Olga Sunneborn Gudnadottir

Of dark mesons and novel methods

A dark sector search in ATLAS data and development of new techniques for challenging final states
Abstract


Studies of the interactions of elementary particles at high energies have been carried out at the Large Hadron Collider (LHC) at CERN for over a decade. Different quantities from the Standard Model (SM) of particle physics have been measured with increasing accuracy without substantial deviations from predictions. Searches for physics beyond the SM are similarly carried out, motivated by the existence of phenomena not yet described by it, such as dark matter. This thesis presents one such search in proton-proton collision data recorded by the ATLAS detector. The search is guided by a new, proposed addition to the SM, where the dark matter candidate arises as a composite particle of a new sector. If this were realised in nature, the same sector would give rise to other composite particles, dark mesons, that would be produced in proton-proton collisions and decay promptly to SM particles. This new model is largely free of previous constraints from searches and measurements. The full analysis targeting pair produced dark pions decaying to top and bottom quarks, $t\bar{t}b\bar{b}$ or $t\bar{t}bb$, in the 1-lepton channel is described.

It is carried out in the full Run 2 dataset of 140 fb$^{-1}$ of proton-proton collisions at $\sqrt{s} = 13$ TeV center-of-mass energy. The analysis is sensitive to large parts of the parameter space of the model, and no significant excess was seen over SM predictions. Based on this, limits on the production cross-section of dark pions were set. By comparing with the theoretical cross sections of the model, these rule out dark pion masses up to 943 GeV in the most sensitive configuration.

Further, several novel techniques that could aid with searches in similar phase-spaces are presented. First, the Extrapolation Engine fast simulation of the inner tracker for the high luminosity upgrade of ATLAS was used in the study of a proposed hardware track trigger (HTT). This could be crucial to retaining efficiency in similar phase-spaces in the extreme conditions at the high luminosity LHC (HL-LHC). Second, the fully scalable multi-dimensional density estimate in SparkDensityTrees was applied on background and signal similar to those in the dark meson analysis and was shown to efficiently find signal-enriched regions. Third, the unsupervised clustering algorithm UCluster which can be trained with any clustering objective, such as signal extraction, anomaly detection or jet tagging was developed to run on multiple cores for arbitrary scalability. Lastly, a Boosted Decision Tree (BDT) was applied for signal and background discrimination in the dark meson analysis, yielding promising results for future iterations of it.

Keywords: High energy physics, dark matter, machine learning, dark mesons, dark sector, ATLAS, LHC
For Viktor
Contents

1 Introduction ................................................................................................ 11
  1.1 A brief history of elementary particle physics ...................................... 11
  1.2 The ATLAS experiment today ............................................................... 12
  1.3 Overview of DM and dark sectors ....................................................... 12
  1.4 Scientific aim ...................................................................................... 15

Part I: Experimental setup and theoretical background ................................. 17

2 The Large Hadron Collider ....................................................................... 19
  2.1 The Run 2 conditions ......................................................................... 19

3 The ATLAS detector ................................................................................. 21
  3.1 The magnet system ........................................................................... 21
  3.2 The inner detector ........................................................................... 21
  3.3 The calorimeters ............................................................................... 23
  3.4 The muon system ............................................................................. 24
  3.5 The luminosity detectors .................................................................. 24
  3.6 The ATLAS coordinate system ......................................................... 24

4 ATLAS datataking ..................................................................................... 26
  4.1 ATLAS trigger and data acquisition system ....................................... 26
  4.1.1 Trigger streams and menus ....................................................... 26
  4.2 Pile-up ................................................................................................ 27
  4.3 The Run 2 dataset ........................................................................... 28

5 Theoretical background ............................................................................. 29
  5.1 The Standard Model ......................................................................... 29
  5.2 Dark mesons .................................................................................... 32

6 Theoretical predictions in ATLAS ............................................................ 36
  6.1 Monte Carlo simulations ................................................................... 36
    6.1.1 MC event simulators .............................................................. 37
    6.1.2 The ATLAS detector simulations ........................................... 37
  6.2 ATLAS simulated samples ............................................................... 38
    6.2.1 SM mismodelling ................................................................... 38
  6.3 Statistical analysis ............................................................................ 38

7 Event reconstruction .................................................................................. 40
  7.1 Objects used in the reconstruction ................................................... 40
    7.1.1 Vertexing ............................................................................... 40
14.1 Summary of Systematics ............................................................... 89
14.2 Instrumental uncertainties ............................................................. 89
  14.2.1 Luminosity uncertainty ................................................... 89
  14.2.2 Pileup reweighting uncertainty ....................................... 89
  14.2.3 Lepton uncertainty ........................................................... 89
  14.2.4 Flavour-tagging uncertainty ............................................ 91
  14.2.5 Jet Energy Scale and Resolution for small-R jets ............. 91
14.3 Theoretical Cross Sections ............................................................ 92
  14.3.1 Top quark $p_T$ reweighting ............................................... 92
  14.3.2 $t\bar{t}$ uncertainty .................................................................... 92
  14.3.3 $t\bar{t}b\bar{b}$ uncertainty ......................................................... 93
  14.3.4 Single-Top uncertainty .................................................... 93
  14.3.5 Minor simulated backgrounds ........................................ 94

15 Results ........................................................................................................ 95

Part III: Looking to the future ....................................................................... 101

16 The UCluster method .............................................................................. 103
  16.1 Distributed training for the particle physics data method
      UCluster ........................................................................................ 103
  16.2 Model and data ............................................................................. 105
  16.3 Modifying the code for scalability and distributed training ...... 106
      16.3.1 Distributed training ........................................................ 107
      16.3.2 Scalability ...................................................................... 107
  16.4 Training and evaluation ............................................................... 107
  16.5 Next steps ..................................................................................... 109
  16.6 Conclusion and outlook ............................................................... 110

17 A Boosted Decision Tree for the Dark Meson analysis ........................ 111
  17.1 Boosted Decision Trees ............................................................... 111
  17.2 Training data and model .............................................................. 114
  17.3 Results ........................................................................................... 115

18 Sparks in the Dark ................................................................................... 116
  18.1 Introduction .................................................................................. 116
  18.2 Datasets and event selection ........................................................ 117
  18.3 Discriminating variables ............................................................. 118
  18.4 Method .......................................................................................... 118
  18.5 Results ........................................................................................... 121
  18.6 Conclusion and Outlook ............................................................... 125

19 A Hardware Tracker for the Trigger ....................................................... 126
  19.1 Introduction .................................................................................. 126
  19.2 Background ................................................................................. 127
      19.2.1 The ITk ........................................................................... 127
1. Introduction

This thesis is a collection of projects, all embedded within the larger context of the general purpose high-energy physics experiment ATLAS at the Large Hadron Collider (LHC) at CERN. The main project deals with one of the outstanding questions in elementary particle physics, "What is dark matter?", through a search for dark mesons made from constituents of a proposed dark sector, in ATLAS proton-proton collision data. The other projects all have in common that with more or less development they might make this analysis, and others like it, more efficient.

1.1 A brief history of elementary particle physics

At the time of conception of the ATLAS experiment in the 1990s, the LHC was already well into being designed. Of it, physicists agreed that the new machine must be capable of exploring the full mass range up to the TeV scale. This was preceded by a time of rapid discovery and development in elementary particle physics. The electron, positron and muon were discovered between 1897 and 1932 [109, 34, 1, 2, 151]. The neutrino was discovered in 1956 after studying beta decays through the lens of conservation laws cite-muon,Neddermeyer:1937md, and in 1962 the electron and muon neutrino were found to be distinct particles [3, 101]. This lead to the theoretical unification of the electromagnetic and weak forces, which in turn lead to the discovery of the W and Z bosons in 1983 [105]. In the meantime, experiments found a virtual zoo of new hadrons (dubbed "The Particle Zoo"). Between 1947 and 1974, the pion, kaon, lambda baryon, antiproton, omega baryon, and the J/Ψ particles were discovered [63]. It also became clear that these particles were not elementary, but rather were made up of other, point-like, particles. This was identified with the quark model which has previously been used to classify the Particle Zoo, and the point-like particles got the name quarks. The properties of the newly discovered quarks lead to the formulation of a quantum field theory of the strong force, quantum chromodynamics (QCD). Combining the electroweak description with QCD gave rise to the Standard Model (SM) of Particle Physics.

The SM was not yet internally consistent however, as it did not account for the gauge boson masses that had already been measured.

This is the basis for the design stipulation that LHC be able to explore masses up to 1 TeV; it corresponds to the point at which the SM unchanged
becomes unphysical. If the SM were indeed valid, it was clear that new physics would be discovered before that.

One theoretical candidate was the Brout–Englert–Higgs mechanism [127, 108] which predicted a new spin-0 boson, the Higgs boson. It already had enough interest in the community that physicists agreed that the matter of the Higgs boson needed to be clarified with the new accelerator and detectors.

Two general-purpose detectors, ATLAS and CMS, were built around the LHC. The physics programme for both was to explore the high-energy frontier of particle physics, and to clarify the question of the existence of the Higgs boson. The LHC took physics data for the first time in 2010 and already in 2012 ATLAS and CMS both announced an observation each in the search for the SM Higgs boson [24, 85]. It has since been studied in detail and nowhere found to deviate from the SM predictions.

With this discovery, the SM took the form that it still has today.

1.2 The ATLAS experiment today

ATLAS still continues the exploration of the TeV range by precision measurements of SM observables, as well as targeted searches for new interactions or particles. The physics programme includes e.g. SM measurements, searches for exotic signatures, searches for Higgs and diboson signatures, Higgs physics, heavy ion physics, searches for supersymmetry and top physics, and has produced over a thousand results to date [9]. One part of the programme, the SM cross section measurements, is summarised in figure 1.1. In all measurements of SM observables at the LHC, no deviations have been found from predictions. Despite the remarkable success of the SM, however, there are still reasons to believe it can be improved upon. A clear example is the neutrino masses, which are explicitly massless in the SM, but which have been measured to have mass by neutrino oscillation experiments [117]. Another reason is the imbalance of matter and anti-matter in the universe, whereas the SM is symmetric with respect to the two. It is also a possibility that the SM can be extended to account for unexplained phenomena in physics in general, which is the case for dark matter.

1.3 Overview of DM and dark sectors

Cosmological mass measurements have shown that there is more mass in the universe than can be accounted for by known particles and current gravitational theories [64]. In summary, the mass measured by gravitational effects is larger than the mass visible by any other type of detection, assuming that Newtonian dynamics and the theory of general relativity are valid descriptions of the masses of the universe.
Nevertheless, the search for evidence of non-gravitational interactions between DM and SM particles is an active field of research, and theoretical models describing it are continuously being developed following the experimental results.

On the experimental side, direct detection (DD) experiments, indirect detection (ID) experiments, and collider experiments complement each other, giving information on interaction cross sections, annihilation cross sections and life-time, and production cross sections and mechanisms, respectively.

Theoretically, any model that includes a particle whose properties fit the experimental results from the mass measurements and searches is a viable candidate. The constraints include that the particle has to be electrically neutral, stable on the time-scale of the universe, and massive. The thermal relic abundance also has to match the amount of dark matter observed in the universe.

Theories that fit this bill can be explicitly constructed in effective field theories, with the assumption that only a lower-energy approximation to a full theory is accessible in DM experiments. EFTs containing a single DM particle with an effective coupling to SM particles have been extensively studied at the LHC as well as in DD and ID experiments.
Figure 1.2. Regions in the parameter space of a benchmark simplified model excluded at 95% confidence level [15]. The shaded areas correspond to the union of the exclusion contours of all searches listed in the legend. The model is a leptophilic axial-vector mediator simplified model, with mediator $Z'_A$ and dark matter $\chi$.

In ATLAS, this has been done primarily with what is known as MET+X signatures and invisible Higgs decays [27]. Both signatures are based on the fact that DM particles would not interact with the detector and would therefore show up in the collision products only as unbalanced energy and momentum equations. MET (missing transverse energy) measures how much energy is missing from the detected decay products, given the initial state energy. The $X$ in the MET+X stands for the SM particle produced in association with the DM particle. An invisible Higgs decay signature consists of two reconstructed quarks and large MET, where the MET comes from a Higgs boson that decays to two DM particles.

Early Run 2 saw the addition of simplified models to the EFTs in the search programme [27]. They completely describe one part of a theory, and often involve adding a mediator particle to the DM candidate, that can be produced resonantly at the LHC. The simplified models make no claim to a full description of reality, but rather describe the full phenomenology of specific mechanisms, that can then be fed back into theory to improve the models. If one specific mediator and search channel plays a dominant role in the DM production, a simplified model might even provide the first evidence of DM production. So far, no new physics has been seen in these searches, and a large part of the probable phase space of the models has been ruled out. As an example, figure 1.2 summarises the excluded regions for a benchmark simplified
model. Several benchmark models have been as thoroughly studied, and more complex scenarios are becoming increasingly interesting. This is the case with dark sectors.

In dark sectors, the very plausible situation that DM is made up of several species of particles is made explicit [60]. In the same way as the SM consist of matter particles and force carriers, with several symmetries and interactions within itself, a hypothetical dark sector has a larger particle content than just the DM particle itself. General-purpose collider experiments like ATLAS make it possible to study several aspects of such models.

Dark sector studies can be challenging, however. The signals are inevitably small, since they have not yet been discovered by the searches described above or precision measurements of SM processes. For the same reason, signal in common phase spaces with common event signatures, have largely been ruled out. This leaves very small parameter spaces still viable for these theories, unless we move to unconventional signatures or phase spaces.

1.4 Scientific aim

It has been shown that dark sectors could give rise to particles that can be produced in proton-proton collisions and then decay promptly back into SM states [136, 135], and that such a model can be constructed to be under almost no constraints from previous searches [136, 135, 70]. The new, dark sector is strongly coupled and confining, and the degrees of freedom at LHC energies are composite particles, dark mesons.

In this thesis, I present a targeted search in proton-proton collision data recorded by the ATLAS detector for dark mesons as described by the model presented in reference [135]. It contains the two lightest states of the full theory, dark pions and dark rhos, and the signal process chosen for the analysis is pair production of dark pions, either resonantly or through Drell-Yan production, where the dark pions subsequently decay into third generation quarks.

This is the first time the model has been probed at the LHC, and the goal of the analysis was to probe as large a parameter space as possible with a simple analysis. The main challenges was the wide spread in kinematic behavior of the signal dark pions over the parameters space probed – dark pion masses ranging from 200 GeV to 1200 GeV, and dark rho masses from two to 10 times heavier than the dark pion – and the modelling of the irreducible background in the analysis.

A secondary aim is the development of different techniques to help in similar final states. The first is an initial exploration of using machine learning in the dark meson analysis, by applying a boosted decision tree (BDT) to signal and background discrimination. The second involves the unsupervised particle clustering method UCluster, which can be trained with different objectives such as signal and background discrimination, anomaly detection or jet tag-
ging. I present a modification to the algorithm that allows for distributed running, making it scalable to any size of input data. In the third, the SparkDensityTree package, which is based on a fast and distributed density estimation, is applied to the problem of finding signal enriched regions in particle collision data. I present a proof-of-concept study where the method is applied in a setting similar to the dark meson analysis. Lastly, I present a study of using the Extrapolation Engine fast simulation package for pattern bank studies for a hardware track trigger.
Part I:
Experimental setup and theoretical background

Data-taking at the LHC has been done in three separate runs with complete shut-downs and upgrades in between. The dark meson analysis presented in part II was done on the full Run 2 proton-proton collision dataset. The experimental setup involves the CERN accelerator complex including the LHC, as well as the ATLAS detector, the ATLAS readout system, and the ATLAS software for reconstructing detected events, monte carlo simulations and data analysis. The signal hypothesis comes from the new, dark sector, while all other theoretical considerations come from the SM. Both the experimental setup of Run 2 and the theoretical foundation of the analysis are summarised in this chapter.
2. The Large Hadron Collider

The LHC is a two-ring circular hadron collider, with the beams in the two rings accelerated in opposite directions and made to collide at four interaction points along the path. At each interaction point a particle detector is located, one of which is the ATLAS detector. The other three are the B-physics experiment LHCb, the heavy ion experiment ALICE and the other general-purpose experiment CMS. A schematic of the LHC is shown in figure 2.1. While the LHC accelerates both protons and heavy ions, all data in this thesis come from proton-proton collisions.

The protons are isolated from hydrogen gas and accelerated through a series of accelerators before injection into the LHC: the linear accelerator Linac2, the Proton Synchrotron Booster (PSB), made up of four superimposed synchrotron rings, and two synchrotrons, Proton Synchrotron (PS) Super Proton Synchrotron (SPS) [10]. These are shown schematically in figure 2.2.

In this injector chain, the protons are arranged into bunches that are separated along the accelerator and make up the beams. Once injected into the LHC, the two beams are accelerated in opposite directions to their target energies, maintaining the bunch pattern. At the collision points, where the beam pipes intersect, one bunch from each beam is made to collide with the other. This is referred to as a bunch crossing, and is the basis for the event view of collision data, described in section 4. The beams are circulated and collided until the instantaneous luminosity decreases enough to motivate dumping the beams and injecting new ones.

