introduced in Chapter 8 and used in the analysis in Part II. This method observes that, for the background process, the average value of τ_{21} is linear as a function of the kinematic variable $\rho^{\text{DDT}} = \log\left(m^2/(p_{\text{T}} \times \mu)\right)$, with $\mu = 1$ GeV, in the range $\rho^{\text{DDT}} \in [1.5, 4.0]$. By performing a linear fit with slope a, and correcting for this dependence as

$$\tau_{21}^{\text{DDT}} = \tau_{21} - a \times (\rho^{\text{DDT}} - 1.5), \tag{15.1}$$

a new mass- and p_{T}-decorrelated jet tagger τ_{21}^{DDT} can be defined. This method only corrects for the mean bias and is limited by the validity of the linear approximation, which breaks down in the low- and high-mass limit. Similarly, the DDT method is limited to τ_{21}, which is the only jet substructure observable known to exhibit this characteristic linear dependence.

15.2 Fixed-efficiency regression

Fixed-efficiency k-NN regression can be considered a non-parametric generalisation of the DDT method. It aims to construct a substructure observable which, for the background process, is decorrelated from the jet mass and p_{T} for a selection with a particular efficiency $\varepsilon_{\mathrm{bkg}}^{\mathrm{rel}}$. This study uses the D_2 substructure observable and a target background efficiency of $\varepsilon_{\mathrm{bkg}}^{\mathrm{rel}} = 16\%$, corresponding to a signal efficiency of $\varepsilon_{\mathrm{sig}}^{\mathrm{rel}} = 50\%$. The 16$^{\mathrm{th}}$ background efficiency percentage of D_2 is measured in bins of (ρ, p_{T}), with $\rho = \log\left(m^2/p_{\mathrm{T}}^2\right)$, and fitted using distance-weighted k-NN regression [223] with $k = 5$, as implemented in the SCIKIT-LEARN (v0.19.1) library [201]. By subtracting the fitted profile $D_2^{(16\%)}(\rho, p_{\mathrm{T}})$ from the D_2 observable

$$D_2^{k\text{-NN}} = D_2 - D_2^{(16\%)}(\rho, p_{\mathrm{T}}), \tag{15.2}$$

the mass- and p_{T}-decorrelated observable $D_2^{k\text{-NN}}$ is obtained. This observable is perfectly decorrelated, within the statistical uncertainties, for the chosen dataset and selection efficiency. Deviations from these may lead to a breakdown of the decorrelation.

15.3 Convolved substructure

The DDT and k-NN methods can be considered first-order corrections, removing the mass- and p_{T}-dependence of the mean or of a particular percentage. The CSS method [224] attempts to also remove the dependence of higher-order moments, *e.g.* the width of a particular substructure observable distribution. In this study, D_2 is used as base observable. The decorrelation is performed by morphing the D_2 distribution in bins of the large-R jet mass to match the distribution at a reference mass m_{ref}. This morphing is done by convolving the D_2 distribution with a Γ-distribution, $F_{\mathrm{CSS}}(D_2 \,|\, \alpha, \Omega_D)$, which results in the mass-decorrelated D_2^{CSS} distribution. Here, α is a shape parameter and Ω_D controls the scale of the morphing, and both parameters are optimised in each mass-bin through a χ^2-minimisation of the morphed D_2^{CSS} distribution to the D_2 distribution at m_{ref}. This method requires sufficient statistics to have smooth distributions suitable for morphing, and the discrete mass-binning may introduce artificial discontinuities. Finally, this method only decorrelates D_2 with respect to the large-R jet mass and not the p_{T}, in contrast to the two above methods.

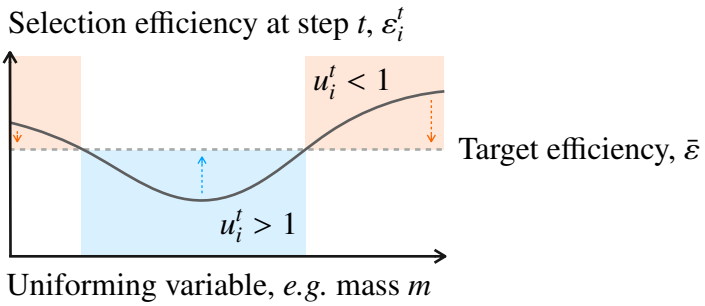


Figure 15.1 Illustration of the value of the uniforming weight u_i^t at boosting step t, depending on the value of the selection efficiency ε_i^t relative to the target efficiency $\bar{\varepsilon}$. The effect of u_i^t is to drive ε_i^t towards $\bar{\varepsilon}$ for increasing t.

15.4 Adaptive boosting for uniform efficiency

The first ML-based mass-decorrelation method builds on adaptive boosting as introduced in Chapter 4 and detailed in Appendix B. At every boosting step t, the standard AdaBoost algorithm [122] updates the relative weight w_i^t of each training example i as

$$w_i^{t+t} = w_i^t \times c_i^t \,, \tag{15.3}$$

where the classification weight c_i^t is based on whether the i^{th} sample was misclassified by the decision tree (DT) at step t. This results in the jet tagger $z_{\text{AdaBoost}} \in [0, 1]$. The uBoost method [226] extends this method by introducing the notion of a target background selection efficiency $\bar{\varepsilon}$. This study uses $\bar{\varepsilon} = 8\%$, corresponding roughly to a signal efficiency of $\varepsilon_{\text{sig}}^{\text{rel}} = 50\%$. The method computes a uniformity weight u_i^t which is less than one if the local background selection efficiency around the i^{th} training example is greater than the target efficiency at boosting step t, *i.e.* $\varepsilon_i^t > \bar{\varepsilon}$; and *vice versa*, *i.e.* $u_i^t > 1$ if $\varepsilon_i^t < \bar{\varepsilon}$ to give increased weight to training examples in regions with a selection efficiency below the target efficiency. This is illustrated in Figure 15.1.

By changing the weight update to

$$w_i^{t+t} = w_i^t \times c_i^t \times u_i^t \,, \tag{15.4}$$

the resulting BDT jet tagger observable z_{uBoost} is trained to provide uniform background selection efficiency. The importance of the uniformity weight u_i^t is controlled by the so-called uniforming rate α, as $\log u_i^t \propto \alpha$, which therefore allows for trading off classification and mass-decorrelation. For $\alpha = 0$, all uniforming weights u_i^t

are equal to 1 and the standard AdaBoost algorithm is recovered. Conversely, for increasing values of α, the uniforming weights u_i^t become more important relative to the classification weights c_i^t in Equation (15.4). The `uBoostBDT` class from the HEP_ML (v0.5.0) library [229] is used for the implementation of both the AdaBoost and uBoost taggers. Both BDT classifiers use the substructure variables listed in Table 13.2 as input features.

15.5 Adversarial neural networks

In this section, ANNs are introduced as a method for training mass-decorrelated NN jet taggers. In Ref. [240], adversarial training was proposed to make NN classifiers independent of certain variables or parametrised systematic uncertainties. Of particular relevance to this study, Ref. [241] then proposed the use of adversarial training to reduce the correlation with the jet mass.

Intuition

First, a classifier NN is be constructed and trained as in Chapter 4. This NN takes N substructure variables as input, outputs a jet tagging variable z in the range $[0, 1]$, and is trained stand-alone to minimise a classification loss L_{clf}. This introduces a correlation with the jet mass, as discussed in Chapter 12. To mitigate this problem, a second NN called the 'adversary' is introduced. The adversary is tasked with inferring the jet mass m from the output z of the classifier by minimising its own loss L_{adv}, which quantifies its ability to perform this task. If the adversary is able to infer the jet mass from the classifier output beyond random guessing, some non-linear correlation must exist between the two.

Optimising the weights θ_{clf} of the classifier according to the classification loss L_{clf}

$$\min_{\theta_{\text{clf}}} L_{\text{clf}}(\theta_{\text{clf}}) \tag{15.5a}$$

is a static problem. This means that there exist stable (local) minima in the $L_{\text{clf}}(\theta_{\text{clf}})$ "landscape" into which the classifier can be led using back-propagation with a stochastic gradient descent algorithm, see Chapter 4. Similarly, for a classifier with fixed weights

θ_{clf}, the weights θ_{adv} of the adversary can be optimised according to L_{adv}

$$\min_{\theta_{\text{adv}}} L_{\text{adv}}(\theta_{\text{clf}}, \theta_{\text{adv}}), \tag{15.5b}$$

which improves the ability of the adversary to infer the jet mass from the classifier output. The adversary loss L_{adv} depends also on the classifier weights θ_{clf}, since the adversary's task is conditional on a specific classifier weight configuration. Nevertheless, the minimisation in Equation (15.5b) is performed only with respect to θ_{adv}, since these are the only weights which the adversary can update. For fixed θ_{clf}, this is also a static optimisation problem.

The combined adversarial training of the classifier and adversary NNs is then designed to proceed by optimising θ_{clf} and θ_{adv} according the joint, effective objective

$$\min_{\theta_{\text{clf}}} \max_{\theta_{\text{adv}}} L_{\text{clf}}(\theta_{\text{clf}}) - \lambda L_{\text{adv}}(\theta_{\text{clf}}, \theta_{\text{adv}}), \tag{15.6}$$

balancing the classification task (first term) with the decorrelation task (second term). Here, λ is a parameter balancing the importance of the two terms in Equation (15.6) — similar to the uniforming rate α for uBoost— which will be discussed further below. The inner optimisation ($\max_{\theta_{\text{adv}}}$) leads the adversary to minimise $L_{\text{adv}}(\theta_{\text{clf}}, \theta_{\text{adv}})$ with respect to the adversary weights θ_{adv}, and is therefore equivalent to Equation (15.5b) up to a multiplicative constant and an offset. The outer optimisation ($\min_{\theta_{\text{clf}}}$) leads the classifier to minimise the effective loss $L_{\text{classifier}}(\theta_{\text{clf}}, \theta_{\text{adv}}) = L_{\text{clf}}(\theta_{\text{clf}}) - \lambda L_{\text{adv}}(\theta_{\text{clf}}, \theta_{\text{adv}})$ with respect to the classifier weights θ_{clf}, and is therefore equivalent to Equation (15.5a) with the addition of a 'penalty term.' The difference in sign between the two terms means that the classifier is trained to maximise L_{adv}, *i.e.* to make it harder for the adversary to infer the jet mass from the classifier output. The trade-off between the two competing objectives is controlled by the parameter λ: for $\lambda \gg 1$, the classifier is allowed only to retain information "orthogonal" to the jet mass; for $\lambda \to 0$, the standard NN classifier is recovered. Since it is otherwise unconstrained, λ can be considered an additional hyperparameter of the ANN, to be optimised according to some use case–specific figure of merit.

The challenge of the so-called min-max optimisation in Equation (15.6) is that it is not static in the sense of Equations (15.5): There are no stable (local) minima for either network to approach, since the optimisation "landscape" seen by each depends on the weights of the other. That is, every change in θ_{clf} may shift the solution(s) to $\min_{\theta_{\text{adv}}} L_{\text{adv}}$; and *vice versa*. This renders the training highly dynamical and means that there is no definitive convergence criteria for the optimisation in Equation (15.6), in contrast to

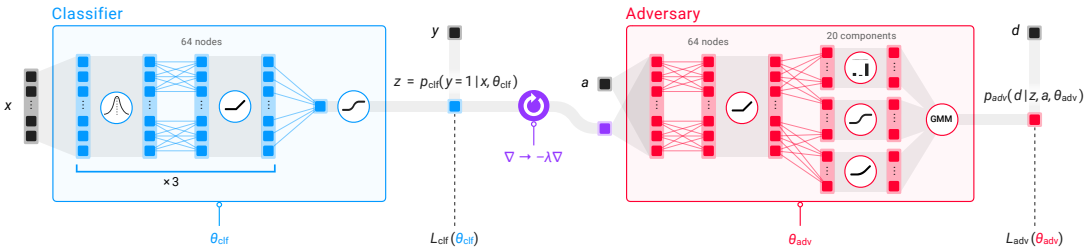


Figure 15.2 Adversarial neural network (ANN) architecture. The classifier network is tasked with predicting jet labels (y) based on some jet substructure variables (x), outputting a tagger variable (z). The adversary network is tasked with inferring the value(s) of the variables from which the classifier is to be decorrelated (d; here, the jet mass m), optionally aided by auxiliary features (a; here, $\log p_{\text{T}}/\mu$ with $\mu = 1$ GeV), by parametrising a posterior probability density function (p.d.f.) as a Gaussian mixture model (GMM). The adversarial training is implemented using a gradient reversal layer, the trade-off between L_{clf} and L_{adv} controlled by the parameter λ.

the ones in Equations (15.5). This challenge is addressed in the implementation of the adversarial architecture.

Implementation

The classifier and adversary NNs are connected in a single architecture as shown in Figure 15.2. Both models are constructed in Keras (v2.1.5) [242] using the TensorFlow (v1.4.1) backend [243]. The project library [244] is open-source and available on GitHub.

The classifier, parametrised by weights θ_{clf}, is trained to perform binary classification of the jet labels Y, taken to be 1 for W jets and 0 for multijets, based on a set of input features X using the binary cross-entropy (BCE) loss in Equation (4.3)

$$L_{\text{clf}}(\theta_{\text{clf}}) = \mathbb{E}_{\mathbf{x}\sim X,\, y\sim Y}\left[-y \log p_{\text{clf}}(\mathbf{x} \mid \theta_{\text{clf}}) - (1-y)\log\left(1 - p_{\text{clf}}(\mathbf{x} \mid \theta_{\text{clf}})\right)\right], \tag{15.7}$$

where $\mathbb{E}$ denotes the average over a coherent batch of jet features $\mathbf{x}$ and associated labels y, drawn from the sample populations X and Y, respectively, and $p_{\text{clf}}(\mathbf{x} \mid \theta_{\text{clf}}) = p_{\text{clf}}(y = 1 \mid \mathbf{x}, \theta_{\text{clf}}) = z$ is the output of the classifier network. Throughout, capitalised variables denote sample populations, and lower-case variables denote elements drawn from those populations, see also Chapter 4 for details on the notation. Minimising L_{clf} is seen to train the classifier output z to tend towards the true label value y.

The task of the adversary to infer the jet mass is implemented by having the adversary

parametrise a mixture density network [245]: It constructs a p.d.f. for the jet mass m, conditional on $z = p_{\text{clf}}(\mathbf{x} \,|\, \theta_{\text{clf}})$, using a Gaussian mixture model (GMM) [240]. That is, the so-called adversary posterior p.d.f. is constructed as a weighted sum of Gaussian distributions with trainable means, widths, and relative normalisation. In order to ease this task, the network can be parametrised as described in Ref. [246], by providing it with a set of auxiliary inputs A from which information about the jet mass can be derived. When the adversary output is parametrised by some auxiliary inputs, better mass-decorrelation can be achieved both inclusively and as a function of these variables. For instance, an adversary NN parametrised by the large-R jet p_{T} will be able to use this knowledge to more easily infer the jet mass m, as the two are correlated, leading to better mass-decorrelation overall and as a function of p_{T}. In practice, this study uses $\log p_{\text{T}}/\mu$ with $\mu = 1$ GeV as the single auxiliary feature. This variable is scaled to the range $[0, 1]$, to match the scale of classifier output. This parametrisation is found to yield relatively robust results as a function of p_{T} and in particular on the lower range of the p_{T}-spectrum, which is the primary regime studied for summary performance.

The output of the adversary is therefore the conditional probability $p_{\text{adv}}(m \,|\, z, a, \theta_{\text{adv}})$, given auxiliary features $a \sim A$, evaluated at the actual jet mass value m for each jet in the training sample. The adversary, tasked with decorrelating the classifier output from the jet mass m, is trained with the negative log-likelihood loss [247]

$$L_{\text{adv}}(\theta_{\text{adv}}) = \mathbb{E}_{z\sim p_{\text{clf}}(X\,|\,\theta_{\text{clf}}),\, m\sim M,\, a\sim A} \left[-\log p_{\text{adv}}(m \,|\, z, a, \theta_{\text{adv}}) \right], \tag{15.8}$$

where $\mathbb{E}$ denotes the average over a coherent batch of classifier outputs z, jet masses, and auxiliary features a (here, $\log p_{\text{T}}/\mu$). This loss, computed only for background jets in order to mitigate the sculpting of the background large-R jet mass distribution, mirrors the decorrelation metric in Equation (14.2). This is what motivated the choice of JSD as the metric for mass-decorrelation.

The challenge of adversarial training of NNs is the non-stable nature of the problem, arising from the joint optimisation of networks with opposing objectives, see Equation (15.6). Ideally, for every parameter update of the classifier, the adversary should be allowed to fully converge, *i.e.* to completely condition itself on the updated classifier outputs. In practice, the optimisation is typically done using alternating weight updates for the classifier and adversary [225], where the inner optimisation in Equation (15.6) is approximated by performing a fixed number of weight updates of the adversary for each classifier weight update. This is sometimes referred to as "nested optimisation." Alternatively, the classifier and adversary networks can be trained simultaneously [248]. This is the approach used in this study, where gradient

reversal [249] is used for the implementation of the joint optimisation. Gradient reversal means that a gradient scaling operation is applied to the connection between the classifier and the adversary, see Figure 15.2. In the forward mode, it acts as the identity operation, outputting just the unmodified classifier output; during back-propagation, the gradient propagating from the adversary back to the classifier is scaled by $-\lambda$. The minus sign means that the gradient flowing backwards from the adversary has the inverse effect for the classifier weights as for the adversary weights. That is, whereas this gradient acts on θ_{adv} to minimise L_{adv}, it will have the effect of maximising L_{adv} for θ_{clf}. The magnitude of the gradient scaling is controlled by λ, leading to exactly the effective behaviour intended in Equation (15.6).

To stabilise the joint convergence, the classifier is trained with a smaller learning rate than the adversary. This resembles the effect of the standard nested optimisation mentioned above. This ratio of learning rates is treated as a hyperparameter to be optimised, see Section E.1. The difference in learning rates also reflects the high dimensionality of the adversary output space compared to the one-dimensional output space for the classifier: The adversary is estimating the parameters of a large number of GMM components, meaning that convergence will necessarily be slower and more complicated than for the classifier. In addition, the learning rate is scaled by $1/(1 + \lambda)$ to avoid excessively large gradients from flowing from the adversary to the classifier. This was found to improve the stability of convergence, since the gradient flowing from the adversary to the classifier is effectively scaled up by a factor of λ, which may lead to unstable gradients for $\lambda \gg 1$. By reducing the overall learning rate, as seen by both networks, this problem is avoided. Finally, the classifier will be pre-trained to an optimal configuration for classification before the adversarial training commences, see Section E.2. Therefore, the adversarial training aims to identify the smallest possible perturbation around the classification optimum, which satisfies the mass-decorrelation requirement for a given value of λ. A small learning rate ratio addresses both of these issues.

Training and hyperparameters

The intuition behind ANNs and their implementation in this thesis were outlined above. Appendix E details the NN training itself, as well as the hyperparameter optimisation for both the classifier and the adversary. The results are summarised below and the chosen ANN architecture is shown in Figure 15.2.

The classifier is constructed as a densely connected NN with the 10 input features listed

in Table 13.2, three hidden layers of 64 nodes equipped with rectified linear unit (ReLU) activation, and a single output node with sigmoid activation, see Appendix B. Batch normalisation is applied before each hidden layer in the classifier to standardise the learned features and speed up the training, see Chapter 4. This network is trained to perform binary jet classification, resulting in an classifier output z_{NN} in the range $[0, 1]$, which will be considered the standard NN jet tagger.

The adversary is similarly constructed as a densely connected NN with two input features — the classifier output z and the large-R jet $\log p_{\text{T}}/\mu$ — and a single hidden layer of 64 nodes equipped with ReLU activation. The adversary infers the jet mass m by parametrising a 20-component GMM by separately outputting the mean, width, and relative normalisation of each GMM component. These outputs are equipped with sigmoid, softplus, and softmax activation, respectively.

Both NNs are trained using the Adam [250] optimiser with a batch size of 8192 training samples. The classifier is trained stand-alone for 200 epochs to yield the standard NN tagger z_{NN}. Then, the adversary is conditioned on the fixed classifier for 10 epochs. Finally, the two NNs are trained simultaneously for 200 epochs with a learning rate ratio of $\ell_{\text{clf}}/\ell_{\text{adv}} = 2 \times 10^{-7}$. This results in the mass-decorrelated ANN tagger z_{ANN}, which can be directly compared to the standard NN jet tagger z_{NN}.

CHAPTER 16

Results

The mass-decorrelation techniques introduced in Chapter 15 all result in observables which classify jets as either W jets or non-resonant multijets. These observables are designed to mitigate sculpting of the background jet mass distribution when imposing a threshold selection. In this study, the decorrelation parameters for the MVA-taggers are chosen to have benchmark values $\lambda = 10$ and $\alpha = 0.3$ for the ANN and uBoost taggers, respectively. These values are chosen since they provide similar levels of large-R jet mass decorrelation, see below.

The distributions of all benchmark jet tagger observables are shown in Figure 16.1. Since the k-NN and CSS methods share D_2 as their base jet substructure observable, the D_2 distribution is shown only once. For the remaining methods, the "standard" and mass-decorrelated taggers are unique and one-to-one. All tagger observables yield distributions for W jets and multijets which are separated to varying degrees. In particular, the standard MVA taggers — NN and AdaBoost— more powerfully separate the two classes of jets than the standard single-variable taggers. The mass-decorrelated MVA taggers also provide clear classification power, although the W jet and multijet distributions become less separated as a result of the mass-decorrelation in both cases. A similar behaviour is not immediately clear for the analytical mass-decorrelation procedures, as discussed in Section 16.1.

As described in Chapter 14, the performance of each W jet tagger is evaluated for both classification and mass-decorrelation. Mass-decorrelated taggers are immediately useful physics tasks such as the analysis in Part II of this thesis. These are typically characterised by p_{T} corresponding to photon and jet trigger thresholds of around 200 and 500 GeV, respectively [1, 73, 155–157, 216], see also Chapter 8. Therefore, to investigate the performance of the various taggers in these different kinematic regimes, results are studied in two bins of p_{T}: $[200, 500]$ GeV and $[500, 1000]$ GeV.

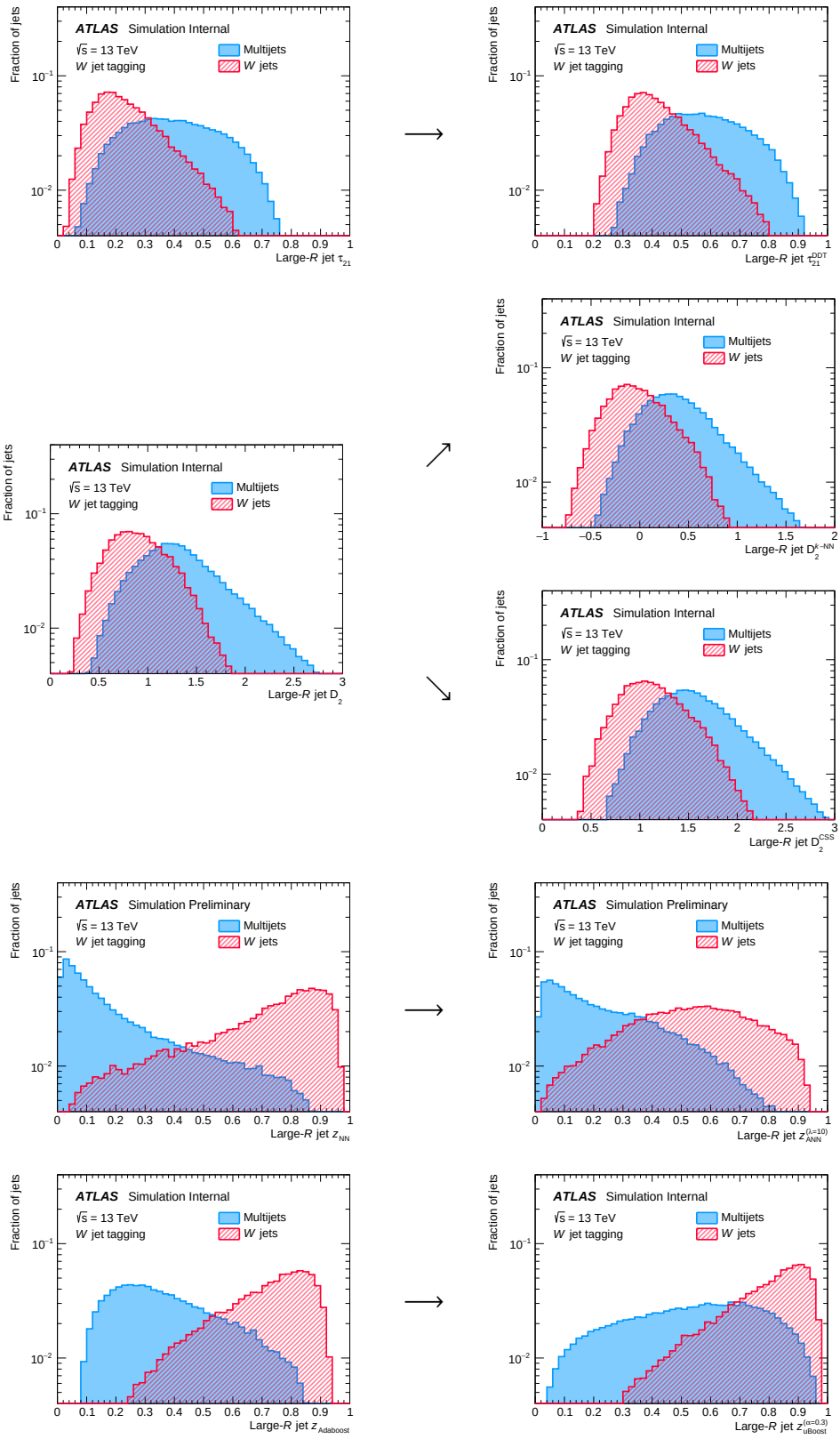
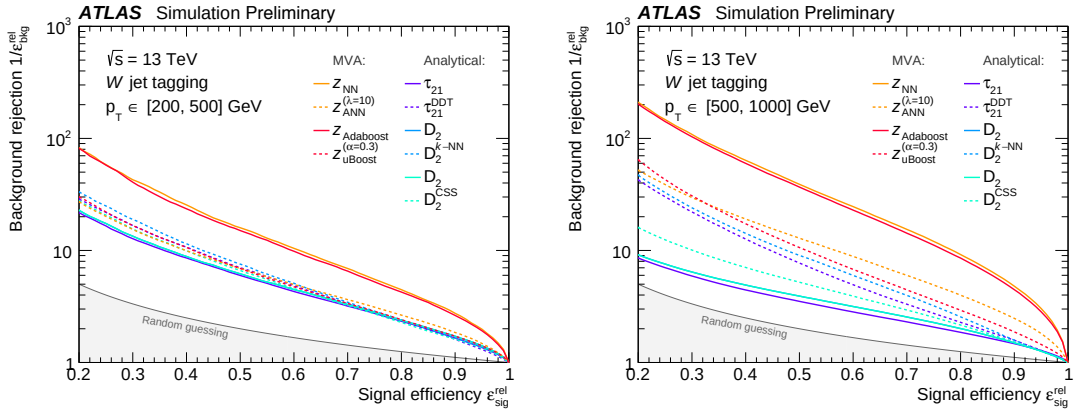


Figure 16.1 Distributions of the jet tagger variables for standard (*left*) and mass-decorrelated (*right*) tagging observables, for multijets and W jets. From the top: Designed decorrelated taggers (DDT), fixed-efficiency k-nearest neighbours (k-NN) regression, convolved substructure (CSS), adversarial neural network (ANN), and uBoost.



(a) Jet $p_{\mathrm{T}} \in [200, 500]$ GeV **(b)** Jet $p_{\mathrm{T}} \in [500, 1000]$ GeV

Figure 16.2 Rejection of multijets (background) as a function of W jet (signal) selection efficiency, for standard and mass-decorrelated version of analytical and multivariate analysis (MVA) jet taggers in two p_{T} bins, without the addition of a jet mass-window selection.

16.1 Classification

The classification performance, measured in terms of so-called receiver operating characteristic (ROC) curves, is shown in Figure 16.2 for each tagger. ROC curves show the background rejection $1/\varepsilon^{\mathrm{rel}}_{\mathrm{bkg}}$ for a threshold selection corresponding to a particular signal efficiency $\varepsilon^{\mathrm{rel}}_{\mathrm{sig}}$. These figures show a deterioration in the discrimination power of the mass-decorrelated MVA taggers with respect to their standard counterparts.