2.1 The Run 2 conditions

The LHC centre-of-mass energy during Run 2 was \( \sqrt{s} = 13 \text{ TeV} \), and the total integrated luminosity delivered was 156 fb\(^{-1}\). The proton-bunch spacing was 25 ns.
Figure 2.1. Schematic of the LHC rings, consisting of 8 arcs and 8 insertion regions (IRs), 4 of which house the detectors (ATLAS, CMS, LHCb and ALICE) and 3 of which are used for beam monitoring and steering (collimation and RF). The figure also shows the beam injection points and the beam dump point. Figure adapted from references [20, 68].

Figure 2.2. The CERN accelerator complex, showing the collider chain used for LHC beam injection during Run 2. See text for details about the LHC injection chain. Figure adapted from reference [147].
3. The ATLAS detector

The ATLAS detector is a general-purpose particle collision detector, built around LHC interaction point 1. It has a near $4\pi$ solid angle coverage, and is forward-backward symmetrical with respect to the interaction point. It consists of several concentric subdetectors, arranged radially outward from the beam pipe, making up the barrel, and in two end-caps orthogonal to the beam pipe, as shown in figure 3.1. The subdetectors each supply specialist information that, when combined, can be used to identify particles and reconstruct the collision kinematics. Another central feature of ATLAS is its powerful magnetic field, which is used to bend the trajectories of charged particles.

3.1 The magnet system

Two types of superconducting magnets are used to create magnetic fields over the detector, as can be seen in figure 3.2. A solenoid surrounds the ID in the barrel region of the detector. It creates a 2T magnetic field along the $z$-direction, while minimally affecting the particles traversing the detector. The 8-coil central toroid around the barrel of the detector as well as the two endcap toroids create magnetic field of up to 3.5 T along the $\phi$-direction.

3.2 The inner detector

Of the subdetectors, the inner detector (ID) lies closest to the beampipe. It consists of three different tracking technologies. Radially outward from the beampipe are the Pixel detector, the Semiconductor Tracker (SCT) and the Transition Radiation Tracker (TRT). The former two are both made up of silicon modules that create electrical pulses when hit by electrically charged particles, referred to in the following as hits. The TRT consists of straw tubes filled with gas, which can be ionised by charged particles, thus creating an electrical signal or hit.

The modules and how they are arranged in the barrel can be seen in figure 3.3 which shows a closer look at a part of the ID. The hits made by a particle in subsequent layers together map out its trajectory through the ID. These tracks are used to measure the sign of the charge and momentum of electrically charged particles, to provide lifetime information for particles that decay in the ID, and for particle identification.
Figure 3.1. The ATLAS detector during Run 2. Figure from reference [153].

Figure 3.2. The ATLAS detector with the magnet system highlighted in blue [106]. Closest to the beampipe is the solenoid magnet. The 8 coils making up the central toroid (only 6 shown for clarity) are placed symmetrically around the $\phi$ axis, and the two endcap toroids symmetrical around the IP.
3.3 The calorimeters

After the ID, the particles come to the calorimeters, arranged as shown in figure 3.1. These are designed to completely stop the incoming particles and, in doing so, measure their energies.

They are all sampling calorimeters, consisting of an active material and a passive material. The passive material is hit by the incoming particles, which initiates a shower of secondary particles. The active material measures the energy of the showers, and the way the calorimeter is segmented provides position information. Closest to the beam pipe, the active material used is liquid argon (LAr). This includes the electromagnetic (EM) calorimeter, the forward calorimeter and the hadronic calorimeter end-caps. The passive material in these is lead, copper (EM) and tungsten (hadronic), and copper, respectively. Radially outward from these is the central hadronic calorimeter, a tile calorimeter with scintillating plastic as active material and steel as passive material.

The EM calorimeter completely stops electrons and photons, while hadrons only deposit part of their energy in the EM calorimeters and are subsequently completely stopped in the hadronic calorimeter.
3.4 The muon system

The muon system (MS) is the outermost part of the ATLAS detector, designed to give tracking information for muons. The precision tracking system consists of monitored drift tube (MDTs) ionisation chambers, reinforced by cathode strip chambers in high-flux regions. When a charged particle goes through the chambers, it generates an electrical signal which is read out as a hit.

The MS also contains a fast-response system, used primarily in the trigger system described below. This consists of resistive place chambers (RPCs) in the barrel and thin gap chambers (TGCs) in the end-caps, both of which detect charged particles through a voltage induced by their ionisation.

3.5 The luminosity detectors

ATLAS has two luminosity detectors, LUCID-2 [54] and the Beam Conditions Monitor (BCM) [87]. LUCID-2 has two stations located at ±17m from the IP, and BCM has two at ±1.84m. These provide information that makes it possible to determine the luminosity measured by ATLAS, which inevitably is lower than the one delivered by the LHC.

3.6 The ATLAS coordinate system

ATLAS uses a right-handed coordinate system with its origin at the nominal interaction point (IP) in the centre of the detector and the z-axis along the beam pipe. The x-axis points from the IP to the centre of the LHC ring, and the y-axis points upwards. Polar coordinates \((r, \phi)\) are used in the transverse plane, \(\phi\) being the azimuthal angle around the z-axis. The pseudorapidity is defined in terms of the polar angle \(\theta\) as \(\eta = -\ln \tan(\theta/2)\) and is equal to the rapidity \(y = \frac{1}{2} \ln \left( \frac{E+p_z}{E-p_z} \right)\) in the relativistic limit. Angular distance is measured in units of \(\Delta R \equiv \sqrt{(\Delta y)^2 + (\Delta \phi)^2}\). This coordinate system is shown in figure 3.4.

---

\(^1\)This description of the coordinate system is written by the ATLAS collaboration, and is used unchanged in every ATLAS publication for consistency and correctness.
Figure 3.4. The coordinate system used in ATLAS (left) and the relationship between \( \eta \) and \( \theta \) (right). Both figures from reference [157].
4. ATLAS datataking

The basis for ATLAS datataking is treating each bunch-crossing as an independent statistical event, and all information pertaining to the same event is treated as one data point. All subdetectors described in the previous chapter have associated readout systems and information from all of them are combined to form a complete picture of the event. Two things complicate this procedure. First, not all proton-proton collisions are interesting to the ATLAS physics programme and the rate of collisions is too high for reading out and storing each event. Second, since each proton bunch contains around one hundred million protons, multiple protons collide each bunch crossing, and products from one bunch-crossing linger until the next is recorded. This is referred to as pile-up.

The former is handled with the trigger system, in which specialised low-latency algorithms are run over detector signals in the event to decide whether to write it to disk or not. To address the latter, information about the pile-up is recorded with the luminosity detectors and taken into account in the data analysis.

4.1 ATLAS trigger and data acquisition system

The trigger system in ATLAS has two levels: the Level 1 (L1) trigger and the High Level Trigger (HLT). L1 is implemented in hardware and uses analogue sums of calorimeter signals as well as signals from the fast-response MS to make the first triggering decision. If an event is accepted, it is sent to the software-based HLT, which makes its decision based on rudimentary event-reconstruction algorithms. If accepted, the readout system sends all event data to a storage disk. The bunch-crossing rate, and L1 trigger input rate, is 40 MHz. The L1 acceptance rate, and HLT input rate, is 100 kHz. The HLT acceptance rate is 1 kHz, and this is the rate at which events are finally written to storage memory.

4.1.1 Trigger streams and menus

The HLT writes out different bytestreams of data for different purposes, called trigger streams. The one most relevant for proton-proton collision analysis is
the main physics stream. This is supported by calibration streams, for calibration and data monitoring, and the express stream for real-time monitoring of detector performance. The debug stream consists of events without a full trigger decision, due to failures in the HLT. In searches, the signal region selection is commonly applied to the debug stream to ascertain that the HLT does not disproportionally crash it the region. If it does, this could hint at a new signature which the current trigger algorithms are unable to work with.

Several trigger algorithms are run in the HLT. They are strung together in trigger chains, each developed with a specific physics goal in mind. The collection of trigger chains is called the trigger menu. Since interesting physics events happen at vastly different rates, some triggers are pre-scaled, meaning they save only a fraction of the triggered events, at random. This frees up space for collecting high statistics also in the less common processes. This is used to find potential systematic problems in the trigger, but also as a cross-check in analysis work. After read-out, the trigger streams are written into Raw Data Object files. These files contain all detector information associated with each event.

4.2 Pile-up

![Figure 4.1. Mean number of proton-proton interactions per bunch crossing during different years of Run 2 [47].](image)
Figure 4.1 shows the distribution of the mean number of interactions per bunch crossing for different years in Run 2, calculated from the per-bunch luminosity. The average ranges from 13.4 for 2015 to 36.1 for 2018. These numbers are used in simulation, to overlay simulated collisions of interest (hard-scattering events) with a realistic description of real data.

4.3 The Run 2 dataset

The Run 2 proton-proton dataset was recorded in 2015–2018 at the LHC Run 2 conditions, and corresponds to an integrated luminosity of 140 fb$^{-1}$. The luminosity was determined using the LUCID-2 detector and has an uncertainty of 1.2 fb$^{-1}$ [23].
5. Theoretical background

With the exception of the signal hypothesis, all theoretical considerations in this thesis come from the SM. Mathematically, the SM is a quantum field theory that describes all known particles and their interactions. As such, its basis is a Lagrangian functional of fields. From the Lagrangian, the particle content of the theory, as well as which interactions exist and how likely they are to take place, can be derived in a mathematical framework. As mentioned in the introduction, it has had remarkable success and is, in particular, able to describe almost everything that happens in proton-proton collisions at the LHC.

In the dark meson model, the SM Lagrangian has been extended by terms that describe the dark mesons, and their phenomenology at the LHC are derived from their interactions with SM fields.

5.1 The Standard Model

The particle content of the SM comprises all known elementary particles, which can be organised into a table as shown in figure 5.1. The leptons come in six flavours: the electron, the muon and the tau lepton, as well as the electron neutrino, the muon neutrino and the tau neutrino. The quarks also come in six flavours: down, up, strange, charm (c), bottom (b) and top. Both groups are arranged into three generations, each having the same charge and spin, but increasing mass. This organisation is shown in figure 5.1. The leptons and quarks are collectively called the fermions, derived from the fact that they follow Fermi-Dirac statistics. The other group of particles in the SM are the bosons, named after following Bose-Einstein statistics. These are the W- and Z-bosons, the photon, the gluon, and the Higgs boson. All particles in the SM have an associated anti-particle, denoted by a bar, as in $\bar{t}$ for the top anti-quark, or by the opposite charge, as in $e^+$ for the positron or anti-electron. In this thesis, this distinction is often omitted in reactions that are symmetrical with respect to an interchange between particle and anti-particle and inclusive selections. Similarly, the $W^\pm$ bosons will simply be referred to as the $W$ boson.

The SM further accounts for the kinematics and interactions of the particles. The three forces described by the SM – the strong interaction, the weak interaction and the electromagnetic interaction – are described as interactions
Figure 5.1. The particles of the SM organised by whether they are fermions or bosons. The fermions are further organised by their generations, by being quarks or leptons and by their electric charge. The bosons are organised by their spin and which particles they interact with, which is also indicated by coloured lines. Figure from [95].

with the photon, the W- and Z-bosons and the gluons, respectively, collectively known as the gauge bosons. Directly from the Lagrangian, Feynman rules [167] can be used to draw Feynman diagrams showing all possible interactions between particles. The quantities that are conserved in interactions are taken into account by symmetries of the Lagrangian, which by Noether's theorem translate to conservation laws in the interactions.

The SM has 19 free parameters that need to be determined by experiment, but once they are fixed, it can be used to make quantitative predictions about particle interactions. It is a probabilistic theory, reflecting the fact that particle interactions are inherently probabilistic.

In this thesis I will explicitly come back to two calculated quantities, the cross-section and the branching ratio.
The cross-section describes the probability of some interaction taking place. Given the cross section $\sigma$ that a particle is produced in a proton-proton collision, the rate of production at the LHC can be written as

$$\text{rate} = \sigma \mathcal{L}$$

where $\mathcal{L}$ is the instantaneous luminosity of the beams. Similarly, the number of times the particle is produced can be obtained by

$$N = \sigma \int \mathcal{L} \, dt$$

where the integral runs over the data-taking time interval. The production cross-section of a particle can be calculated by identifying all SM interactions that can produce the particle, and then summing the cross-sections for all individual contributions.

Similarly, all ways the particle can decay can be identified, and is referred to as a decay channel. The probability of each decay channel is identified with the branching ratio, which is defined as the ratio between the number of times the particle decays through a specific decay and the total decay. When a decay happens in multiple steps, the probabilities can be combined to obtain the branching ratio for the full decay chain.

Theoretical predictions that will not explicitly be discussed in this thesis include angular distributions, energy and momenta, and cross-sections for other types of interactions. They are however implicitly used in the Monte Carlo simulations described in chapter 5. The SM predictions cannot be exactly calculated, but instead rely on a series expansion approximation. In practice, the Feynman rules give a prescription for how to write down all contributing diagrams at each order in the expansion. Thus, a prediction at leading order (LO) includes all diagrams at leading order in the series approximation, a prediction at next-to-leading order (NLO) also includes the ones form the next-to-leading order, and so on for $N^k\text{NLO}$, $k = 1, 2, 3, \ldots$. They are all associated with integrals that can be calculated and summed to get the full probability, and the underlying assumption is that each order in the expansion is a perturbative correction to the sum at the previous order.

This perturbation theory approach works well for processes having to do with the electroweak (EW) sector in the SM. In the strongly interacting part, however, described by quantum chromodynamics (QCD) it works well only at high energies. This is because the strong force gets stronger at lower energies, contrary to the EW force. At sufficiently low energies, chiral perturbation theory gives a good description of the QCD sector by viewing hadrons, composite particles of quarks, as the degrees of freedom. The intermediate regime is tackled by different approximate or modelling computational techniques.

The number of integrals to solve grows with each order in the perturbation expansion, and it is an active field of research to calculate the predictions at increasing accuracy. At the same time, the free parameters of the SM are being
measured at increasing accuracy, leading to an overall more precise knowledge of the elementary particles.

### 5.2 Dark mesons

Reference [136] presents the phenomenological study that lead to the formulation of the dark meson model used in this thesis. In it, several dark sector theories are studied. They all contain a new, strongly coupled, confining sector containing fermions that transform under both the new dark sector and the EW sector of the SM. Similarly to the quarks of QCD, the constituents of the new dark sector form composite particles, dark mesons. When the theories are mapped onto a lower-energy effective theory, the dark mesons become the degrees of freedom.

The production and phenomenology of dark mesons at the LHC have been studied in detail [135]. The particle species relevant to LHC phenomenology were determined to be the two lowest-mass states, the dark pions and the dark rhos. These are direct analogues to the SM hadrons with the same names, at a higher energy. The sector also contains a stable dark baryon which is a viable candidate for dark matter.

The new dark sector does not interact at all with the SM QCD-sector. It does transform under and mix with the EW sector, however, which allows both for the production of dark rhos at the LHC and prompt decays of dark pions into SM states. Because of this, if dark mesons were realised in nature, they could already be produced at the LHC, and it would be possible to detect their decay-products directly.

The model provided [135] is the basis for the analysis signal hypothesis, but two choices still have to be made. First, two possibilities for gauging the underlying theory exist, with the categories referred to as SU(2)_L and SU(2)_R, following the original model. This affects the rate of production of the dark rhos, as can be seen in figure 5.2 which shows the production cross-section at \( \sqrt{s} = 13\, \text{TeV} \) of dark rhos as a function of their mass, \( m_{\rho_D} \).

The ratio between the masses is an important parameter in the signal modelling, and we introduce \( \eta = m_{\pi_D}/m_{\rho_D} \) where \( m_{\pi_D} \) is the mass of the dark pion. Once the decay of the dark rho to a pair of pions is kinematically open, i.e. when \( \eta < 0.5 \), it has a branching ratio of almost 1, as can be seen from figure 5.4. When \( \eta > 0.5 \), the dark rho decays directly back to SM states. In the latter case, dilepton resonant searches have ruled out dark pion masses up to 1150 GeV for SU(2)_L and 800 GeV for SU(2)_R. In the former, however, only weak limits exist on the model when \( \eta < 0.5 \) [135, 70]. The search for dark mesons is therefore made when this is the case, and the target is dark pion pair production.

Figure 5.5 shows the Feynman diagrams of the three processes that contribute to it at the LHC which are the basis for the signal hypothesis pre-
Figure 5.2. The production cross section of dark rhos at $\sqrt{s} = 13$ TeV, for both $SU(2)_L$ and $SU(2)_R$ models, as a function of the dark rho mass. From reference [135]

Figure 5.3. Branching ratios of the charged (top) and neutral (bottom) dark pions, for both the gaugephilic (left) and gaugephobic (right) models. From reference [135]
Figure 5.4. The branching ratio of dark rho decays to a dark pion pair, a quark-antiquark pair and a lepton-antilepton pair, shown both for $\eta < 0.5$, where the decay into two dark pions is kinematically allowed, and $\eta > 0.5$, where it is not. Figure from reference [135]

dictions. Two of them proceed through the production of a dark rho, which promptly decays into two dark pions. The third is pair-production through an intermediate gauge boson, in analogy to Drell-Yan production in the SM.