Classification power is measured as the background rejection rate at the benchmark signal efficiency of $\varepsilon^{\mathrm{rel}}_{\mathrm{sig}} = 50\%$. For 200 GeV $< p_{\mathrm{T}} <$ 500 GeV, shown in Figure 16.2a, among all taggers, the best classification is achieved by the standard MVA taggers, NN and AdaBoost. The two standard MVA taggers yield background rejection rates which are more than twice those of the analytical single-variable taggers. In this kinematic region, all mass-decorrelated taggers — both MVA-based and analytical — perform similarly in terms of classification.

For 500 GeV $< p_{\mathrm{T}} <$ 1000 GeV, shown in Figure 16.2b, the standard MVA taggers still perform better than the single-variable taggers by a factor of approx. 10 in terms of $1/\varepsilon^{\mathrm{rel}}_{\mathrm{bkg}}$ at $\varepsilon^{\mathrm{rel}}_{\mathrm{sig}} = 50\%$. However, considerable variation is observed between the mass-decorrelated taggers. The mass-decorrelated MVA taggers in Figure 16.2b report greater background rejection than the mass-decorrelated single-variable taggers, by a factor of approx. 2 for $\varepsilon^{\mathrm{rel}}_{\mathrm{sig}} = 50\%$. Among the analytical mass-decorrelation methods, k-NN provides the best classification; also better than CSS, both based on the D_2 substructure

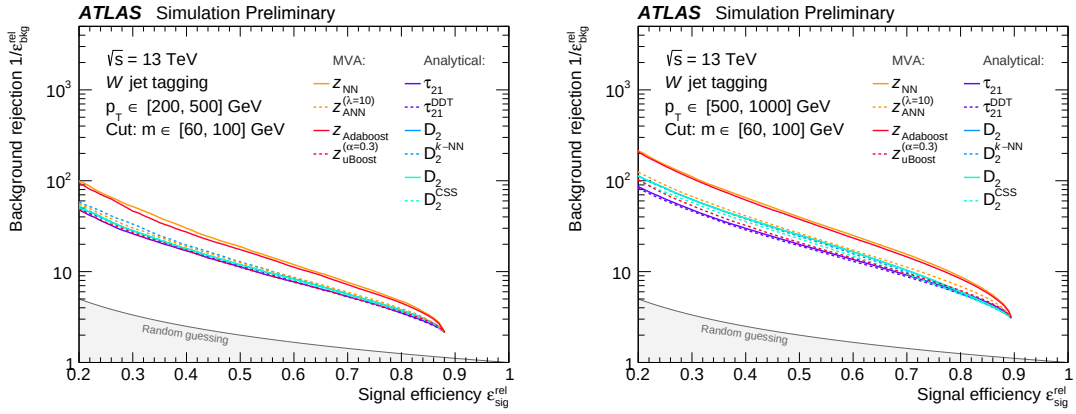
variable.

For all single-variable taggers, the background rejection after mass-decorrelation is greater than for the original taggers. This is true in both bins of p_{T}, but is particularly evident in the high-p_{T} bin in Figure 16.2b. This may be attributed to the standard single-variable taggers losing classification power with p_{T}, in the absence of an additional selection on the jet mass. This is an expression of relative shifts of the jet substructure observable distributions in question, for multijets and W jets, as a function of p_{T}. The DDT and fixed-efficiency k-NN regression methods both decorrelate the substructure variable not only from mass, but also from p_{T}, through ρ^{DDT} and ρ. The fact that these methods recover lost performance exactly by removing p_{T}-dependence is consistent with this observation. In particular, the background rejections rates for τ_{21}^{DDT} and $D_2^{k\text{-NN}}$ are close to equal in the two p_{T} bins. By contrast, the CSS method does not decorrelate from p_{T} and therefore experiences the smallest relative increase in classification power. The same effect is also evident in the low-p_{T} bin in Figure 16.2a, although to a much smaller extent.

The above effect was studied in the case of DDT. Here, it was found that the linear transform acts similarly to a Fisher discriminant transform in the $(\rho^{\mathrm{DDT}}, \tau_{21})$-plane. That is, the linear transformation that decorrelates τ_{21} from ρ^{DDT} also turns out to be almost identical to the linear transform that optimally separates signal and background jets in the $(\rho^{\mathrm{DDT}}, \tau_{21})$-plane. In this way, the DDT transform extracts additional information from ρ^{DDT}, which slightly improves classification.

The classification performance of the standard MVA taggers in Figure 16.2 improve with p_{T}, in contrast to the standard single-variable taggers. This is a result of the standard MVA taggers being trained to provide equal attention across p_{T}, see Chapter 13. Therefore, these taggers should provide robust classification for all p_{T} by construction.

In contrast to the mass-decorrelated single-variable taggers, the mass-decorrelated MVA taggers exhibit a considerable decrease in classification performance relative to their standard variants. This is due to the fact that the standard MVA taggers rely heavily on mass information, extracted from the 10 jet substructure inputs listed in Table 13.2. Effectively learning a proxy for the jet mass is what enables the powerful classification in Figure 16.2. The standard single-variable taggers are not correlated with the jet mass to the same extent. Therefore, the mass-decorrelation procedure has a greater potential for degrading the classification power for the MVA taggers than for the single-variable taggers, since there is more correlation to undo. This effect is particularly evident in Figure 16.2, which does not include an additional selection on



(a) Jet $p_{\mathrm{T}} \in [200, 500]$ GeV **(b)** Jet $p_{\mathrm{T}} \in [500, 1000]$ GeV

Figure 16.3 Rejection of multijets (background) as a function of W jet (signal) selection efficiency, for standard and mass-decorrelated versions of analytical and multivariate (MVA) jet taggers with the addition of a jet mass-window selection $m \in [60, 100]$ GeV, in two p_{T} bins.

the jet mass, thereby exacerbating the impact of the correlation with the jet mass on the classification. Finally, the observed deterioration of MVA tagger classification power due to the mass-decorrelation procedures is expected and will be acceptable to relevant analyses as part of a trade-off between these two competing objectives.

Overall, it is emphasised that the relative background efficiencies in Figure 16.2 are computed with respect to jets with invariant masses in the range $m \in [50, 300]$ GeV. Therefore, improvements in classification following analytical mass-decorrelation are dominated by jets which are far from the W boson mass. In general, computing the background rejection with respect to the full multijet sample passing the baseline selection, see Chapter 13, is well-motivated, since no window selection on the jet mass is envisioned for the most obvious physics use cases [1, 73, 155–157, 216]. However, for the specific case of tagging known resonances — *e.g.* W jet tagging — an additional selection on the jet mass of $m \in [60, 100]$ GeV is typically used to further increase the non-resonant background rejection. The ROC curves for the various taggers in the two p_{T} bins, with the addition of such a jet mass window selection is shown in Figure 16.3.

Comparing to Figure 16.2, this selection leads to increases in background rejection at similar signal efficiencies for all taggers. However, the relative performance of the various taggers is generally the same. Furthermore, the effect of the analytical mass-decorrelation methods improving classification power, noted in Figure 16.2b, mostly disappears after imposing the mass-window selection. In this way, the effect of W jet classification *per se* is separated from the effect of non-uniform selection efficiencies

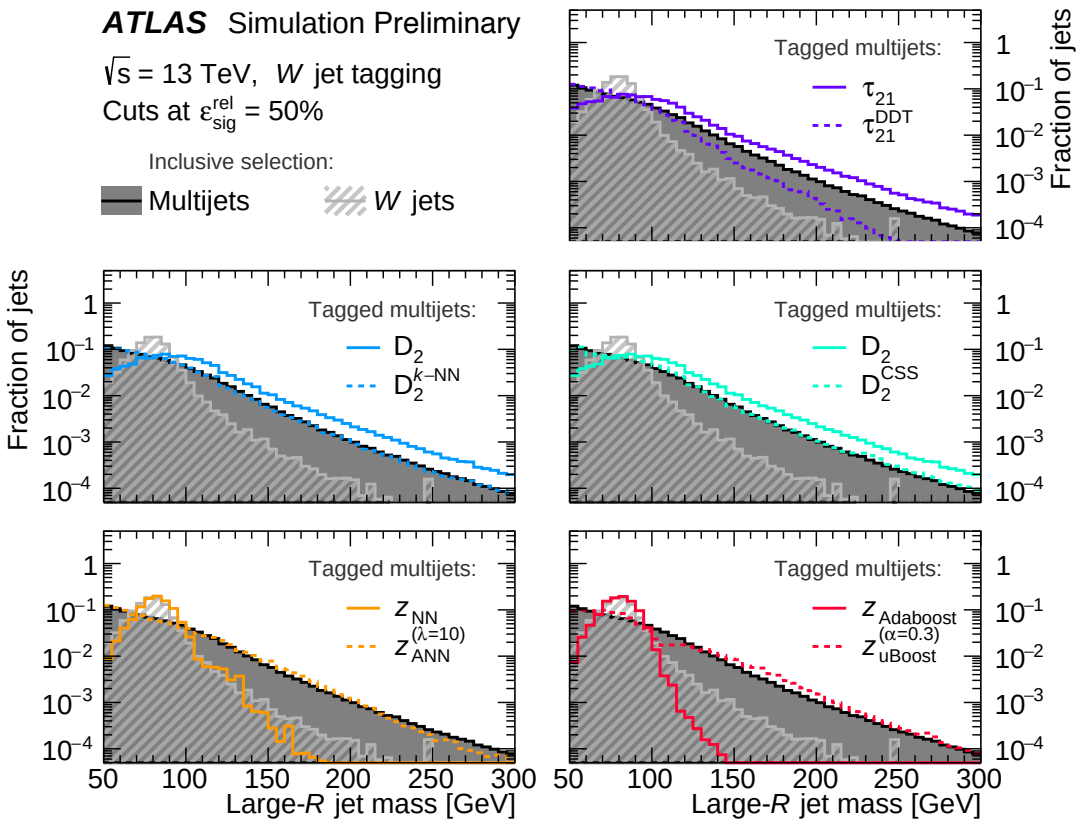


Figure 16.4 Normalised jet mass distribution for inclusive multijets before selections, compared to the same distributions after selections on the studied jet tagging observables. Also shown for reference is the jet mass distribution for W jets before tagging. Selections are chosen to correspond to a W jet (signal) selection efficiency of $\varepsilon_{\text{sig}}^{\text{rel}} = 50\%$.

outside the window around the W boson mass.

16.2 Mass-decorrelation

To study the mass-decorrelation of various taggers, the most direct measure is the inspection of the normalised multijet mass distribution before and after the application of a selection on the tagging observables. Such comparisons of jet mass distributions are shown in Figure 16.4.

Performing a threshold selection on each of the standard taggers sculpts the background jet mass distribution to varying degrees, thereby introducing artificial structures not present in the original spectrum. Such sculpting may directly impact the sensitivity of physics searches by reducing the ability to constrain systematic uncertainties on *e.g.*

multijet backgrounds, as was found in Chapter 10. In particular, the standard MVA taggers sculpt the background jet mass distribution to resemble the W jet mass peak, thereby rendering the use of side-bands in the jet mass distribution, as control regions for constraining systematic uncertainties, virtually impossible.

Each of the mass-decorrelation procedures serves to mitigate such sculpting. For the mass-decorrelated single-variable taggers, CSS and particularly k-NN both remove most sculpting effects. The DDT transform removes the sculpting in the lower jet mass region, while some disagreement persists at higher masses. This is a result of the limitations of the assumption of linearity underlying the method as shown in Figure D.1. It could be mitigated by restricting the jet kinematic phase space to the region of validity of the method, as was done with the boosted topology selection in Chapter 8, at the cost of discarding potentially valuable data.

The sculpting of both MVA taggers is considerably mitigated following the application of their respective training methods for decorrelation. In particular, the ANN tagger yields a largely smooth, un-sculpted distribution of multijet events passing the selection on $z_{\text{ANN}}^{(\lambda=10)}$. In contrast, the multijet distribution passing the selection on $z_{\text{uBoost}}^{(\alpha=0.3)}$ retains some residual sculpting, particularly around $m \approx 100$ GeV, as a consequence of persisting non-uniformity in the selection efficiency.

The details of such non-uniformities in the background selection efficiency can be studied by inspecting the local background efficiency as a function of the jet mass. This is shown in Figure 16.5 for a range of inclusive background efficiencies.

The sculpting of the multijet background around the W jet mass peak for the standard MVA taggers is evident. Similarly, both MVA-based mass-decorrelation procedures result in background efficiency profiles which are roughly uniform as a function of the jet mass. Some localised non-uniformity of the uBoost tagger persists in the region around the W boson mass at the relevant inclusive background selection efficiency (approx. 5%). The ANN tagger does not exhibit a similar degree of residual sculpting, although some deviation from uniformity remains between jet masses of 150 and 200 GeV, reflecting the residual correlation with the jet mass. This may be reduced by more fine-tuned optimisation, longer adversarial training, providing the adversary NN with more information, or not starting the adversarial training from a fully pre-trained classifier network. For the analytical taggers, the local background efficiencies exhibit similar behaviour: the initial, non-uniform background efficiency profiles are transformed to be less dependent on the jet mass as a result of the decorrelation procedures, at the relevant inclusive background selection efficiency (approx. 20%). Due to the generally

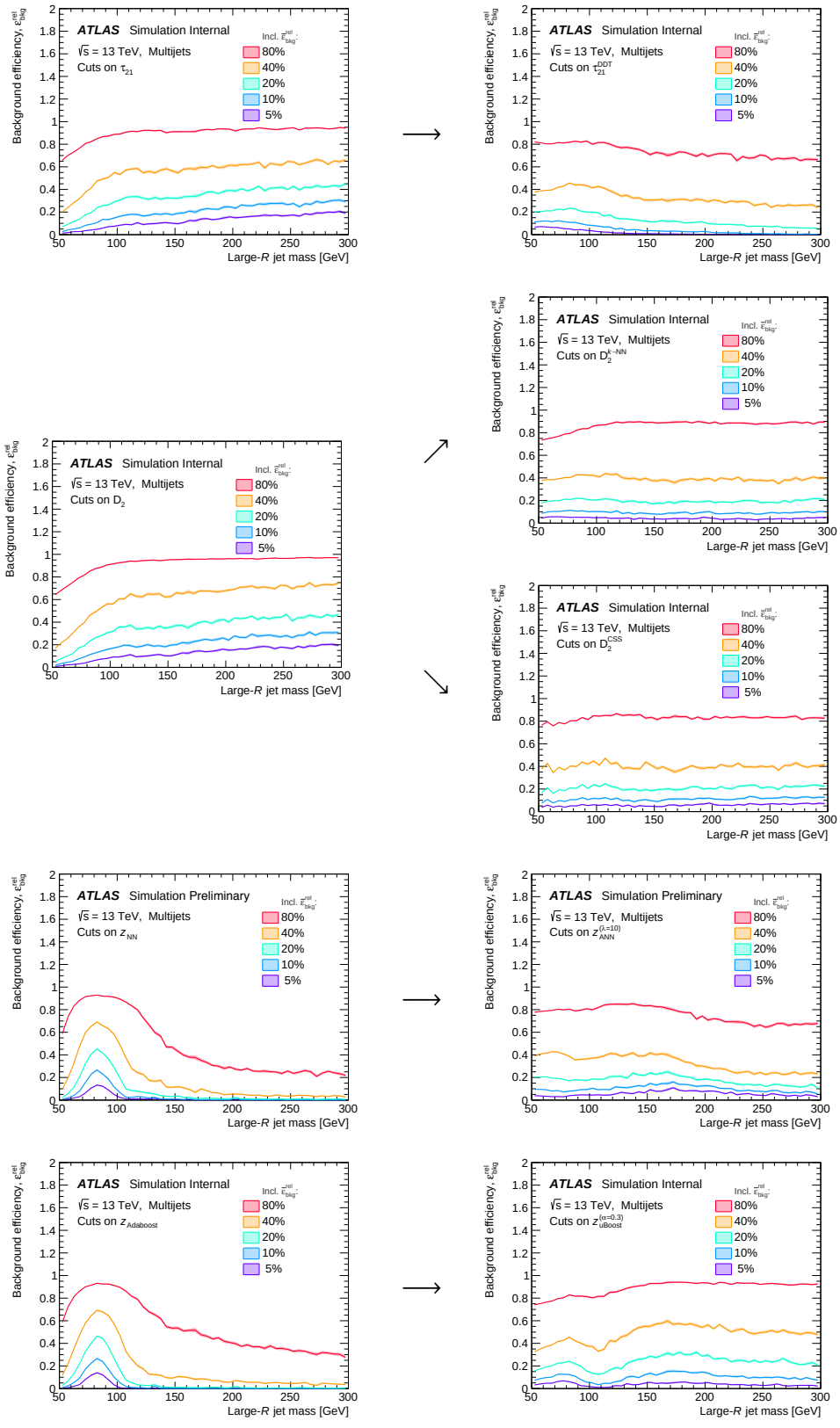


Figure 16.5 Jet mass-dependent multijet selection efficiencies for various inclusive efficiencies for standard (*left*) and mass-decorrelated (*right*) tagging observables. From the top: Designed decorrelated taggers (DDT), fixed-efficiency k-nearest neighbours (k-NN) regression, convolved substructure (CSS), adversarial neural network (ANN), and uBoost.

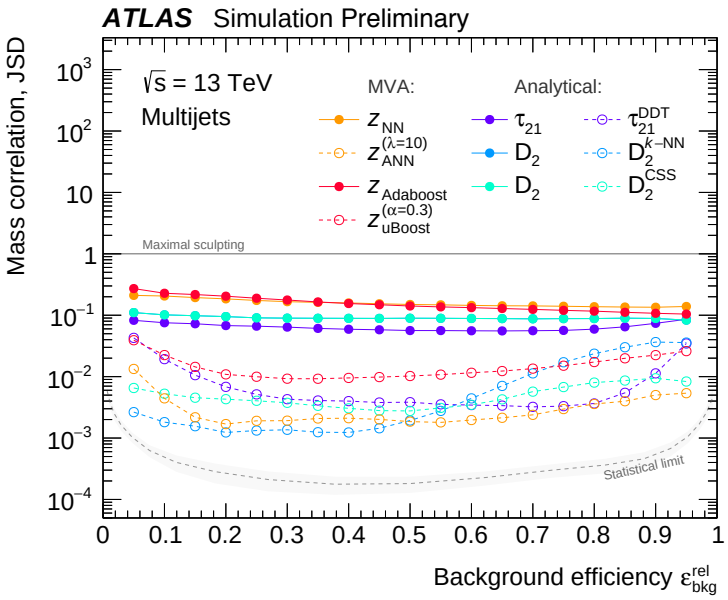


Figure 16.6 Profiles of the Jensen-Shannon divergence (JSD) for selections corresponding to various multijet (background) efficiencies. Standard classifiers are indicated with filled markers. Mass-decorrelated classifiers indicated with open markers. The shaded grey band indicates the statistical limit on JSD from the finite number of simulated jets. See text for details.

simpler nature of the analytical mass-decorrelation procedures, these are also less prone to introduce local non-uniformities than the more complex MVA techniques.

As the quantitative summary metric for the mass-decorrelation, Figure 16.6 shows the JSD, defined in Chapter 14, as a function of the background selection efficiency. As expected, the JSD values for the mass-decorrelated taggers are consistently and considerably lower — *i.e.* have lower degrees of correlation with the jet mass — than the standard ones.

The k-NN method leads to the greatest degree of mass-decorrelation, especially in the vicinity of the background efficiency at which the regression is performed (16%). The other methods exhibit more uniform mass-decorrelation across $\varepsilon_{\text{bkg}}^{\text{rel}}$. Among the MVA taggers, the ANN performs considerably better than uBoost, for the chosen values of λ and α, according to the JSD metric.

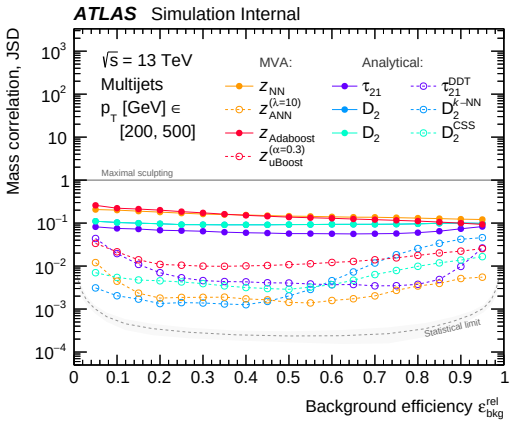
The fact that JSD is computed from histograms with finite statistics imposes a lower bound on the mass-decorrelation given the chosen testing dataset: Drawing two finite samples from the same underlying distribution will still result in non-zero JSD when evaluated on the two sampled distribution. The statistical limit on JSD can be estimated using bootstrap sampling [251]. Given a dataset D comprising N jets and a background

efficiency ε, two bootstrapped samples $S_{1,2}$ of jet masses are found by drawing samples of $\varepsilon \times N$ and $(1-\varepsilon) \times N$ jets, respectively, with replacement from the original dataset D. As the sampled distributions $S_{1,2}$ are drawn at random from the same underlying distribution, the value $\mathrm{JSD}(S_1 \,||\, S_2)$ is a measure of the lower limit on JSD in the ideal case of no sculpting. Since the bootstrapping method is stochastic, it is repeated 10 times to get a distribution of sampled statistical limits on JSD. Properties of the bootstrap sampled distribution, *e.g.* the standard deviation, are unbiased estimators of the corresponding properties of the underlying distribution [251]. In addition to estimating the uncertainty band on the statistical JSD limit, the same procedure can therefore also be employed to estimate uncertainties on the classification and mass-decorrelation metrics for individual tagger observables.

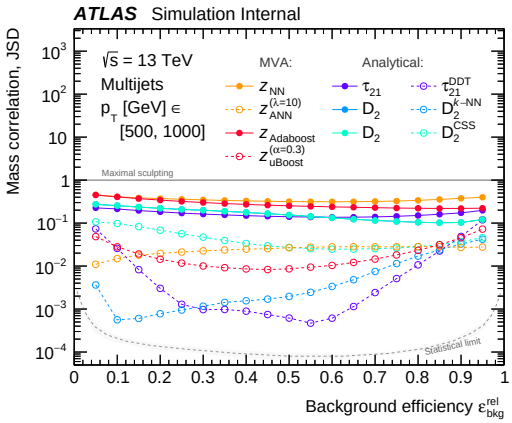
The bootstrapped statistical limit on JSD is shown as a function of the background selection efficiency in Figure 16.6 as a smoothed, dashed line and shaded band, indicating the mean and standard deviation, respectively, of the bootstrap distribution. Additionally, for a given physics task, full mass-decorrelation might not be necessary or optimal. For instance, depending on the level of systematic uncertainty in a given search, some moderate correlation with the jet mass might ultimately lead to better sensitivity to new physics than full mass-decorrelation, at the cost of worse classification.

Finally, since Figure 16.6 show JSD profiles for the entire multijet sample, variations at high p_{T} may not be clear due to the weighting of jets according to cross-section in the testing sample. Figure 16.7 shows similar profiles in three relevant bins of p_{T}.

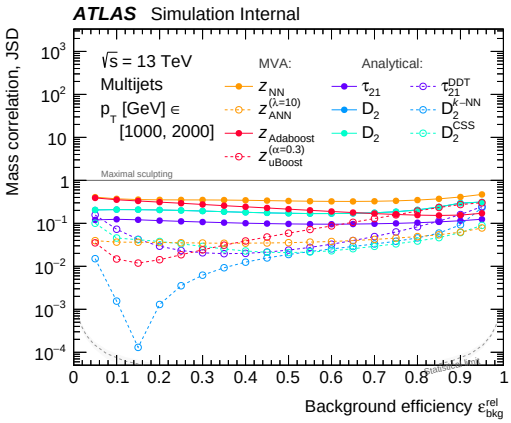
Figure 16.7a is all but identical to Figure 16.6, covering the high-population region in p_{T} just above the selection threshold of 200 GeV, see Figure 13.1. A deterioration in the mass-decorrelation for the ANN tagger is seen in Figure 16.7b, which is discussed in Section 16.4. In this p_{T}-bin, τ_{21}^{DDT} does somewhat better across $\varepsilon_{\mathrm{bkg}}^{\mathrm{rel}}$, and k-NN retains a large degree of mass-decorrelation in the vicinity of its target background selection efficiency of 16%. Finally, Figure 16.7c shows how all methods but k-NN fail to provide substantial mass-decorrelation for $p_{\mathrm{T}} \gtrsim 1$ TeV. k-NN reports the best performance in this p_{T}-bin among the methods considered in this study. However, it also provides mass-decorrelation only specifically and narrowly at the target background selection efficiency. This is the method working as intended, but it illustrates the importance of sticking to the tagging value for which the mass-decorrelated tagger is designed (*i.e.* $D_2^{k\text{-NN}} < 0$). Conversely, for threshold selections at values even modestly away from this point, the mass-decorrelation quickly degrades. As a result, a substructure observable constructed using k-NN in MC simulated data may perform unreliably in recorded data samples.



(a) Jet $p_{\mathrm{T}} \in [200, 500]$ GeV



(b) Jet $p_{\mathrm{T}} \in [500, 1000]$ GeV



(c) Jet $p_{\mathrm{T}} \in [1000, 2000]$ GeV

Figure 16.7 Profiles of the Jensen-Shannon divergence (JSD) in three bins of jet p_{T}, for selections corresponding to various multijet (background) selection efficiencies. Standard classifiers are indicated with filled markers. Mass-decorrelated classifiers indicated with open markers. The shaded grey band indicates the statistical limit on JSD from the finite number of simulated jets. See text for details.

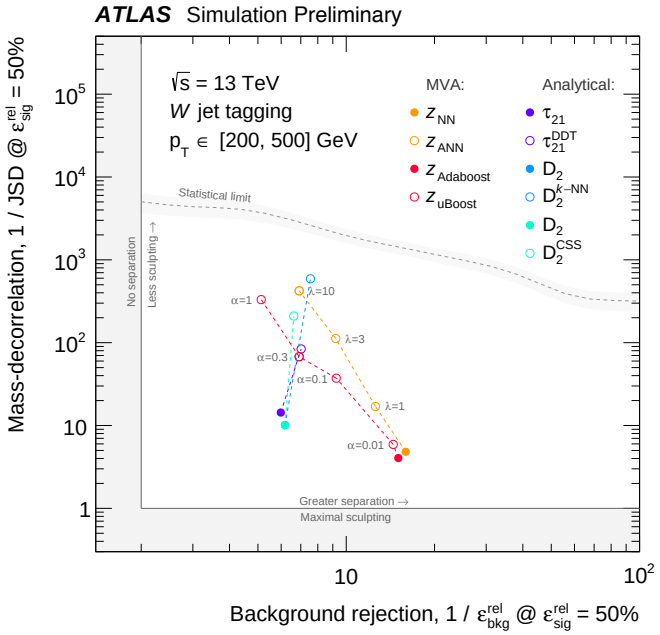
16.3 Combined metric

A simultaneous study of the metrics for classification and mass-decorrelation is necessary to assess the trade-offs balanced by each of the studied techniques. This can be done by plotting the two metrics together: Figure 16.8 shows the mass-decorrelation (1/JSD) versus the background rejection ($1/\varepsilon_{\text{bkg}}^{\text{rel}}$) for tagger selections at $\varepsilon_{\text{sig}}^{\text{rel}} = 50\%$, in two p_{T} bins. The x-axis measures classification and the y-axis measures mass-decorrelation, with larger values along each indicating better performance. For any given physics task, some specific direction in the plane of Figure 16.8 will correspond to the optimal trade-off. The dashed line and shaded band at high 1/JSD indicate the statistical limit of the mass-decorrelation, estimated using bootstrap sampling.

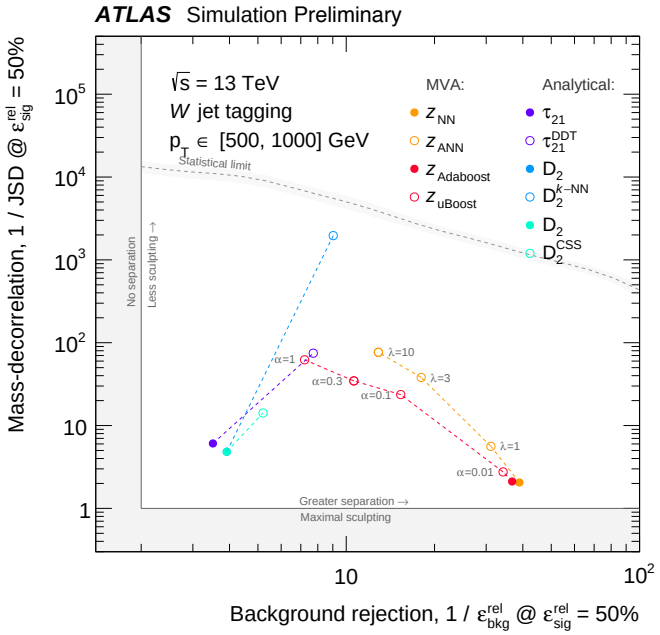
For the mass-decorrelated MVA taggers, several working points are evaluated by scanning λ for the ANN tagger and α for uBoost. For high values of λ ($\gtrsim 10$), the ANN method starts to saturate given the chosen network configuration, training procedure, and datasets.

Figure 16.8 shows that for equal levels of mass-decorrelation, the ANN tagger generally provides the greatest background rejection. The BDT-based MVA taggers have comparable performance to the NN-based taggers for the standard variants. However, the adversarial training method for mass-decorrelation is seen to perform better than the uBoost method for the chosen configurations and performance metrics. From Figure 16.8b, the effect of the analytical mass-decorrelation methods on improving the classification while simultaneously decorrelating from the jet mass, as discussed above, is particularly evident.

The k-NN method is the most effective analytical decorrelation method, leading to a tagger variable which is close to fully decorrelated from the jet mass. The saturation of the ANN tagger at high λ means that an upper limit on 1/JSD exists for $\lambda \gtrsim 10$ with the chosen configuration. This is a reasonable observation, considering the complexity of the ANN training procedure and the fact that the tagger is evaluated on a dataset 10 times larger than the training dataset. Therefore, complex decorrelation methods like the ANN are likely decorrelated at, or close to, the statistical limit of the training dataset, but not necessarily at the statistical limit for the much larger evaluation dataset. The simplicity of *e.g.* the k-NN method means that it will be more robust to such differences in statistics. For these reasons, raising the upper limit on 1/JSD for ANN will likely be possible by increasing training statistics, performing a more fine-tuned model architecture optimisation, providing the adversary with more information, *etc.*.

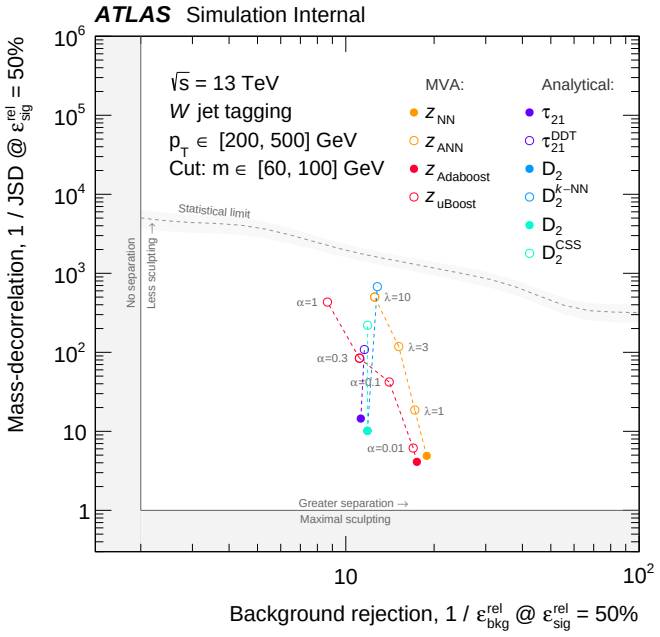


(a) Jet $p_{\mathrm{T}} \in [200, 500]$ GeV

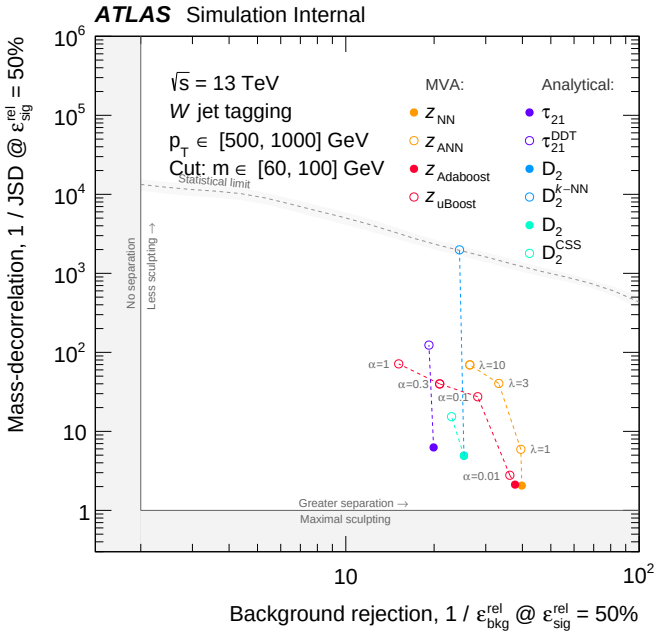


(b) Jet $p_{\mathrm{T}} \in [500, 1000]$ GeV

Figure 16.8 Unified plot of the metrics for classification (background rejection, $1/\varepsilon_{\mathrm{bkg}}^{\mathrm{rel}}$) and mass-decorrelation (inverse Jensen-Shannon divergence, $1/\mathrm{JSD}$), for selections on each tagger observable corresponding to $\varepsilon_{\mathrm{sig}}^{\mathrm{rel}} = 50\%$, in two bins of p_{T}, without a selection on the jet mass. The additional jet mass selection is applied only for classification, such that $1/\mathrm{JSD}$ is always calculated for the full jet mass spectrum. Greater values along each axis indicate better performance. Standard classifiers are indicated with filled markers. Mass-decorrelated classifiers are indicated with open markers, with parameter scans traced out by dashed lines. The shaded grey band indicates the statistical limit on $1/\mathrm{JSD}$ from the finite number of simulated jets.



(a) Jet $p_{\mathrm{T}} \in [200, 500]$ GeV



(b) Jet $p_{\mathrm{T}} \in [500, 1000]$ GeV

Figure 16.9 Unified plot of the metrics for classification (background rejection, $1/\varepsilon_{\mathrm{bkg}}^{\mathrm{rel}}$) and mass-decorrelation (inverse Jensen-Shannon divergence, 1/JSD), for selections on each tagger observable corresponding to $\varepsilon_{\mathrm{sig}}^{\mathrm{rel}} = 50\%$, in two bins of p_{T}, with an additional selection on the jet mass of $m \in [60, 100]$ GeV. The additional jet mass selection is applied only for classification, such that 1/JSD is always calculated for the full jet mass spectrum. Greater values along each axis indicate better performance. Standard classifiers are indicated with filled markers. Mass-decorrelated classifiers are indicated with open markers, with parameter scans traced out by dashed lines. The shaded grey band indicates the statistical limit on 1/JSD from the finite number of simulated jets.

Figure 16.9 shows the unified plot of performance metrics with the addition of a window selection on the jet mass. This selection is only applied for the calculation of the background rejection, since it is not meaningful when calculating the mass-decorrelation metric. The relative performance of the substructure taggers is roughly unchanged with respect to Figure 16.8, since the addition of the window selection on the jet mass coherently increases the background rejection but is not used directly in the calculation of $1/\text{JSD}$. After the jet mass selection, the analytical mass-decorrelation procedures no longer improve background rejection, as most evident in Figure 16.9b, in accordance with Figure 16.3. Also here, the k-NN method provides near-perfect mass-decorrelation, while the CSS method — which does not take the jet p_{T} into account — provides minimal improvement at high p_{T}.

16.4 Robustness

For the chosen metrics, an important consideration is their robustness to variations in jet kinematics and event conditions. The background rejection and $1/\text{JSD}$ as a function of jet p_{T} and the number of reconstructed primary vertices N_{PV} is shown in Figure 16.10. In each p_{T} bin the background rejection is computed for a selection corresponding to $\varepsilon_{\text{sig}}^{\text{rel}} = 50\%$, thereby isolating the measurement of background rejection from the uniformity of the $\varepsilon_{\text{sig}}^{\text{rel}}$ as a function of p_{T} for a threshold selection at a fixed tagger value.

The statistical limit on $1/\text{JSD}$ is estimated separately in each p_{T} bin using bootstrap sampling of the multijet mass distribution, similar to the procedure in Sections 16.2 and 16.3. However, because the background rejection $\varepsilon_{\text{bkg}}^{\text{rel}}$ at $\varepsilon_{\text{sig}}^{\text{rel}} = 50\%$ will be different for each tagger in each p_{T} bin, no unique background selection efficiency can be used in the bootstrap sampling. Instead the average background selection efficiency across all taggers is used to estimate the mean statistical limit on $1/\text{JSD}$ within each p_{T} bin, indicated as the dashed line in Figure 16.10. The statistical uncertainty on this estimate is taken to be standard deviation of statistical limits on $1/\text{JSD}$ across bootstrap samples. A systematic uncertainty is assigned as half the absolute difference between the mean statistical limit on $1/\text{JSD}$ at the maximal and minimal background selection efficiency, respectively, within each p_{T} bin, thereby taking the variation in $\varepsilon_{\text{bkg}}^{\text{rel}}$ at $\varepsilon_{\text{sig}}^{\text{rel}} = 50\%$ into account. The shaded band in Figure 16.10 is the sum in quadrature of these two uncertainty components.

Across p_{T}, the standard MVA taggers yield the largest background rejection as well as the greatest correlation with the jet mass (*i.e.* lowest values of $1/\text{JSD}$). In the absence

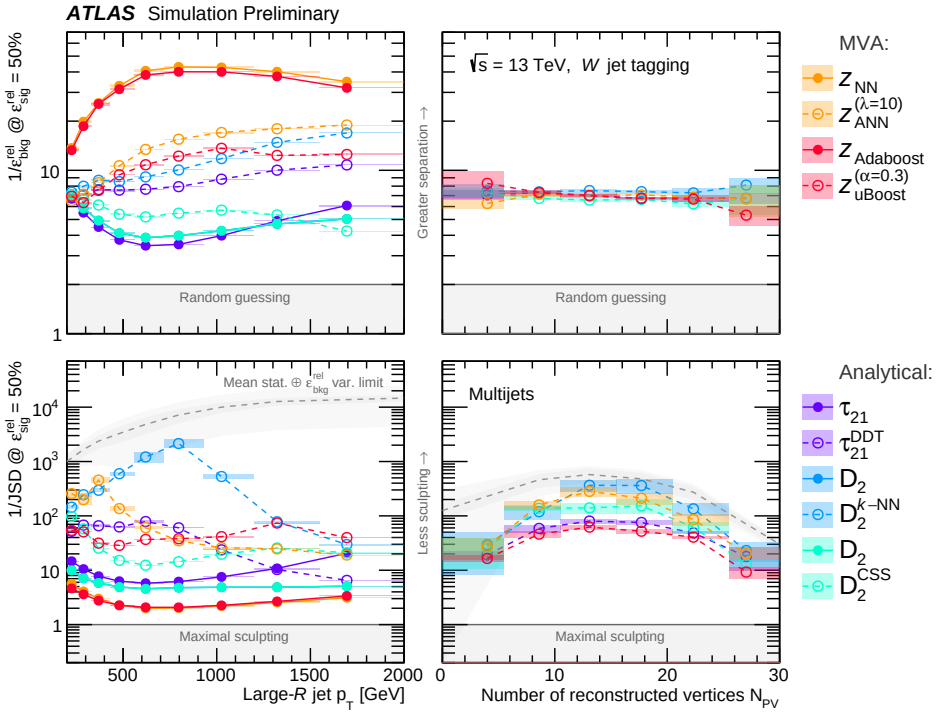


Figure 16.10 Plot of the metrics for classification (background rejection, $1/\varepsilon_{\text{bkg}}^{\text{rel}}$; *top*) and mass-decorrelation (inverse Jensen-Shannon divergence, 1/JSD; *bottom*), for selections corresponding to $\varepsilon_{\text{sig}}^{\text{rel}} = 50\%$, as a function of the reconstructed jet p_{T} (*left*) and the number of reconstructed vertices, N_{PV} (*right*). Standard classifiers are indicated with filled markers. Mass-decorrelated classifiers indicated with open markers. Statistical uncertainties are indicated with shaded boxes, derived using bootstrap sampling. The statistical limit on 1/JSD, also accounting for variation in $\varepsilon_{\text{bkg}}^{\text{rel}}$ for different taggers within the same bin, is shown as a shaded grey band (*bottom*). Only mass-decorrelated taggers are shown for N_{PV} (*right*).

of an additional selection on the jet mass, the standard single-variable taggers all show the best performance for p_{T} close to the lower selection threshold of 200 GeV and decreasing with p_{T}, regaining performance again towards $p_{\text{T}} \sim 2$ TeV. Since the single-variable taggers are highly physics-motivated and constructed from physically weighted distributions, the fact that their performance is optimal in the high-population low-p_{T} end of the spectrum might not be surprising. However, as also noted in Section 16.1, this behaviour is driven by jets outside the region around the W boson mass.

For the analytical taggers, the improvement in classification power arising from the mass-decorrelation procedures is seen to increase with p_{T}. As no jet mass-window selection is performed in Figure 16.10, this is due to the mass- and p_{T}-dependence of τ_{21} and D_2. The CSS methods, which removes mass-dependence in a way which is

biased towards $p_{\mathrm{T}} \sim 200$ GeV, leads to the smallest improvement of the three analytical mass-decorrelation techniques; the DDT and fixed-efficiency k-NN regression methods, which decorrelate from the jet p_{T} in addition to the jet mass, yield bigger improvements across p_{T} by removing this additional dependency. In particular, k-NN yields a powerful single-variable tagger which is close to fully decorrelation from the jet mass, within statistical uncertainties, up to $p_{\mathrm{T}} \approx 1$ TeV.

However, for $p_{\mathrm{T}} \gtrsim 1$ TeV, $D_2^{k\text{-NN}}$ experiences a significant drop in $1/\mathrm{JSD}$. This is due to the fact that the k-NN method is designed to decorrelate D_2 from the jet mass at $\varepsilon_{\mathrm{bkg}}^{\mathrm{rel}} = 16\%$, which corresponds to $\varepsilon_{\mathrm{sig}}^{\mathrm{rel}} = 50\%$ for the inclusive sample. However, the selections in Figure 16.10 are calculated to correspond to $\varepsilon_{\mathrm{sig}}^{\mathrm{rel}} = 50\%$ in each p_{T} bin, and since the background rejection of $D_2^{k\text{-NN}}$ increases with p_{T}, the selection at high p_{T} will correspond $\varepsilon_{\mathrm{bkg}}^{\mathrm{rel}}$ considerable smaller than 16%. From Figure 16.7c, it is seen that the k-NN method's capacity for mass-decorrelation at $p_{\mathrm{T}} \gtrsim 1$ TeV is narrowly centred on selections with $\varepsilon_{\mathrm{bkg}}^{\mathrm{rel}} = 16\%$. Therefore, since that mass-decorrelation for k-NN is tuned to $\varepsilon_{\mathrm{bkg}}^{\mathrm{rel}} = 16\%$, and since the decrease in $1/\mathrm{JSD}$ around this value is found to become more prominent with increasing p_{T}, this explains the behaviour seen for k-NN in Figure 16.10. If a threshold selection at a fixed value were used, *i.e.* $D_2^{k\text{-NN}} < 0$, k-NN would be able to decorrelate D_2 from the jet mass to roughly within statistical uncertainty across p_{T}, but without a constant signal efficiency as a function of p_{T}.

The mass-decorrelated MVA taggers have relatively robust performance across p_{T}, with the ANN outperforming uBoost in mass-decorrelation at low p_{T} and uBoost performing slightly better at the highest p_{T}. In the lowest p_{T} bins, z_{ANN} is seen to be as mass-decorrelated as $D_2^{k\text{-NN}}$. However, for $p_{\mathrm{T}} \gtrsim 400$ GeV the mass-decorrelation for ANN degrades. To some extent, this is an effect of the parametrisation of the adversary network in terms of $\log p_{\mathrm{T}}/\mu$ with $\mu = 1$ GeV. The parametrisation serves to guide the attention of the adversary, and using the logarithm of p_{T} leads to competitive mass-decorrelation as a function of p_{T}, but with an emphasis on low p_{T}. Studies have indicated that an adversary parametrised by p_{T} would have slightly more robust performance, but report comparatively worse summary metrics, see Figure 16.8, since the testing dataset is dominated by jet with p_{T} just above the lower selection threshold.

However, this effect persists across $\varepsilon_{\mathrm{bkg}}^{\mathrm{rel}}$ and λ, and is more likely related to the inability of the adversary to fully reverse the sculpting around the W jet peak. Therefore, a more likely cause is the difference in the training procedure, compared to uBoost. Where the uBoost classifier starts from a random initialisation and is then trained by adaptive boosting, balancing competing objectives, the ANN tagger is treated as a minimal perturbation around the standard NN classifier, by starting from a fully pre-trained

NN classifier, see Chapter 15. The ANN tagger has to reverse a large degree of mass-sculpting; the uBoost tagger is trained not to learn it in the first place. This choice works well in some regimes, particularly at moderately low p_{T}, but appears to work less well at high p_{T}, where the initial mass sculpting of the standard NN tagger is more dramatic see Figure 16.10. The fact that $1/\mathrm{JSD}$ for uBoost is roughly constant across p_{T} can be seen as an expression of this difference in training. Therefore, performing the adversarial training by starting from a classifier without pre-training may lead to more robust mass-decorrelation as a function of p_{T}. The corresponding impact on classification power is not clear. Alternative training procedures and improved architecture design may be able to optimally reconcile p_{T}-robustness and summary performance. Additionally, having the adversary decorrelate the classifier variable from the jet mass mass and p_{T} simultaneously, similar to what is done by k-NN and to a lesser extent DDT, rather than just treating p_{T} as auxiliary information, may also lead to more robust performance.

Finally, the mass-decorrelated taggers are found to be robust as a function of the number of reconstructed vertices. The background rejection for all mass-decorrelated taggers exhibits a regular, linear relationship with N_{PV}, with a slight negative slope. The behaviour of $1/\mathrm{JSD}$ as a function of N_{PV} is less regular, due to its dependence on the statistics in each bin. Since the simulated data samples are generated with a bell-shaped distribution of N_{PV}, centered around 15, the statistical limit on $1/\mathrm{JSD}$ will be maximal in this region, and decreasing with statistics towards lower and higher values of N_{PV}. However, the mass-decorrelated taggers all exhibit a regular behaviour across N_{PV} in terms of N_{PV}. Therefore, although the absolute performance of each tagger changes with N_{PV}, the relative performance of the taggers is largely unaffected by pile-up.

Finally, the classification metric can also be studied as a function of the jet p_{T} with the addition of a window selection on the jet mass, shown in Figure 16.11.

The qualitative behaviour of $1/\varepsilon_{\mathrm{bkg}}^{\mathrm{rel}}$ as a function of N_{PV} is unchanged. The application of the jet mass selection, however, does qualitatively affect the background rejection as a function of p_{T}. Applying the jet mass selection increases the classification power of all single-variable taggers as well as mass-decorrelated MVA taggers. In contrast, the application of the jet mass selection has little impact on $\varepsilon_{\mathrm{bkg}}^{\mathrm{rel}}$ for the standard MVA taggers, as these already utilise information about the mass close to optimally. With the addition of a jet mass selection, all mass-decorrelation procedures are seen to reduce the background rejection relative to their standard variants, in contrast to Figure 16.10. This is another representation of the effect seen in *e.g.* Figures 16.8b and 16.9b. By focusing on jets with masses close to the W boson mass, the identification of a known type of jets is emphasised over the shifts in the underlying substructure observable

distributions arising from correlations with the jet mass and p_{T}. This could otherwise give the appearance of the mass-decorrelation methods improving classification, due to shifts in substructure observable distributions outside of the W mass region.

The robustness of $1/\mathrm{JSD}$ as a function of p_{T} and N_{PV} is largely unchanged, as the additional jet mass window selection is only applied for classification, such that mass-decorrelation is always computed on the full jet mass spectrum, although with a background selection efficiency corresponding to that found with the addition of the jet mass selection. The main difference for the $1/\mathrm{JSD}$ plots is the statistical limits. These limits are computed using the average of background selection efficiencies across taggers, and since the addition of the jet mass selection results in these becoming smaller and more coherent, the systematic uncertainty assigned to the statistical limit on $1/\mathrm{JSD}$ is reduced substantially. Finally, although the background rejection for $D_2^{k\text{-NN}}$ increases with the addition of the jet mass selection, the behaviour of the $1/\mathrm{JSD}$ profile is qualitatively unchanged. This means that the turn-around at $p_{\mathrm{T}} \sim 1$ TeV is driven by the sudden emergence of the peaked structure in Figure 16.7c.

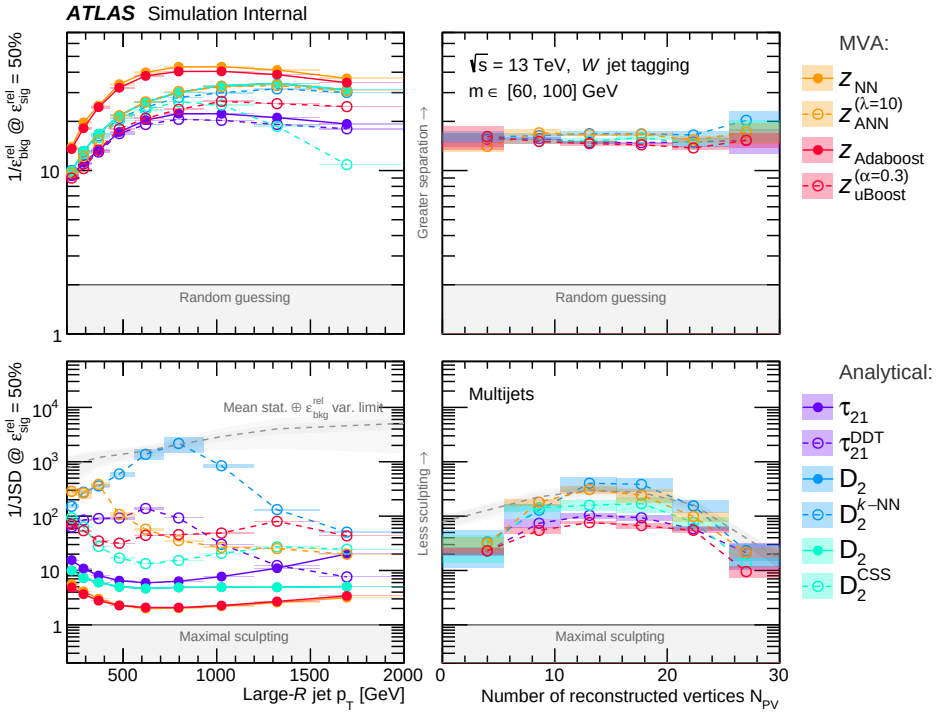


Figure 16.11 Plot of the metrics for classification (background rejection, $1/\varepsilon_{\text{bkg}}^{\text{rel}}$, with an additional jet mass selection of $m \in [60, 100]$ GeV; *top*) and mass-decorrelation (inverse Jensen-Shannon divergence, 1/JSD, calculated for the full jet mass distribution; *bottom*), for selections corresponding to $\varepsilon_{\text{sig}}^{\text{rel}} = 50\%$, as a function of the reconstructed jet p_{T} (*left*) and the number of reconstructed vertices, N_{PV} (*right*). The additional jet mass selection is applied only for classification, such that 1/JSD is always calculated for the full jet mass spectrum. Standard classifiers are indicated with filled markers. Mass-decorrelated classifiers indicated with open markers. Statistical uncertainties are indicated with shaded boxes, derived using bootstrap sampling. The statistical limit on 1/JSD, also accounting for variation in $\varepsilon_{\text{bkg}}^{\text{rel}}$ for different taggers within the same bin, is shown as a shaded grey band (*bottom*). Only mass-decorrelated taggers are shown for N_{PV} (*right*).

CHAPTER 17

Conclusion and outlook

This part has presented a study of various techniques for the construction of mass-decorrelated jet taggers for two-body hadronic resonance decays. Jets from the hadronic decay of high-p_{T} W boson are used to demonstrate the usefulness of such taggers, but the mass-decorrelation techniques should be applicable to hadronic two-body decays with different or unknown resonance masses. DDT, CSS, and fixed-efficiency k-NN regression are used to decorrelate single jet substructure observables from the jet mass. These are compared with MVA techniques, where the mass-decorrelation is performed using ANNs and adaptive boosting for uniform efficiency (uBoost), applied to BDT taggers. The multijet background rejection rate and JSD are proposed as metrics for evaluating classification and mass-decorrelation, respectively, and are studied in MC simulated data samples. Standard MVA taggers are found to yield superior classification performance compared to single-variable taggers, but exhibit strong non-linear correlations with the jet mass, potentially reducing sensitivity in searches for new physics. Fixed-efficiency k-NN regression is able to decorrelate single jet substructure observables to within statistical uncertainties. The ANN tagger is generally found to have better classification power for similar levels of mass-decorrelation than the remaining mass-decorrelated taggers in the primary kinematic regime of interest.

This study has found that, in particular, the k-NN and ANN methods for constructing mass-decorrelated jet taggers have the potential to benefit searches for new physics in the invariant mass spectrum of large-R jets, such as the search for low-mass leptophobic DM mediators presented in Part II of this thesis. By improving the rejection of large-R jets from the leading background process, the ANN method can improve the purity of jets from two-body resonance decays relative to the current generation of analyses, relying on single jet substructure observables for jet classification [1, 73, 155–157, 215]. By increasing signal purity in the final data samples, these methods can increase the

sensitivity of searches for BSM physics. Similarly, by mitigating the sculpting effects in the large-R jet mass spectrum, methods such as k-NN and ANN can reduce one of the leading systematic uncertainties in such searches, namely the uncertainty associated with the leading background estimate [1]. In these searches, a selection on the chosen jet tagger is used to create a signal-enriched region, in which the search for new physics is performed, along with one or more signal-depleted regions, used to estimate the leading background contribution in the signal-enriched region. By removing the mass-dependence of the jet tagger, a more robust leading background estimate can be performed, which in turn will reduce the systematic uncertainty on this estimate. The ATLAS search presented in Part II of this thesis used the DDT method, and the CMS Collaboration has employed the k-NN method in multiple searches [73, 156, 157]. Therefore, this study can hopefully contribute to an improved understanding of the relative merits of different approaches to the development of mass-decorrelated jet taggers, and thereby help improve future generations of searches for new physics in the large-R jet mass spectrum.

In the context of the search for hadronically decaying DM mediators presented in Part II, other analyses suggest that *e.g.* improvement large-R jet mass decorrelation may improve the existing coupling limits by as much as 30% [73, 214, 215]. Apart from this search, a number of other efforts in ATLAS — ranging from Higgs measurements to dark jets searches — are exploring the potential for using the mass-decorrelated jet taggers studied in this thesis in full Run 2 analyses. At present, no public results exist on the direct impact of these techniques for mass-decorrelated jet tagging on BSM searches. In addition, the results of this study are currently being adopted in the ATLAS Collaboration, with plans to develop and commission common mass-decorrelated jet tagging algorithm. Another promising use-case is trigger-level classification, where taggers decorrelated from *e.g.* mass may reduce the potential for a biased selection of data, and thereby support searches for anomalous or unexpected processes. This may be possible due to ongoing efforts to enable real-time inference of deep neural networks at the trigger level [252]. As proposed in Ref. [240], adversarial training of neural network classifiers may also be used more broadly to *e.g.* train them to be robust against parametrised systematic uncertainties. This could *e.g.* be used to mitigate the impact of the large-R jet uncertainties on the signal and W/Z processes in the analysis in Part II. Finally, these properties of decorrelated neural networks may also find use outside of physics, where they can be used to combat bias and discrimination, *e.g.* in automated approval and recommendation systems, which is known to occur on the basis of *e.g.* age, sex, gender, and ethnicity.

Appendices

APPENDIX A

Jet substructure observables

The N-subjettiness jet substructure observables were introduced in Chapter 7. The remaining jet substructure observables used in the technical study in Part III are detailed below.