The second choice to be made affects the subsequent decay channels of the dark rho. Two possibilities are explored in reference [135]: the gaugephobic and the gaugephilic. As the names suggest, in the gaugephilic model, the dominant decay modes are to SM gauge bosons, and in the gaugephobic, to SM fermions. The most common ones can be seen in figure 5.3 for both the gaugephobic and gaugephilic models. Motivated by these branching fractions, the target for the analysis is the gaugephobic model, when charged dark pions decaying to $tb$ and neutral dark pions decaying to $tt$. It covers both the $SU(2)_L$ and $SU(2)_R$ models.

Once these choices have been fixed, the only free parameters of the model are $m_{\pi_D}$ and $\eta$. 
Figure 5.5. Feynman diagrams showing the production interactions of dark pions, through resonant dark rho production (top two) and Drell-Yan-type production (bottom). Figure by J.J. Heinrich.
6. Theoretical predictions in ATLAS

The data taken by ATLAS is recorded event-by-event, and with enough statistics, distributions of observables emerge. To show that a specific theory agrees with data, therefore, it must be shown that the theory gives rise to the same distributions. In ATLAS, this is done by simulating the entire chain from the proton-proton collision to the detector readout system. This is a task that requires some consideration.

Below, I list some of the subtleties in this approach, and then summarise how the theoretical predictions are obtained in ATLAS, and how they are compared to data.

First, the protons collided in LHC are not point-like particles, but made up of quarks and gluons, collectively called partons. In inelastic proton-proton collisions, the collision takes place between individual partons, referred to as a hard collision. The position of the partons within the proton is in itself modeled by a probabilistic function, the parton distribution function (PDF).

Second, while the cross-section for the hard interaction can be perturbatively calculated, the full collision includes strong-interaction processes at non-perturbative energies. These are the hadronisation of the partons after the collision, i.e. when they combine to form colourless hadrons, which in turn can decay to produce even more particles in a jet-like structure. The picture is complicated by QCD and electroweak radiation before and after the hard interaction. Before, the radiation is referred to as initial state radiation (ISR) and after it is referred to as final state radiation (FSR).

Third, partons in the protons may also scatter elastically, i.e. without undergoing any changes other than in their momenta. These partons may also take part in the hadronisation process.

Fourth, the recorded data correspond to particles that have not only been created in a proton-proton interaction, but have also traversed the detector. To compare the theoretical predictions to the data, the predictions therefore have to include the detector response to the predicted particles. In ATLAS, dedicated simulations of the detector are used for this purpose.

6.1 Monte Carlo simulations

Monte Carlo (MC) simulations are a versatile class of algorithms which rely on random sampling from given distributions to solve a problem. They are...
ideal for modelling the time-evolution of a probabilistic system and for solving high-dimensional integrals. This makes it possible to also simulate systems with several probabilistic interactions and combine the predictions. MC methods are used in all steps of the simulation chain within ATLAS, and have become almost synonymous with the simulations themselves.

### 6.1.1 MC event simulators

Several software packages exist to do the MC simulations needed for hadron colliders. These need to take into account the PDFs, hadronisation and showering, and correctly connect these effects to the hard-scattering interaction. The probability amplitudes for the hard-scattering event are given by matrix element\(^1\) (ME) generators. The hadronisation and showering cannot be perturbatively calculated but can be modeled. What remains is then to match these two descriptions to each other. Some common generators for the hadronisation, showering and matching are Herwig, Pythia and Sherpa. ME generators include MadGraph5\_aMC@NLO and PowHeg Box. The choices that can be made when running an event generator includes the choice of PDFs, parameters having to do with the matching and tuning parameters taken from data. In addition to the above, some processes also require the use of dedicated decay simulations, for example to conserve spin structures or for the decay of heavy flavour quarks.

### 6.1.2 The ATLAS detector simulations

In Run 2, two different detector simulations were used in ATLAS, the full ATLAS simulation (FullSim) and the (second version of the) fast ATLAS simulation (AFII). The full simulation is a complete description of the ATLAS detector geometry, implemented in GEANT4. This implementation involves mapping out the boundaries of the different materials in the detector. The particle transport through the detector is then simulated with yet another MC method. From the interactions of the particles with the detector material, it is possible to simulate the electrical signals in the detector, and from there treat the readout exactly the same as data.

The fast simulation is a simplification of the full simulation, in which the calorimeters have been replaced by a detector response function.

Finally, the effects of the readout system are accounted for, and the final product of the simulation chain is a sample in the exact same format as the detector data.

\(^1\)The matrix element refers to an element in the S-matrix, used to calculate scattering amplitudes, which squared gives the probability of obtaining the final state
6.2 ATLAS simulated samples

As is clear from the discussion above, simulating proton-proton collisions is a complicated process which requires a deep knowledge of both data and theoretical processes. In ATLAS, exactly which generators and parameters to use for which processes is studied by a dedicated group, the Physics Modeling Group (PMG), which makes recommendations to the rest of the collaboration. This is done for every SM process that might show up as background in any analysis. These samples are then produced centrally, meaning that there is an infrastructure in place for simulating each sample, and the samples are then available to the full collaboration. In the dark meson analysis we found no need to deviate from the PMG recommendations.

6.2.1 SM mismodelling

The main background processes that enter into the dark meson analysis come from top quark pair production, the theoretical prediction of which does not fully agree with data. This is in part a consequence of the perturbation calculations – the NLO predictions currently recommended by the PMG are not accurate enough. One known issue with the $t\bar{t}$ modelling is the overestimation of data in the tail of the top quark T-spectrum [164]. The agreement is approved in NNLO calculations, and though the full samples for this predictions are not yet available within ATLAS, the NLO samples can be reweighted to yield better agreement with data.

The situation becomes even worse when the top quark pair is produced in association with additional heavy flavour (HF) quarks. Modeling this includes modelling both radiation and hadronisation. The $t\bar{t} + $HF cross-section is overall underestimated by current MC samples in the analysis phase-space [48], and the relative fractions of different HF events is badly modelled. This needs to be taken into account when doing an analysis in this phase-space.

6.3 Statistical analysis

Once the theoretical predictions are simulated, they can be compared to the data with statistical hypothesis testing. This is commonly used in searches for new physics, where any deviation from the SM expectations would be interesting.

A signal region, enriched in signal events, is defined, and data in this region are tested against the background-only hypothesis, i.e. that the region contains only events that do not come from the signal process. In analyses that rely on MC simulations for the background estimation, it is the MC samples that represent the background-only hypothesis. It is therefore imperative that the simulations correctly describe background data. It is also important to take all systematic uncertainties into account.
A common method for this is to use a maximum profile log likelihood fit and a subsequent hypothesis test on a test-statistic based on the fit, as described in detail in [96]. The fitted parameters are the signal strength and nuisance parameters, which describe the systematic uncertainties. It can also be used to fit the background estimate to data, which requires the definition of control regions, which should be similar enough to the signal region that the modelling can be carried over, but depleted in signal events.

A test-statistic is defined as a function of fitted parameter values, and the hypothesis test is performed. If the background-only hypothesis cannot be rejected, the signal-plus-background hypothesis is tested against it, the hypothesis that the signal exists in the collision data and would show up as an addition to the background events in the signal region. If the signal-plus-background hypothesis can be rejected, we can say that the model does not describe data. In practice, the hypotheses tested often depend on some free parameters in the model, and multiple hypotheses are tested by varying these parameters. Rejecting the signal-plus-background hypothesis for a range of parameters is referred to as excluding the model with parameters within these ranges.

The profile log-likelihood method quickly becomes involved when applied to multiple signal regions and even histograms of sensitive variables, and several frameworks exist within ATLAS to apply it.
All data recorded by ATLAS are saved in their raw detector readout format, and stored indefinitely, so that no information is lost. The raw data objects coming directly from the bytestreams, however, contain only the direct detector information, i.e. which modules gave which signals in each event. These are combined with algorithms that encode the existing knowledge about the detector response to different particles, energies and decays. They are used to reconstruct the event, including which particles were created in the primary interaction, what their properties were and how they decayed. This information is written in the analysis object data (xAOD) format, which underlies most analyses in ATLAS. This processing is done in the same way for both data and MC samples.

7.1 Objects used in the reconstruction

From the raw detector hits described in section 3, tracks, vertices and clusters are created as a first step in the reconstruction of the event. These are later used in the specific particle reconstruction algorithms described below.

7.1.1 Vertexing

A vertex corresponds to a point in the beam-pipe or the detector in which an interaction or a decay has taken place. The primary vertex (PV) is the one associated with the hard-scattering interaction. Vertices are found from the tracks that originate at them, described below.

7.1.2 Tracking

When a charged particle goes through the ID, it leaves (ideally) one hit in each layer of modules, which together map out its trajectory through the subdetector, a track. The ATLAS magnetic field will bend the trajectory of any electrically charged particle, following a helical segment. A track is thus uniquely defined by five parameters: the polar and azimuthal angles $\theta$ and $\phi$, the impact parameters $d_0$ and $z_0$ and the ratio of the charge of the particle and its momentum $q/p$ [173]. The transverse impact parameter, $d_0$, is defined as the shortest distance between a track and the beam line in the transverse plane and the
longitudinal impact parameter, $z_0$, is defined as the distance in $z$ between the primary vertex and the point on the track used to evaluate $d_0$ [19]. The main challenge in tracking is to find out which hits belong to which track and to reconstruct it. To tackle this, both seeded track-finding algorithms, which start with some hit and then iteratively add more hits to it until a track is formed, and pattern matching algorithms, which use a pattern-bank that the patterns in the ID are compared with, are used [18]. Additionally, ambiguity arising from overlapping tracks or in general the dense environment, are solved algorithmically.

7.1.3 Calorimeter clusters
The calorimeters are designed to completely stop the particles that traverse them, and so the particles that interact with them deposit all their energy in them. Electron and photons deposit all their energy in the EM calorimeter while hadrons deposit some energy in the EM calorimeter and the rest in the hadronic calorimeter. The energy deposits are measured by several adjacent calorimeter cells. To capture the readout of all cells corresponding to the same energy deposit, dynamic, variable size clusters, superclusters, are built algorithmically, and their energy readouts are added.

7.2 Reconstruction and identification of particles
The analysis objects used in the analysis correspond to electrons, muons, neutrinos and quarks, where the latter two are reconstructed as $E_T^{\text{miss}}$ and jets, respectively. The algorithms used to reconstruct them are summarised below.

Two central figures of merit for reconstruction are efficiency, i.e. the fraction of true particles that are captured and reconstructed, and purity, i.e. the fraction of the reconstructed objects that correspond to true particles. To increase the purity of the reconstructed objects, background-suppression techniques can be applied. These are studied and optimised by dedicated combined performance groups in ATLAS, who make them available to the rest of the collaboration. How to optimise such techniques can differ between different analysis scenarios, and they are therefore maintained at different working points (WP), with different trade-offs on e.g. efficiency and purity. This is known as identification.

For some particles, isolation WPs are also maintained, requiring the particle to be more or less isolated from other objects in the same event.

The reconstruction algorithms are run independently of each other and might end up using the same detector hits. Before they objects are used in analysis, therefore, the overlap between them is removed.
7.2.1 Electrons

The algorithm for reconstructing electrons uses tracks in the ID and clusters in (primarily) the EM calorimeter [25]. An electron is defined as an object consisting of a cluster and a matched track.

A detailed energy calibration described in reference [21, 25] is applied to all candidates.

The charge of the particles can be inferred by the existence and bending of the tracks in the inner detector, and the bending of the track and detector response in the calorimeter can be used to measure momentum and energy.

A likelihood (LH) based discriminant built with several discriminating variables as described in reference [22, 25] is used for electron identification. An electron candidate is identified as an electron if it gives a discriminant value above a certain threshold, and the electron identification WPs are defined by different threshold values and different requirements on the tracks used for identification.

The identification WPs are described in detail in reference [25], where they are referred to as operating points. It also describes the identification WPs, which use a measure of the activity in a cone around the reconstructed object to determine how isolated it is. A characteristic signature of electrons from hard-scattering processes is that they are isolated, i.e., have little activity in the cone. The isolation is done either on the tracks or the calorimeter clusters.

7.2.2 Jets

Jets are objects designed to capture the detector response to the particles coming from the hadronisation of quarks and gluons. They are commonly used when there are quarks in the final state of the process of interest. The process going from the interaction through hadronisation to the detector response can be seen in figure 7.1. While jets can be created with different algorithms, only the anti-kT algorithm is relevant for this thesis. It defines a distance parameter between objects in the event that is inversely proportional to the user-supplied R-parameter. Once the jet-finding algorithm is seeded with an object, it iteratively adds the next closest object in the event to the jet, until it reaches the beam axis, at which point the algorithm is stopped and the jet is returned. The distance parameter is also inversely proportional to the square of the minimum transverse momentum of the two objects being considered. This creates a bias for high-momentum particles, valuable in the LHC environment where QCD effects create an abundance of low-energy hadronic objects.

The objects used in ATLAS in the anti-kT algorithm are topological clusters in the calorimeter. In the Particle Flow (PFlow) algorithm used in the dark meson analysis, they are first combined with tracks from the ID, giving additional information about energy and momentum [44]. This suppresses energy deposits originating from charged pileup particles, and allows for tak-
Figure 7.1. Illustration of how the hadronisation gives rise to jet signatures in the detector. Figure from reference [78].
ing momentum measurements from tracking information whenever the tracker resolution outperforms the calorimeter resolution.

At the end, soft objects that have been clustered into the jet are dropped from it with the soft-drop algorithm.

**Flavour tagging of jets**
The jet reconstruction in ATLAS makes no difference between hadronic jets originating from different particles, and they all end up in the jet collection. The jets can be still be identified as coming from a certain flavour of jet using auxiliary algorithms. These are called flavour tagging algorithms and a jet that passes the criteria is called a flavour-tagged jet, or more precisely a b-tagged jet, c-tagged jet or top-tagged jet. In the dark meson analysis, b-tagged jets play an important role. A characteristic signature of b-quarks in the detector is their slightly longer lifetime. Algorithms which consider track displacement and secondary and tertiary decay vertices within the volume of the reconstructed jet cone are therefore used to identify them. As with other identification algorithms, the flavour tagging is maintained at different WPs, with an efficiency/purity trade-off.

### 7.2.3 Muons

The muons are predominantly reconstructed using tracks in the ID and in the MS [52]. A reconstructed track in the MS is matched to a reconstructed track in the ID, and then a track reconstruction is done on the constituent hits of both simultaneously. This is supplemented with muons found in the MS that are compatible with coming from the interaction point. In the region where the MS does not have full acceptance, $|\eta| < 0.1$, it can be further supplemented by calorimeter information, and in cases where the muon does not traverse all layers of the MS, a track in the ID can be matched to only a segment of a track in the MS.

Muon identification is applied to suppress background, mainly from pion and kaon decays, and to select muons with high efficiency and ensure a robust momentum measurement. The WPs include different requirements on the track reconstruction parameters and the underlying hits, as well as the method of reconstruction of the muon candidates.

The isolation requirements are similar to the electron isolation requirements, with the WPs listed in reference [52].

### 7.2.4 Missing transverse energy

Missing transverse energy, $E_T^{\text{miss}}$, is an object that does not correspond directly to a specific particle. Rather, it is a measure of how much energy leaves the detector after the collision. It is defined as the negative sum of the transverse
components of the energy-momentum vector of all collision products. Before the collision, this transverse component of the energy-momentum vector is 0, so by conservation of energy and momentum, it can be concluded that the energy measured by $E_{T}^{\text{miss}}$ has been carried out of the detector by particles that have not been detected. Of the SM particles, the only one that would give rise to $E_{T}^{\text{miss}}$ is the neutrino, that does not interact with the detector.
Part II: 
Search for Dark Mesons

In the dark meson analysis that is the main subject of this thesis, I have analysed a phase-space inspired by the phenomenology of the model introduced in 5.2 for signs of new physics, i.e. any processes not already described by the SM.

As this is the first time such an analysis is performed on LHC data it is a simple analysis by design. Even though the final analysis strategy is simple, however, the considerations that went into defining it were not. The vastly different kinematic signatures of the model for different particle masses make it difficult to define a single region most sensitive to the model, and the high jet multiplicity phase space makes background modelling challenging.

The analysis was done simultaneously in two channels: the 1 lepton channel and the all-hadronic channel. While this thesis only describes the 1 lepton channel, a detailed account of the full analysis is available within ATLAS in reference [119].
8. The Dark Meson analysis

Previous limits have been set on the dark meson model through reinterpretation of other analyses, and are shown in figure 8.1.

![Figure 8.1. Existing limits from previous searches conducted in ATLAS and CMS as described in reference [70]. The white solid line corresponds to the 95\% CL exclusion contour, with excluded area above the line, and the white dashed line corresponds to the 68\% exclusion. The colored boxes indicate which search gave the sensitivity in the area, the details of which will not be repeated here.](image)

The current work is a dedicated search at different signal points, defined by the values of $\eta$ and $m_{\pi_D}$, covering dark pion masses up to 1200 GeV in SU(2)$_L$ and 700 GeV in SU(2)$_R$, and values below 0.5 in $\eta$. The signal model was set up in the ATLAS event generation system and validated before the signal samples were produced centrally.