Energy correction function ratios

An approach similar to N-subjettiness underlies the $C_2^{(\beta)}$ and $D_2^{(\beta)}$ variables [230], which are ratios of the two- and three-point energy correlation functions (ECFs) $e_2^{(\beta)}$ and $e_3^{(\beta)}$,

$$e_2^{(\beta)} = \sum_{1\leq i<j\leq n_J} z_i z_j \Delta R_{ij}^{\beta} \quad \text{and} \quad e_3^{(\beta)} = \sum_{1\leq i<j<k\leq n_J} z_i z_j z_k \Delta R_{ij}^{\beta} \Delta R_{ik}^{\beta} \Delta R_{jk}^{\beta} \tag{A.1}$$

specifically as

$$C_2^{(\beta)} = \frac{e_3^{(\beta)}}{\left(e_2^{(\beta)}\right)^2} \quad \text{and} \quad D_2^{(\beta)} = \frac{e_3^{(\beta)}}{\left(e_2^{(\beta)}\right)^3}, \tag{A.2}$$

respectively. Here, $z_i = p_{\mathrm{T}i}/p_{\mathrm{T}J}$ is the transverse momentum fraction carried by the i^{th} jet constituent, n_J is the number of jet constituents, ΔR_{ij} is the distance between the i^{th} and j^{th} jet constituents in the $\eta - \phi$ plane, and β is an angular exponent taken to be 1 in the following. For brevity, the notation $D_2 = D_2^{(\beta=1)}$ and $C_2 = C_2^{(\beta=1)}$ is used. The two-point ECF $e_2^{(\beta)}$ is the simplest such function which is able to probe the radiation structure inside the jet. However, in the case of a jet which is dominated by two symmetric, energetic prongs, as is the case of a jet reconstructing the hadronic two-body decay of a high-p_{T} W boson, the two-point ECF simply reduces to $e_2^{(\beta)} \sim \Delta R_{12}^{\beta} \approx (m_J/p_{\mathrm{T}J})^{\beta}$, where ΔR_{12} is the distance in $\eta - \phi$ between the two quarks from the W boson decay, see Equation (1.5). Therefore, both the two- and three-point ECFs are needed to resolve

the substructure inside the jets, and Ref. [230] finds the ECF ratios in Equation (A.2) to provide the optimal separation of jets dominated by one and two energetic prongs, respectively.

Angularity

The angularity variable a_3 used in ATLAS is defined as [231, 253]

$$a_3 = \frac{1}{m_J} \sum_{i=1}^{n_J} E_i \sin^{-2} \theta_i \, (1 - \cos \theta_i)^3, \tag{A.3}$$

where m_J is the invariant mass of the large-radius (large-R) jet, n_J is the number of constituents of the jet, E_i is the energy of the i^{th} constituent of the jet, and θ_i is the angle of the i^{th} jet constituent with respect to the large-R jet axis. The choice of exponents of the angular functions emphasises radiation near the edge of the large-R jet — *i.e.* large values of θ_i — and therefore measures the degree to which the jet is dominated by wide-angle radiation. Therefore, this definition of angularity leads to small values of a_3 centred around $a_3 \approx (m_J/2p_{\text{T}})^3$ for central, symmetric, energetic two-body decays of *e.g.* W bosons reconstructed as large-R jets with mass m_J and transverse momentum p_{T}, while the diffuse radiation from non-resonant jets initiated by single quarks or gluons will tend to produce distributions with longer tails towards large values of a_3.

Aplanarity

The aplanarity of a jet is a shape variable defined in the centre-of-mass frame of the large-R jet [232]. Boosting the large-R jet constituents to this frame is achieved by performing a general Lorentz transform of four-momenta $p = (E, \mathbf{p})$ with boost $-\boldsymbol{\beta}$, *i.e.*

$$\tilde{E} = \gamma(E + \boldsymbol{\beta} \cdot \mathbf{p}) \tag{A.4a}$$

$$\tilde{\mathbf{p}} = \mathbf{p} + \frac{\gamma - 1}{|\boldsymbol{\beta}|^2}(\boldsymbol{\beta} \cdot \mathbf{p})\boldsymbol{\beta} + \gamma E \boldsymbol{\beta}, \tag{A.4b}$$

where $\boldsymbol{\beta} = \mathbf{p}_J/E_J$ is the ratio of the three-velocity of the large-R jet to the speed of light in vacuum and γ is the corresponding Lorentz factor. With this transform, the 3×3

sphericity matrix is defined as

$$S^{k,l} = \frac{\sum\limits_{1 \leq i < j \leq n_J} \tilde{p}_i^k \tilde{p}_j^l}{\sum_{i=1}^{n_J} |\tilde{\mathbf{p}}_i|^2}, \qquad k, l \in \{x, y, z\}, \tag{A.5}$$

where n_J is the number of constituents in the large-R jet, $\tilde{\mathbf{p}}_i$ and $\tilde{p}_i$ are the three- and four-momenta, respectively, of the i^{th} constituent of the large-R jet in the centre-of-mass frame of the large-R jet, and k and l denote one of the three spatial components of the constituents' momenta. The sphericity matrix is diagonalised to yield three eigenvalues $\lambda_1 \geq \lambda_2 \geq \lambda_3$ which sum to 1. The aplanarity is defined as

$$A = \frac{3\lambda_3}{2}, \tag{A.6}$$

with $0 \leq A \leq 1/2$. For the isotropic or diffuse radiation found inside non-resonant jets, $\lambda_1 \approx \lambda_2 \approx \lambda_3$ and $A \approx 1/2$, whereas a highly directional radiation — *e.g.* from the decay of a heavy resonance to a back-to-back quark pair — results in $A \approx 0$.

Planar flow

Planar flow [231] measures the degree to which the energy of a large-R jet is spread evenly across the plane transverse to the jet axis (called planar radiation, corresponding to a large planar flow) or linearly along some axis in the transverse plane (called linear radiation, corresponding to a small planar flow). Such radiation patterns are studied using the components of the constituents' momenta transverse to the direction of the large-R jet. These are found by rotating the large-R jet, with three-momentum $\mathbf{P}$, and all constituents, with three-momenta $\mathbf{p}_i$, by $-\phi$ around the z-axis using the rotation matrix

$$\mathbf{R}_\phi = \begin{bmatrix} \cos\phi & \sin\phi & 0 \\ -\sin\phi & \cos\phi & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{A.7}$$

and then rotating them by $-\theta$ along the y-axis using the rotation matrix

$$\mathbf{R}_\eta = \begin{bmatrix} \cos\theta & 0 & -\sin\theta \\ 0 & 1 & 0 \\ \sin\theta & 0 & \cos\theta \end{bmatrix} = \begin{bmatrix} \tanh\eta & 0 & -\operatorname{sech}\eta \\ 0 & 1 & 0 \\ \operatorname{sech}\eta & 0 & \tanh\eta \end{bmatrix}, \tag{A.8}$$

using the definition of pseudorapidity in Equation (1.1). Here, ϕ and η are the azimuthal angle and pseudorapidity, respectively, of the large-R jet. This transform leaves

the large-R jet three-momentum pointing along the positive z-axis with unchanged magnitude, *i.e.*

$$\tilde{\mathbf{P}} = \mathbf{R}_\eta \, \mathbf{R}_\phi \, \mathbf{P} = \begin{bmatrix} \tanh\eta\cos\phi & \tanh\eta\sin\phi & -\operatorname{sech}\eta \\ -\sin\phi & \cos\phi & 0 \\ \operatorname{sech}\eta\cos\phi & \operatorname{sech}\eta\sin\phi & \tanh\eta \end{bmatrix} \begin{bmatrix} \mathbf{P}_x \\ \mathbf{P}_y \\ \mathbf{P}_z \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ |\mathbf{P}| \end{bmatrix}, \tag{A.9}$$

and therefore allows for studying the momenta of jet constituents transverse to the large-R jet axis. Using these matrices to transform $\tilde{\mathbf{p}}_i = \mathbf{R}_\eta \, \mathbf{R}_\phi \, \mathbf{p}_i$, where $\mathbf{p}_i$ is the three-momentum of the i^{th} large-R jet constituent, a 2×2 momentum matrix I_E is constructed as

$$I_E^{kl} = \frac{1}{m_J} \sum_{i=1}^{n_J} \frac{\tilde{\mathbf{p}}_i^k \tilde{\mathbf{p}}_i^l}{E_i}, \qquad k, l \in \{x, y\}, \tag{A.10}$$

where m_J is the mass of the large-R jet, n_J is the number of constituents in the large-R jet, E_i is the energy of each jet constituent, and $\tilde{\mathbf{p}}_i^k$ is the k^{th} component of the i^{th} constituent's momentum transverse to the large-R jet axis. The planar flow variable is then defined as

$$\mathcal{P} = \frac{3\det(I_E)}{\operatorname{tr}(I_E)^2} = \frac{4\lambda_1\lambda_2}{(\lambda_1 + \lambda_2)^2}, \tag{A.11}$$

where $\lambda_{1,2}$ are the eigenvalues of I_E. A planar flow of energy in the plane transverse to the large-R jet axis will yield $\lambda_1 \approx \lambda_2 \approx 1/2$ resulting in $\mathcal{P} \approx 1$. Conversely, a linear flow will yield $\lambda_1 \gg \lambda_2$, resulting in $\mathcal{P} \approx 0$. Since resonant two-body decays will tend to produce a linear flow of energy from the colour connection between the quarks originating from the decay, whereas non-resonant jets will tend to produce diffuse, isotropic radiation, planar flow is well suited to discriminate the two.

Fox-Wolfram moment

R_2^{FW} is the ratio of the second to zeroth order Fox-Wolfram moments [233]. The Fox-Wolfram moments H_l were proposed to calculate event shapes in e^+e^- collisions, but may be adapted to study large-R jet shapes by boosting the jet constituents into the centre-of-mass frame of the large-R jet using Equations (A.4). These moments are then

given as

$$H_l = \sum_{1 \leq i < j \leq n_J} \frac{|\mathbf{p}_i||\mathbf{p}_j|}{s} P_l(\cos\theta_{ij}), \qquad \text{(in centre-of-mass frame)} \tag{A.12}$$

where $\mathbf{p}_i$ is the three-momentum vector of the i^{th} large-R jet, $s = (\sum_i E_i)^2$ is the squared sum of energies of the jet constituents, P_l is the l^{th} Legendre polynomial, and θ_{ij} is the angle between the three-momentum vectors of the i^{th} and j^{th} large-R jet constituents in the centre-of-mass frame. The normalised second Fox-Wolfram moment is then defined as

$$R_2^{\text{FW}} = \frac{H_2}{H_0} = \frac{\sum\limits_{1 \leq i < j \leq n_J} |\mathbf{p}_i||\mathbf{p}_j|(3\cos^2\theta_{ij} - 1)/2}{\sum\limits_{1 \leq i < j \leq n_J} |\mathbf{p}_i||\mathbf{p}_j|}\,. \tag{A.13}$$

The normalised second moment is sensitive to symmetric, back-to-back radiation in the centre-of-mass frame — *i.e.* where $\cos^2\theta_{ij} \approx 1$ for all i, j — making R_2^{FW} a suitable observable to identify two-prong jets.

k_t-subjet ΔR

The so-called *KtDR* variable is found by re-clustering the constituents of the large-R jet into exactly two subjets. This is done using the k_t jet reconstruction algorithm, introduced in Chapter 1.3, with a radius parameter of $R = 0.4$. The *KtDR* variable is then defined as the distance ΔR between the two subjets in the $\eta - \phi$ plane. The k_t algorithm predominantly clusters soft and collinear constituents early in the clustering sequence, meaning that the last clustering step will generally be of the hardest splitting inside the jet. In the context of a two-body decay, this will generally correspond to the initial hard splitting of the resonance into two quarks. Therefore, this variable will have a characteristic value of $KtDR \approx 2m/p_{\text{T}}$ for a symmetric two-body decay of resonance with mass m and transverse momentum p_{T}, see Equation (1.5).

Splitting scales

The final two substructure variables used in this study are closely related and attempt to directly probe the energetic scale of an assumed two-body structure inside the jet.

The first variable, the k_t splitting scale $\sqrt{d_{ij}}$ [234] is found by clustering the constituents

of the large-R jets using the k_t algorithm and identifying the step where $j = i + 1$ subjets are recombined to i subjet(s). This splitting scale is based on the distance measure d_{ij} of k_t jet clustering algorithm, see Equation (1.4a), and is given by

$$d_{ij} = \min\left(p_{\mathrm{T}i}^2, p_{\mathrm{T}j}^2\right) \Delta R_{ij}^2 \,, \tag{A.14}$$

where $p_{\mathrm{T}i}$ is the transverse momentum of the i^{th} (partially recombined) subjet, and ΔR_{ij} is the distance in the $\eta - \phi$ plane between the i^{th} and j^{th} subjet. Specifically, the splitting scale $\sqrt{d_{12}}$ is a dimensionful measure of the energy in the final recombination step in the k_t clustering very similar to the *KtDR* variable defined above. Equation (A.14) shows that the k_t algorithm aims to primarily cluster soft and collinear subjets, which means that hard symmetric splittings like those found in resonant two-body decays, will generally be clustered last when using the k_t algorithm. $\sqrt{d_{12}}$ is therefore a proxy for the energy in the hardest splitting in the jet, which will tend to be much higher for symmetric tow-body decays than for the generally soft and collinear radiation in non-resonant jets.

A related variable is z_{ij} [235], defined as

$$z_{ij} = \frac{d_{ij}}{d_{ij} + m^2} \,, \tag{A.15}$$

where d_{ij} is the squared k_t splitting scale defined in Equation (A.14) and m is the mass of the subjet resulting from the recombination of subjets i and j. For the $2 \to 1$ recombination step used here, the recombined subjet will equal the full, large-R jet and consequently the subjet mass m will equal the large-R jet mass m_J. Specifically for two-body decays, z_{12} is an appropriate substructure variable for the same reasons as $\sqrt{d_{12}}$ above, and for two-body decays m is a proxy for the mass of the decaying resonance. By normalising to the (sub)jet mass m, z_{12} measures the symmetry of the hardest decay in the jet in a similar way to $\sqrt{d_{12}}$, but in a way which is less dependent on the (sub)jet mass.

Thrust

The jet thrust variables are defined as [254]

$$(\mathbf{v}_1), T = (\arg)\max_{|\mathbf{n}|=1} \frac{\sum_{i=1}^{n_J} |\mathbf{n} \cdot \tilde{\mathbf{p}}_i|}{\sum_{i=1}^{n_J} |\tilde{\mathbf{p}}_i|}, \tag{A.16a}$$

$$(\mathbf{v}_2), T_{\text{maj}} = (\arg)\max_{|\mathbf{n}|=1,\, \mathbf{n}\perp\mathbf{v}_1} \frac{\sum_{i=1}^{n_J} |\mathbf{n} \cdot \tilde{\mathbf{p}}_i|}{\sum_{i=1}^{n_J} |\tilde{\mathbf{p}}_i|}, \quad \text{and} \tag{A.16b}$$

$$T_{\text{min}} = \frac{\sum_{i=1}^{n_J} |\mathbf{v}_3 \cdot \tilde{\mathbf{p}}_i|}{\sum_{i=1}^{n_J} |\tilde{\mathbf{p}}_i|}, \qquad \text{for } \mathbf{v}_3 = \pm(\mathbf{v}_1 \times \mathbf{v}_2) \tag{A.16c}$$

where T is the thrust variable, n_J is the number of large-R jet constituents and $\tilde{\mathbf{p}}_i$ is the three-momentum of the i^{th} jet constituent in the large-R jet centre-of-mass frame.

ATLAS-specific jet taggers

ATLAS uses a number of non–machine learning (ML) jet taggers in addition to single analytical jet substructure observables. These include more advanced methods, shown in Figure 12.1, such as:

- two-variable optimised taggers, where a threshold selection on the substructure observable and a window selection on the large-R jet mass are jointly optimised in bins of jet p_{T} to yield maximum background jet rejection at a fixed W/top signal selection efficiency;
- the HEPTopTagger [255, 256], which reconstructs top candidates as Cambridge/Aachen (CA) jets with $R = 1.5$, applies the trimming algorithm to the constituents, and tests the resulting set of which are tested against a hadronic three-body top decay hypothesis. Jets passing this selection with reconstructed masses close to the top quark mass are considered tagged;
- and finally Shower Deconstruction [257, 258], where a large-R jet compatible with a three-body top decay hypothesis is reclustered into 3–6 exclusive k_t subjets which are considered proxies for outgoing hard process partons, and the likelihood of obtaining this configuration based on signal (top) and background (non-resonant parton emission) hypotheses, respectively, is evaluated using a set of simplified parton shower histories.

In Ref. [39] it was also found that a deep neural network (DNN)-based top tagger

using the kinematic properties (p_T, η, ϕ) of the 10 leading jet constituents as inputs [259] was able to improve classification, compared to multivariate analysis (MVA) taggers using high-level features as input, in the kinematic regime in which it was trained. This is consistent with the experience from *e.g.* Ref. [260] that MVA methods with sufficient capacity acting on low-level inputs (*e.g.* the kinematic properties of jet constituents) can provide more powerful classification than a simple MVA combination of physics-motivated variables derived from such low-level inputs. This is an indication that there is typically additional information in the higher-dimensional set of low-level variables, that analytical observables are unable to extract, unlike sufficiently complex MVA models. A similar approach has been employed by CMS through DeepJet and DeepBoostedJet [261], which use kinematic information from both charged and neutral particles as well as secondary vertices to perform small-radius (small-R) and large-R jet classification.

Quark separation in two-body decays

Finally, the rule of thumb in Equation (1.5), describing the decay of a massive resonance to two quarks, can be derived as follows. To do this, the definition of the invariant mass in Equation (1.6) is used. The invariant mass m of the decay products, and therefore also of the resonance itself, is approximately given by

$$\begin{aligned} m^2 = (p_1 + p_2)^2 &= \cancel{p_1^2} + \cancel{p_2^2} + 2p_1 \cdot p_2 \approx 2E_1E_2(1 - \cos\theta_{12}) \\ &\approx E^2 z(1-z)\,\theta_{12}^2 \qquad \text{for } \theta_{12} \ll 1 \iff m \ll E \end{aligned}$$

which gives

$$R_{12} = \frac{m}{E}\frac{1}{\sqrt{z(1-z)}} \approx \frac{2m}{p_T} \qquad \text{for } z \approx 1/2 \text{ and } E \approx p_T, \tag{A.17}$$

where $p_{1,2}$ and $E_{1,2}$ are the four-momenta and energies, respectively, of the two approximately massless quarks produced in the resonance decay; s θ_{12} is the angle between the two quarks with respect to the direction of motion of the resonance, which corresponds roughly to the distance R_{12} between the two quarks in the $\eta - \phi$ plane for a central decay; and finally z and $(1 - z)$ are the fractions of the energy E of the original resonance carried by each of the two quarks. For symmetric decays ($z \approx 1/2$) in the boosted kinematic regime ($p_T \gg m$), the last approximation holds.

APPENDIX B

Machine learning fundamentals

This appendix provides some additional background material on the basics of the two machine learning (ML) algorithms used in this thesis, neural networks (NNs) and boosted decision trees (BDTs). This is intended to complement material in Chapter 4.

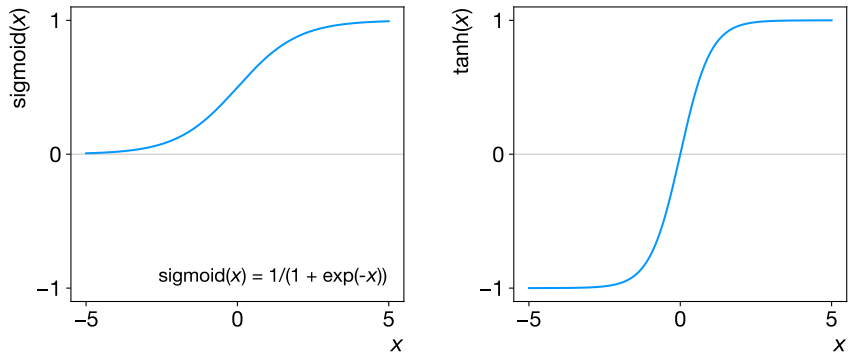
Activation functions

Typical examples of NN activation functions h are shown in Figure B.1.

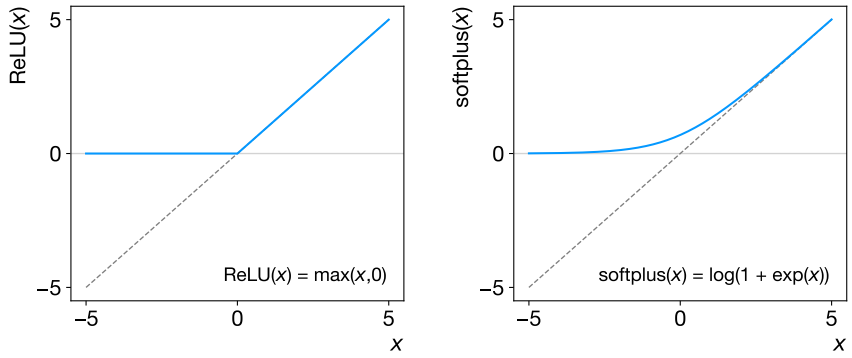
In principle, h may be taken to be the identity map for all layers in the NN. However, in that case, the connections in Equation (4.1) would result in the outputs $\mathbf{y}$ simply being a direct linear combination of the inputs $\mathbf{x}$. Non-linear activation functions breaks this would-be redundancy of hidden layers and is what allows the NN to model complex, non-linear relations. In general, different activation functions may be applied for different layers. Similarly, the activation of the output layer is generally chosen to suit the specific classification or regression task at hand. For instance, for binary classification of target labels 0 or 1, the sigmoid activation is suitable. Alternatively, for regression of targets which may assume positive or negative values of indefinite magnitude, the identity activation may be the most appropriate choice.

Back-propagation

To train the NN according to such losses such as those in Equations (4.2) and 4.3, a prescription for estimating the gradient $\partial L/\partial w_{ij}^{(l)}$, that does not require individually varying each parameter in the network weight space, is required. Instead, the gradient may be estimated using the chain rule of differentiation, since a standard NN with D



(a) Sigmoid **(b)** Hyperbolic tangent (tanh)



(c) Rectified linear unit (ReLU) **(d)** Softplus

Figure B.1 Four standard neural network (NN) activation functions, or non-linearities: **(a)** sigmoid, **(b)** hyperbolic tangent (tanh), **(c)** ReLU, and **(d)** softplus.

hidden layers is simply as a composition of $D + 1$ non-linear differentiable functions, see Equation (4.1). This will allow for the training of the NN using an approximate gradient descent algorithm.

In a densely connected network, the variations in $a^{(\ell-1)}$ give rise to changes in L_{MSE} only through their connection to $a^{(\ell)}$, and the resulting variations therein, see Equation (4.1). Similarly, the outputs $a^{(\ell)}$ depend directly on the networks weights feeding into the l^{th} layer $w_{ij}^{(l-1)}$, see Figure 4.1a. Since all connections in the network, as well as the loss function, are differentiable with respect to the network weights, the chain rule of differentiation implies that

$$\left.\frac{\partial L}{\partial w_{ij}^{(l)}}\right|_{\{\mathbf{x},\mathbf{y}\}} \equiv \frac{\partial L}{\partial w_{ij}^{(l)}} = \underbrace{\frac{\partial L}{\partial a_i^{(l+1)}}}_{\equiv \delta_i^{(l+1)}} \frac{\partial a_i^{(l+1)}}{\partial w_{ij}^{(l)}} = \delta_i^{(l+1)} z_j^{(l)}\,, \tag{B.1}$$

where the so-called error $\delta_i^{(l)}$ has been introduced as short-hand, and where Equation (4.1) has been used in the second factor. That is, the gradient of the loss with respect to a particular network weight is uniquely determined by the activation of the incoming node and some error on the output node to which the weight is connected, see Figure 4.1b. This gradient calculation is valid for a given set of inputs $\mathbf{x}$ and targets $\mathbf{y}$, giving rise to the activation and error values in Equation (B.1). However, this distinction will be made implicit below for convenience. The error in layer l can be further decomposed, again using the chain rule of differentiation

$$\delta_i^{(l)} = \frac{\partial L}{\partial a_i^{(l)}} = \sum_k \underbrace{\frac{\partial L}{\partial a_k^{(l+1)}}}_{=\delta_k^{(l+1)}} \underbrace{\frac{\partial a_k^{(l+1)}}{\partial a_i^{(l)}}}_{=w_{ki}^{(l)} h'\left(a_i^{(l)}\right)} \approx h'\left(a_i^{(l)}\right) \sum_k \delta_k^{(l+1)} w_{ki}^{(l)} \,, \tag{B.2}$$

where the definition of the error and Equation (4.1) have been used again, and where h' denotes the first derivative of the activation function. Equation (B.2) relates the error associated with each node in layer l to the errors associated with the nodes in layer $l+1$. The starting condition is given by the errors on the output, at layer $l = D+1$ for a NN with D hidden layers

$$\delta_i^{(D+1)} = \frac{\partial L}{\partial a_i^{(D+1)}} = \frac{\partial L}{\partial z_i^{(D+1)}} \frac{\partial z_i^{(D+1)}}{\partial a_i^{(D+1)}} = \frac{\partial L}{\partial p_i} h'(a_i^{(D+1)}). \tag{B.3}$$

Recursive application of Equation (B.2) therefore allows for the back-propagation of errors, from the output layer ($l = D+1$) to the input ($l = 0$). Given these errors, the approximate derivative of the loss with respect to the individual weights throughout the network is given by Equation (B.1). For each set of inputs $\mathbf{x}$ and associated targets $\mathbf{y}$, the optimisation loss can therefore, in the simplest case, be minimised by updating the network weights according to

$$w_{ij}^{(l)} \leftarrow w_{ij}^{(l)} - \eta \left.\frac{\partial L}{\partial w_{ij}^{(l)}}\right|_{\{\mathbf{x},\mathbf{y}\}} = w_{ij}^{(l)} - \eta\, \delta_i^{(l+1)} z_j^{(l)} \,, \tag{B.4}$$

where the learning rate η controls the size of each weight update, and thereby the rate of the so-called stochastic gradient descent. By iteratively performing weight updates according to Equation (B.4), the network is trained to minimise $L(\theta)$.

Regularisation

Cross-validation may be used to test for over-fitting and under-fitting. The former manifests as a large difference between the optimisation loss L when evaluated on the training and validation datasets, respectively, indicating that the NN is tuned to perform well on the training data at the cost of a poor ability to generalise to unseen data. The latter manifests as comparable but large training and validation losses. Over-fitting may be addressed by reducing the capacity of the network by choosing a more restrictive architecture; alternatively so-called regularisation may be introduced. Regularisation refers to a set of techniques for preventing over-fitting, *e.g.* by penalising large network weights $w_{ij}^{(l)}$ through an additional regularisation term in the loss L. An alternative method is dropout regularisation [262], where in every forward pass during the training phase, a random subset of network connections are set to zero, or "dropped out." This way, the NN is forced to learn "broader" and more robust paths of information flow as it cannot rely on individual high-weight connections, since in every forward pass they may be disabled. The fraction of connections that are dropped out is part of the so-called hyperparameters of the NN, which also include the architecture, learning rate, choice of activation function, *etc.*. These are typically subject to optimisation using the cross-validation method described above.

Decision tree node splitting

For a decision tree (DT), the Gini impurity $I(\mathcal{N})$ for each child node $\mathcal{N}$ resulting from a potential split on the root node is defined as

$$I(\mathcal{N}) = \sum_{1 \leq c \leq n_{\text{classes}}} p_c(1 - p_c), \tag{B.5}$$

where c enumerates class labels, and p_c is the proportion of samples of class c assigned to node $\mathcal{N}$ after the split. The Gini impurity is the product of the probability for correctly selecting a sample of class c on the node in question, and the probability for misclassification. Therefore, as the name implies, it measures the degree of impurity of the dominant class when evaluated on node $\mathcal{N}$. In the case of binary classification of classes A and B, the Gini impurity is given by $I(\mathcal{N}) = 1 - p_A^2 - p_B^2$. It is minimised when $p_c = 1$ and $p_{k \neq c} = 0$, *i.e.* when a node perfectly classifies samples of class c, and maximised for $p_c = 0.5$. Since classification and regression tree (CART) trees are binary, each child node will contain samples with feature values either smaller or greater

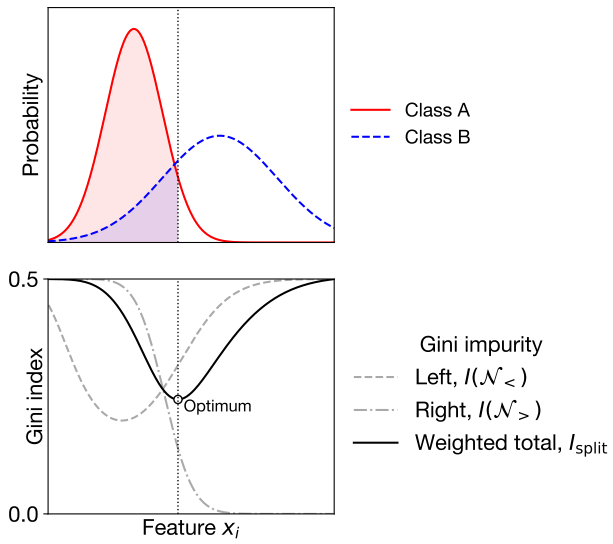


Figure B.2 Example of the procedure for determining the optimal split using the Gini impurity $I(\mathcal{N})$. Top panel: distributions of classes A and B for a feature x_i which provides separation. Bottom panel: Gini impurity as a function of the corresponding threshold value on x_i for the potential left and right node as well as the weighted average of the two. The vertical line indicates the optimal split, determined as the minimum of the weighted Gini impurity.

than some threshold value θ. For this reason, it is convenient to refer to the two child nodes as the left and the right one, or $\mathcal{N}_<$ and $\mathcal{N}_>$, respectively. The overall metric for a particular split is found as the average of the Gini impurities for each of the potential child nodes weighted by the proportion of samples assigned to each node after the split

$$I_{\text{split}} = \frac{I(\mathcal{N}_<) \times w_< + I(\mathcal{N}_>) \times w_<}{w_< + w_>}\,, \tag{B.6}$$

where $w_{<,>}$ are the weighted number of samples on the left and right child node. An example of this calculation is shown in Figure B.2.