As previously mentioned, this search targets the $tttb$ and $ttbb$ final states. Top quarks decay with a branching fraction consistent with 1 to a W boson and a b-quark [175]. The Feynman diagrams showing the decays can be seen in figure 8.2.

The W-boson can subsequently decay to either a quark-antiquark pair, or to a lepton and its corresponding neutrino. Quarks other than the top-quark hadronise before they decay and are identified as jets. Electrons and muons are directly identified as such, but tau leptons decay further.

Figure 8.3 shows the combined decay channels of the $tttb$ and $ttbb$ final states. In both cases, the channel with the largest branching ratio is the all-hadronic channel, i.e. when all top-quarks decay hadronically, to a quark-antiquark pair. The second largest comes from the 1-lepton channel, which combines the single electron and single muon channels. This corresponds to
the events in which exactly one of the top quarks decays leptonically, excluding tau decays. Tau decays are indirectly taken into account in the 1-lepton channel when it includes either of the other leptons.

Vastly different kinematic behaviour was seen as the two free parameters were varied, as well as between the resonant and nonresonant production modes. In resonant pair production, a small $\eta$-parameter will lead to high dark pion $p_T$, which in turn will give rise to collimated top- and b-quarks. Large dark pion masses will give high $p_T$ to top- and b-quarks. The non-resonant production becomes important only at higher dark pion masses and lower $\eta$ values, as seen in figure 8.4, and will not give rise to such a boost.

8.1 Already published results
The work in both the 1-lepton and the all-hadronic channel was done in collaboration and the same analysis framework was used. Furthermore, the analysis choices in both channels were made with a final combination of the results
in mind. The results from the all-hadronic channel have already been made public in reference [16].

Its signature includes 8 or 10 quarks, out of which at least 4 are b-tagged, and no leptons. The analysis in the all-hadronic final state explicitly reconstructs the dark pions by reclustering the R=0.4 anti-KT jets, using the same algorithm with R=1.2. The main challenge of the analysis is suppressing the QCD background, manifesting as soft jets. No excess over SM expectation was seen in data, and it was possible to exclude values of the dark pions with masses between 280 GeV and 522 GeV for \( \eta = 0.25 \). The full exclusion contour can be seen in figure 8.5, in which the lined area shows the parameter space excluded by the analysis, and the shaded area shows the parameter space excluded by reinterpretations from reference [135].

At the time of writing, the 1 lepton channel is going through ATLAS internal review, and the next step is publishing a combined paper for both analyses.

8.2 Overview of the method

The bulk of the analysis work is explorative, with the goal to understand the data and theoretical predictions well enough to construct a maximum likelihood fit that accounts for uncertainties and MC mismodelling without obscuring any unknown but real data features, as well as a statistical test that has the power to distinguish between the signal-plus-background and the background-only hypotheses for as many of the signal points as possible.

The entire procedure up until the final statistical fit is done without looking at data in any sensitive region, referred to as blinding. This is done by automatically applying a filter every time data is plotted, by which the datapoint is
set to 0 when the expected number of signal events in the bin exceeds 15% of the expected number of background events.

A preselection of events is then defined based on the 1-lepton signal signature. This determines the processes that make up the background – any SM process expected to give rise to events that would pass the preselection are included. The largest background component is $t\bar{t}$. In each step, the background is plotted together with the data in all relevant distributions, to assess to what degree the simulations describe the data. As the phase-space is subject to known mismodelling, some deviations were expected.

The signal simulations are used to find the region in which signal and background can be discriminated. This required the definition of additional discriminating variables, based on the kinematic signature of the signal.

Once a signal region was defined based on the study, the mismodelling of the background became more pronounced. In order to take this into account in the final fit, control and validation regions were defined. The $t\bar{t}$ background was also iteratively reweighted to match higher-order theoretical predictions.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure85.png}
\caption{Region in the $\eta-m_{\pi_0}$-plane for $SU(2)_L$ models excluded by the all-hadronic channel.}
\end{figure}
The statistical fit is made to the spectrum of a sensitive variable in the signal and control regions, further subdivided depending on their jet and b-jet multiplicities. The fit includes all uncertainties.

No excess of events over the background expectation is observed, and the final result is given as upper limits on the production cross-section, which are also translated to exclusion limits of the model parameter space.

This process is incremental, iterative and non-linear, and, for clarity, will not be chronologically described in what follows. Instead, the information is divided by topic for each of which I describe the main considerations and the reader is advised that they all depend on results from each other.
9. Data and Simulated Samples

9.1 Data Sample
The analysis is based on the full Run 2 proton–proton collision dataset. The data are collected in the main physics stream and processed following the procedure outlined in chapter 7.

9.2 Monte Carlo Samples
Both the signal and background samples are generated and processed as outlined in sections 5 and 7 and the specific samples used are described below, following recommendations from PMG. All events are passed through either the full or fast ATLAS detector simulation.

9.2.1 Signal samples
Signal samples are produced for points on a regular grid in the $\eta$-$m_{\pi^0}$ plane, as shown in figure 9.1. They were simulated in MADGRAPH5_aMC@NLO v2.4.3 [32], where the matrix elements are calculated at NLO, interfaced with PYTHIA8.212 [166] for hadronisation and showering or MADGRAPH5_aMC@NLO v2.9.9 interfaced with PYTHIA8.306 for hadronisation and showering. The bulk of the studies have been made with the original signal grid, shown in black in figure 9.1, but the final statistical analysis also includes new samples, shown in purple, which were added to increase statistics in the all-hadronic channel.

9.2.2 Background Monte Carlo Samples
The final states contributing to the background in the analysis, in order of decreasing importance, are the production of $t\bar{t}$ pairs, single top quarks, W or Z bosons in association with jets, $t\bar{t}$ in association with other particles, and multiple bosons. The processes making up each one of them are detailed below, as well as the generators and generator settings used. A summary can be found in table 9.1.

The production of $t\bar{t}$ events is modelled using the POWHEGBOX v2 [116, 150, 115, 31] generator which provides matrix elements at next-to-leading order (NLO) in the strong coupling constant $\alpha_S$ with
Figure 9.1. Distribution of simulated samples in $\eta-m_{\pi_D}$ space. The left grid corresponds to SU(2)$_L$, the right to SU(2)$_R$ models. The black markers indicate the original signal grid, while the purple markers show the extension of the SU(2)$_L$ signal grid for the all-hadronic channel.

The NNPDF3.0nlo [58] parton distribution function (PDF) and the $h_{\text{damp}}$ parameter, which controls the matching in POWHEG and effectively regulates the high-$p_T$ radiation against which the $t\bar{t}$ system recoils, set to 1.5 $m_{\text{top}}$ [42]. The functional form of the renormalisation and factorisation scale is set to the default scale $\sqrt{m_{\text{top}}^2 + p_T^2}$. The events are interfaced with PYTHIA8.230 [165] for the parton shower and hadronisation, using the A14 set of tuned parameters [37] and the NNPDF2.3lo set of PDFs [57]. The decays of bottom and charm hadrons are simulated using the EVTGEN v1.6.0 program [138].

The $t\bar{t}$ sample is normalised to the cross-section prediction at next-to-next-to-leading order (NNLO) in QCD including the resummation of next-to-next-to-leading logarithmic (NNLL) soft-gluon terms calculated using TOP++2.0 [62, 72, 59, 98, 99, 97, 100]. For proton–proton collisions at a centre-of-mass energy of $\sqrt{s} = 13$ TeV, this cross section corresponds to $\sigma(t\bar{t})_{\text{NNLO+NNLL}} = 832 \pm 51$ pb using a top-quark mass of $m_{\text{top}} = 172.5$ GeV. The uncertainties on the cross-section due to PDF and $\alpha_S$ are calculated using the PDF4LHC prescription [71] with the MSTW2008 68% CL NNLO [139, 140], CT10 NNLO [137, 118] and NNPDF2.3 5f FFN [57] PDF sets, and are added in quadrature to the scale uncertainty.

The uncertainty due to initial-state radiation (ISR) is estimated by comparing the nominal $t\bar{t}$ sample with additional samples [46]. To simulate higher parton radiation, the factorisation and renormalisation scales are varied by a factor of 0.5 while simultaneously increasing the $h_{\text{damp}}$ value to 3.0 $m_{\text{top}}$ and using the Var3c up variation from the A14 tune. For lower parton radiation, $\mu_r$ and $\mu_F$ are varied by a factor of two while keeping the $h_{\text{damp}}$ value to 1.5 $m_{\text{top}}$ and using the Var3c down variation in the parton shower. The Var3c A14 tune...
variation [37] largely corresponds to the variation of $\alpha_S$ for ISR in the A14 tune. The impact of final-state radiation (FSR) is evaluated by varying the renormalisation scale for emissions from the parton shower up and down by a factor of two.

The NNPDF3.0lo replicas are used to evaluate the PDF uncertainties for the nominal PDF. The central value of this PDF is further compared with the central values of the CT14nnlo [107] and MMHT2014nnlo [123] PDF sets.

Samples for $t\bar{t}$+HF processes were produced with the POWHEG BOX RES [133] generator and OPENLOOPS 1 [69, 80, 104], using a pre-release of the implementation of this process in POWHEG BOX RES provided by the authors [132], with the NNPDF3.0nlo [58] PDF set. It was interfaced with PYTHIA8.240 [165], using the A14 set of tuned parameters [37] and the NNPDF2.3lo PDF set. The four-flavour scheme was used with the $b$-quark mass set to 4.95 GeV.

The factorisation scale was set to $0.5 \times \sum_{i=t,\bar{t},b,\bar{b}} m_{T,i}$, the renormalisation scale was set to $4 \sqrt{m_T(t) \cdot m_T(\bar{t}) \cdot m_T(b) \cdot m_T(\bar{b})}$, and the $h_{damp}$ parameter was set to $0.5 \times \sum_{i=t,\bar{t},b,\bar{b}} m_{T,i}$.

The production of $W/Z + \text{jets}$ is simulated with the SHERPA v2.2.11 [65] generator using next-to-leading order (NLO) matrix elements (ME) for up to two partons, and leading order (LO) matrix elements for up to five partons calculated with the Comix [120] and OPENLOOPS 1 [69, 80, 104] libraries. They are matched with the SHERPA parton shower [162] using the MEPS@NLO prescription [128, 129, 81, 130] using the set of tuned parameters developed by the SHERPA authors. The HESSIAN NNPDF3.0nnlo set of PDFs [58, 79] is used and the samples are normalised to a next-to-next-to-leading order (NNLO) prediction [33].

The associated production of top quarks with $W$ bosons ($tW$) is modelled using the POWHEGBOX v2 [156, 150, 115, 31] generator at NLO in QCD using the five-flavour scheme and the NNPDF3.0nlo set of PDFs [58]. The diagram removal scheme [114] is used to remove interference and overlap with $t\bar{t}$ production. The related uncertainty is estimated by comparing with an alternative sample generated using the diagram subtraction scheme [114, 42]. The events are interfaced to PYTHIA8.230 [165] using the A14 tune [37] and the NNPDF2.3lo set of PDFs [57]. Single-top $t$-channel production is modelled using the POWHEGBOX v2 [112, 150, 115, 31] generator at NLO in QCD using the four-flavour scheme and the corresponding NNPDF3.0nlo set of PDFs [58]. The events are interfaced with PYTHIA8.230 [165] using the A14 tune [37] and the NNPDF2.3lo set of PDFs [57]. Single-top $s$-channel production is modelled using the POWHEGBOX v2 [30, 150, 115, 31] generator at NLO in QCD in the five-flavour scheme with the NNPDF3.0nlo [58] parton distribution function (PDF) set. The events are interfaced with PYTHIA8.230 [165] using the A14 tune [37] and the NNPDF2.3lo PDF set.
The production of $t\bar{t}t\bar{t}$ events is modelled using the MADGRAPH5_aMC@NLO v2.4.3 [32] generator which provides matrix elements at next-to-leading order (NLO) in the strong coupling constant $\alpha_S$ with the NNPDF3.1nlo [58] parton distribution function (PDF). The functional form of the renormalisation and factorisation scales are set to $0.25 \times \sum_i \sqrt{m_i^2 + p_{T,i}^2}$, where the sum runs over all the particles generated from the matrix element calculation, following the Ref. [111]. Top quarks are decayed at LO using MADSPIN [113, 35] to preserve all spin correlations. The events are interfaced with PYTHIA 8.230 [165] for the parton shower and hadronisation, using the A14 set of tuned parameters [37] and the NNPDF2.3lo [58] PDF set. The decays of bottom and charm hadrons are simulated using the EVTGEN v1.6.0 program [138].

The production of $t\bar{t} + W/Z$ events is modelled using the MADGRAPH5_aMC@NLO v2.3.3 [32] generator at NLO with the NNPDF3.0nlo [58] parton distribution function (PDF). The events are interfaced to PYTHIA8.210 [165] using the A14 tune [37] and the NNPDF2.3lo [58] PDF set. The decays of bottom and charm hadrons are performed using the EVTGEN v1.2.0 program [138].

The production of $t\bar{t}H$ events is modelled using the POWHEG BOX v2 [116, 150, 115, 31, 125] generator at NLO with the NNPDF3.0nlo [58] PDF set. The events are interfaced to PYTHIA8.230 [165] using the A14 tune [37] and the NNPDF2.3lo [58] PDF set. The decays of bottom and charm hadrons are performed by EVTGEN v1.6.0 [138].

The further rare backgrounds $ttt$, $t\bar{t}ZZ$, $t\bar{t}WW$, $t\bar{t}WZ$, $t\bar{t}WH$ and $t\bar{t}HH$ are all produced using the LO MADGRAPH5_aMC@NLO generator interfaced with PYTHIA8 using the A14 set of tuned parameters and scaled to NLO cross sections [103].

Samples of diboson final states ($VV$) are simulated with the SHERPA v2.2.1 or v2.2.2 [65] generator depending on the process, including off-shell effects and Higgs-boson contributions, where appropriate. Semileptonic final states, where one boson decays leptonically and the other hadronically, are generated using matrix elements at NLO accuracy in QCD for up to one additional parton and at LO accuracy for up to three additional parton emissions. Samples for the loop-induced processes $gg \rightarrow VV$ are generated using LO-accurate matrix elements for up to one additional parton emission. The matrix element calculations are matched and merged with the SHERPA parton shower based on Catani-Seymour dipole factorisation [120, 162] using the MEPS@NLO prescription [128, 129, 81, 130]. The virtual QCD correction are provided by the OPENLOOPS1 library [69, 80, 104]. The NNPDF3.0nlo set of PDFs is used [58], along with the dedicated set of tuned parton-shower parameters developed by the SHERPA authors.
9.2.3 Pileup Reweighting

To mimic the pileup conditions during data taking, all simulated samples have been overlaid with simulated minimum-bias collisions, where the number of additional collisions approximate the pileup distributions observed in data. The minimum-bias overlay events are generated with PYTHIA 8.186 [166] using the NNPDF2.3lo set of parton distribution functions (PDF) [56] and the A3 set of tuned parameters [11]. The simulated pile-up profile is compared to actual data, and scale factors to increase the agreement are added as event weights to the MC samples. The $\mu$ profile in data is scaled by $1/1.03$ before weight calculation, following ATLAS recommendations.

9.2.4 $t\bar{t}$+HF overlap removal

As described in section 9.2.2, a dedicated sample in which a pair of top quarks is produced in association with two $b$-quarks can be used to supplement statistics in the phase space relevant for this analysis. Since the same diagrams are also contained within the inclusive $t\bar{t}$ samples an event overlap removal procedure is needed to avoid an overestimation of $t\bar{t}$+HF events. For this purpose, an event classification strategy analogous to [40] has been implemented. The flavours of additional jets in $t\bar{t}$ events are identified from generator-level particle jets reconstructed with the anti-$k_T$ algorithm using a radius parameter of $R = 0.4$ and are required to be central in the detector where $|\eta| < 2.5$ and to have $p_T > 15$ GeV. Truth $b$-jets are identified and categorised by counting the number of final $b$-quarks within the cone of the jet that do not arise from the decay of top quarks, $W/Z$, or Higgs bosons. A jet matched to exactly one $b$-quark with $p_T > 5$ GeV is labelled a $b$-jet. Jets matched to more than one $b$-quark where at least one of them satisfies the $p_T$ requirement of 5 GeV are labelled $B$-jets. Similarly we can determine $c$-jets from additional $c$-quarks in the event. For $c$-jets we do not distinguish whether they are matched to more than one $c$-quark.

Events can now be categorised based on the number of $b$-, $B$- and $c$-jets. If there is exactly one $b$-jet ($B$-jet) in the event, it is categorised as $t\bar{t} + 1b$ ($t\bar{t} + 1B$). If the sum of $b$- and $B$-jets is exactly two, the event is categorised as $t\bar{t} + 2b$. Events containing more than three $b$- or $B$-jets are labelled $t\bar{t} + \geq 3b$. For events without any additional $b$- or $B$-jets two different categories are created. If there is at least one $c$-jet in the event, it is classified as $t\bar{t} + \geq 1c$. All other events, i.e. events without any additional $b$-, $B$- or $c$-jets, are categorised as $t\bar{t} + \text{light}$.