Adaboost

The AdaBoost algorithm proceeds by training a set of DTs, considered weak learners, each tree DT^t enumerated by the boosting step counter t. Initially, a single $\text{DT}^{t=0}$ is constructed for the training dataset with some initial example weights $\{w_i^{t=0}\}$, possibly

all ones. Then it is determined whether this DT misclassified the training data

$$\gamma_i^t = \begin{cases} 1 & \text{if example } i \text{ was misclassified by DT}^t \\ 0 & \text{otherwise} \end{cases} . \tag{B.7}$$

Using this, the DT error for boosting step t is calculated as the weighted average of misclassifications

$$e^t = \frac{\sum_i w_i^t \gamma_i^t}{\sum_i w_i^t} . \tag{B.8}$$

For a perfect classifier, e^t will go to zero. Using this DT error, the DT weight is given by

$$a^t = \log\left[(1 - e^t)/e^t\right] . \tag{B.9}$$

Finally, the classification boosting weight is computed as

$$c_i^t = \exp(a^t \gamma_i^t) \tag{B.10}$$

and is used to update the training weights as

$$w_i^{t+1} = w_i^t \times c_i^t . \tag{B.11}$$

These training examples weights are then used to train the next weak learner DT^{t+1} for boosting step $t + 1$, and so on. The boosting proceeds for a fixed number of boosting steps, similar to the number of NN training epochs, where at step $t + 1$ AdaBoost assigns increased importance to training examples that were misclassified by DT^t through Equation (B.11). The full set of boosted, weak learners is then combined as the average over the DTs weighed by a^t.

APPENDIX C

Gaussian process techniques

In this thesis, Gaussian process (GP)-based techniques are used both in Part II for the transfer factor (TF) estimate of the leading background process and in Part III for Bayesian optimisation of neural network (NN) hyperparameters. This appendix provides some additional details on these techniques.

Gaussian process regression

GP regression [199] is based on the covariance of function values y at different measurement sites x, expressed as a relation between the inputs through a so-called kernel, K: $\mathrm{cov}(y_1, y_2) = K(x_1, x_2)$. This analysis uses the squared-exponential, or Gaussian, kernel[1]

$$K(x_1, x_2) = \exp\left(-\frac{|x_1 - x_2|^2}{2\ell^2}\right). \tag{C.1}$$

Here, ℓ is a length scale controlling the characteristic range within which functional values are correlated. Generally, inputs $\mathbf{x}$ may be d-dimensional, with one characteristic length scale ℓ_d per dimension. Below, X will denote a list of n inputs $\mathbf{x}$, and $\mathbf{y}$ a list of measurements y. In the analysis in Part II, the inputs to the GP regression will be pairs of $\mathbf{x} = \left(\rho^{\mathrm{DDT}}, \log(p_{\mathrm{T}}/\mu)\right)$, taken to be the centre of each bin in histograms like the one in Figure 9.2, and the associated measurement y will be the value of $\mathrm{TF}_{\mathrm{meas}}$ in the corresponding bin.

Being non-parametric, the only free parameters in the GP regressions are the length scales $\{\ell_d\}$. Given these, and a set of n measurements $\{X, \mathbf{y}\}$ (*i.e.* the large-radius (large-R)

[1] The following expressions are simplified by supposing that the inputs $\mathbf{x}$ and targets $\mathbf{y}$ are standardised to zero mean and unit variance [199, 200].

jet mass sidebands), the mean and variance function for the value of the underlying function at a possibly new set of measurement points X^* (*i.e.* the signal region (SR) window) are given by [199]

$$\mu(X^*) = K(X^*, X)^\top \left[K(X, X) + \sigma_n^2 \mathbb{I}\right]^{-1} \mathbf{y} \qquad \text{and} \tag{C.2a}$$

$$\text{var}(X^*) = K(X^*, X^*) - K(X^*, X)^\top \left[K(X, X) + \sigma_n^2 \mathbb{I}\right]^{-1} K(X, X^*), \tag{C.2b}$$

where σ_n is the uncertainty associated with each of the n measurements $\mathbf{y}$ and $\mathbb{I}$ is the $n \times n$ identity matrix. The expression in Equation (C.2a) corresponds to a linear combination of function values $\mathbf{y}$, weighted by the proximity of the test sites X^* to the measurement sites X. In Equation (C.2b), the first term is the uniform covariance prior and the second term accounts for updates to this prior through the provided measurements.

In addition to a best-fit value for the regression to the data, $\mu(X^*)$, the GP regression also provides an estimate of the uncertainty on this value, given by $\sqrt{\text{var}(X^*)}$. GP regression therefore provides a natural notion of the uncertainty inherent in the regression, which will be used directly as a systematic uncertainty associated with the data-driven background estimation procedure: The best-fit estimate of the large-R jet mass spectrum of the leading background in the pass region is given by Equation (9.2) with $\text{TF}_{\text{pred}}(X^*) = \mu(X^*)$, and systematic variations of this estimate are found by varying TF_{pred} up and down by $\pm\sqrt{\text{var}(X^*)}$. The resulting variation in the TF estimate of the leading background in the large-R jet mass spectrum will reflect the underlying uncertainty in the GP regression.

The GP length scales along the ρ^{DDT} and $\log(p_{\text{T}}/\mu)$ axes in the analysis in Part II are not given *a priori*. Therefore, they are determined by maximising the log-likelihood given in Ref. [200] modulo a constant term [201]

$$\log L(\mathbf{y} \,|\, X, \{\ell_d\}) = -\frac{n}{2} \log\left[\mathbf{y}^\top \left(K(X, X) + \sigma_n^2 \mathbb{I}\right)^{-1} \mathbf{y}\right] - \frac{1}{2} \log\left|K(X, X) + \sigma_n^2 \mathbb{I}\right|, \tag{C.3}$$

where n is the number of measurements $\{(\mathbf{x}, y)\} \sim X, \mathbf{y}$ and $|\cdot|$ denotes the matrix determinant. The kernel is implicitly parametrised by the characteristic length scales $\{\ell_d\}$ through Equation (C.1). Figure C.1 provides an example of a log-likelihood similar to Equation C.3, shown as a function of some arbitrary squared-exponential kernel length scale ℓ.

The first term in Equation (C.3) quantifies the quality of the regression to the measurement data, and the second term penalises model complexity. For large length

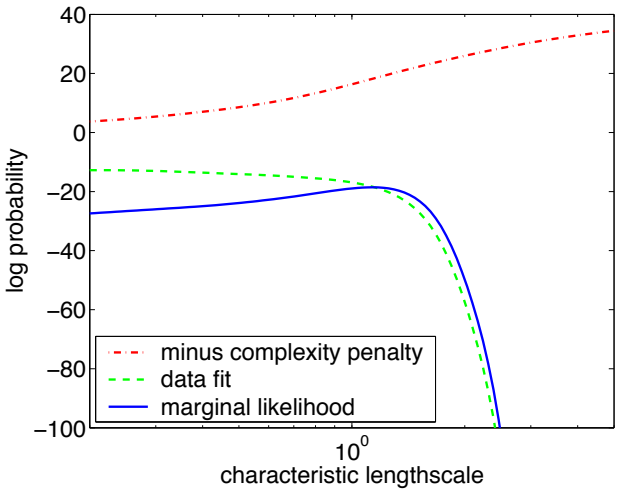


Figure C.1 Example of the decomposition of a log-likelihood similar to Equation (C.3) into the data-fit term (first term in Eq. (C.3)) and the term penalising model complexity (second term in Eq. (C.3)), as a function of the characteristic Gaussian process (GP) squared-exponential length scale. Figure from Ref. [199].

scales, the covariance matrix $K(X, X)$ approaches the $n \times n$ matrix of all ones, maximising the second term in Equation (C.3) by simplifying the model, at the cost of a poor fit to the data. This is called under-fitting. For short length scales, $K(X, X)$ approaches the identity $n \times n$ matrix, maximising the first term in Equation (C.3) by allowing for a more flexible regression, at the cost of increased model complexity. This is called over-fitting. The length scales $\{\ell_d^*\}$ found by optimising the log-likelihood in Equation (C.3) balance these two requirements, yielding a regression with the largest possible length scales that still provide an acceptable fit to the data.

Bayesian optimisation

Bayesian hyperparameter optimisation is based on GP regression, as presented above. The optimisation of NN hyperparameters is performed in a generally high-dimensional parameter space, and each evaluation of the optimisation metric is typically very compute-expensive and time-consuming, since it requires the full training of one or more NNs. This means that an exhaustive grid search of hyperparameters is not feasible. By using GP regression, it is possible to efficiently determine which unseen hyperparameter configurations have the largest potential to yield an optimum, making this type of regression a good tool for this type of optimisation problem.

Given a set of evaluated hyperparameter configurations $\mathcal{D} = \{h_i\}$, the corresponding

optimisation metric values $\{f(h) \,|\, h \in \mathcal{D}\}$ can be fitted using GP regression. The GP posterior provides a mean function $\mu(h \,|\, \mathcal{D})$ and a variance function $\sigma^2(h \,|\, \mathcal{D})$ for the optimisation metric $f(h)$, given evaluations $\mathcal{D}$. For a given hyperparameter configuration h, the true optimisation metric $f(h)$ is then expected to lie within $\mu(h \,|\, \mathcal{D}) \pm \sigma(h \,|\, \mathcal{D})$ with a coverage of approx. 68%. For a minimisation problem, the function

$$\gamma(h \,|\, \mathcal{D}) = \frac{f(h') - \mu(h \,|\, \mathcal{D})}{\sigma(h \,|\, \mathcal{D})}, \tag{C.4}$$

where $h' = \operatorname{argmin}_{h \in \mathcal{D}} f(h)$ is the current best hyperparameter configuration, provides a useful heuristic by which new hyperparameter configurations can be selected: for $\gamma(h \,|\, \mathcal{D}) \approx 0$, h is expected to yield an optimisation metric value which is comparable to the current best value; for $\gamma(h \,|\, \mathcal{D}) > 0$, h is expected to yield an optimisation metric value which is better than the current best value; and *vice versa*. To select new hyperparameter configurations to query, an acquisition function is needed to guide the sampling of new h values to evaluate in a possibly vast hyperparameter space. The basis of the acquisition function used in Spearmint, the library used in Appendix E, is the 'expected improvement' utility function

$$u_{\mathrm{EI}}(h \,|\, \mathcal{D}) = \max\left(0, f(h') - \mu(h \,|\, \mathcal{D})\right) \quad \Longleftrightarrow \quad u_{\mathrm{EI}}(y) = \sigma(h \,|\, \mathcal{D}) \max(0, \gamma' - \gamma), \tag{C.5}$$

which measures the average, expected improvement at h relative to the current best value h'. In the second equation, $\gamma' = \gamma(h' \,|\, \mathcal{D})$ and the dependence of γ on h and $\mathcal{D}$, and γ' on $\mathcal{D}$, is implicit. Since the GP regression provides an uncertainty estimate on $f(h)$, in the form of $\sigma(h)$, in addition to the mean, best-fit value $\mu(h)$, an acquisition function based on the expected improvement can be defined as

$$\begin{aligned}
\alpha_{\mathrm{EI}}(h \,|\, \mathcal{D}) = \mathbb{E}[u(\gamma \,|\, \mathcal{D})] &= \int_{-\infty}^{\infty} u(\gamma \,|\, \mathcal{D}) \mathcal{N}(\gamma \,|\, 0, 1) \,\mathrm{d}\gamma = \sigma(h \,|\, \mathcal{D}) \int_{-\infty}^{\gamma'} (\gamma' - \gamma) \mathcal{N}(\gamma \,|\, 0, 1) \,\mathrm{d}\gamma \\
&= \sigma(h \,|\, \mathcal{D}) \left[\gamma' \Phi(\gamma' \,|\, 0, 1) + \mathcal{N}(\gamma' \,|\, 0, 1)\right] \\
&= \quad \left[f(h') - \mu(h \,|\, \mathcal{D})\right] \Phi(f(h') \,|\, \mu(h \,|\, \mathcal{D}), \sigma(h \,|\, \mathcal{D})) \\
&\quad + \sigma(h \,|\, \mathcal{D}) \, \mathcal{N}(f(h') \,|\, \mu(h \,|\, \mathcal{D}), \sigma(h \,|\, \mathcal{D})) \,,
\end{aligned} \tag{C.6}$$

where $\mathbb{E}$ denotes the expectation value, $\mathcal{N}(x \,|\, \mu, \sigma)$ is the normal distribution function in x with mean μ and width σ, and Φ is the associated cumulative distribution function. Here, the identity

$$\int_{-\infty}^{x} t \mathcal{N}(t \,|\, 0, 1) \,\mathrm{d}t = -\mathcal{N}(x \,|\, 0, 1) \tag{C.7}$$

has been used. New hyperparameter configurations h^* to be evaluated can then be queried iteratively as

$$h^* = \underset{h}{\operatorname{argmax}}\, \alpha_{\mathrm{EI}}(h \mid \mathcal{D}), \tag{C.8}$$

where $\{f(h) \mid h \in \mathcal{D}\}$ is re-fitted upon each new evaluation $f(h^*)$, which updates $\mu(h \mid \mathcal{D})$ and $\sigma(h \mid \mathcal{D})$ and, in turn, $\alpha_{\mathrm{EI}}(h \mid \mathcal{D})$. From Equation (C.6), it is seen that two factors contribute to a large expected improvement: best-fit mean values which are smaller than the current best value (first term; called 'exploitation') and large uncertainty bands (second term; called 'exploration'). Since the acquisition function $\alpha_{\mathrm{EI}}(h \mid \mathcal{D})$ is fast to evaluate compared to $f(h)$, it can be used as an efficient surrogate function to be optimised through Equation (C.8). These features allow the Bayesian optimisation to efficiently probe a large parameter space by only evaluating new hyperparameter configurations with the largest expected improvement.

APPENDIX D

Alternative mass-decorrelation techniques

Chapter 15 briefly introduced the five methods for constructing mass-decorrelated jet taggers studied in this thesis. This appendix provides a more in-depth description of the first four mass-decorrelation methods, from simple linear transforms to specialised boosting of decision tree (DT) classifiers. Additional details on adversarial neural networks (ANNs) are given in Appendix E.

D.1 Designed decorrelated taggers

A simple approach to substructure decorrelation is provided by designed decorrelated taggers (DDT) [189], which was introduced and used also in Chapter 6. The original method relies on the jet scaling variable $\rho = \log\left(m^2/p_{\mathrm{T}}^2\right)$, where m is the mass of the jet, and p_{T} is the transverse momentum. It is observed empirically that profiling the jet substructure observable τ_{21}, defined in Chapter 1.3, as a function of ρ exposes a linear relationship. This can be exploited to perform a linear transform, removing the mean bias of τ_{21} with respect to ρ.

In practice [1, 155] it is found that the jet scaling variable

$$\rho^{\mathrm{DDT}} = \log\left(\frac{m^2}{p_{\mathrm{T}} \times \mu}\right) \tag{D.1}$$

more robustly removes residual dependence on the jet p_{T}, and therefore leads to better decorrelation across the jet kinematic phase space. Here, the parameter μ balances the dimensions and is taken to be $\mu = 1$ GeV.

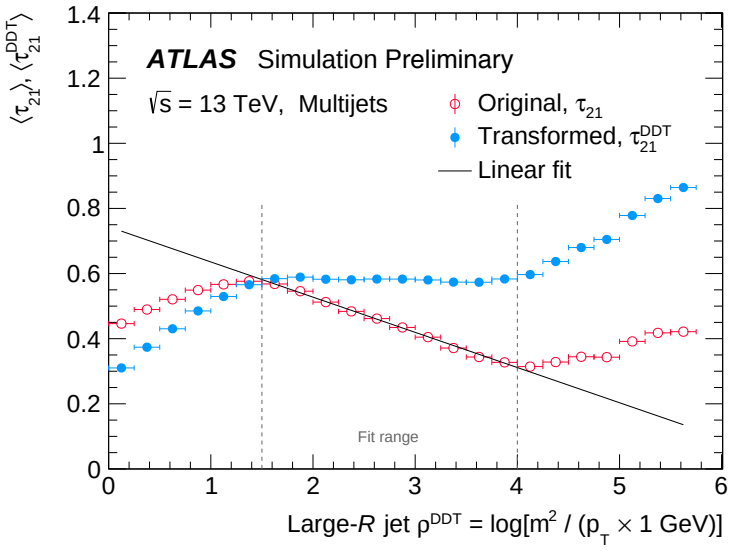


Figure D.1 Mean values of τ_{21} and τ_{21}^{DDT} as functions of ρ^{DDT} for the multijet background. A linear fit to the τ_{21}-profile, used in the definition of τ_{21}^{DDT}, is performed on the indicated range.

To perform the decorrelation, the mean value of τ_{21} for the multijet training sample is plotted as a function of ρ^{DDT}. This is shown in Figure D.1. A linear relationship between the two variables is observed, roughly in the range $\rho^{\text{DDT}} \in [1.5, 4.0]$. As explained in Section 8.2, the linearity breaks down towards higher values due to effects arising from the fixed-radius jet clustering algorithm used. Towards low values of ρ^{DDT}, the linearity breaks down due to soft QCD effects becoming dominant.

A linear fit is performed to the τ_{21} profile in the range $\rho^{\text{DDT}} \in [1.5, 4.0]$. From this fit, the transform $\tau_{21} \mapsto \tau_{21}^{\text{DDT}}$ is defined as

$$\tau_{21}^{\text{DDT}} = \tau_{21} - a \times (\rho^{\text{DDT}} - 1.5), \tag{D.2}$$

where $a = -0.108 \pm 0.002$ is the measured slope of the fit in Figure D.1, which shows how the DDT transformation removes the linear correlation of τ_{21} with ρ^{DDT}. Consequently, since ρ^{DDT} encodes information about both the jet mass and p_{T}, the DDT transform yields a jet substructure discriminant which is decorrelated from both of these kinematic variables in a specific region of phase space. Jets with $\rho^{\text{DDT}} \notin [1.5, 4.0]$ are kept and the transform in Equation (D.2) is applied to these as well to have a common basis for comparison with other mass-decorrelation methods.

D.2 Fixed-efficiency regression

The DDT transform requires the existence of a linear relationship between a substructure variable and kinematic variable(s) in order to remove the mean bias, which means that the method is only applicable to substructure observables with exhibit this property. The more general strategy of fixed-efficiency regression does not have such a requirement, which allows it to decorrelate a larger set of substructure observables from the jet mass. In this study, D_2, defined in Appendix A, is used as the base variable for mass-decorrelation based on fixed-efficiency regression.

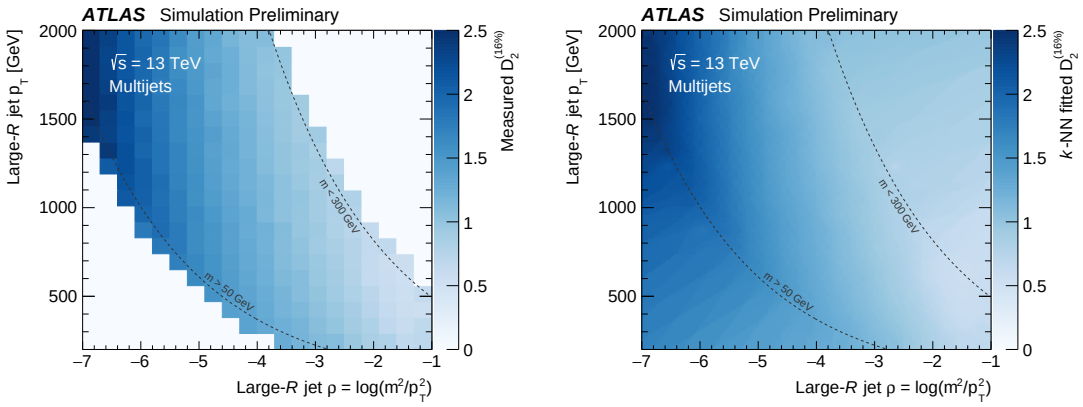
First, the value of a selection threshold on D_2 corresponding to a certain percentage of background efficiency $\varepsilon_{\text{bkg}}^{\text{rel}}$ is computed for multijets in bins of $\rho = \log(m^2/p_{\text{T}}^2)$ and p_{T}. This results in the two-dimensional profile shown in Figure D.2a. The target background selection efficiency used in this case is $\varepsilon_{\text{bkg}}^{\text{rel}} = 16\%$, which is found to correspond roughly to a signal efficiency of $\varepsilon_{\text{sig}}^{\text{rel}} = 50\%$. Second, a two-dimensional non-parametriec regression to this measured profile is performed using the distance-weighted k-nearest neighbours (k-NN) algorithm [223] with $k = 5$. The k-NN fit is performed, yielding the fitted profile $D_2^{(16\%)}(\rho, p_{\text{T}})$ shown in Figure D.2b. Finally, for each jet a new observable $D_2 \mapsto D_2^{k\text{-NN}}$ is constructed by subtracting the fit values from D_2,

$$D_2^{k\text{-NN}} = D_2 - D_2^{(16\%)}. \tag{D.3}$$

The fixed-efficiency regression generalises the central concept behind DDT, thereby making the method admissible to a more general class of substructure variables. Crucially, the k-NN method removes the dependence of a particular substructure observable on the jet mass at a selection threshold value corresponding to a specific background efficiency, whereas DDT removes the mean bias.

D.3 Convolved substructure

The DDT and k-NN methods introduced above perform the mass-decorrelation by subtracting a fitted prediction from a base substructure observable on a jet-by-jet basis, in order to remove an overall bias. These methods effectively make the first-order moment of the distribution independent of the jet mass (and p_{T}), but do not address the higher-order moments of this distribution. Using D_2 as the base substructure observable to be decorrelated from the jet mass, Ref. [224] proposes convolved substructure (CSS)



(a) Measured **(b)** k-NN fitted

Figure D.2 Profiles of the $\varepsilon_{\text{bkg}}^{\text{rel}} = 16\%$ profile of D_2 for multijets, $D_2^{(16\%)}(\rho, p_{\text{T}})$, as measured in the training dataset and as fitted using k-nearest neighbours (k-NN) regression. Dashed lines indicate the phase space limits arising from the jet mass selection, see Chapter 13.

as a way to make also higher-order moments of the D_2 distribution independent of the jet mass through a convolution. The shape function used in the convolution, $F_{\text{CSS}}(D_2 \,|\, \alpha, \Omega_D)$, is taken to be a Gamma distribution

$$F_{\text{CSS}}(D_2 \,|\, \alpha, \Omega_D) = \left(\frac{\alpha}{\Omega_D}\right)^{\alpha} \frac{1}{\Gamma(\alpha)} D_2^{\alpha-1} e^{-\alpha D_2/\Omega_D}, \tag{D.4}$$

where the mean of the distribution is given by Ω_D and α is a shape parameter. The distribution convolved jet substructure observable D_2^{CSS} is therefore given by

$$\frac{1}{\sigma}\frac{\mathrm{d}\sigma}{\mathrm{d}D_2} \mapsto \frac{1}{\sigma}\frac{\mathrm{d}\sigma}{\mathrm{d}D_2^{\text{CSS}}} = \frac{1}{\sigma}\frac{\mathrm{d}\sigma}{\mathrm{d}D_2} \otimes F_{\text{CSS}}(D_2 \,|\, \alpha, \Omega_D) \tag{D.5}$$

$$= \int_0^{\infty} \mathrm{d}x \, \frac{1}{\sigma}\frac{\mathrm{d}\sigma}{\mathrm{d}D_2}(D_2 - x) \otimes F_{\text{CSS}}(x \,|\, \alpha, \Omega_D). \tag{D.6}$$

The CSS method, through Equation (D.5), performs the mass-decorrelation of the entire D_2 distribution by morphing the distribution of D_2 at one mass m to the distribution at a (lower) reference mass m_{ref}. This is in contrast to the two methods described above, which performed the mass-decorrelation directly at the level of individual jets. Therefore, in practice, the jet-by-jet mass-decorrelation of D_2 through CSS is implemented as the transform $D_2 \mapsto D_2^{\text{CSS}} = G^{-1}(C(D_2) \,|\, \alpha, \Omega_D)$, where C and G are the cumulative distribution functions of D_2 and D_2^{CSS}, respectively. This procedure first maps the D_2 value for a particular jet to the corresponding percentage along the D_2 distribution, and the maps this percentage onto the new D_2^{CSS} observable. The shape function F_{CSS} is

used to convolve the D_2 distribution into a D_2^{CSS} distribution, from which G can then be measured directly and used for the jet-by-jet transform.

The two parameters characterising the shape function, α and Ω_D, are optimised empirically to yield the optimal transform. A single value of α is used throughout and Ω_D is optimised in bins of the jet mass. This choice means that the shape of the convolution F_{CSS} is the same for all masses, but the average shift of the D_2 distribution changes with the jet mass. The binning of the jet mass for the optimisation of Ω_D is chosen to have the shape of the D_2 be roughly unchanged between neighbouring bins while still retaining sufficient statistics in each bin to be able to reliably construct the D_2 distribution. In this study, jet mass bins between 50 and 300 GeV in increments of 10 GeV are used. Following Ref. [224], the lowest jet mass bin is used as the reference throughout this study. The parameter selection is performed by optimising Ω_D in each mass bin for a certain α, by performing a χ^2-minimisation of the transformed D_2^{CSS} distribution with respect to the target reference distribution. This procedure is then repeated for a range of values of α. α is scanned between 0.5 and 3.0 in increments of 0.5 while Ω_D is scanned between 0 and 1 in steps of 0.01. The optimal value of α is determined to be 1.5, found by minimising the χ^2 across mass bins. The corresponding profile of optimal Ω_D values in each jet mass bin are shown in Figure D.3. In order to reduce fluctuations of the best-fit Ω_D values found for the optimal α, the measurements of Ω_D as a function of the jet mass for $\alpha = 1.5$ are fitted using the functional form proposed in Ref. [224, Eq. (3.8)]

$$\Omega_D(m, m_{\text{ref}}) = a\left(\frac{1}{m_{\text{ref}}} - \frac{1}{m}\right) + b\log\left(\frac{m}{m_{\text{ref}}}\right), \tag{D.7}$$

where m_{ref} is the reference mass, taken to be the midpoint of the first jet mass bin, and m is the centre of the mass bins being fitted. The fitted profile is also shown in Figure D.3.

Additionally, in order to mitigate the effect of limited statistics for the training dataset, a kernel-density estimation (KDE) based smoothing with a length scale of 0.15 is applied to all training distributions. For the D_2^{CSS} distribution, the smoothing is performed after the convolution. The distributions of D_2 and D_2^{CSS} in two jet mass bins are shown in Figure D.4, along with the target reference distribution of D_2 in the low mass bin $m \in [50, 60]$ GeV.

In the next-to-lowest mass bin in Figure D.4a, the D_2 distribution is approximately unchanged relative to the reference distribution, since the lowest mass bin is used for the reference. As a result, the effect of the CSS transform is minor, but the D_2^{CSS} distribution is coherently closer to the reference distribution than D_2 in the same bin.