The dedicated $t\bar{t}bb$ sample is used for all events categorised as other than $t\bar{t} + \text{light}$. This means all $t\bar{t} + \text{light}$ events are rejected from this sample, while similarly all events classified as $t\bar{t} + 1b$, $t\bar{t} + 1B$ or $t\bar{t} + \geq 2b$ are vetoed from the inclusive $t\bar{t}$ samples.
Table 9.1. Summary of the event generation configuration for all nominal background samples used in the analysis.

<table>
<thead>
<tr>
<th>Process</th>
<th>Generator</th>
<th>PDF</th>
<th>Showering</th>
<th>Tune</th>
<th>Cross section</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t\bar{t}$</td>
<td>POWHEG BOX v2</td>
<td>NNPDF3.0nlo</td>
<td>PYTHIA8</td>
<td>A14</td>
<td>NNLO+NNLL</td>
</tr>
<tr>
<td>$t\bar{t}$+HF</td>
<td>POWHEG BOX RES</td>
<td>NNPDF3.0nlo</td>
<td>PYTHIA8</td>
<td>A14</td>
<td>NNLO</td>
</tr>
<tr>
<td>$W/Z + \text{jets}$</td>
<td>SHERPA v2.2.11</td>
<td>NNPDF3.0nn1o</td>
<td>SHERPA Def.</td>
<td>NLO</td>
<td></td>
</tr>
<tr>
<td>Single top</td>
<td>POWHEG BOX v2</td>
<td>NNPDF3.0nlo</td>
<td>PYTHIA8</td>
<td>A14</td>
<td>NLO+NNLL</td>
</tr>
<tr>
<td>$t\bar{t}t\bar{t}$</td>
<td>MADGRAPH5_aMC@NLO v2.4.3</td>
<td>NNPDF3.1nn1o</td>
<td>PYTHIA8</td>
<td>A14</td>
<td>NLO</td>
</tr>
<tr>
<td>$t\bar{t}$+$W/Z$</td>
<td>MADGRAPH5_aMC@NLO v2.3.3</td>
<td>NNPDF3.0nlo</td>
<td>PYTHIA8</td>
<td>A14</td>
<td>NLO</td>
</tr>
<tr>
<td>$t\bar{t}H$</td>
<td>POWHEG BOX v2</td>
<td>NNPDF3.0nlo</td>
<td>PYTHIA8</td>
<td>A14</td>
<td>NLO</td>
</tr>
<tr>
<td>Other $t\bar{t} + X$</td>
<td>MADGRAPH5_aMC@NLO v2.3.3</td>
<td>NNPDF2.3nn1o</td>
<td>PYTHIA8</td>
<td>A14</td>
<td>NLO</td>
</tr>
<tr>
<td>Multiboson</td>
<td>SHERPA v2.2.1/v2.2.2</td>
<td>NNPDF3.0nn1o</td>
<td>SHERPA Def.</td>
<td>NLO</td>
<td></td>
</tr>
</tbody>
</table>
10. Analysis objects

10.1 Electrons

Electrons are identified using the LH electron identification algorithm. They are further required to fulfil the ATLAS standard electron track to vertex association requirements and to have $|\eta| < 1.37$ or $1.52 < |\eta| < 2.47$, corresponding to the central region of the detector excluding the transition region between barrel and endcap regions, where the reconstruction efficiency is low [38]. Signal electrons are identified using the Tight working point as it contains variables that provide good rejection of heavy-flavour jets and photon conversions. The isolation WP used is the fixed tight, track only WP, which corresponds to a transverse momentum dependent variable-cone cut on the isolation of the electron track. These electrons are further required to have $p_T > 28$ GeV. The electrons that fulfil these criteria, summarised in table 10.1, are referred to as tight.

Table 10.1. Selection criteria for tight electrons, used in the 1-lepton channel for selecting the signal lepton within each event.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Criterion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pseudorapidity range</td>
<td>$</td>
</tr>
<tr>
<td>Energy calibration</td>
<td>es2018_R21_v0 (ESModel)</td>
</tr>
<tr>
<td>Transverse momentum</td>
<td>$p_T &gt; 28$ GeV</td>
</tr>
<tr>
<td>Track to vertex association</td>
<td>$</td>
</tr>
<tr>
<td></td>
<td>$</td>
</tr>
<tr>
<td>Identification</td>
<td>TightLH</td>
</tr>
<tr>
<td>Isolation</td>
<td>TightTrackOnly</td>
</tr>
</tbody>
</table>

In addition to the tight electrons, loose electrons are identified with the Medium working point and have no isolation requirement. They are used to veto events with any additional leptons. They are required to have $p_T > 10$ GeV. A summary of the loose electron requirements can be found in table 10.2.

To account for working point reconstruction and selection efficiency differences between Monte Carlo and data, scale factors are applied as weights in the Monte Carlo events.
Table 10.2. Selection criteria for loose electrons used for vetoing events with more than one lepton in the 1-lepton channel.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Criterion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pseudorapidity range</td>
<td>$(</td>
</tr>
<tr>
<td>Energy calibration</td>
<td>\texttt{es2018_R21_v0} (ESModel)</td>
</tr>
<tr>
<td>Transverse momentum</td>
<td>$p_T &gt; 10$ GeV</td>
</tr>
</tbody>
</table>
| Track to vertex association | $|d^B_0/(\sigma)| < 5 $  \\
|                          | $|\Delta z^B_0 \sin(\theta)| < 0.5$ mm                              |
| Identification           | MediumLH (1-lepton), LooseAndBLayerLH (all-hadronic)                     |
| Isolation                | None (1-lepton), FCLoose (all-hadronic)                                  |

10.2 Muons

All muons in this analysis are identified with the Medium WP. They are further required to be found within the acceptance region of the ID, $|\eta| < 2.5$, and to satisfy the ATLAS recommended vertex association criteria, $|d^B_0/(\sigma(d^B_0))| < 3$ and $\Delta z^B_0 \sin(\theta) < 0.5$ mm, as defined in [52].

Signal muons are required to have $p_T > 28$ GeV and to fulfil the TightTrack-Only isolation WP with varying cone radius, which is similar to the isolation definition used for electrons. The muons defined as described here are referred to as tight. A summary of the definition can be found in table 10.3.

Table 10.3. Selection criteria for tight muons, used to select the signal lepton in the event.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Criterion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Selection working point</td>
<td>Medium</td>
</tr>
<tr>
<td>Isolation working point</td>
<td>TightTrackOnly VarRad</td>
</tr>
<tr>
<td>$p_T$</td>
<td>$&gt; 28$ GeV</td>
</tr>
<tr>
<td>$</td>
<td>\eta</td>
</tr>
<tr>
<td>$</td>
<td>d^B_0/(\sigma(d^B_0))</td>
</tr>
<tr>
<td>$\Delta z^B_0 \sin(\theta)$</td>
<td>$&lt; 0.5$ mm</td>
</tr>
</tbody>
</table>

Loose muons have no isolation criteria. They have to satisfy $p_T > 10$ GeV. The full definition can be found in table 10.4. Loose muons are used for vetoing events with more than one lepton.

To account for muon reconstruction and working point selection efficiency differences between Monte Carlo and data, scale factors are applied to the weights of the Monte Carlo events.
### Table 10.4. Selection criteria for loose muons, used to veto events containing more than one lepton.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Criterion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Selection working point Medium</td>
<td></td>
</tr>
<tr>
<td>Isolation working point None</td>
<td></td>
</tr>
<tr>
<td>$p_T$</td>
<td>$&gt; 10 \text{GeV}$</td>
</tr>
<tr>
<td>$</td>
<td>\eta</td>
</tr>
<tr>
<td>$</td>
<td>d_0^{BL}/\sigma(d_0^{BL})</td>
</tr>
<tr>
<td>$\Delta z_0^{BL} \sin(\theta)$</td>
<td>$&lt; 0.5 \text{mm}$</td>
</tr>
</tbody>
</table>

### 10.3 Small-R jets

The jets used are anti-$k_T$ jets with $R = 0.4$, identified using the PFlow algorithm. The most up-to-date recommendations are applied to calibrate all jets [28]. After calibration all jets are required to have $p_T > 20 \text{GeV}$ and to satisfy $|\eta| < 2.5$. In order to minimise the effects of pileup, a jet vertex tagger (JVT) [41] is used to make sure matched inner detector tracks are consistent with the primary vertex. The tight working point is used.

### Table 10.5. Jet reconstruction criteria.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Criterion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Algorithm</td>
<td>anti-$k_T$</td>
</tr>
<tr>
<td>$R$-parameter</td>
<td>0.4</td>
</tr>
<tr>
<td>Input constituent</td>
<td>EMPFlow</td>
</tr>
<tr>
<td>Analysis release number</td>
<td>21.2.200</td>
</tr>
<tr>
<td>CalibArea tag</td>
<td>00-04-82</td>
</tr>
<tr>
<td>Calibration configuration</td>
<td>JES_MC16Recommendation_Consolidated_PFlow_Apr2019_Rel21.config</td>
</tr>
<tr>
<td>Calibration sequence (Data)</td>
<td>JetArea_Residual_EtaJES_GSC_Insitu</td>
</tr>
<tr>
<td>Calibration sequence (MC)</td>
<td>JetArea_Residual_EtaJES_GSC_Smear</td>
</tr>
</tbody>
</table>

### Selection requirements

<table>
<thead>
<tr>
<th>Observable</th>
<th>requirement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jet cleaning</td>
<td>LooseBad</td>
</tr>
<tr>
<td>BatMan cleaning</td>
<td>No</td>
</tr>
<tr>
<td>$p_T$</td>
<td>$&gt; 20 \text{GeV}$</td>
</tr>
<tr>
<td>$</td>
<td>\eta</td>
</tr>
</tbody>
</table>
10.4 Large-R Reclustered Jets

Large-R jets are reclustered from calibrated R=0.4 PFlow jets using the anti-$k_T$ algorithm with an R parameter of $R = 1.2$. In the all-hadronic channel, these jets reconstruct dark pion candidates, as shown in figure 10.1.

**Figure 10.1.** Jet reclustering for a range of $R$ parameters for an $SU(2)_L$ signal point with $\eta = 0.25$ and $m_{\pi_D} = 400$ GeV on the left and on the right for an $SU(2)_L$ sample with $\eta = 0.25$ and $m_{\pi_D} = 500$ GeV after preselection of the all-hadronic channel.

For the 1-lepton channel the signal lepton is added to the $R = 0.4$ jet collection before the reclustering, which then proceeds in the same way as for the all-hadronic channel. After reclustering, we call the large-R jet containing the lepton $J^{lep}$ and the leading fully hadronic large-R jet $J^{had}$. Note that both come from the same reclustering, ensuring there is no overlap between the two.

10.5 b-tagging

This analysis uses the DL1r algorithm based on an artificial deep neural network trained on a simulated hybrid sample composed of $t\bar{t}$ and $Z'$ events [49, 45] for b-tagging. The algorithm has a multidimensional output corresponding to the probabilities for a jet to originate from a $b$ quark, $c$ quark, or any light flavour. The flavour probabilities are then used to define a single threshold value on the $b$-jet probability. The ATLAS Flavour Tagging Group maintains various operating points to provide a specific $b$-jet tagging efficiency of 60%, 70%, 77% or 85% in a simulated $t\bar{t}$ sample. In this analysis we make use of the 77% working point, which has a rejection factor of 5 and 170 on charm and light-flavoured jets, respectively.
10.6 Overlap removal

The metric used in the analysis to evaluate whether objects overlap is defined as $\Delta R = \sqrt{\Delta y^2 + \Delta \phi^2}$, where $\Delta y$ is the rapidity distance between two objects. The sequence of removal of overlapping objects is summarised in table 10.6.

**Table 10.6. The object overlap removal procedure used in the analysis.**

<table>
<thead>
<tr>
<th>Reject</th>
<th>Against</th>
<th>Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>Electron</td>
<td>Electron</td>
<td>Shared track, higher-$p_T$ electron kept</td>
</tr>
<tr>
<td>Muon</td>
<td>Electron</td>
<td>Calo-tagged muon and shared ID track</td>
</tr>
<tr>
<td>Electron</td>
<td>Muon</td>
<td>Shared ID track</td>
</tr>
<tr>
<td>Photon</td>
<td>Electron</td>
<td>$\Delta R &lt; 0.4$</td>
</tr>
<tr>
<td>Photon</td>
<td>Muon</td>
<td>$\Delta R &lt; 0.4$</td>
</tr>
<tr>
<td>Jet</td>
<td>Electron</td>
<td>$\Delta R &lt; 0.2$</td>
</tr>
<tr>
<td>Electron</td>
<td>Jet</td>
<td>$\Delta R &lt; 0.4$</td>
</tr>
<tr>
<td>Jet</td>
<td>Muon</td>
<td>numTrack &lt; 3 and (ghost-associated OR $\Delta R &lt; 0.2$)</td>
</tr>
<tr>
<td>Muon</td>
<td>Jet</td>
<td>$\Delta R &lt; 0.4$</td>
</tr>
<tr>
<td>Photon</td>
<td>Jet</td>
<td>$\Delta R &lt; 0.4$</td>
</tr>
<tr>
<td>Fat Jet</td>
<td>Electron</td>
<td>$\Delta R &lt; 0.1$</td>
</tr>
<tr>
<td>Jet</td>
<td>Fat Jet</td>
<td>$\Delta R &lt; 1.0$</td>
</tr>
</tbody>
</table>
11. Event and Object Selection

11.1 Online Event Selection

Because of their sharp onsets and generally high efficiency to accept events containing above-threshold leptons single-lepton triggers, where the lepton is either an electron or a muon, are used to select events in data.

To retain maximum efficiency the analysis uses a logical OR between the lowest unprescaled triggers with a lepton isolation requirement and a higher lepton-$p_T$ trigger without a lepton isolation requirement. Since the thresholds for the lowest-unprescaled triggers have changed multiple times throughout the data-taking period, it uses exactly one configuration of isolated and non-isolated triggers per year, per ATLAS recommendations. Tables 11.2 and 11.1 give a full list of the single-lepton triggers used per data-taking period, as well as the corresponding integrated luminosities.

The efficiencies of single-lepton triggers for all signal samples are plotted in figure 11.1. They range from 20% to 50% for all samples of all models with lower values in $\eta$ boosting the efficiency due to the higher lepton $p_T$ in these topologies. This trend is only broken by the $\eta = 0.15$ efficiencies in the $SU(2)_L$ models which are slightly lower than the corresponding $\eta = 0.25$ mass points, due to the larger fraction of dark pions produced via Drell-Yan-type processes at these signal points.

Table 11.1. List of all lowest unprescaled isolated and non-isolated single-muon triggers alongside the data-taking periods and the corresponding integrated luminosity. A logical OR of isolated and non-isolated triggers is used in the analysis.

<table>
<thead>
<tr>
<th>Year</th>
<th>run numbers</th>
<th>isolated trigger</th>
<th>non-isolated trigger</th>
<th>Int. Lum. [fb$^{-1}$]</th>
</tr>
</thead>
<tbody>
<tr>
<td>2015</td>
<td>276262-284484</td>
<td>HLT_mu20_iloose_L1MU15</td>
<td>HLT_mu50</td>
<td>3.2</td>
</tr>
<tr>
<td>2016</td>
<td>297730-311481</td>
<td>HLT_mu26_iwarmdium</td>
<td>HLT_mu50</td>
<td>33.0</td>
</tr>
<tr>
<td>2017</td>
<td>325713-340453</td>
<td>HLT_mu26_iwarmdium</td>
<td>HLT_mu50</td>
<td>44.3</td>
</tr>
<tr>
<td>2018</td>
<td>349169-364292</td>
<td>HLT_mu26_iwarmdium</td>
<td>HLT_mu50</td>
<td>58.5</td>
</tr>
</tbody>
</table>
Table 11.2. List of all lowest unprescaled isolated and non-isolated single-electron triggers alongside the data-taking periods during which they were used and the corresponding integrated luminosity. A logical OR of isolated and non-isolated triggers is used in the analysis.

<table>
<thead>
<tr>
<th>Year</th>
<th>run numbers</th>
<th>isolated triggers</th>
<th>non-isolated triggers</th>
<th>Int. Lum. [fb⁻¹]</th>
</tr>
</thead>
<tbody>
<tr>
<td>2015</td>
<td>276262-284484</td>
<td>HLT_e24_lhmedium_L1EM20VH</td>
<td>HLT_e60_lhmedium</td>
<td>3.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>HLT_e120_lhloose</td>
<td></td>
</tr>
<tr>
<td>2016</td>
<td>297730-311481</td>
<td>HLT_e26_lhtight_nod0_ivarloose</td>
<td>HLT_e60_lhmedium_nod0</td>
<td>33.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>HLT_e140_lhloose_nod0</td>
<td></td>
</tr>
<tr>
<td>2017</td>
<td>325713-340453</td>
<td>HLT_e26_lhtight_nod0_ivarloose</td>
<td>HLT_e60_lhmedium_nod0</td>
<td>44.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>HLT_e140_lhloose_nod0</td>
<td></td>
</tr>
<tr>
<td>2018</td>
<td>349169-364292</td>
<td>HLT_e26_lhtight_nod0_ivarloose</td>
<td>HLT_e60_lhmedium_nod0</td>
<td>58.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>HLT_e140_lhloose_nod0</td>
<td></td>
</tr>
</tbody>
</table>
Further, events that fail the reconstruction of a primary vertex are rejected. To account for small differences in the single-lepton trigger behaviour between data and simulation, all triggered simulated events receive an additional event weight to match data. Only events recorded in the physics `Main` stream are considered for this search. No events from the debug stream pass the pre-selection.

![Figure 11.1. Single-lepton trigger efficiencies as function of dark pion mass and η parameter. All simulated SU(2)$_L$ signal points are plotted on the left, the SU(2)$_R$ efficiencies are shown on the right.](image)

### 11.2 Offline Event Selection

All data events are required to pass the standard good run list (GRL) selection, which specifies the recorded events deemed good for physics analysis by the Data Preparation Group [4].

Furthermore, a number of event-level vetos are applied to reject bad or corrupted events. This includes checks for LAr noise bursts, corruptions in the Tile calorimeter, events affected by the recovery procedure for single-event upsets in the SCT, and incomplete events due to a Timing, Trigger and Control restart. Jet cleaning and bad muon vetos, as recommended by respective CP groups, is also applied.

### 11.3 Discriminating Variables

The analysis uses a collection of discriminating variables, listed and described below:

- $H_T$ is the scalar sum of transverse momentum for jets in an event.
• \(N_{jets}\) is the number of jets in the event, reconstructed and identified as described in section 10.3
• \(N_{b-jets}\) is the number of jets identified as coming from a b-decay using a 77% efficiency working point, following section 10.5
• \(m_{jet,R=1.2}\) is the re-clustered anti-\(k_T\) \(R = 1.2\) jet mass
• \(\not{p}_T^{lep}\) and \(\not{p}_T^{had}\) are \(R = 1.2\) jets reclustered with the inclusion of a lepton as described in 10.4. \(\not{p}_T^{lep}\) is the leading reclustered jet containing a lepton, and \(\not{p}_T^{had}\) is the leading all hadronic reclustered jet.
• \(m_{bb,\Delta R_{min}}\) is the invariant mass of the two b-jets in the event that are closest to each other.
• \(\Delta R(l, b_2)\) is the delta R between the lepton and the second closest b-jet. These are further discussed in sections 11.4 and 12.

11.4 Object Selection

In addition to the trigger, good run list, jet/muon cleaning, and derivation requirements, a preselection based on the 1-lepton signature is applied. The final state comprises exactly one lepton (muon or electron), one neutrino, eight (\(tttb\)) or six (\(ttbb\)) quarks out of which four are b-quarks, and possibly an additional neutrino from the tau decay.

Based on this, events are preselected if they satisfy the following criteria: First they have to have exactly one tight and no additional loose leptons (electrons or muons), as defined in sections 10.1 and 10.2. Additionally, they are required to have at least four jets, as defined in section 10.3, out of which at least three also have to be identified as b-jets, following section 10.5.

Finally, the \(H_T\) of the event is required to be larger than 300 GeV to both increase the signal-to-background ratio and reduce the effects of mismodelling in the low end of the spectrum. The preselection is summarised in table 11.3.

<table>
<thead>
<tr>
<th>Table 11.3. Summary of the pre-selection used.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of leptons</td>
</tr>
<tr>
<td>Number of additional loose leptons</td>
</tr>
<tr>
<td>Number of jets</td>
</tr>
<tr>
<td>Number of b-tagged jets</td>
</tr>
<tr>
<td>(H_T)</td>
</tr>
</tbody>
</table>

Figure 11.2 shows the agreement between data and simulation after the pre-selection for six kinematic distributions: leading jet \(p_T\), \(m_T\), \(H_T\), and lepton \(p_T\). Given the known mismodelling of \(t\bar{t} + \geq 1b\) events, the ratio of \(t\bar{t}\) events with associated b-quarks to \(t\bar{t}\) events without is also shown.
Figure 11.2. Comparison between data and Monte Carlo prediction after preselection for 140 fb$^{-1}$ for: Leading jet $p_T$ (top left), $m_T$ (top right), $H_T$ (bottom left), lepton $p_T$ (bottom right). The red line represents the sum of predicted background yields, and six benchmark signal samples are shown with dashed lines. The bottom panel shows the ratio of data to Monte Carlo events in each bin with black dots, as well as the ratio of $t\bar{t}$ produced in association with heavy flavour quarks ($t\bar{t}+=>1b$) and $t\bar{t}$ without associated heavy flavour quarks ($tt+0b$) as a blue line.
After the preselection, three analysis regions are defined in terms of requirements on $N_{\text{jets}}$, $\Delta R(l,b_2)$, and $m_{bb\Delta R_{\text{min}}}$. The signal region (SR), the control region (CR), and the validation region (VR). These variables are shown after the preselection in figure 12.1. A detailed description of each of the analysis regions can be found in the following sections.

**Figure 12.1.** Distributions of the variables used to define the analysis regions: (Top left): $N_{\text{jets}}$; (Top right) $\Delta R(l,b_2)$; and (Bottom) $m_{bb\Delta R_{\text{min}}}$. The background samples are shown as stacked solid histograms and the signal samples as dashed lines. All variables are shown after the preselection described in section 11.4, for an integrated luminosity of 140 $\text{fb}^{-1}$.
12.1 Signal region

After the preselection described in 11.4 is applied, $3.195 \times 10^5$ background events remain for a luminosity of $140 \, \text{fb}^{-1}$. The expected signal yield at this point ranges from its largest value of $1.7 \times 10^4$ events for $SU(2)_L$ ($m_{\pi_D} = 300 \, \text{GeV}$, $\eta = 0.45$) through different orders of magnitude depending on the considered model, with a general decrease in expected yields with increasing dark pion masses and decreasing $\eta$, and with lower event yields generally expected from $SU(2)_R$ models than $SU(2)_L$ models. This includes a small set of signal points to which this channel will not be sensitive: e.g. $SU(2)_L$ with $m_{\pi_D} = 1000 \, \text{GeV}$, $\eta = 0.25$ with less than 2 events expected at the preselection step.

In order to further isolate the signal, the requirement $N_{\text{jets}} \geq 5$ is made, which reduces $t\bar{t}$ +light background by 25% and background in general by 21% while retaining between 77% and 98% of signal events for all signal points.

<table>
<thead>
<tr>
<th>Table 12.1. Expected event yields for a luminosity of $140 , \text{fb}^{-1}$, for background and a selection of signal samples, in the 1-lepton channel after each selection step requirement.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Process</td>
</tr>
<tr>
<td>$t\bar{t}$light</td>
</tr>
<tr>
<td>$t\bar{t}+b$</td>
</tr>
<tr>
<td>$t\bar{t}c$</td>
</tr>
<tr>
<td>Single top</td>
</tr>
<tr>
<td>$t\bar{t}$PlusX</td>
</tr>
<tr>
<td>$V + jets$</td>
</tr>
<tr>
<td>Multiboson</td>
</tr>
<tr>
<td>Total predicted background</td>
</tr>
</tbody>
</table>

Table 12.1. Expected event yields for a luminosity of $140 \, \text{fb}^{-1}$, for background and a selection of signal samples, in the 1-lepton channel after each selection step requirement.

After this, several potentially discriminating variables were defined, which take advantage of the kinematic differences between signal and background events. The full list of variables studied can be found in appendix B. To finalize the signal region, different combinations of these variables were studied. The expected significance of observing the signal plus background yields in the region, was compared between regions for all signal points simultaneously. The significance was estimated by $s/\sqrt{b}$, where $s$ is the number of expected signal events and $b$ is the number of expected background events. Throughout, we found the highest significances to be in the low dark pion mass, high $\eta$ region. Rather than optimising for a specific signal point, we maximised the number of signal points with high significances, while still keeping the number of signal events that pass the selection high enough to not limit the analysis sensitivity by the statistical uncertainty. This optimisation was done both across the different combinations of discriminating variables and across the values of the cuts in each variable.
The optimisation described above leads to the additional requirements
\[ \Delta R(l, b_2) < 2.7 \ \text{and} \ m_{bb\Delta R_{\text{min}}^2} > 100 \text{GeV} \] (defined in section 11.3), after which the predicted number of background events is reduced by 72\% w.r.p. the preselection. \( \Delta R(l, b_2) \) takes advantage of the fact that decay products of the dark pion that decays leptonically always include two b-quarks, one coming from the same top-quark as the lepton and the other coming either directly from the dark pion decay (tb decay) or from the other top quark (tt decay).

The distribution of this variable peaks at lower values than the background distribution for all signal points, and in general, signal points with lower \( \eta \) values and higher dark pion masses peak at lower values than the signals with relatively less boosted systems. This trend is only broken by the \( \eta = 0.15 \) points in which the dominating production mode are Drell-Yan-type processes.

\( m_{bb\Delta R_{\text{min}}} \) is especially effective for distinguishing high dark pion mass signal points, in which the two b-quarks closest to each other come from the same heavy dark pion, from background where the two b-quarks closest to each other come from less heavy resonances, such as the Z boson, or from two separate collision products or radiation. Normalised plots of the signal and background distributions of these two variables can be seen in figure 12.2 and the expected significance estimated by \( s/\sqrt{b} \) as well as the number of unweighted signal events in the SR for each of the original signal points can be seen in figure 12.3. The expected yields for 140 fb\(^{-1}\) for background and six benchmark signal points after each subsequent requirement is applied can be seen in table 12.1. The remaining signal yields, as well as the signal to background ratios, can be found in appendix A. The masses of the reclustered jets, described in section 11.3, were also promising variables considered for the signal region definition. In the end they were combined by taking the (scalar) sum between them, \( m_{\text{had}} + m_{\text{lep}} \). This variable has distributions that differ visibly between background and almost all signal points, but also differs a lot between different signal points. It was therefore chosen as the sensitive variable to make the binned-likelihood fit with in the final statistical analysis. It is also shown in figure 12.2.
Figure 12.2. Normalised distributions of \( \Delta R(l, b_2) \) (top, left) and \( m_{bb\Delta R_{\min}} \) (top, right) and the sensitive variable \( m_{\text{had}} + m_{\text{lep}} \) after the preselection and cut on number of jets have been made. The background samples are shown as stacked solid histograms and the signal samples as dashed lines. The bottom panel shows the ratio of data to Monte Carlo events in each bin with black dots, as well as the ratio of \( t\bar{t} \) produced in association with heavy flavour quarks \( (t\bar{t}+>1b) \) and \( t\bar{t} \) without associated heavy flavour quarks \( (t\bar{t}+0b) \) as a blue line.
Figure 12.3. The significances for each signal point in the $SU(2)_L$ (top, left) and $SU(2)_R$ (top, right) signal grids of observing the signal plus background yields predicted in the SR for luminosity $140 \text{ fb}^{-1}$ as estimated by $s/\sqrt{b}$. The number of simulated signal events (unweighted) that pass the SR selection for each signal point for $SU(2)_L$ (bottom, left) and $SU(2)_R$ (bottom, right).

Figure 12.4. The signal contamination as defined by the number of signal events, $s$, over the number of background events, $b$, for each signal point in the $SU(2)_L$ (left) and $SU(2)_R$ (right) signal grids in the CR.
12.2 Control region

As the final state consists of multiple top and bottom quarks, the analysis is dependent on a good estimate of the background from $t\bar{t}$ produced with additional heavy-flavour quarks ($t\bar{t}$ +HF). A control region dedicated to this was therefore defined by requiring events to satisfy preselection, $N_{\text{jets}} \geq 5$, $2.7 < \Delta R(l, b_2) < 3.5$ and $40 \text{GeV} < m_{bb_{\Delta R_{\text{min}}}} < 100 \text{GeV}$. The first two requirements ensure that the region is enriched in $t\bar{t}$ +HF similarly to the SR, while the latter two ensure that it is depleted of signal and orthogonal to the SR. The lower limit on $m_{bb_{\Delta R_{\text{min}}}}$ and upper limit on $\Delta R(l, b_2)$ are set to avoid the regions where the ratio between data and MC predictions goes from being a constant displacement into a slope as can be seen in figure 12.1.

The discrepancy between data and MC simulation in this region, which can be seen for several variables in figure 12.5, lies between 10 and 20% for most bins in most variables, and for some reach as high as 40%. This mismodelling is further discussed in section 13.

The signal contamination for this region can be seen in figure 12.4. The signal points with the highest signal-to-background ratio have already been excluded by reinterpretation (see figure 8.1). All other points have a signal-to-background ratio of less than 0.01.
Figure 12.5. Comparison between data and Monte Carlo prediction in the CR for 140 fb$^{-1}$ for: Leading jet $p_T$ (top left), $m_T$ (top right), $H_T$ (mid left), lepton $p_T$ (mid, right), $\Delta R(l, b_2)$ (bottom, left) and $m_{bb_{\text{min}}}$ (bottom, right). The red line represents the sum of predicted background yields, and six benchmark signal samples are shown with dashed lines. The bottom panel shows the ratio of data to Monte Carlo events in each bin with black dots, as well as the ratio of $t\bar{t}$ produced in association with heavy flavour quarks ($tt+=> 1b$) and $t\bar{t}$ without associated heavy flavour quarks ($tt+0b$) as a blue line.
s = 13 TeV, 140 fb⁻¹

Data / SM ratio

<table>
<thead>
<tr>
<th></th>
<th>tt+0b</th>
<th>tt+1b</th>
<th>tt+2b</th>
<th>tt+3b</th>
</tr>
</thead>
<tbody>
<tr>
<td>SM</td>
<td>0.005</td>
<td>0.010</td>
<td>0.015</td>
<td>0.020</td>
</tr>
<tr>
<td>Data</td>
<td>0.015</td>
<td>0.020</td>
<td>0.025</td>
<td>0.030</td>
</tr>
</tbody>
</table>

Number of events

<table>
<thead>
<tr>
<th></th>
<th>ttlight</th>
<th>ttge1b</th>
<th>tt1c</th>
<th>SingleTop</th>
<th>VJets</th>
<th>ttPlusX</th>
<th>Multiboson</th>
</tr>
</thead>
<tbody>
<tr>
<td>SM</td>
<td>25013.9</td>
<td>27395</td>
<td>210</td>
<td>310</td>
<td>410</td>
<td>510</td>
<td>610</td>
</tr>
<tr>
<td>Data</td>
<td>27494</td>
<td>27514</td>
<td>210</td>
<td>310</td>
<td>410</td>
<td>510</td>
<td>610</td>
</tr>
</tbody>
</table>

VR

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data</td>
<td>0.05</td>
<td>0.1</td>
<td>0.15</td>
<td>0.2</td>
<td>0.25</td>
<td>0.3</td>
<td>0.35</td>
<td>0.4</td>
<td>0.45</td>
<td>0.5</td>
<td>0.55</td>
</tr>
</tbody>
</table>

p_T [GeV]

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>100</th>
<th>200</th>
<th>300</th>
<th>400</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data</td>
<td>0.005</td>
<td>0.010</td>
<td>0.015</td>
<td>0.020</td>
<td>0.025</td>
</tr>
<tr>
<td>SM</td>
<td>0.005</td>
<td>0.010</td>
<td>0.015</td>
<td>0.020</td>
<td>0.025</td>
</tr>
</tbody>
</table>
Figure 12.6. Comparison between data and Monte Carlo prediction in the VR for 140 fb$^{-1}$ for: Leading jet $p_T$ (top left), $m_T$ (top right), $H_T$ (mid left), lepton $p_T$ (mid, right), $\Delta R(l, b_2)$ (bottom, left) and $m_{bb_{\text{min}}}$ (bottom, right). The red line represents the sum of predicted background yields, and six benchmark signal samples are shown with dashed lines. The bottom panel shows the ratio of data to Monte Carlo events in each bin with black dots, as well as the ratio of $t\bar{t}$ produced in association with heavy flavour quarks ($tt+=>1b$) and $t\bar{t}$ without associated heavy flavour quarks ($tt+0b$) as a blue line.
12.3 Validation region

To validate the corrections to the background estimate obtained in the final statistical fit, a validation region which shows the same mismodelling as the SR is used. It is defined by requiring events to satisfy preselection, $N_{\text{jets}} \geq 5$, $2.7 < \Delta R(l, b_2) < 3.5$ and $m_{b\bar{b}\Delta R_{\text{min}}} > 100\text{GeV}$. The requirement on $\Delta R(l, b_2)$ is inverted w.r.t. the SR, while the requirement on $m_{b\bar{b}\Delta R_{\text{min}}}$ is inverted w.r.t. the CR. This makes it both orthogonal to and similar to both the SR and CR. The data-to-MC agreement can be seen in figure 12.6.