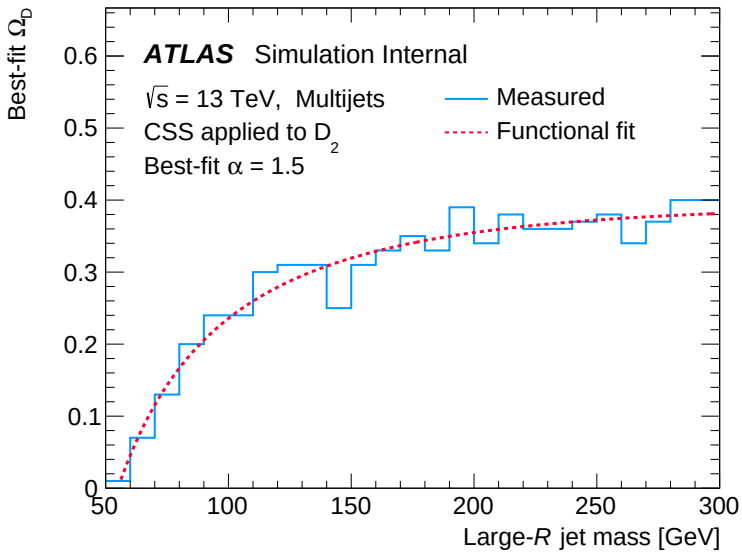
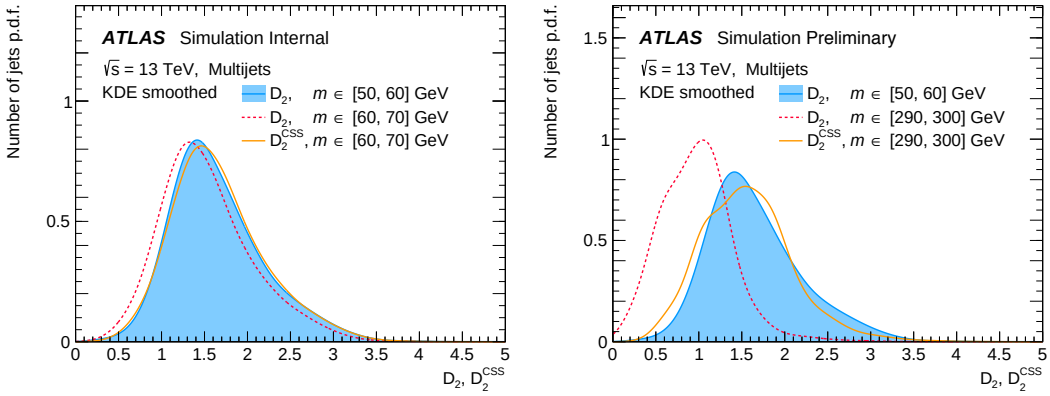


Figure D.3 Evolution of the optimal values for Ω_D in each jet mass bin for $\alpha = 1.5$, as well as the functional fit to this profile used for smoothing the convolved substructure (CSS) transform of D_2.



(a) Next-to-lowest mass bin **(b)** Highest mass bin

Figure D.4 Distributions of D_2 and D_2^{CSS} for multijets in the next-to-lowest and highest mass bin, respectively, along with the D_2 reference distribution. A kernel-density estimation (KDE) based smoothing is applied to all training distributions.

In the highest mass bin in Figure D.4b, the D_2 distribution has shifted substantially relative to the reference. The CSS transform has the effect of shifting the mean value of the D_2^{CSS} distribution towards that of the reference, and as well as morph the shape to arguably make it closer to the reference distribution.

Using the CSS technique, the D_2 distribution is transformed to be similar for all jet mass bins, thereby directly reducing correlation with the jet mass. In principle, CSS leads to a more complete decorrelation than the DDT and k-NN methods by decorrelating the entire jet substructure observable distribution from the jet mass, rather than just a single moment or percentage. However, Figure D.4 shows that while CSS works well for small transforms between adjacent mass bins, notable differences in the shape of the D_2 distribution arise at larger jet masses, which the method is unable to mitigate. Additionally, unlike the DDT and k-NN methods, CSS does not include or account for the jet p_{T} in the mass-decorrelation, neither indirectly nor explicitly. This means that the CSS method is not able to account for changes in the D_2 distribution as a function of p_{T}. The CSS transform is optimised on the inclusive, cross-section–weighted training sample which means that its mass-decorrelation is biased towards jets in the populous region just above the lower p_{T} bound of 200 GeV, see Chapter 13. Therefore, if the CSS transform is applied and evaluated in a region of p_{T} which is substantially higher than 200 GeV, it may not lead to the same mass-decorrelation effect as observed in the inclusive training sample. This is indeed what is observed in Chapter 16.

D.4 Adaptive boosting for uniform efficiency

When used for classification of hadronic resonance decays, boosted decision tree (BDT) algorithms like AdaBoost [122], introduced in Chapter 4, also yield selection efficiencies which are non-uniform with respect to the jet mass [39, 222]. The uBoost method [226] seeks to mitigate this non-uniformity by updating the training weights for each jet based on both classification error and the uniformity of the background selection efficiency with respect to the jet mass, at a fixed target selection efficiency $\bar{\varepsilon}$.

Standard adaptive boosting is based on the misclassification measure of performance γ_i, see Equation B.7, which has a value of 1 if jet i of N is misclassified by a given DT estimator, and 0 otherwise. For uBoost, the performance measure for uniformity of the background selection is taken to be $\bar{\varepsilon} - \varepsilon_i^t$, where ε_i^t is the approximate, local background selection efficiency of the DT estimator in the vicinity of jet i along the jet mass axis at boosting step t. This local selection efficiency is calculated by first finding the value

z_{cut} of the DT score that would yield a global background selection efficiency of $\bar{\varepsilon}$ if a threshold selection were applied at this value. Then, the local selection efficiency is found by computing a k-NN average of $\hat{z} > z_{\text{cut}}$, where $\hat{z}$ is the output score from the DT classifier, across the 50 jets nearest to the i^{th} jet along the jet mass.

The non-uniformity measure is defined as $\delta_i^t = \bar{\varepsilon} - \varepsilon_i^t$. The uniformity error, in analogy with Equation (B.8), is defined then as

$$f^t = \frac{\sum_i w_i^t |\delta_i^t|}{\sum_i w_i} \,. \tag{D.8}$$

This is used to calculate the DT weight for uniformity $b^t = \log\left[1/f^t\right]$ such that the uniformity boosting weight is given by

$$u_i^t = \exp(\alpha\, b^t \delta_i^t), \tag{D.9}$$

where α is a hyperparameter of the uBoost method called the uniforming rate. Finally, the AdaBoost weight update in Equation (B.11) is modified as

$$w_i^{t+1} = w_i^t \times c_i^t \times u_i^t \,. \tag{D.10}$$

Here it is seen how α controls the relative contributions of the classification and uniformity boosting weights at each weight update. The same method as in Equation (4.5) is used for combining the individual DTs into a single BDT.

The structure of the uniformity boosting weights in Equation (D.9) means that jets with masses in regions where the selection efficiency ε_i^t is lower than the target efficiency $\bar{\varepsilon}$ are boosted to have larger training weights w_i^{t+1}; and *vice versa*. The uniformity boosting weights u_i^t are only applied to the class for which uniformity is desired, *i.e.* the multijet background in this study; Equation (B.11) is used for the signal process. In this way, using adaptive training weights, uBoost balances classification power and uniformity of the background selection efficiency in the mass observable during training.

The uniforming rate balances the trade-off between classification performance and mass-decorrelation for the BDT jet taggers, similarly to the regularisation parameter λ for the ANN tagger. For $\alpha \to 0$, the adaptive boosting only takes the classification loss into account, and the standard AdaBoost classifier is recovered. Conversely, for larger α, the boosting for uniform background selection efficiency becomes gradually more important.

The hyperparameter configuration adopted for AdaBoost is the same as the one used for

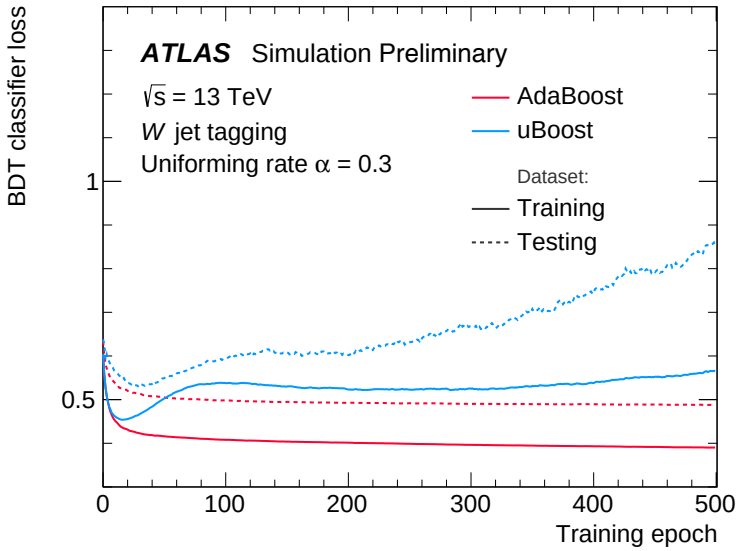


Figure D.5 Binary cross-entropy classification objective during training of AdaBoost and uBoost classifiers for training and testing datasets, with $\alpha = 0.3$ for the uBoost classifier. Figure from Ref. [2].

the BDT classifier in Ref. [222]. For the remaining uBoost hyperparameters, the default values in Ref. [263] are used. For comparison with other taggers, a value of $\alpha = 0.3$ is chosen, since it leads to roughly the same level of background rejection as the ANN for the chosen default value of $\lambda = 10$ and the chosen BDT configuration.

The binary cross-entropy classification loss during training, providing a measure of classification performance similar to the neural networks classifier loss in Figure E.3 for the AdaBoost and uBoost classifiers, is shown in Figure D.5.

The AdaBoost classification training loss is seen to decrease monotonically and reach a plateau for the testing dataset after 500 epochs of training. Similarly, the classification loss as computed on the testing dataset also decreases monotonically, indicating that the classifier does not over-fit the training data. In contrast, the classification objective for uBoost initially decreases due to improved discriminating power, and then rebounds as the adaptive boosting for uniform efficiency takes effect. The binary cross-entropy classification loss does not provide a meaningful convergence criteria for the uBoost training, as was also the case for the ANN training in Chapter 15. Instead, a fixed duration of 500 training epochs is used for all BDT-based models. This constitutes a well-defined procedure which yields a collection of consistently trained jet classifiers with varying degrees of mass-decorrelation. For these, the level of mass-decorrelation is given by the degree of divergence at the end of the fixed training duration, which in turn is controlled by rate at which the uniformity boosting takes effect, as determined by α.

APPENDIX E

Adversarial neural network details

The construction and training of standard neural networks (NNs) was detailed in Chapter 4 and Appendix B. This appendix provide additional details regarding some of technical aspects of the NN hyperparameter optimisation and the adversarial training.

E.1 Hyperparameter optimisation

As described in Chapter 15, both the classifier and adversary networks are constructed as standard NNs. The architecture and learning configuration for each is referred to collectively as the hyperparameters of the network. The study in Part III of this thesis uses Bayesian optimisation, implemented in the Spearmint library [264, 265], to optimise these hyperparameter. A brief overview of this method is given in Appendix C.

Classifier network

All tested classifier NNs are constructed as densely connected networks with the 10 input features listed in Table 13.2, a number of hidden layers all with the same number of nodes, and a single output node. The output node is equipped with sigmoid activation, see Figure B.1, to produce an output z in the range $[0, 1]$ corresponding to the probability which the classifier assigns for a given jet to belong to the W jet class. This property is due to the choice of classification loss in Equation (15.7) and the class-balanced training weighting discussed in Chapter 13.

The training of the classifier network is performed with the Adam [250] optimiser. Each neural network weight update is performed on coherent batches of features and

Parameter	Range	Scale	Chosen value
Learning rate	$[10^{-5}, 10^{-1}]$	Logarithmic	10^{-2}
Learning rate decay	$[10^{-6}, 10^{-2}]$	Logarithmic	10^{-3}
Hidden layers	$[1, 6]$	Linear	3
Nodes per hidden layers	$[2, 512]$	Logarithmic	64
Dropout regularisation	$[0, 0.5]$	Linear	0
Hidden layer activation	{ReLU, tanh}	Choice	ReLU

Table E.1 Neural network (NN) classifier hyperparameters optimised with SPEARMINT, the parameter range searched, the scale of the space samples, and the chosen hyperparameter configuration. See Chapter 4 for details.

labels, each with a fixed size of 8192 samples, found to balance high computational throughput and memory requirements on the NVIDIA Tesla K80 graphics processing units (GPUs) used in this study. Similarly, to accelerate the training of the classifier, batch normalisation is applied before each hidden layer in the network to standardise the learned features, see Chapter 4. The parameters considered in the optimisation, the range of the parameters, the scale of the parameter space, and the chosen hyperparameter configuration are listed in Table E.1. Since Bayesian optimisation does not require the hyperparameter space to be discretised, in the way that an exhaustive grid search would, no binning of the parameter search ranges in Table E.1 is necessary. A 'logarithmic' scale in Table E.1 means that the parameter in question is transformed as $p \to \log p$ before being used in the Gaussian process (GP) regression, thereby enforcing a uniform prior in $\log p$ for parameters which are strictly positive and can span several orders of magnitude. A 'choice' scale simply means a choice between multiple definite alternatives.

The classifier is optimised according to the loss L_{clf} in Equation (15.7) with the flat-p_{T} training weights discussed in Chapter 13. During optimisation, 3-fold stratified cross-validation is employed, see Chapter 4, to obtain the mean and variation of the optimisation metric across 3 independent samples of unseen data, called the validation splits. This provides an estimate of the ability of the classifier to generalise well to the testing dataset. If the optimisation were performed with respect to the training loss ($L_{\text{clf}}^{\text{train}}$), the classifier network would be "rewarded" for over-fitting the training data, *i.e.* exploiting features in the training dataset which are not representative of the broader population from which the training dataset is drawn. As this is not desirable, the optimisation is performed with respect to the validation loss ($L_{\text{clf}}^{\text{val}}$). An NN classifier is trained on each cross-validation fold for 50 epochs, *i.e.* passes through the full training dataset. The order of the jets in the training dataset is shuffled between each epoch. In order to ensure stability of the result, the optimisation metric is chosen to be the

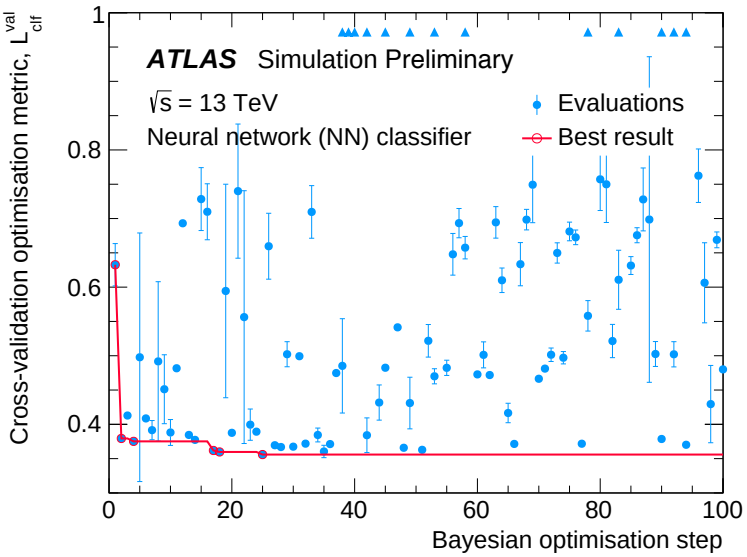


Figure E.1 Neural network (NN) classifier Bayesian hyperparameter optimisation. Blue markers indicate the mean and standard deviation for each evaluation. The red line indicates the running optimisation metric minimum. Open red markers indicate improvements. Triangular markers indicate evaluation metrics outside of the axis range.

mean classification loss across the validation splits plus one standard deviation, this value taking the place of $f(h)$ in Equation (C.6). The optimisation process, run for 100 Bayesian optimisation steps, see Equation (C.8), is shown in Figure E.1.

Based on the best optimisation metric values identified by SPEARMINT, a classifier with three hidden layers, each with 64 nodes with rectified linear unit (ReLU) activation, see Figure B.1, and no dropout regularisation is chosen, see Appendix B. Bayesian optimisation has no convergence criteria and is not guaranteed to yield the true, global, optimal configuration in any finite number of iterations. However, it is capable of efficiently probing a large parameter space with few evaluations, as discussed above. The balance between exploration and exploitation, see Equation (C.6), is evident in the large spread of evaluations in Figure E.1. Exploration of extreme regions of the parameter space leads either to insufficient or excessive capacity of the classifier network, resulting in either under-fitting or over-fitting of the training data, see Chapter 4, and poor classification performance across the validation splits. Conversely, exploitation of identified minimal regions of parameter space leads to minimal values of $L_{\text{clf}}^{\text{val}}$ in relatively few iterations. Therefore, the chosen hyperparameter configuration is deemed to be performant but is not guaranteed to be optimal.

Adversary network

All tested adversary NNs are constructed as a densely connected network with two input features (the classifier output and the auxiliary input $\log p_{\mathrm{T}}/\mu$ with $\mu = 1$ GeV), a number of hidden layers with the same number of nodes, and outputting the parameters for the posterior probability density function (p.d.f.) Batch normalisation is not used in the adversary, as it is found to yield unstable results. Similarly, no regularisation is necessary in the adversary, since over-fitting in not a concern.

Mass-decorrelation and robustness with respect to the jet p_{T} is implemented by having the adversary parametrise a p.d.f. in m conditional on the auxiliary input $\log p_{\mathrm{T}}/\mu$. For convenience, the jet mass is scaled to the unit interval to allow for better use of output activations in the adversary. The adversary posterior Gaussian mixture model (GMM) is constructed by N_{GMM} components, *i.e.*

$$p_{\mathrm{adv}}(\tilde{m} \,|\, z, \log p_{\mathrm{T}}/\mu, \theta_{\mathrm{adv}}) = \sum_{i=1}^{N_{\mathrm{GMM}}} c_i \mathcal{N}_{[0,1]}(\tilde{m} \,|\, \mu_i, \sigma_i), \tag{E.1}$$

where $\tilde{m}$ is the large-radius (large-R) jet mass scaled to the range $[0, 1]$, z is the output from the classifier network, p_{T} is the large-R jet transverse momentum, θ_{adv} are the weights of the adversary neural network, $\mathcal{N}_{[0,1]}$ is the normal distribution function normalised to unity on the range $[0, 1]$, and c_i, μ_i, and σ_i are the N_{GMM} normalisation coefficients, means, and widths, respectively, for the GMM. This means that the adversary is tasked with parametrising N_{GMM} Gaussian means, widths, and $N_{\mathrm{GMM}} - 1$ normalisation coefficients. The nodes in the adversary network corresponding to the GMM means μ_i are equipped with sigmoid activation to ensure that they are constrained to the range $[0, 1]$, see Figure B.1. Similarly, the GMM widths σ_i are equipped with softplus activation, to ensure that they are strictly positive, see Figure B.1. Finally, the GMM normalisation coefficients c_i are equipped with softmax activation, which is defined as

$$c_i = \mathrm{softmax}\,(\tilde{c}_i) = \frac{\exp(\tilde{c}_i)}{\sum_{j=1}^{N_{\mathrm{GMM}}} \exp(\tilde{c}_j)}\,, \tag{E.2}$$

where $\tilde{c}_i$ are the values of the corresponding nodes prior to applying the activation. The softmax activation ensures that $\{c_i\}$ sum to one, thereby preserving the normalisation of the GMM in Equation (E.1). Similar to the classifier, the adversary is trained with flat-p_{T} training jet weights, constructed to retain physical distributions in m for all p_{T}. This jet mass distribution conditional on $\log p_{\mathrm{T}}/\mu$, weighted in this way, plays the role

of the adversary prior, *i.e.* the distribution to which the adversary p.d.f. should default in the absence of additional information in the classifier output z, since it amounts to the optimal "random guessing." Finally, the fact that the decorrelation variable is scaled to be confined to the range $\tilde{m} \in [0, 1]$ and the fact that the adversary posterior p.d.f. in Equation (E.1) is normalised to unit integral on the same range allows for an easier admission of the prior.

The hyperparameters of the adversary network are also optimised using Spearmint. In contrast with the stand-alone classifier optimisation, however, the adversary cannot be meaningfully optimised according to the loss in Equation (15.8) alone. This loss measures only the capacity of the adversary to construct the posterior p.d.f. for the jet mass, not the quality of the resulting mass-decorrelated tagger. Similarly, optimising according to a combination of Equations (15.7) and (15.8) as in Equation (15.6) is vulnerable to breakdowns of the adversarial training procedure, where the inability of the adversary to infer the jet mass is due to an unbalanced, joint optimisation rather than the absence of correlation with the jet mass. Therefore, to tune the adversarially trained neural network tagger according to expected performance, the optimisation is performed by maximising the metric $1/\varepsilon_{\text{bkg}}^{\text{rel}} + \lambda/\text{JSD}$ computed at $\varepsilon_{\text{sig}}^{\text{rel}} = 50\%$ for a fixed value of λ, chosen to be $\lambda = 10$ since it is found to yield robust results for mass-decorrelation, as discussed in Chapter 16. The choice to multiply the Jensen-Shannon divergence (JSD) term by λ is intended to emulate Equation (15.6), but other linear combinations of the metrics classification and mass-decorrelation could also be used. Specifically, to obtain a robust result, the optimisation is performed to maximise the mean value of the metric across cross-validation folds minus one standard deviation, similarly to the classifier optimisation.

Throughout the adversarial optimisation, a classifier pre-trained with the chosen hyperparameters in Table E.1 is used. In order to provide the adversary with reasonable initial conditions, the adversarial training starts with an adversary-only pre-training period of 20 epochs for the optimisation, where the adversary is allowed to condition its posterior on the pre-trained classifier, the weights of which are kept fixed during this pre-training.

The adversarial optimisation is performed by 3-fold stratified cross-validation. Training is performed for 200 epochs following the adversary-only pre-training using the Adam optimiser with a batch size of 8192 samples and between-epoch shuffling, similar to the stand-alone classifier training. The parameters considered in the hyperparameter optimisation, the range of the parameters, the scale of the parameter space, and the optimal hyperparameter configuration are listed in Table E.2.

Parameter	Range	Scale	Chosen value
Learning rate	$[10^{-5}, 10^{-1}]$	Logarithmic	5×10^{-2}
Learning rate decay	$[10^{-6}, 10^{-2}]$	Logarithmic	10^{-2}
Hidden layers	$[1, 6]$	Linear	1
Nodes per hidden layers	$[2, 128]$	Logarithmic	64
GMM components, N_{GMM}	$[0, 20]$	Linear	20
Learning rate ratio, $\ell_{\text{clf}}/\ell_{\text{adv}}$	$[10^{-8}, 10^{-1}]$	Logarithmic	2×10^{-7}
Hidden layer activation	{ReLU, tanh}	Choice	ReLU

Table E.2 Adversary network hyperparameters optimised with SPEARMINT, the parameter range searched, the scale of the space samples, and the chosen hyperparameter configuration.

The evolution of the optimisation metric with the Bayesian optimisation steps for the combined adversarial neural network (ANN) classifier is shown in Figure E.2.

An adversary with a single hidden layer comprising 64 nodes with ReLU activation, parametrising a GMM posterior p.d.f. with $N_{\text{GMM}} = 20$ components is found to have sufficient capacity to perform the mass-decorrelation. The combined adversarial neural network architecture with chosen hyperparameters was shown in Figure 15.2.

E.2 Training characteristics

Classifier

The training of both classifier and adversary is performed on a cluster of NVIDIA Tesla K80 GPUs, both for hyperparameter optimisation and for the subsequent training using the chosen hyperparameter configuration. To ensure that the training converges and that no over-training is observed, the classifier is first trained stand-alone with the chosen hyperparameter configuration using 3-fold cross-validation for 200 epochs, *i.e.* four times the number used for optimisation. The evolution of the classifier loss L_{clf} as a function of the number of training epochs is shown in Figure E.3.

L_{clf} decreases monotonically during training, for both training and cross-validation splits, indicating no over-training for the chosen hyperparameter configuration. In addition, the final classifier loss is comparable to Ref. [222], suggesting that the chosen classifier network architecture is performant.

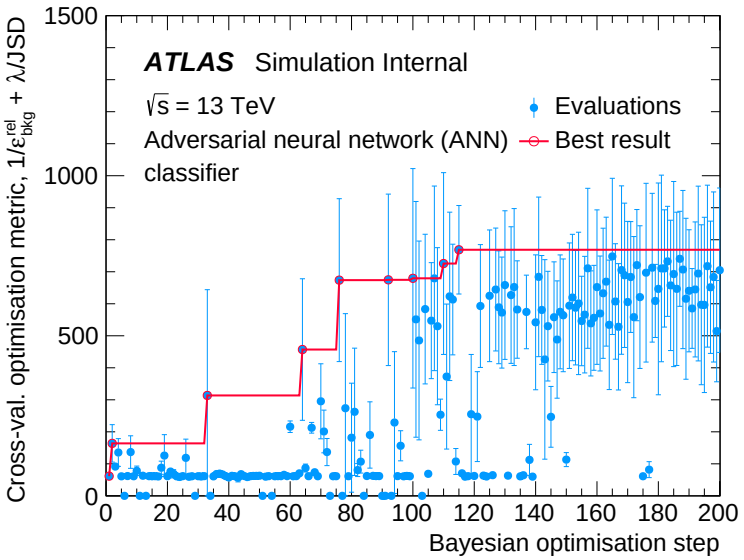


Figure E.2 Adversarial neural network (ANN) classifier Bayesian hyperparameter optimisation. Blue markers indicate the mean and standard deviation for each evaluation. The red line indicates the running optimisation metric maximum. Open red markers indicate improvements.

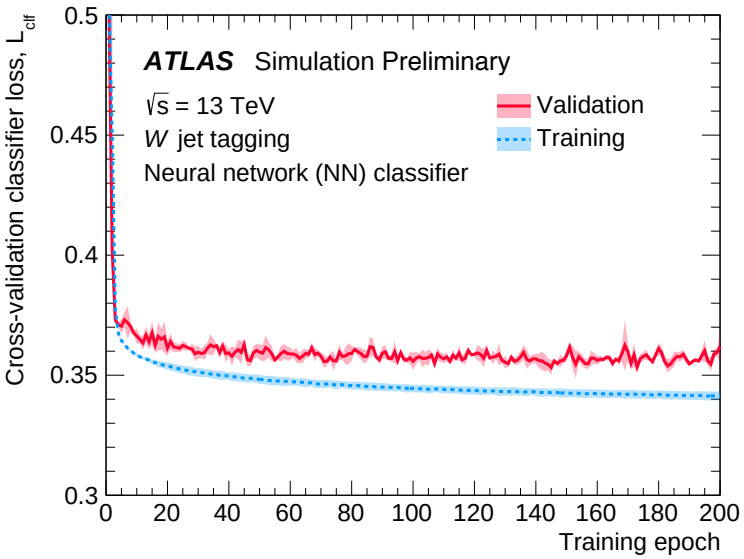


Figure E.3 Stand-alone neural network (NN) classifier loss during 3-fold cross validation training as a function of the number of training epochs, for training and validation splits. Lines indicate mean loss across folds. Shaded bands indicate standard deviation of losses across folds.

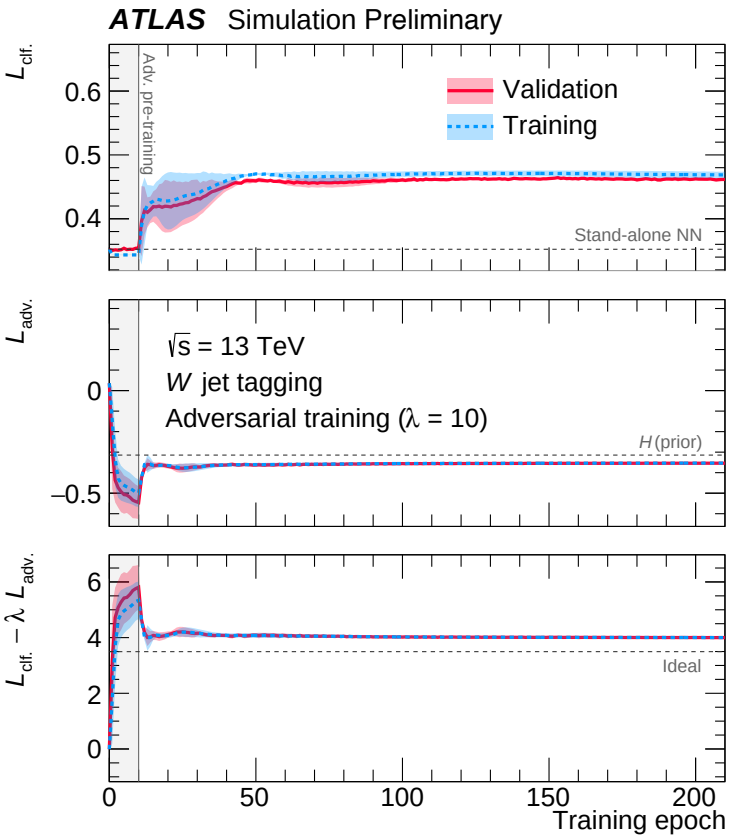


Figure E.4 Classification (*top*), adversary (*middle*), and effective classifier losses (*bottom*) associated with the adversarial neural network tagger during 3-fold cross validation training as a function of the number of epochs, for training and validation splits. The first 10 epochs are spent on pre-training the adversary network. Lines indicate mean loss across folds, shaded bands indicate standard deviation of losses across folds. References for the stand-alone neural network (NN) classifier loss, the entropy H of the adversary prior, and the ideal, effective loss are shown.