12.4 Summary of analysis regions

The analysis regions thus defined are orthogonal to each other and similar in background composition, which is important for the conclusions drawn about the background in the CR to be applicable in the final statistical analysis, when the SR data are fitted alongside it. To further study the background composition in the regions, the distributions of the minimum $\Delta R$ between b-jets is shown for our three largest background samples in figure 12.7 in all analysis regions. This is the angle of the two b-jets used to calculate $m_{b\bar{b}\Delta R_{\text{min}}}$. The CR requirements results in a sample with b-jets that are closer to each other, while the SR and VR choose slightly larger $\Delta R$ values. Nevertheless, there is a large overlap between the distributions, and it is therefore not identified as a problem for the fit. This was confirmed by swapping the roles of VRs and CRs in a test run of the final fit, which did not change the results.

The CR and VR are depleted in signal and the SR is enriched. They are summarised in table 12.2.

<table>
<thead>
<tr>
<th>Region</th>
<th>Preselection</th>
<th>$N_{\text{jets}}$</th>
<th>$\Delta R(l, b_2)$</th>
<th>$m_{b\bar{b}\Delta R_{\text{min}}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>SR</td>
<td>See table 11.3</td>
<td>$\geq 5$</td>
<td>$&lt; 2.7$</td>
<td>$&gt; 100\text{GeV}$</td>
</tr>
<tr>
<td>CR</td>
<td>See table 11.3</td>
<td>$\geq 5$</td>
<td>$&gt; 2.7, &lt; 3.5$</td>
<td>$&gt; 40\text{GeV}, &lt; 100\text{GeV}$</td>
</tr>
<tr>
<td>VR</td>
<td>See table 11.3</td>
<td>$\geq 5$</td>
<td>$&gt; 2.7, &lt; 3.5$</td>
<td>$&gt; 100\text{GeV}$</td>
</tr>
</tbody>
</table>
Figure 12.7. Analysis region comparison of the minimum $\Delta R$ between b-jets shown in different $t\bar{t}bb$ (top row) and $t\bar{t}$ inclusive (bottom row) samples, with heavy flavour overlap removal already applied. The distributions are normalised to 1.
13. Background Estimation

The background in the 1-lepton channel is estimated primarily from MC simulations, using the samples described in section 9.2.2. The largest background contribution in the SR is from $t\bar{t}$, often produced in association with additional heavy quarks. Lesser backgrounds, in order of declining signal region yields, are single top production, $t\bar{t}$ produced in association with additional bosons or quarks, bosons produced in association with jets, and multiboson production.

13.1 $t\bar{t}$ +HF correction

Out of the $t\bar{t}$ events in the SR, one fourth comes from the dedicated $t\bar{t}b\bar{b}$ sample described in 9.2.2 and 9.2.4. The underestimation of $t\bar{t}$ +HF events described in section 6.2.1 is evident in all analysis regions, which are high in jet and b-jet multiplicity. It is addressed in the final statistical fit by fitting normalisation factors of different components of the $t\bar{t}$ +HF background independently of each other. This method was developed in Run-1 [40] and has been used also in more recent ATLAS analyses e.g. in reference [53]. The method has been adapted for use with our analysis strategy, and proceeds as follows. First, all $t\bar{t}$ MC events are classified into different categories depending on which (if any) additional quarks are present with the $t\bar{t}$ pair, using truth information. This is the same classification used for overlap removal between the inclusive $t\bar{t}$ sample and the dedicated $t\bar{t}b\bar{b}$ sample, described in section 9.2.4. Two normalisation factors are then defined: The first associated with the $t\bar{t}+ \geq 1b$ category, while the second one combines the $t\bar{t}+\text{ light}$ and $t\bar{t}+ \geq 1c$ categories. Both are included as a parameter in the final statistical fit. The control region was defined for this purpose, and is divided into six jet and b-jet multiplicity bins in the fit, where in each bin we have 5, 6 or $\geq 7$ jets and 3 or $\geq 4$ b-jets. The fractions of different $t\bar{t}$ +HF components in the different multiplicity bins of the CR, the SR and the VR can be seen in figure 13.1, figure 13.2 and figure 13.3, respectively. and a comparison of the sensitive variable $m_{3\text{had}} + m_{3\text{lep}}$ showing the HF content in each region is shown in figure 13.4. The procedure outlined here is used in the final statistical fit summarised in section 9.2.4.

13.2 Top $p_T$ correction

The top $p_T$ mismodelling has been addressed by iteratively reweighting top $p_T$, anti-top $p_T$ and $t\bar{t}$ mass to their NNLO QCD + NLO EW predictions us-
Figure 13.1. Percentages of the different HF categories in the different CR bins included in the fit.
Figure 13.2. Percentages of the different HF categories in the different SR bins included in the fit.
Figure 13.3. Percentages of the different HF categories in the different VR bins included in the fit.
Figure 13.4. The sensitive variable $m_{\text{had}} + m_{\text{lep}}$ shown in the SR (top, left), CR (top, right), and VR (bottom).
ing the 2D reweighting implemented in reference [5]. Figure 13.5 shows the distribution of several variables before and after reweighting.
Figure 13.5. Leading jet $p_T$, $E_T^{\text{miss}}$ and $H_T$ (left) and after (right) reweighting the $\bar{t}t$ samples to NNLO calculations.
14. Systematic Uncertainties

14.1 Summary of Systematics

There are a multitude of different sources of potential systematic variations to consider in this analysis. They have been split up into two groups. Uncertainties arising from instrumental origins are discussed in section 14.2, systematic uncertainties from theoretical considerations are detailed in 14.3.

For a better overview a list of all considered uncertainties is given in table 14.1.

14.2 Instrumental uncertainties

14.2.1 Luminosity uncertainty

The integrated luminosity of the full dataset used in this analysis was determined with the LUCID-2 detector [54] following the method detailed in [51]. The uncertainty on this value is determined to be 0.83% by using beam-separation scans in $x$ and $y$. The uncertainty value is applied to all simulated samples in the analysis.

14.2.2 Pileup reweighting uncertainty

As described in section 9.2.3, all simulated samples are reweighted to match the pileup profile observed in data. To estimate a systematic variation that might be introduced by this procedure the scale factor applied to the pileup distributions is varied from its nominal value of 1.0/1.03 to 1.0/0.99 to account for the up systematics and 1.0/1.07 for the down systematic.

14.2.3 Lepton uncertainty

Uncertainties on leptons arise from multiple sources having to do with identification, isolation, reconstruction, trigger as well as momentum scale and resolution.

Slight performance differences between data and simulation in lepton reconstruction, identification, isolation and triggering are corrected by the application of scale factors that have been estimated from tag-and-probe experiments in $Z \rightarrow l^+l^-$ events in data and simulation [50, 52]. The scale factors
Table 14.1. *Overview of all considered systematic uncertainties.*

<table>
<thead>
<tr>
<th>Source of uncertainty</th>
<th>Instrumental uncertainties</th>
<th>Theoretical uncertainties</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Luminosity</td>
<td>( t\bar{t} ) modelling</td>
</tr>
<tr>
<td></td>
<td>Pileup reweighting</td>
<td>( t\bar{t}b\bar{b} ) modelling</td>
</tr>
<tr>
<td></td>
<td><em>Leptons</em></td>
<td>single-top modelling</td>
</tr>
<tr>
<td></td>
<td>Electron scale factors</td>
<td>Flat uncertainty</td>
</tr>
<tr>
<td></td>
<td>Electron resolution and scale</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Muon scale factors</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Muon resolution and scale</td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>Jets</em></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Jet vertex tagging</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Flavour-tagging</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Jet energy scale</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Jet energy resolution</td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>Scale variations</em></td>
<td></td>
</tr>
<tr>
<td></td>
<td>PDF variations</td>
<td></td>
</tr>
<tr>
<td></td>
<td>( \alpha_5 ) variations</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ISR variations</td>
<td></td>
</tr>
<tr>
<td></td>
<td>FSR variations</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Parton shower</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Choice of ( h_{\text{damp}} )</td>
<td></td>
</tr>
<tr>
<td></td>
<td>NNLO reweighting</td>
<td></td>
</tr>
</tbody>
</table>

90
constitute a potential source of systematic variation. The uncertainties on the scale factors can be propagated to the analysis in form of a set of alternative event weights. For electrons this results in four individual variations. For muons eight components arise from ID, isolation, track-to-vertex association and trigger. Each of these four is in itself split up into a statistical and a systematic component.

In simulation the lepton momentum scale and resolution is corrected to match the distribution in data. This is a source of a potential systematic variation. To evaluate the impact of scale systematics the lepton energy or momentum is varied by $\pm \sigma$ and the signal selection redone. For resolution uncertainties the lepton energy or momentum is smeared and the signal selection redone. For electrons this results in three individual components while for muons four separate components exist.

14.2.4 Flavour-tagging uncertainty

Flavour-tagging efficiency differences between data and simulation are corrected by a reweighting of events. This introduces a source of potential systematic variations. Uncertainties on the corrections are derived from dedicated flavour-enriched subsets of the data which are propagated to the analysis in form of alternative sets of event weights. The flavour tagging uncertainties comprises of nine independent NPs for $b$-jets, four parameters for $c$-jets and four parameters from light jets. Additionally there are two more NPs to extrapolate the measured uncertainties to the high-$p_T$ region.

14.2.5 Jet Energy Scale and Resolution for small-R jets

The determination of jet energy scale (JES) and jet energy resolution (JER) is done by combining information from actual collision data, test beam data and simulation as described in [39, 43]. The Jet/EtMiss group provides different types of JES and JER systematics configurations where either all nuisance parameters (NP) are kept independently or whether some of them are grouped [6]. For JES a total of nearly 100 NPs exists that can be reduced by category to roughly 30, by a global reduction to about 20, or through a strong reduction to 6 or 7. The 1-lepton channel is expected to be somewhat sensitive to systematic variations of JES and JER, so the medium conservative category reduction scheme has been used.

JER has a total of 34 individual NPs that can be reduced to 13 in the FullJER scheme, where the smearing is done both (pseudo-)data and simulation, or 8 NPs in the SimpleJER scheme, where the smearing is done exclusively in simulation. The more complete FullJER scheme is used for the same reason stated in the paragraph above.
14.3 Theoretical Cross Sections

14.3.1 Top quark $p_T$ reweighting

The systematic uncertainties associated with the top quark $p_T$ reweighting described in section 13.2 are calculated using the same method used for the nominal reweighting factors. For details, see reference [155]. We use the maximum and minimum scale variations according to 7-point variations independently for the top/antitop $p_T$ and the $t\bar{t}$ mass. The variations are taken into account in the final statistical fit by including them as scale variations on the $t\bar{t}$ background. Furthermore, we use the nominal top quark $p_T$ reweighting to also correct the systematic variation samples described in section 14.3.2 to accurately estimate the modelling systematics of $t\bar{t}$ independently of this reweighting procedure.

14.3.2 $t\bar{t}$ uncertainty

The leading simulated background in the analysis stems from $t\bar{t}$. It is affected by a number of systematic uncertainties that impact both scale and shape of distributions. The $t\bar{t}$ theory systematics apply only on the $t\bar{t} + \text{light}$ and $t\bar{t} + \geq 1c$ background components as they are estimated from the bulk $t\bar{t}$ sample, while the $t\bar{t} + \geq 1b$ background has dedicated systematics which are detailed in section 14.3.3. All $t\bar{t}$ theory uncertainties are treated as completely uncorrelated between the $t\bar{t} + \text{light}$ and the $t\bar{t} + \geq 1c$ background components.

Missing higher order contributions in perturbative expansion of the $t\bar{t}$ production cross-section are estimated by adding in quadrature contributions from renormalisation and factorisation scale variations, which are obtained by independently varying the parameters $\mu_R$ and $\mu_F$ by a factor 0.5 and 2.0 and taking the envelope.

Uncertainties on the choice of the PDF set used for event simulation are estimated by using the PDF4LHC and NNPDF error sets following the typical PDF4LHC treatment [66] and taking the envelope. The uncertainty on the strong coupling constant $\alpha_S$ is evaluated by using the same PDF set with two different choices of $\alpha_S$. In the fit model both parameters are added in quadrature and treated as correlated parameters.

The initial state radiation (ISR) modelling, i.e. the amount of predicted ISR in an event by the parton showering algorithm can be estimated by making use of the built-in up and down variations of the Var3c parameter [7].

The amount of final state radiation (FSR) in an event is estimated by varying the factorisation scale by factors 0.5 and 2.0 inside PYTHIA8.

The impact of using a different parton shower and hadronisation model was evaluated by comparing the nominal $t\bar{t}$ sample with another event sample produced with the POWHEGBOX v2 [116, 150, 115, 31] generator using the NNPDF3.0nlo [58] parton distribution function (PDF). Events in the latter sample were interfaced with HERWIG 7.04 [55, 61], using the H7UE set of
tuned parameters [61] and the MMHT2014LO PDF set [123]. The decays of bottom and charm hadrons were simulated using the EVTGEN 1.6.0 program [138]. The matching uncertainty is evaluated by comparing the nominal sample with an alternative sample obtained setting the pthard PYTHIA8 parameter to 1 (the default is 0). This parameter regulates the definition of the vetoed region of the showering, important to avoid holes and overlaps in the phase space filled by POWHEG and PYTHIA8. This recommendation follows the description included in [131]. The alternative sample was produced using POWHEG interfaced with PYTHIA8.306 and EVTGEN V1.7.0 using the NNPDF2.3lo PDF set and the A14 set of tuned parameters.

The impact of a variation of the $h_{\text{damp}}$ parameter is assessed by comparing the nominal samples to an alternative set of samples for which $h_{\text{damp}}$ has been set to $3.0 \cdot m_{\text{top}}$.

14.3.3 $t\bar{t}b\bar{b}$ uncertainty

Theory uncertainties on the $t\bar{t}b\bar{b}$ sample only apply to the $t\bar{t} + \geq 1b$ component of the background as all other $t\bar{t}$ components are estimated from the bulk $t\bar{t}$ sample. Their dedicated uncertainties have been described in section 14.3.2.

The $t\bar{t}b\bar{b}$ sample is treated the same way as the bulk $t\bar{t}$ sample.

Just like for the $t\bar{t}$ background, the impact of using a different parton shower and hadronisation model was evaluated by comparing the nominal sample with another event sample produced with the POWHEGBOX v2 generator using the NNPDF3.0nlo parton distribution function. Events in the latter sample were interfaced with HERWIG 7.1, using the H7.1-Default set of tuned parameters and the MMHT2014LO PDF set [123]. The decays of bottom and charm hadrons were simulated using the EVTGEN 1.7.0 program.

The matching uncertainty is evaluated by comparing the nominal sample with an alternative sample obtained setting the pthard PYTHIA8 parameter to 1. The alternative sample was produced using POWHEG interfaced with PYTHIA8.307 and EVTGEN V1.7.0 using the NNPDF2.3lo PDF set and the A14 set of tuned parameters.

14.3.4 Single-Top uncertainty

The background arising from single-top production is a small background in all signal and control regions. Significant contributions of the order of 10% of the total background appear only in a few bins at large values of $m_{\text{had}} + m_{\text{lep}}$. A simplified approach to estimating the single-top uncertainty is therefore chosen. The single-top background is vastly dominated ($\sim 90\%$) by the $tW$-channel where the by far dominant systematic arises from the single-top interference scheme (diagram removal vs diagram subtraction). Other analyses working in similar phase spaces as the present search, e.g. the SM $t\bar{t}t\bar{t}$ analysis [36], have
conservatively estimated the total uncertainty on the single-top background at 30%. A flat uncertainty is therefore applied of the same 30% on the entire single-top background. In order to verify this estimation we have repeated the fit while setting the single-top uncertainty to 100% and found that the results were unchanged by the increase in systematic uncertainty.

14.3.5 Minor simulated backgrounds
As can be seen from tables 12.1 all other simulated backgrounds are negligible and we therefore assign no systematic uncertainty on them.
15. Results

A profile maximum likelihood fit was performed to the spectrum of $m_{\text{had}} + m_{\text{lep}}$ in the SR and CR, further subdivided into jet- and b-jet multiplicity bins, as shown in figure 15.1, where the 7 jet and 4 b-jet bins are upwardly inclusive. The fit was made both for $SU(2)_L$ and $SU(2)_R$ models, but showed no sensitivity to signal in $SU(2)_R$. All results presented in this chapter are therefore for the $SU(2)_L$ models.

The parameter of interest was the signal strength, and the nuisance parameters included both correction factors to the $t\bar{t} + \text{HF}$ background and the nuisance parameters for all systematic uncertainties. A summary plot of all fit regions is shown in figure 15.1, both pre- and post-fit. The signal strength parameter was consistent with 0, and the correction factors derived from the fit clearly increase the MC-to-data agreement.

In the hypothesis test we failed to reject the background-only hypothesis for all signal points. Figure 15.2 shows the postfit distributions of some kinematic variables, including the signal hypothesis. No excess in data over SM expectation can be seen.

Figure 15.2 shows the 95% CL limits set by data on dark pi0 production cross-section for the original $\eta$ signal points in the $SU(2)_L$ model. They are overlaid on the limits expected in the background-only scenario, and agree to within one or, less commonly, two standard deviations. The theoretical cross-section from the dark meson effective theory is also shown, and is larger than the limits in large intervals, excluding at 95% CL the possibility that dark pions from the specific signal points are realised in nature. The largest interval is in the $\eta = 0.45$ configuration, where dark pion masses up to 943 GeV are excluded.