Adversary

The adversary's training configuration is chosen so as to find a stable, minimal perturbation around the stand-alone NN classifier. Therefore, the training starts from the pre-trained NN classifier with chosen hyperparameters as listed in Table E.1. After an adversary pre-training for 10 epochs during the final training, the two networks are trained simultaneously for 200 epochs with a small effective learning for the classifier, emulating full convergence on the inner optimisation in Equation (15.6). The evolution of the classifier, adversary, and effective losses is shown in Figure E.4.

During the adversary pre-training, the adversary loss L_{adv}, see Equation (15.8), decreases to a minimum, showing the convergence of the adversary posterior towards the

$p(\tilde{m} \,|\, z, \log p_{\mathrm{T}}/\mu, \theta_{\mathrm{adv}})$ for z drawn from the pre-trained classifier. In this portion of the training, the classifier is kept fixed, and thus the classifier loss L_{clf}, see Equation (15.7), remains constant.

After the adversary pre-training, L_{clf} is seen to rise in sync with a rise in L_{adv}, illustrating the classifier balancing the two competing objectives. The balance is such that the effective loss seen by the classifier, $L_{\mathrm{clf}} - \lambda L_{\mathrm{adv}}$, is minimised. As a result of the mass-decorrelation, the adversary's task becomes more difficult, and in the limit of full mass-decorrelation the adversary posterior is equal to the prior, corresponding to the multijet mass distribution with training weights, assuming an adversary with sufficient capacity and full convergence of the inner optimisation in Equation (15.6). In this limit, the value of L_{adv} will tend towards the entropy H of the prior [240], see also Section 14.2. A deviation from this asymptotic limit indicates a balance between classification and mass-decorrelation for a given λ.

Relevant reference values for each loss are shown on Figure E.4. In particular, the "Ideal" value of $L_{\mathrm{clf}} - \lambda L_{\mathrm{adv}}$ corresponds to the case where the stand-alone NN classification power is retained along with full mass-decorrelation ($L_{\mathrm{adv}} \sim H(\mathrm{prior})$). Deviations from the ideal case reflect the information lost in order to prioritise the mass-decorrelation.

Bibliography

[1] ATLAS Collaboration, *Search for light resonances decaying to boosted quark pairs and produced in association with a photon or a jet in proton–proton collisions at* $\sqrt{s} = 13$ *TeV with the ATLAS detector*, Phys. Lett. B **788** (2019) 316, arXiv:1801.08769 [hep-ex].

[2] ATLAS Collaboration, *Performance of mass-decorrelated jet substructure observables for hadronic two-body decay tagging in ATLAS*, ATL-PHYS-PUB-2018-014, 2018, https://cds.cern.ch/record/2630973.

[3] S. L. Glashow, *Partial Symmetries of Weak Interactions*, Nucl. Phys. **22** (1961) 579–588.

[4] A. Salam and J. C. Ward, *Electromagnetic and weak interactions*, Phys. Lett. **13** (1964) 168–171.

[5] S. Weinberg, *A Model of Leptons*, Phys. Rev. Lett. **19** (1967) 1264–1266.

[6] H. D. Politzer, *Reliable Perturbative Results for Strong Interactions?*, Phys. Rev. Lett. **30** (1973) 1346–1349, [,274(1973)].

[7] S. Weinberg, *Nonabelian Gauge Theories of the Strong Interactions*, Phys. Rev. Lett. **31** (1973) 494–497.

[8] D. J. Gross and F. Wilczek, *Ultraviolet Behavior of Nonabelian Gauge Theories*, Phys. Rev. Lett. **30** (1973) 1343–1346, [,271(1973)].

[9] D. Galbraith and C. Burgard, *Standard Model Standard Infographic*, http://davidgalbraith.org/portfolio/ux-standard-model-of-the-standard-model/.

[10] C. Burgard, *Example: Standard model of physics*, http://www.texample.net/tikz/examples/model-physics/.

[11] Particle Data Group Collaboration, Tanabashi, M. and others, *Review of Particle Physics*, Phys. Rev. **D98** (2018) 030001.

[12] R. K. Ellis, W. J. Stirling, and B. R. Webber, *QCD and collider physics*, Camb. Monogr. Part. Phys. Nucl. Phys. Cosmol. **8** (1996) 1–435.

[13] P. W. Higgs, *Broken Symmetries and the Masses of Gauge Bosons*, Phys. Rev. Lett. **13** (1964) 508–509.

[14] F. Englert and R. Brout, *Broken Symmetry and the Mass of Gauge Vector Mesons*, Phys. Rev. Lett. **13** (1964) 321–323.

[15] ATLAS Collaboration, *Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC*, Phys. Lett. **B716** (2012) 1–29, arXiv:1207.7214 [hep-ex].

[16] CMS Collaboration, *Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC*, Phys. Lett. **B716** (2012) 30–61, arXiv:1207.7235 [hep-ex].

[17] H. Wiedemann, *Particle Accelerator Physics*. Springer Berlin Heidelberg, 2007.

[18] A. Buckley *et al.*, *General-purpose event generators for LHC physics*, Phys. Rept. **504** (2011) 145–233, arXiv:1101.2599 [hep-ph].

[19] J. C. Collins and D. E. Soper, *The Theorems of Perturbative QCD*, Ann. Rev. Nucl. Part. Sci. **37** (1987) 383–409.

[20] Y. L. Dokshitzer, *Calculation of the Structure Functions for Deep Inelastic Scattering and e+ e- Annihilation by Perturbation Theory in Quantum Chromodynamics.*, Sov. Phys. JETP **46** (1977) 641–653, [Zh. Eksp. Teor. Fiz.73,1216(1977)].

[21] V. N. Gribov and L. N. Lipatov, *Deep inelastic e p scattering in perturbation theory*, Sov. J. Nucl. Phys. **15** (1972) 438–450, [Yad. Fiz.15,781(1972)].

[22] G. Altarelli and G. Parisi, *Asymptotic Freedom in Parton Language*, Nucl. Phys. **B126** (1977) 298–318.

[23] ATLAS Collaboration, *Performance of pile-up mitigation techniques for jets in pp collisions at* $\sqrt{s} = 8$ *TeV using the ATLAS detector*, Eur. Phys. J. C **76** (2016) 581, arXiv:1510.03823 [hep-ex].

[24] G. Sterman and S. Weinberg, *Jets from Quantum Chromodynamics*, Phys. Rev. Lett. **39** (1977) 1436–1439.

[25] G. P. Salam, *Towards Jetography*, Eur. Phys. J. **C67** (2010) 637–686, arXiv:0906.1833 [hep-ph].

[26] A. J. Larkoski, I. Moult, and B. Nachman, *Jet Substructure at the Large Hadron Collider: A Review of Recent Advances in Theory and Machine Learning*, arXiv:1709.04464 [hep-ph].

[27] L. Asquith *et al.*, *Jet Substructure at the Large Hadron Collider : Experimental Review*, arXiv:1803.06991 [hep-ex].

[28] S. Catani, Y. L. Dokshitzer, M. Seymour, and B. Webber, *Longitudinally-invariant k_t-clustering algorithms for hadron-hadron collisions*, Nucl. Phys. **B406** (1993) 187–224.

[29] S. D. Ellis and D. E. Soper, *Successive combination jet algorithm for hadron collisions*, Phys. Rev. **D48** (1993) 3160–3166, arXiv:hep-ph/9305266 [hep-ph].

[30] M. Cacciari, G. P. Salam, and G. Soyez, *The anti-k_t jet clustering algorithm*, JHEP **04** (2008) 063, arXiv:0802.1189 [hep-ph].

[31] Y. L. Dokshitzer, G. D. Leder, S. Moretti, and B. R. Webber, *Better jet clustering algorithms*, JHEP **08** (1997) 001, arXiv:hep-ph/9707323 [hep-ph].

[32] M. Wobisch and T. Wengler, *Hadronization corrections to jet cross-sections in deep inelastic scattering*, arXiv:hep-ph/9907280 [hep-ph].

[33] ATLAS Collaboration, *Jet mass reconstruction with the ATLAS Detector in early Run 2 data*, ATLAS-CONF-2016-035, 2016, https://cds.cern.ch/record/2200211.

[34] J. Thaler and K. Van Tilburg, *Identifying Boosted Objects with N-subjettiness*, JHEP **03** (2011) 015, arXiv:1011.2268 [hep-ph].

[35] J. M. Butterworth, A. R. Davison, M. Rubin, and G. P. Salam, *Jet substructure as a new Higgs search channel at the LHC*, Phys. Rev. Lett. **100** (2008) 242001, arXiv:0802.2470 [hep-ph].

[36] D. Krohn, J. Thaler, and L.-T. Wang, *Jet Trimming*, JHEP **02** (2010) 084, arXiv:0912.1342 [hep-ph].

[37] S. D. Ellis, C. K. Vermilion, and J. R. Walsh, *Recombination Algorithms and Jet Substructure: Pruning as a Tool for Heavy Particle Searches*, Phys. Rev. **D81** (2010) 094023, arXiv:0912.0033 [hep-ph].

[38] A. J. Larkoski, S. Marzani, G. Soyez, and J. Thaler, *Soft Drop*, JHEP **05** (2014) 146, arXiv:1402.2657 [hep-ph].

[39] ATLAS Collaboration, *Performance of top-quark and W-boson tagging with ATLAS in Run 2 of the LHC*, arXiv:1808.07858 [hep-ex].

[40] ATLAS Collaboration, *Performance of jet substructure techniques for large-R jets in proton–proton collisions at $\sqrt{s} = 7$ TeV using the ATLAS detector*, JHEP **09** (2013) 076, arXiv:1306.4945 [hep-ex].

[41] F. Zwicky, *Spectral displacement of extra galactic nebulae*, Helv. Phys. Acta **6** (1933) 110–127.

[42] F. Zwicky, *On the Masses of Nebulae and of Clusters of Nebulae*, Astrophys. J. **86** (1937) 217.

[43] V. C. Rubin and W. K. Ford, Jr., *Rotation of the Andromeda Nebula from a Spectroscopic Survey of Emission Regions*, Astrophys. J. **159** (1970) 379–403.

[44] V. C. Rubin, N. Thonnard, and W. K. Ford, Jr., *Rotational properties of 21 SC galaxies with a large range of luminosities and radii, from NGC 4605 /R = 4kpc/ to UGC 2885 /R = 122 kpc/*, Astrophys. J. **238** (1980) 471.

[45] T. S. van Albada, J. N. Bahcall, K. Begeman, and R. Sancisi, *The Distribution of Dark Matter in the Spiral Galaxy NGC-3198*, Astrophys. J. **295** (1985) 305–313.

[46] K. G. Begeman, A. H. Broeils, and R. H. Sanders, *Extended rotation curves of spiral galaxies: Dark haloes and modified dynamics*, Mon. Not. Roy. Astron. Soc. **249** (1991) 523.

[47] G. Bertone, ed., *Particle Dark Matter: Observations, Models and Searches*. Cambridge University Press, 2010.

[48] A. A. Penzias and R. W. Wilson, *A Measurement of excess antenna temperature at 4080-Mc/s*, Astrophys. J. **142** (1965) 419–421.

[49] W. Hu and S. Dodelson, *Cosmic microwave background anisotropies*, Ann. Rev. Astron. Astrophys. **40** (2002) 171–216, arXiv:astro-ph/0110414 [astro-ph].

[50] Planck Collaboration, *Planck 2018 results. VI. Cosmological parameters*, arXiv:1807.06209 [astro-ph.CO].

[51] G. R. Blumenthal, S. M. Faber, J. R. Primack, and M. J. Rees, *Formation of Galaxies and Large Scale Structure with Cold Dark Matter*, Nature **311** (1984) 517–525, [,96(1984)].

[52] G. Bertone, D. Hooper, and J. Silk, *Particle dark matter: evidence, candidates and constraints*, Physics Reports **405** (2005) 279–390.

[53] D. Clowe, A. Gonzalez, and M. Markevitch, *Weak lensing mass reconstruction of the interacting cluster 1E0657-558: Direct evidence for the existence of dark matter*, Astrophys. J. **604** (2004) 596–603, arXiv:astro-ph/0312273 [astro-ph].

[54] M. Markevitch *et al.*, *Direct constraints on the dark matter self-interaction cross-section from the merging galaxy cluster 1E0657-56*, Astrophys. J. **606** (2004) 819–824, arXiv:astro-ph/0309303 [astro-ph].

[55] J. Lesgourgues and S. Pastor, *Neutrino mass from Cosmology*, Adv. High Energy Phys. **2012** (2012) 608515, arXiv:1212.6154 [hep-ph].

[56] G. Steigman and M. S. Turner, *Cosmological Constraints on the Properties of Weakly Interacting Massive Particles*, Nucl. Phys. **B253** (1985) 375–386.

[57] E. W. Kolb and M. S. Turner, *The Early Universe*, Front. Phys. **69** (1990) 1–547.

[58] J. L. Feng and J. Kumar, *The WIMPless Miracle: Dark-Matter Particles without Weak-Scale Masses or Weak Interactions*, Phys. Rev. Lett. **101** (2008) 231301, arXiv:0803.4196 [hep-ph].

[59] J. Conrad and O. Reimer, *Indirect dark matter searches in gamma and cosmic rays*, Nature Phys. **13** (2017) 224–231, arXiv:1705.11165 [astro-ph.HE].

[60] J. Goodman, M. Ibe, A. Rajaraman, W. Shepherd, T. M. P. Tait, and H.-B. Yu, *Constraints on Dark Matter from Colliders*, Phys. Rev. **D82** (2010) 116010, arXiv:1008.1783 [hep-ph].

[61] S. Kanemura, S. Matsumoto, T. Nabeshima, and N. Okada, *Can WIMP Dark Matter overcome the Nightmare Scenario?*, Phys. Rev. **D82** (2010) 055026, arXiv:1005.5651 [hep-ph].

[62] A. Djouadi, O. Lebedev, Y. Mambrini, and J. Quevillon, *Implications of LHC searches for Higgs–portal dark matter*, Phys. Lett. **B709** (2012) 65–69, arXiv:1112.3299 [hep-ph].

[63] ATLAS Collaboration, *Combination of searches for invisible Higgs boson decays with the ATLAS experiment*, ATLAS-CONF-2018-054, 2018, https://cds.cern.ch/record/2649407.

[64] CMS Collaboration, *Searches for invisible decays of the Higgs boson in pp collisions at* $\sqrt{s} = 7, 8$, *and* 13 *TeV*, JHEP **02** (2017) 135, arXiv:1610.09218 [hep-ex].

[65] D. Abercrombie *et al.*, *Dark Matter Benchmark Models for Early LHC Run-2 Searches: Report of the ATLAS/CMS Dark Matter Forum*, arXiv:1507.00966 [hep-ex].

[66] O. Buchmueller, M. J. Dolan, and C. McCabe, *Beyond Effective Field Theory for Dark Matter Searches at the LHC*, JHEP **01** (2014) 025, arXiv:1308.6799 [hep-ph].

[67] O. Buchmueller, M. J. Dolan, S. A. Malik, and C. McCabe, *Characterising dark matter searches at colliders and direct detection experiments: Vector mediators*, JHEP **01** (2015) 037, arXiv:1407.8257 [hep-ph].

[68] LUX Collaboration, *Results from a search for dark matter in the complete LUX exposure*, Phys. Rev. Lett. **118** (2017) 021303, arXiv:1608.07648 [astro-ph.CO].

[69] XENON Collaboration, *Dark Matter Search Results from a One Ton-Year Exposure of XENON1T*, Phys. Rev. Lett. **121** (2018) 111302, arXiv:1805.12562 [astro-ph.CO].

[70] G. D'Ambrosio, G. F. Giudice, G. Isidori, and A. Strumia, *Minimal flavor violation: An Effective field theory approach*, Nucl. Phys. **B645** (2002) 155–187, arXiv:hep-ph/0207036 [hep-ph].

[71] ATLAS Collaboration, *Search for dark matter produced in association with bottom or top quarks in $\sqrt{s} = 13$ TeV pp collisions with the ATLAS detector*, Eur. Phys. J. C **78** (2018) 18, arXiv:1710.11412 [hep-ex].

[72] A. Boveia and C. Doglioni, *Dark Matter Searches at Colliders*, Ann. Rev. Nucl. Part. Sci. **68** (2018) 429–459, arXiv:1810.12238 [hep-ex].

[73] CMS Collaboration, *Search for low mass vector resonances decaying into quark–antiquark pairs in proton–proton collisions at $\sqrt{s} = 13$ TeV*, JHEP **01** (2018) 097, arXiv:1710.00159 [hep-ex].

[74] M. Backovic, K. Kong, and M. McCaskey, *MadDM v.1.0: Computation of Dark Matter Relic Abundance Using MadGraph5*, Physics of the Dark Universe **5-6** (2014) 18–28, arXiv:1308.4955 [hep-ph].

[75] F. Kahlhoefer, K. Schmidt-Hoberg, T. Schwetz, and S. Vogl, *Implications of unitarity and gauge invariance for simplified dark matter models*, JHEP **02** (2016) 016, arXiv:1510.02110 [hep-ph].

[76] B. W. Lee, C. Quigg, and H. B. Thacker, *Weak Interactions at Very High-Energies: The Role of the Higgs Boson Mass*, Phys. Rev. **D16** (1977) 1519.

[77] E. Mobs, *The CERN accelerator complex. Complexe des accélérateurs du CERN*, https://cds.cern.ch/record/2197559, © 2016 CERN.

[78] *LHC Machine*, JINST **3** (2008) S08001.

[79] D. J. Warner, *Project study for a new 50 MeV linear accelerator for the C. P. S*, Tech. Rep. CERN-MPS-LINP-73-1, CERN, Geneva, Oct, 1973. http://cds.cern.ch/record/414071.

[80] ATLAS Collaboration, *The ATLAS Experiment at the CERN Large Hadron Collider*, JINST **3** (2008) S08003.

[81] CMS Collaboration, *The CMS experiment at the CERN LHC*, JINST **3** (2008) S08004.

[82] ALICE Collaboration, *The ALICE experiment at the CERN LHC*, JINST **3** (2008) S08002.

[83] LHCb Collaboration, *The LHCb Detector at the LHC*, JINST **3** (2008) S08005.

[84] ATLAS Collaboration, *Luminosity Public Results for Run 2 – Multiple Year Collision Plots*, https://twiki.cern.ch/twiki/bin/view/AtlasPublic/LuminosityPublicResultsRun2. Accessed August 2019.

[85] J. Pequenao, *Computer generated image of the whole ATLAS detector*, https://cds.cern.ch/record/1095924, © 2008 CERN.

[86] ATLAS Collaboration, *ATLAS Insertable B-Layer Technical Design Report*, Tech. Rep. CERN-LHCC-2010-013. ATLAS-TDR-19, Sep, 2010. https://cds.cern.ch/record/1291633.

[87] ATLAS Collaboration, *Combined measurements of Higgs boson production and decay using up to* $80\,fb^{-1}$ *of proton–proton collision data at* $\sqrt{s} = 13$ *TeV collected with the ATLAS experiment*, ATLAS-CONF-2018-031, 2018, https://cds.cern.ch/record/2629412.

[88] A. Yamamoto *et al.*, *The ATLAS central solenoid*, Nucl. Instrum. Meth. **A584** (2008) 53–74.

[89] ATLAS Collaboration, *Track Reconstruction Performance of the ATLAS Inner Detector at* $\sqrt{s} = 13$ *TeV*, ATL-PHYS-PUB-2015-018, 2015, https://cds.cern.ch/record/2037683.

[90] ATLAS Collaboration, *ATLAS pixel detector electronics and sensors*, JINST **3** (2008) P07007.

[91] ATLAS TRT Collaboration, *The ATLAS Transition Radiation Tracker (TRT) proportional drift tube: Design and performance*, JINST **3** (2008) P02013.

[92] ATLAS Collaboration, B. Mindur, *ATLAS Transition Radiation Tracker (TRT): Straw tubes for tracking and particle identification at the Large Hadron Collider*, Nucl. Instrum. Meth. **A845** (2017) 257–261.

[93] J. Pequenao, *Computer generated image of the ATLAS calorimeter*, https://cds.cern.ch/record/1095927, © 2008 CERN.

[94] R. Wigmans, *Calorimetry: Energy Measurement in Particle Physics*. Oxford University Press, 2017.

[95] ATLAS Collaboration, *Measurement of the photon identification efficiencies with the ATLAS detector using LHC Run-1 data*, Eur. Phys. J. C **76** (2016) 666, arXiv:1606.01813 [hep-ex].

[96] ATLAS Collaboration, *Electron and photon energy calibration with the ATLAS detector using LHC Run 1 data*, Eur. Phys. J. C **74** (2014) 3071, arXiv:1407.5063 [hep-ex].

[97] ATLAS Collaboration, *Jet energy measurement and its systematic uncertainty in proton–proton collisions at* $\sqrt{s} = 7$ *TeV with the ATLAS detector*, Eur. Phys. J. C **75** (2015) 17, arXiv:1406.0076 [hep-ex].

[98] ATLAS Collaboration, *Jet energy resolution in 2017 data and simulation*, https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PLOTS/JETM-2018-005/. Accessed August 2019.

[99] ATLAS Collaboration, *Jet energy resolution in proton–proton collisions at* $\sqrt{s} = 7$ *TeV recorded in 2010 with the ATLAS detector*, Eur. Phys. J. C **73** (2013) 2306, arXiv:1210.6210 [hep-ex].

[100] ATLAS Collaboration, *Jet reconstruction and performance using particle flow with the ATLAS Detector*, Eur. Phys. J. C **77** (2017) 466, arXiv:1703.10485 [hep-ex].

[101] ATLAS Collaboration, *In situ calibration of large-radius jet energy and mass in 13 TeV proton–proton collisions with the ATLAS detector*, Eur. Phys. J. C **79** (2019) 135, arXiv:1807.09477 [hep-ex].

[102] ATLAS Collaboration, *In situ large-R jet energy scale calibration and uncertainties in 2015-2017 data*, https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PLOTS/JETM-2019-05/. Accessed August 2019.

[103] J. Pequenao, *Computer generated image of the ATLAS Muons subsystem*, https://cds.cern.ch/record/1095929, © 2008 CERN.

[104] ATLAS Collaboration, W. A. Leight *et al.*, *New Fitting Concept in ATLAS muon tracking for the LHC Run-2*, Tech. Rep. ATL-SOFT-PROC-2018-052, 2018. https://cds.cern.ch/record/2649597.

[105] ATLAS Collaboration, *Momentum resolution improvements with the inclusion of the Alignment Errors On Track*, https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PLOTS/MUON-2018-003/. Accessed August 2019.

[106] ATLAS Collaboration, *2015 start-up trigger menu and initial performance assessment of the ATLAS trigger using Run-2 data*, ATL-DAQ-PUB-2016-001, 2016, https://cds.cern.ch/record/2136007.

[107] LHC Coordination, *Schedules and luminosity forecasts – HL-LHC*, https://lhc-commissioning.web.cern.ch/lhc-commissioning/schedule/HL-LHC-plots.htm. Accessed August 2019.

[108] ATLAS Collaboration, *New Small Wheel Technical Design Report*, tech. rep., 2013. https://cds.cern.ch/record/1552862.

[109] ATLAS Collaboration, *ATLAS Liquid Argon Calorimeter Phase-I Upgrade Technical Design Report*, Tech. Rep. CERN-LHCC-2013-017. ATLAS-TDR-022, Sep, 2013. https://cds.cern.ch/record/1602230.

[110] ATLAS Collaboration, E. Simioni *et al.*, *The Topological Processor for the future ATLAS Level-1 Trigger: from design to commissioning*, arXiv:1406.4316 [physics.ins-det].

[111] ATLAS Collaboration, *Fast TracKer (FTK) Technical Design Report*, Tech. Rep. CERN-LHCC-2013-007. ATLAS-TDR-021, Jun, 2013. https://cds.cern.ch/record/1552953.

[112] ATLAS Collaboration, *Technical Design Report for the ATLAS Inner Tracker Strip Detector*, Tech. Rep. CERN-LHCC-2017-005. ATLAS-TDR-025, Apr, 2017. https://cds.cern.ch/record/2257755.

[113] ATLAS Collaboration, *Technical Design Report for the ATLAS Inner Tracker Pixel Detector*, Tech. Rep. CERN-LHCC-2017-021. ATLAS-TDR-030, Sep, 2017. https://cds.cern.ch/record/2285585.

[114] ATLAS Collaboration, *Technical Design Report for the Phase-II Upgrade of the ATLAS TDAQ System*, Tech. Rep. CERN-LHCC-2017-020. ATLAS-TDR-029, Sep, 2017. https://cds.cern.ch/record/2285584.

[115] K. Albertsson *et al.*, *Machine Learning in High Energy Physics Community White Paper*, J. Phys. Conf. Ser. **1085** (2018) 022008, arXiv:1807.02876 [physics.comp-ph].

[116] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, 2006.

[117] W. S. McCulloch and W. Pitts, *A logical calculus of the ideas immanent in nervous activity*, Bull. Math. Biophys. **5** (1943) 115–133.

[118] *The Loss Surfaces of Multilayer Networks*, J. Mach. Lear. Res. **38** (2015) 192–204, arXiv:1412.0233 [cs.LG].

[119] S. Ioffe and C. Szegedy, *Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift*, CoRR (2015), arXiv:1502.03167.

[120] L. Breiman *et al.*, *Classification and Regression Trees*. Wadsworth International Group, 1984.

[121] Y. Freund and R. E. Schapire, *A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting*, J. Comput. Syst. Sci. **55** (1997) 119–139.

[122] Y. Freund and R. E. Schapire, *A Short Introduction to Boosting*, Artif. Intell. **14** (1999) 771–780.

[123] ATLAS Collaboration, *Summary plots from the ATLAS Experiment Exotic Physics group*, https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/CombinedSummaryPlots/EXOTICS/. Accessed January 2019.

[124] ATLAS Collaboration, *Search for dark matter and other new phenomena in events with an energetic jet and large missing transverse momentum using the ATLAS detector*, JHEP **01** (2018) 126, arXiv:1711.03301 [hep-ex].

[125] CMS Collaboration, *Search for dark matter, extra dimensions, and unparticles in monojet events in proton–proton collisions at* $\sqrt{s} = 8$ *TeV*, Eur. Phys. J. C **75** (2015) 235, arXiv:1408.3583 [hep-ex].

[126] CMS Collaboration, *Search for physics beyond the standard model in final states with a lepton and missing transverse energy in proton–proton collisions at* $\sqrt{s} = 8$ *TeV*, Phys. Rev. D **91** (2015) 092005, arXiv:1408.2745 [hep-ex].

[127] ATLAS Collaboration, *Search for an invisibly decaying Higgs boson or dark matter candidates produced in association with a Z boson in pp collisions at* $\sqrt{s} = 13$ *TeV with the ATLAS detector*, Phys. Lett. B **776** (2018) 318, arXiv:1708.09624 [hep-ex].

[128] CMS Collaboration, *Search for dark matter and unparticles produced in association with a Z boson in proton–proton collisions at* $\sqrt{s} = 8$ *TeV*, Phys. Rev. D **93** (2016) 052011, arXiv:1511.09375 [hep-ex].

[129] ATLAS Collaboration, *Search for dark matter at* $\sqrt{s} = 13$ *TeV in final states containing an energetic photon and large missing transverse momentum with the ATLAS detector*, Eur. Phys. J. C **77** (2017) 393, arXiv:1704.03848 [hep-ex].

[130] CMS Collaboration, *Search for new phenomena in monophoton final states in proton–proton collisions at* $\sqrt{s} = 8$ *TeV*, Phys. Lett. B **755** (2016) 102, arXiv:1410.8812 [hep-ex].

[131] ATLAS Collaboration, *Search for dark matter in association with a Higgs boson decaying to two photons at* $\sqrt{s} = 13$ *TeV with the ATLAS detector*, Phys. Rev. D **96** (2017) 112004, arXiv:1706.03948 [hep-ex].

[132] ATLAS Collaboration, *Search for Dark Matter Produced in Association with a Higgs Boson Decaying to* $b\bar{b}$ *using* $36\,fb^{-1}$ *of* pp *collisions at* $\sqrt{s} = 13$ *TeV with the ATLAS Detector*, Phys. Rev. Lett. **119** (2017) 181804, arXiv:1707.01302 [hep-ex].

[133] ATLAS Collaboration, *Search for Dark Matter Produced in Association with a Higgs Boson decaying to* $b\bar{b}$ *at* $\sqrt{s} = 13$ *TeV with the ATLAS Detector using* $79.8\,fb^{-1}$ *of proton–proton collision data*, ATLAS-CONF-2018-039, 2018, https://cds.cern.ch/record/2632344.

[134] ATLAS Collaboration, *Search for invisible Higgs boson decays in vector boson fusion at* $\sqrt{s} = 13$ *TeV with the ATLAS detector*, Phys. Lett. (2018), arXiv:1809.06682 [hep-ex].

[135] T. Ohl, *Drawing Feynman diagrams with FX340-1 and METAFONT*, Comput. Phys. Commun. **90** (1995) 340–354, arXiv:hep-ph/9505351.

[136] ATLAS Collaboration, *Muon reconstruction performance in early* $\sqrt{s} = 13$ *TeV data*, ATL-PHYS-PUB-2015-037, 2015, https://cds.cern.ch/record/2047831.

[137] CMS Collaboration, *Dimuon spectrum 2016*, CMS-DP-2016/059, 2016, https://cds.cern.ch/record/2212114.

[138] ATLAS Collaboration, *Search for new high-mass phenomena in the dilepton final state using* $36\,fb^{-1}$ *of proton–proton collision data at* $\sqrt{s} = 13$ *TeV with the ATLAS detector*, JHEP **10** (2017) 182, arXiv:1707.02424 [hep-ex].

[139] CMS Collaboration, *Search for high-mass resonances in dilepton final states in proton–proton collisions at* $\sqrt{s} = 13$ *TeV*, JHEP **06** (2018) 120, arXiv:1803.06292 [hep-ex].

[140] UA1 Collaboration, *Two Jet Mass Distributions at the CERN Proton - Anti-Proton Collider*, Phys. Lett. **B209** (1988) 127–134.

[141] UA2 Collaboration, *A Search for new intermediate vector mesons and excited quarks decaying to two jets at the CERN $\bar{p}p$ collider*, Nucl. Phys. **B400** (1993) 3–24.

[142] CDF Collaboration, *Search for new particles decaying into dijets in proton-antiproton collisions at s**(1/2) = 1.96-TeV*, Phys. Rev. **D79** (2009) 112002, arXiv:0812.4036 [hep-ex].

[143] D0 Collaboration, *Search for new particles in the two jet decay channel with the D0 detector*, Phys. Rev. **D69** (2004) 111101, arXiv:hep-ex/0308033 [hep-ex].

[144] B. A. Dobrescu and F. Yu, *Coupling-mass mapping of dijet peak searches*, Phys. Rev. **D88** (2013) 035021, arXiv:1306.2629 [hep-ph], [Erratum: Phys. Rev.D90,no.7,079901(2014)].

[145] ATLAS Collaboration, *Search for new phenomena in dijet events using 37 fb^{-1} of pp collision data collected at $\sqrt{s}$ = 13 TeV with the ATLAS detector*, Phys. Rev. D **96** (2017) 052004, arXiv:1703.09127 [hep-ex].

[146] CMS Collaboration, *Search for narrow and broad dijet resonances in proton–proton collisions at $\sqrt{s}$ = 13 TeV and constraints on dark matter mediators and other new particles*, JHEP **08** (2018) 130, arXiv:1806.00843 [hep-ex].

[147] ATLAS Collaboration, *Search for resonances in the mass distribution of jet pairs with one or two jets identified as b-jets in proton–proton collisions at $\sqrt{s}$ = 13 TeV with the ATLAS detector*, Phys. Rev. D **98** (2018) 032016, arXiv:1805.09299 [hep-ex].

[148] ATLAS Collaboration, *Search for Low-Mass Dijet Resonances Using Trigger-Level Jets with the ATLAS Detector in pp Collisions at $\sqrt{s}$ = 13 TeV*, Phys. Rev. Lett. **121** (2018) 081801, arXiv:1804.03496 [hep-ex].

[149] CMS Collaboration, *Search for new physics in dijet angular distributions using proton–proton collisions at $\sqrt{s}$ = 13 TeV and constraints on dark matter and other models*, Eur. Phys. J. C **78** (2018) 789, arXiv:1803.08030 [hep-ex].

[150] ATLAS Collaboration, *Performance of the ATLAS trigger system in 2015*, Eur. Phys. J. C **77** (2017) 317, arXiv:1611.09661 [hep-ex].

[151] H. An, R. Huo, and L.-T. Wang, *Searching for Low Mass Dark Portal at the LHC*, Phys. Dark Univ. **2** (2013) 50–57, arXiv:1212.2221 [hep-ph].

[152] C. Shimmin and D. Whiteson, *Boosting low-mass hadronic resonances*, Phys. Rev. **D94** (2016) 055001, arXiv:1602.07727 [hep-ph].

[153] ATLAS Collaboration, *Search for low-mass resonances decaying into two jets and produced in association with a photon using pp collisions at $\sqrt{s}$ = 13 TeV with the ATLAS detector*, Phys. Lett. (2019), arXiv:1901.10917 [hep-ex].

[154] ATLAS Collaboration, *Search for new light resonances decaying to jet pairs and produced in association with a photon or a jet in proton–proton collisions at* $\sqrt{s}$ = 13 *TeV with the ATLAS detector*, ATLAS-CONF-2016-070, 2016, https://cds.cern.ch/record/2206221.

[155] CMS Collaboration, *Search for low mass vector resonances decaying to quark–antiquark pairs in proton–proton collisions at* $\sqrt{s}$ = 13 *TeV*, Phys. Rev. Lett. **119** (2017) 111802, arXiv:1705.10532 [hep-ex].

[156] CMS Collaboration, *Inclusive Search for a Highly Boosted Higgs Boson Decaying to a Bottom Quark–Antiquark Pair*, Phys. Rev. Lett. **120** (2018) 071802, arXiv:1709.05543 [hep-ex].

[157] CMS Collaboration, *Search for low-mass resonances decaying into bottom quark–antiquark pairs in proton–proton collisions at* $\sqrt{s}$ = 13 *TeV*, Phys. Rev. D **99** (2019) 012005, arXiv:1810.11822 [hep-ex].

[158] ATLAS Collaboration, *Search for low mass di-jet resonances using proton–proton collisions at* $\sqrt{s}$ = 13 *TeV with the ATLAS detector*, tech. rep., CERN, 2016. ATL-COM-PHYS-2016-1786.

[159] ATLAS Collaboration, *Trigger Menu in 2016*, ATL-DAQ-PUB-2017-001, 2017, https://cds.cern.ch/record/2242069.

[160] ATLAS Collaboration, *Electron and photon energy calibration with the ATLAS detector using data collected in 2015 at* $\sqrt{s}$ = 13 *TeV*, ATL-PHYS-PUB-2016-015, 2016, https://cds.cern.ch/record/2203514.

[161] ATLAS Collaboration, P. J. Laycock *et al.*, *ATLAS data preparation in run 2*, J. Phys. Conf. Ser. **898** (2017) 042050.

[162] ATLAS Collaboration, *Search for New Phenomena in Dijet Mass and Angular Distributions from pp Collisions at* $\sqrt{s}$ = 13 *TeV with the ATLAS Detector*, Phys. Lett. B **754** (2016) 302, arXiv:1512.01530 [hep-ex].

[163] J. Alwall *et al.*, *The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations*, JHEP **07** (2014) 079, arXiv:1405.0301 [hep-ph].

[164] R. D. Ball *et al.*, *Parton distributions with LHC data*, Nucl. Phys. **B867** (2013) 244–289, arXiv:1207.1303 [hep-ph].

[165] T. Sjostrand, S. Mrenna, and P. Z. Skands, *A Brief Introduction to PYTHIA 8.1*, Comput. Phys. Commun. **178** (2008) 852–867, arXiv:0710.3820 [hep-ph].

[166] ATLAS Collaboration, *ATLAS Pythia 8 tunes to* 7 *TeV data*, ATL-PHYS-PUB-2014-021, 2014, https://cds.cern.ch/record/1966419.

[167] T. Gleisberg, S. Hoeche, F. Krauss, M. Schonherr, S. Schumann, F. Siegert, and J. Winter, *Event generation with SHERPA 1.1*, JHEP **02** (2009) 007, arXiv:0811.4622 [hep-ph].

[168] H.-L. Lai *et al.*, *New parton distributions for collider physics*, Phys. Rev. **D82** (2010) 074024, arXiv:1007.2241 [hep-ph].

[169] S. Schumann and F. Krauss, *A Parton shower algorithm based on Catani-Seymour dipole factorisation*, JHEP **03** (2008) 038, arXiv:0709.1027 [hep-ph].

[170] S. Hoeche, F. Krauss, S. Schumann, and F. Siegert, *QCD matrix elements and truncated showers*, JHEP **05** (2009) 053, arXiv:0903.1219 [hep-ph].

[171] ATLAS Collaboration, *The ATLAS Simulation Infrastructure*, Eur. Phys. J. C **70** (2010) 823, arXiv:1005.4568 [physics.ins-det].

[172] GEANT4 Collaboration, S. Agostinelli *et al.*, *GEANT4: A Simulation toolkit*, Nucl. Instrum. Meth. **A506** (2003) 250–303.

[173] ATLAS Collaboration, *Summary of ATLAS Pythia 8 tunes*, ATL-PHYS-PUB-2012-003, 2012, https://cds.cern.ch/record/1474107.

[174] A. D. Martin, W. J. Stirling, R. S. Thorne, and G. Watt, *Parton distributions for the LHC*, Eur. Phys. J. C **63** (2009) 189–285, arXiv:0901.0002 [hep-ph].

[175] ATLAS Collaboration, *Photon identification in 2015 ATLAS data*, ATL-PHYS-PUB-2016-014, 2016, https://cds.cern.ch/record/2203125.

[176] M. Cacciari, G. P. Salam, and G. Soyez, *The Catchment Area of Jets*, JHEP **04** (2008) 005, arXiv:0802.1188 [hep-ph].

[177] M. Cacciari, G. P. Salam, and S. Sapeta, *On the characterisation of the underlying event*, JHEP **04** (2010) 065, arXiv:0912.4926 [hep-ph].

[178] ATLAS Collaboration, *ATLAS electron, photon and muon isolation in Run 2*, tech. rep., CERN, 2017. ATL-COM-PHYS-2017-290.

[179] ATLAS Collaboration, *Monitoring and data quality assessment of the ATLAS liquid argon calorimeter*, JINST **9** (2014) P07024, arXiv:1405.3768 [hep-ex].

[180] ATLAS Collaboration, *Topological cell clustering in the ATLAS calorimeters and its performance in LHC Run 1*, Eur. Phys. J. C **77** (2017) 490, arXiv:1603.02934 [hep-ex].

[181] M. Cacciari, G. P. Salam, and G. Soyez, *FastJet User Manual*, Eur. Phys. J. C **72** (2012) 1896, arXiv:1111.6097 [hep-ph].

[182] ATLAS Collaboration, *Jet energy scale measurements and their systematic uncertainties in proton–proton collisions at* $\sqrt{s} = 13$ *TeV with the ATLAS detector*, Phys. Rev. D **96** (2017) 072002, arXiv:1703.09665 [hep-ex].

[183] ATLAS Collaboration, *Monte Carlo Calibration and Combination of In-situ Measurements of Jet Energy Scale, Jet Energy Resolution and Jet Mass in ATLAS*, ATLAS-CONF-2015-037, 2015, https://cds.cern.ch/record/2044941.

[184] ATLAS Collaboration, *Selection of jets produced in* 13 *TeV proton–proton collisions with the ATLAS detector*, ATLAS-CONF-2015-029, 2015, https://cds.cern.ch/record/2037702.

[185] ATLAS Collaboration, *Jet energy measurement with the ATLAS detector in proton–proton collisions at* $\sqrt{s} = 7$ *TeV*, Eur. Phys. J. C **73** (2013) 2304, arXiv:1112.6426 [hep-ex].

[186] A. J. Larkoski, D. Neill, and J. Thaler, *Jet Shapes with the Broadening Axis*, JHEP **04** (2014) 017, arXiv:1401.2158 [hep-ph].

[187] ATLAS Collaboration, *Reconstruction of primary vertices at the ATLAS experiment in Run 1 proton–proton collisions at the LHC*, Eur. Phys. J. C **77** (2017) 332, arXiv:1611.10235 [hep-ex].

[188] G. Kasieczka, T. Plehn, T. Schell, T. Strebler, and G. P. Salam, *Resonance Searches with an Updated Top Tagger*, JHEP **06** (2015) 203, arXiv:1503.05921 [hep-ph].

[189] J. Dolen, P. Harris, S. Marzani, S. Rappoccio, and N. Tran, *Thinking outside the ROCs: Designing Decorrelated Taggers (DDT) for jet substructure*, JHEP **05** (2016) 156, arXiv:1603.00027 [hep-ph].

[190] M. Dasgupta, A. Fregoso, S. Marzani, and G. P. Salam, *Towards an understanding of jet substructure*, JHEP **09** (2013) 029, arXiv:1307.0007 [hep-ph].

[191] ATLAS Collaboration, *Improving jet substructure performance in ATLAS using Track-CaloClusters*, ATL-PHYS-PUB-2017-015, 2017, https://cds.cern.ch/record/2275636.

[192] ATLAS Collaboration, *Variable Radius, Exclusive-*k_T*, and Center-of-Mass Subjet Reconstruction for Higgs(*$\to b\bar{b}$*) Tagging in ATLAS*, ATL-PHYS-PUB-2017-010, 2017, https://cds.cern.ch/record/2268678.

[193] S. Baker and R. D. Cousins, *Clarification of the Use of Chi Square and Likelihood Functions in Fits to Histograms*, Nucl. Instrum. Meth. **221** (1984) 437–442.

[194] G. Cowan, K. Cranmer, E. Gross, and O. Vitells, *Asymptotic formulae for likelihood-based tests of new physics*, Eur. Phys. J. **C71** (2011) 1554, arXiv:1007.1727 [physics.data-an], [Erratum: Eur. Phys. J. **C73** (2013) 2501].

[195] L. Kaplan. Private communication, 2017.

[196] ATLAS Collaboration, *Measurement of the cross section for isolated-photon plus jet production in pp collisions at* $\sqrt{s} = 13$ *TeV using the ATLAS detector*, Phys. Lett. B **780** (2018) 578, arXiv:1801.00112 [hep-ex].

[197] ATLAS Collaboration, *Search for new phenomena with photon+jet events in proton–proton collisions at* $\sqrt{s} = 13$ *TeV with the ATLAS detector*, JHEP **03** (2016) 041, arXiv:1512.05910 [hep-ex].

[198] M. Bahr *et al.*, *Herwig++ Physics and Manual*, Eur. Phys. J. **C58** (2008) 639–707, arXiv:0803.0883 [hep-ph].

[199] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005.

[200] S. N. Lophaven, H. B. Nielsen, and J. Søndergaard, *Aspects of the Matlab toolbox DACE*, tech. rep., Informatics and Mathematical Modelling, Technical University of Denmark, DTU, 2002. http://www2.imm.dtu.dk/pubdb/p.php?1050.

[201] F. Pedregosa *et al.*, *Scikit-learn: Machine Learning in Python*, J. Mach. Learn. Res. **12** (2011) 2825–2830, arXiv:1201.0490 [cs.LG].

[202] W. Verkerke and D. P. Kirkby, *The RooFit toolkit for data modeling*, eConf **C0303241** (2003) MOLT007, arXiv:physics/0306116 [physics].

[203] ATLAS Collaboration, *Luminosity determination in pp collisions at* $\sqrt{s} = 8$ *TeV using the ATLAS detector at the LHC*, Eur. Phys. J. C **76** (2016) 653, arXiv:1608.03953 [hep-ex].

[204] G. Avoni *et al.*, *The new LUCID-2 detector for luminosity measurement and monitoring in ATLAS*, JINST **13** (2018) P07017.

[205] M. Baak, S. Gadatsch, R. Harrington, and W. Verkerke, *Interpolation between multi-dimensional histograms using a new non-linear moment morphing method*, Nucl. Instrum. Meth. **A771** (2015) 39–48, arXiv:1410.7388 [physics.data-an].

[206] J. Neyman and E. S. Pearson, *On the Problem of the Most Efficient Tests of Statistical Hypotheses*, Philosophical Transactions of the Royal Society of London Series A **231** (1933) 289–337.

[207] S. S. Wilks, *The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses*, Annals Math. Statist. **9** (1938) 60–62.

[208] A. Wald, *Tests of Statistical Hypotheses Concerning Several Parameters When the Number of Observations is Large*, Transactions of the American Mathematical Society **54** (1943) 426–482.

[209] A. L. Read, *Presentation of search results: the* CL_s *technique*, J. Phys. **G28** (2002) 2693.

[210] R. Brun and F. Rademakers, *ROOT — An object oriented data analysis framework*, Nucl. Instrum. Methods Phys. Res. **A389** (1997) 81–86.

[211] L. Moneta *et al.*, *The RooStats Project*, PoS **ACAT2010** (2010) 057, arXiv:1009.1003 [physics.data-an].

[212] M. Baak *et al.*, *HistFitter software framework for statistical data analysis*, Eur. Phys. J. **C75** (2015) 153, arXiv:1410.1280 [hep-ex].

[213] E. Gross and O. Vitells, *Trial factors for the look elsewhere effect in high energy physics*, Eur. Phys. J. **C70** (2010) 525–530, arXiv:1005.1891 [physics.data-an].

[214] CMS Collaboration, *Search for low-mass quark-antiquark resonances produced in association with a photon at* $\sqrt{s}$ *= 13 TeV*, arXiv:1905.10331 [hep-ex].

[215] ATLAS Collaboration, *Search for boosted resonances decaying to two b-quarks and produced in association with a jet at* $\sqrt{s}$ = 13 *TeV with the ATLAS detector*, ATLAS-CONF-2018-052, 2018, https://cds.cern.ch/record/2649081.

[216] ATLAS Collaboration, *Search for dark matter in events with a hadronically decaying vector boson and missing transverse momentum in pp collisions at* $\sqrt{s}$ = 13 *TeV with the ATLAS detector*, JHEP **10** (2018) 180, arXiv:1807.11471 [hep-ex].

[217] ATLAS Collaboration, *A new method to distinguish hadronically decaying boosted Z bosons from W bosons using the ATLAS detector*, Eur. Phys. J. C **76** (2016) 238, arXiv:1509.04939 [hep-ex].

[218] ATLAS Collaboration, *Identification of high transverse momentum top quarks in pp collisions at* $\sqrt{s}$ = 8 *TeV with the ATLAS detector*, JHEP **06** (2016) 093, arXiv:1603.03127 [hep-ex].

[219] ATLAS Collaboration, *Identification of boosted, hadronically decaying W bosons and comparisons with ATLAS data taken at* $\sqrt{s}$ = 8 *TeV*, Eur. Phys. J. C **76** (2016) 154, arXiv:1510.05821 [hep-ex].

[220] ATLAS Collaboration, *Boosted hadronic top identification at ATLAS for early* 13 *TeV data*, ATL-PHYS-PUB-2015-053, 2015, https://cds.cern.ch/record/2116351.

[221] ATLAS Collaboration, *Identification of Boosted, Hadronically-Decaying W and Z Bosons in* $\sqrt{s}$ = 13 *TeV Monte Carlo Simulations for ATLAS*, ATL-PHYS-PUB-2015-033, 2015, https://cds.cern.ch/record/2041461.

[222] ATLAS Collaboration, *Identification of Hadronically-Decaying W Bosons and Top Quarks Using High-Level Features as Input to Boosted Decision Trees and Deep Neural Networks in ATLAS at* $\sqrt{s} = 13$ *TeV*, ATL-PHYS-PUB-2017-004, 2017, https://cds.cern.ch/record/2259646.

[223] S. A. Dudani, *The Distance-Weighted k-Nearest-Neighbor Rule*, IEEE Transactions on Systems, Man, and Cybernetics **SMC-6** (1976) 325–327.

[224] I. Moult, B. Nachman, and D. Neill, *Convolved Substructure: Analytically Decorrelating Jet Substructure Observables*, JHEP **05** (2018) 002, arXiv:1710.06859 [hep-ph].

[225] I. Goodfellow *et al.*, *Generative Adversarial Nets*, Advances in Neural Information Processing Systems **27** (2014) 2672–2680, arXiv:1406.2661 [stat.ML].

[226] J. Stevens and M. Williams, *uBoost: A boosting method for producing uniform selection efficiencies from multivariate classifiers*, JINST **8** (2013) P12013, arXiv:1305.7248 [nucl-ex].

[227] NNPDF Collaboration, R. Ball *et al.*, *Parton distributions with QED corrections*, Nucl. Phys. B **877** (2013) 290–320, arXiv:1308.0598 [hep-ph].

[228] ATLAS Collaboration, *Performance of Top Quark and W Boson Tagging in Run 2 with ATLAS*, ATLAS-CONF-2017-064, 2017, https://cds.cern.ch/record/2281054.

[229] A. Rogozhnikov, *hep_ml*, https://arogozhnikov.github.io/hep_ml/index.html.

[230] A. J. Larkoski, I. Moult, and D. Neill, *Power Counting to Better Jet Observables*, JHEP **12** (2014) 009, arXiv:1409.6298 [hep-ph].

[231] L. G. Almeida *et al.*, *Substructure of high-*p_T *Jets at the LHC*, Phys. Rev. **D79** (2009) 074017, arXiv:0807.0234 [hep-ph].

[232] C. Chen, *New approach to identifying boosted hadronically-decaying particle using jet substructure in its center-of-mass frame*, Phys. Rev. D **85** (2012) 034007, arXiv:1112.2567 [hep-ph].

[233] G. C. Fox and S. Wolfram, *Observables for the Analysis of Event Shapes in* e^+e^- *Annihilation and Other Processes*, Phys. Rev. Lett. **41** (1978) 1581–1585.

[234] ATLAS Collaboration, *Measurement of* k_T *splitting scales in* $W \to \ell\nu$ *events at* $\sqrt{s} = 7$ *TeV with the ATLAS detector*, Eur. Phys. J. C **73** (2013) 2432, arXiv:1302.1415 [hep-ex].

[235] J. Thaler and L.-T. Wang, *Strategies to Identify Boosted Tops*, JHEP **07** (2008) 092, arXiv:0806.0023 [hep-ph].

[236] A. Kolmogorov, *Sulla determinazione empirica di una lgge di distribuzione*, Inst. Ital. Attuari, Giorn. **4** (1933) 83–91.

[237] N. V. Smirnov, *Estimate of deviation between empirical distribution functions in two independent samples*, Bulletin Moscow University **2** (1939) 3–16.

[238] S. Kullback and R. A. Leibler, *On Information and Sufficiency*, Ann. Math. Statist. **22** (1951) 79–86, https://doi.org/10.1214/aoms/1177729694.

[239] J. Lin, *Divergence measures based on the Shannon entropy*, IEEE Transactions on Information Theory **37** (1991) 145–151.

[240] G. Louppe, M. Kagan, and K. Cranmer, *Learning to Pivot with Adversarial Networks*, Advances in Neural Information Processing Systems **30** (2017) 981–990, arXiv:1611.01046 [stat.ML].

[241] C. Shimmin *et al.*, *Decorrelated Jet Substructure Tagging using Adversarial Neural Networks*, Phys. Rev. **D96** (2017) 074034, arXiv:1703.03507 [hep-ex].

[242] F. Chollet *et al.*, *Keras*, https://github.com/fchollet/keras, 2018.

[243] M. Abadi *et al.*, *TensorFlow: A System for Large-scale Machine Learning*, Proc. OSDI (2016) 265–283, arXiv:1605.08695 [cs.DC].

[244] A. Søgaard, *adversarial*, https://github.com/asogaard/adversarial/tree/PUBNOTE, 2018.

[245] C. M. Bishop, *Mixture density networks*, https://publications.aston.ac.uk/373/1/NCRG_94_004.pdf, 1994.

[246] P. Baldi, K. Cranmer, T. Faucett, P. Sadowski, and D. Whiteson, *Parameterized neural networks for high-energy physics*, Eur. Phys. J. C **76** (2016) 235, arXiv:1601.07913 [hep-ex].

[247] I. J. Goodfellow, *NIPS 2016 Tutorial: Generative Adversarial Networks*, CoRR (2017), arXiv:1701.00160.

[248] V. Nagarajan and J. Z. Kolter, *Gradient descent GAN optimization is locally stable*, CoRR (2017), arXiv:1706.04156.

[249] Y. Ganin *et al.*, *Domain-Adversarial Training of Neural Networks*, J. Mach. Learn. Res. **17** (2016) 1–35, arXiv:1505.07818 [stat.ML].

[250] D. P. Kingma and J. Ba, *Adam: A Method for Stochastic Optimization*, Proceedings of the 3rd International Conference on Learning Representations (ICLR) (2015), arXiv:1412.6980 [cs.LG].

[251] B. Efron, *Bootstrap Methods: Another Look at the Jackknife*, Ann. Statist. **7** (1979) 1–26, https://doi.org/10.1214/aos/1176344552.

[252] J. Duarte *et al.*, *Fast inference of deep neural networks in FPGAs for particle physics*, JINST **13** (2018) P07027, arXiv:1804.06913 [physics.ins-det].

[253] ATLAS Collaboration, *ATLAS measurements of the properties of jets for boosted particle searches*, Phys. Rev. D **86** (2012) 072006, arXiv:1206.5369 [hep-ex].

[254] T. Sjostrand, S. Mrenna, and P. Z. Skands, *PYTHIA 6.4 Physics and Manual*, JHEP **05** (2006) 026, arXiv:hep-ph/0603175 [hep-ph].

[255] T. Plehn, G. P. Salam, and M. Spannowsky, *Fat Jets for a Light Higgs*, Phys. Rev. Lett. **104** (2010) 111801, arXiv:0910.5472 [hep-ph].

[256] T. Plehn, M. Spannowsky, M. Takeuchi, and D. Zerwas, *Stop Reconstruction with Tagged Tops*, JHEP **10** (2010) 078, arXiv:1006.2833 [hep-ph].

[257] D. E. Soper and M. Spannowsky, *Finding physics signals with shower deconstruction*, Phys. Rev. D **84** (2011) 074002, arXiv:1102.3480 [hep-ph].

[258] D. E. Soper and M. Spannowsky, *Finding top quarks with shower deconstruction*, Phys. Rev. D **87** (2013) 054012, arXiv:1211.3140 [hep-ph].

[259] J. Pearkes, W. Fedorko, A. Lister, and C. Gay, *Jet Constituents for Deep Neural Network Based Top Quark Tagging*, arXiv:1704.02124 [hep-ex].

[260] P. Baldi, P. Sadowski, and D. Whiteson, *Searching for Exotic Particles in High-Energy Physics with Deep Learning*, Nature Commun. **5** (2014) 4308.

[261] M. Stoye, J. Kieseler, H. Qu, L. Gouskos, and M. Verzetti, *DeepJet: Generic physics object based jet multiclass classification for LHC experiments*, Workshop on Deep Learning for Physical Sciences (DLPS2017), NIPS **10** (2017), https://dl4physicalsciences.github.io/files/nips_dlps_2017_10.pdfk.

[262] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, *Improving neural networks by preventing co-adaptation of feature detectors*, arXiv:1207.0580 [cs.NE].

[263] A. Rogozhnikov *et al.*, *hep_ml: Machine Learning for High Energy Physics*, https://github.com/arogozhnikov/hep_ml, 2017.

[264] J. Snoek, H. Larochelle, and R. P. Adams, *Practical Bayesian Optimization of Machine Learning Algorithms*, Advances in Neural Information Processing Systems **25** (2012) 2951–2959, arXiv:1206.2944 [stat.ML].

[265] J. Snoek *et al.*, *Spearmint*, https://github.com/HIPS/Spearmint, accessed 2018. Commit: ffbab6653ae785c9acdcf2abb01c63127be40c2f.