Figure 15.3 shows the dark pion parameter space that has been excluded by comparing the theoretical cross section to the limits set by the analysis. It shows a large improvement with respect to reinterpretations, and even the dedicated analysis in the all-hadronic channel.
Figure 15.1. The number of data events in each bin (black dots) overlaid on the number expected from all components of the background distribution (solid, stacked histograms), as well as expected signal yields (solid lines). The bottom panel shows the ratio between the number of data events to the number of expected background events, as well as how that ratio would be expected to change were the signal process realised in nature. The top plot shows the situation before the maximum likelihood fit of signal strength, background normalisation factors and nuisance parameters, while the bottom one shows the situation after.
Figure 15.2. The 95% CL limits on the dark pion cross section obtained by the statistical analysis for the $SU(2)_L$ models. The limits observed from data are shown as black dots interpolated by black lines. The expected limits are shown as dashed lines and the green and yellow bands show the 1 and 2 standard deviations, respectively, from the expected limits. The theoretical cross-section is shown as a solid line.
Figure 15.2. Comparison of five distributions in the VR before and after the background-only fit. In each pair, the left shows the data (black dots) and expected background (stacked, solid histograms) before the fit and the right shows the same after the normalisation factors from the fit have been applied to the background estimate and the background has been constrained.
Figure 15.3. The region in the $m_{\pi_0}$ and $\eta$ place excluded by the limits set on the cross section of dark pions. The exclusion contour obtained from data is shown as a red solid line, while the expected exclusion contour is shown as a dashed black line, with one standard deviation from it shown in solid yellow. The dashed region is excluded by the 1 lepton channel, while the two greyed out regions had already been excluded by reinterpretation or the all-hadronic channel of this analysis.
Part III: Looking to the future

While the data analysis described in the previous part constitutes a landmark result in the exploration of these specific dark sectors, it is clear that to fully explore them, they need to be tackled from additional angles. In this part of the thesis, I present the four auxiliary projects that I have been a part of, that in different ways supports this endeavour.
16. The UCluster method

The following section describes the unsupervised particle physics data clustering method UCluster, and the work done to modify it so that the calculations can be distributed over multiple machines. I presented the work at the CHEP conference, and the following paper was published as part of the proceedings. The main body is unchanged with respect to what was published in reference [170], with the exception of layout editing.

16.1 Distributed training for the particle physics data method UCluster

Although machine-learning methods have been used in high energy physics for more than 50 years, recent years have seen a substantial increase in their variety and prevalence (see e.g. [67]). This can be connected to different factors, such as, the increased need for more precise methods in experiments, the fast-paced development of novel machine-learning methods coming from both within the academic setting and the private sector, and improvements in computer hardware leading to a larger computing capacity.

Within the Large Hadron Collider (LHC) experiments, several research problems are already solved with machine-learning; neural networks and boosted decision trees are used for e.g. flavour tagging jets, separating signal from background in analysis, and particle identification [29]. Improving these methods, extending their reach and developing new ones is currently an active field of study. In addition to the above, machine-learning has been proposed to solve such diverse problems as collision and detector simulation, trigger decision-making, model-independent searches for new physics, jet substructure studies and much more [126].

A lot of the development in machine-learning methods for High Energy Physics is done on data sets that have already undergone a great deal of event selection, and are small enough to allow for training on a single Graphical Processing Unit (GPU) used as the compute node. For more complex models, such as graph networks, generative adversarial networks, or any large enough model, this type of training is limited by its lack of scalability, i.e. its capacity to handle growing amounts of data. This problem can be ameliorated by distributing the training over multiple workers in a cluster of several compute nodes, and thereby increasing the capacity for data ingestion by utilising the distinct storage devices and the working parallel random access memory of
several GPU nodes. Using distributed training algorithms also reduces the overall training time by effectively increasing the batch size, which in turn mitigates the problem of prohibitively long training times. In this paper, we apply distributed training to the recently proposed UCluster method [146] for unsupervised clustering of particle physics data, with the goal of both speeding up the training and making it scalable to arbitrarily large data sets.

UCluster is a neural network for unsupervised clustering on particle collider data. It creates a latent space using a classification network and then clusters particles, that are close in the latent space, together.

Depending on the desired properties of the clusters, different classification objectives can be used to create different latent spaces. By choosing jet mass classification, the model produces clusters of jets with the same mass, by choosing event classification the model produces clusters of events with similar properties, etc. This makes UCluster general and highly adaptable, and it has the potential to be useful for several physics problems relating to particle or event classification. So far, it has shown promising results when applied to multi-class classification of jets and anomaly detection at the event level. However, with all training and evaluation executed on a single GPU, the size of the input data is limited to the number of events that can be loaded onto one GPU memory simultaneously. This excludes modifying UCluster, as implemented in a single machine setup [146], to any task that requires training on bigger sample sizes, e.g. full data samples from the LHC experiments.

Unsupervised multi-class classification is something that could be of interest in precision measurements, e.g. in cases where simulated data are not precise enough to be used for background estimation. In this case, data-driven methods, i.e. methods that use real data to estimate the background, are commonly used. This approach can quickly become involved if it has to be done for more than one background process. Instead, an unsupervised multi-class classification method could be applied directly to data, labelling the processes without the need for multiple background fits. UCluster is reported to have an 81% clustering accuracy when applied to the problem of classifying the fat jets of the HLS4ML LHC Jet data set [154] into clusters in which the member jets all originate from the same particle, making use of particle mass classification. However, for data-driven background estimation, the data sets could become several orders of magnitudes bigger.

Anomaly detection is interesting in particle physics, since it is model-independent by nature and can be used to find deviations from the Standard Model, which can then be used as the basis for new studies. UCluster has been applied to the R&D Dataset for the LHC Olympics 2020 Anomaly Detection Challenge [134] and reports an increase in signal-to-background ratio from 1% to 28%, in which the signal represents the anomalies. This is accomplished through all anomalies ending up in the same cluster. Making this setup scalable would open up the possibilities of looking at larger data sets or
even the full experimental data samples of the large LHC collaborations, as has been proposed in reference [83].

To achieve distributed and scalable training, we made use of Apache Spark [176], an open-source distributed general-purpose cluster-computing framework, which creates an architecture over a cluster of multiple compute nodes for distributed data processing through a distributed file system, that splits and stores the data for processing in a fault-tolerant manner. We set this up on the Databricks [14] platform, which allows for easy creation of Apache Spark clusters. This setup bypasses many of the challenges of processing large data sets such as cluster management, unreliable hardware or running out of memory on a single GPU. We used the distributed deep learning framework Horovod [163], which contains a wrapper to run distributed training in Spark clusters, to run the training. Once a training algorithm has been set up with Horovod, it can be run on any number of GPUs (including only one) without any changes in the code. Distributing the training across multiple GPUs means an effective increase in the batch size of the underlying stochastic gradient descent optimisation algorithm, leading to a faster convergence of the optimisation functions. We expect the training time to be in inverse proportion to the number of GPUs.

16.2 Model and data

The details of the UCluster model can be found in the original reference [146], but some elements needed to understand the rest of this paper will be repeated here. Consider a data set from a particle collider, which has already gone through digitisation and object reconstruction. The reconstructed objects are then represented as nodes in the graph-based neural network known as ABC-net [145], a classification net that aids in the over-arching classification problem.

This classification net needs to be optimised to create a suitable latent space on the overarching clustering problem of the UCluster model, i.e. a space in which particles with similar properties are close to each other, making clustering possible. The ABCnet is pre-trained for a number of epochs and then a classical k-means algorithm is applied to the latent space. The resulting clusters are used to initialise the cluster centroids in the full model, which is a Deep k-means algorithm that combines the classification net with clustering [110]. The full model is trained end-to-end, with the combined classification and clustering loss, and the trained model assigns every data-point to a cluster. The code is written in TensorFlow v.1 [26].

In this paper, we will mimic the first use case demonstrated in the original paper – multi-class classification of fat jets from the HLS4ML LHC Jets data set – using distributed training. The classification objective is mass clas-
Table 16.1. Data features used as input in the UCluster method.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Delta \eta$</td>
<td>Difference in pseudorapidity $\eta$ between jet constituent and jet.</td>
</tr>
<tr>
<td>$\Delta \phi$</td>
<td>Difference in azimuthal angle $\phi$ between jet constituent and jet.</td>
</tr>
<tr>
<td>$\log(p_T)$</td>
<td>Logarithm of the constituent transverse momentum $p_T$.</td>
</tr>
<tr>
<td>$\log(E)$</td>
<td>Logarithm of the constituent energy $E$.</td>
</tr>
<tr>
<td>$\log\left(\frac{p_T}{p_T^{(jet)}}\right)$</td>
<td>Logarithm of the constituent transverse momentum relative to the jet transverse momentum.</td>
</tr>
<tr>
<td>$\log\left(\frac{E}{E^{(jet)}}\right)$</td>
<td>Logarithm of the constituent energy relative to the jet energy.</td>
</tr>
<tr>
<td>$\Delta R$</td>
<td>Distance defined as $\sqrt{\Delta \eta^2 + \Delta \phi^2}$</td>
</tr>
<tr>
<td>PID</td>
<td>Particle ID [171]</td>
</tr>
</tbody>
</table>

The data are stored in two HDF5 files [172]: one training data set and one validation data set. Following the unconventional nomenclature of the HLS4ML challenge [154], we call the two data sets, used for development, the training and testing data sets and the data set we test our final model on as the validation data set.

16.3 Modifying the code for scalability and distributed training

To make use of the Horovod framework, we needed to migrate the code to Tensorflow v2. We then extended the model to allow for distributed training. We thus have three incarnations of the code:

- **The original code** which is obtained from the GitHub of the original authors [144]. This was used to validate our setup against.
- **The original code migrated to TensorFlow v2**. This was done with an automated function supplied by Tensorflow and relies on TensorFlow 1 compatible functions. This code was validated against the original code and shows comparable results in both training behaviours, results and execution time.
- **The distributed model**. This is described in the next section.
16.3.1 Distributed training

With the Tensorflow v2 code as a starting point, we used the HorovodRunner, the Horovod Databricks Application Programming Interface (API), to be able to run distributed training. This was done by creating a HorovodRunner instance and passing it the training function. The training function had to be modified for use with Horovod, including changes, such as, increasing the learning rate to compensate for the bigger effective batch size and handling of checkpoints to ensure consistent saving to and initialisation from them. With our data on the local driver, we initialise the HorovodRunner instance, which copies the data to each GPU. The Horovod framework takes care of the distributed training using a distributed optimiser to accumulate gradients across multiple GPUs, and saves model checkpoints at regular intervals. The learning rate is scaled by the number of GPUs used. Currently, all data is copied onto each GPU, which limits the scalability to what can be fit into a single GPU memory. This will be addressed by copying only fractions of data to each GPU at a time.

16.3.2 Scalability

The UCluster repository includes a pre-processing script specifically for the HLS4ML data set which we use to extract the relevant features from it (see Table 16.1). After pre-processing, the data are contained in a single file and will have to be loaded in their entirety onto the local driver before training. This currently puts limits on the scalability of our setup, and will be addressed in future work by writing data loaders directly into the distributed file system.

16.4 Training and evaluation

The model was trained with the same hyper-parameters as found in the original paper, summarised in Table 16.2. It was first pre-trained for 20 epochs with only the classification net, before being trained end-to-end (with the combined classification and clustering loss) for a total of 100 epochs. The training was done on GPU clusters with either 2, 4 or 8 NVIDIA T4 GPUs, each with 16 GB of memory, running Apache Spark 3.0.1 (Amazon EC2 G4dn.xlarge instance). The training time per epoch for different number of GPUs can be seen in Table 16.3, compared to the training time for the single machine codes. The training time can be seen to inversely scale with the number of GPUs. Each of the models, the distributed code as well as the single machine codes, has been trained several times, and the testing accuracy during training has been recorded. A representative plot for each model can be seen in Figure 16.1. The models display very similar training behaviour, which is to be expected if the distributed training works as it should: For all models, the accuracy lies in
Table 16.2. *Hyper-parameters used in training.*

<table>
<thead>
<tr>
<th>Hyper-parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Batch size</td>
<td>1024</td>
</tr>
<tr>
<td>Inverse temperature $\alpha$</td>
<td>Starting at 1, increasing linearly by 2 every following epoch</td>
</tr>
<tr>
<td>Proportionality constant between</td>
<td>10</td>
</tr>
<tr>
<td>classification loss and clustering loss $\beta$</td>
<td>2</td>
</tr>
<tr>
<td>Focal loss hyperparameter $\gamma$</td>
<td>Starting at 0.001 and decreases by 2 every three epochs until it has reached $10^{-5}$. Multiplied by a factor equal to the number of GPUs.</td>
</tr>
<tr>
<td>Learning rate</td>
<td></td>
</tr>
<tr>
<td>Optimizer</td>
<td>Adam</td>
</tr>
</tbody>
</table>

Table 16.3. *Training times for the single machine codes and the distributed code.*

<table>
<thead>
<tr>
<th>Code</th>
<th>Training time per epoch</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original code</td>
<td>4 minutes</td>
</tr>
<tr>
<td>TensorFlow 2 version of original code</td>
<td>4 minutes</td>
</tr>
<tr>
<td>Distributed training, 2 GPUs</td>
<td>2 minutes</td>
</tr>
<tr>
<td>Distributed training, 4 GPUs</td>
<td>1 minute</td>
</tr>
<tr>
<td>Distributed training, 8 GPUs</td>
<td>30 seconds</td>
</tr>
</tbody>
</table>

the interval 0.5 to 0.6 throughout the training, usually increasing slightly with training time and occasionally decreasing over a number of epochs before returning to the stable behaviour. It can be noted that the testing accuracy is far from the 81% validation accuracy obtained in the original paper for all models, including the original one. Since there is virtually no improvement with training, the optimisation algorithms might have found a local minimum. Figure 16.2 shows the clustering loss during training of the distributed model on 4 GPUs. Here we can see a minimum at around epoch 28, and then a slight increase until epoch 65. Some hyper-parameter optimisation was done, testing $\alpha$ (inverse temperature) values that changed by a factor of (1,2,5,10) every epoch following the starting value of 1, changing the batch size to 512, changing the proportionality constant $\beta$ between the classification loss and the clustering loss to $(10^{-1},1,10,100)$ and lowering the learning rate. All trials gave worse or comparable results to those in Table 16.2. Furthermore, using the same hyper-parameters consistently shows the same behaviour as displayed in Figure 16.1 across a number of trials, independent on the number of GPUs used for training.
Testing accuracy of single machine training with original code.  

Testing accuracy of single machine training with TensorFlow v2 code.  

Testing accuracy of distributed training, 8 GPUs. 

Figure 16.1. Testing accuracies. The first 20 epochs were pre-training without any clustering, so the clustering accuracy is set to 0.

16.5 Next steps

To make the model fully scalable, we need both the data ingestion and the training to be scalable. We will accomplish this by loading data directly into the distributed file system, bypassing the memory limitations of the main memory, as well as copying only subsets of data onto each GPU. For the data loading, we will use file formats designed for distributed processing of large data sets such as the open-source column-oriented data storage format Parquet or the to high energy physicists well-known ROOT format. For the data distribution, we will compare the Horovod framework to the Maggy [143] framework to see if there are already tools implemented that can be used for this purpose.

We are also actively looking into how the accuracy can be improved to that which was shown in the original paper. Since the optimisation might be stuck in a local minimum, a natural path forward would be to investigate the inverse temperature and potential energy surface defined by the optimisation problem. If the weights of the fully trained, accurate model can be made
available to us, this could provide valuable insight without having to look into the inverse temperature. We could initialise our weights with the fully trained parameters, or slight perturbations there of, and see if our training reaches the same parameters again.

16.6 Conclusion and outlook

We have implemented distributed training for the UCluster method and are in the process of making it scalable to any input data size. We migrated the UCluster model to TensorFlow v2 and added distributed training using the HorovodRunner. After this, the training behaviour of the model is very similar to that of the original code, and this behaviour is consistent over a large number of trials. However, we see a significantly lower accuracy than that reported in the original paper. We are in the process of making the setup fully scalable, as well as troubleshooting the lowered accuracy. The UCluster method is a very general method in the sense that it can be modified for any task in which unsupervised clustering of particles could be used. The generality of the original model together with the scalability added in this project has the potential to be very powerful in processing large amounts of data for a wide variety of tasks at the LHC. The distributed model can already be used as is for very fast training with HDF5 data, a format commonly used in particle physics machine-learning challenges open to researchers outside of the big LHC collaborations. Once it has been made fully scalable, it will be able to train directly on both experimental data and simulated data from the LHC – possibly requiring some pre-processing of the files.
Section 5.2 describes the models under study in the Dark Meson analysis, as well as the parameters, \( \eta \) and \( m_{\pi D} \), that they depend on. During the initial data exploration in the analysis, it became clear that the large ranges of \( \eta \) and \( m_{\pi D} \) values that we chose to simulate signal samples with lead to very large variations in the kinematic behaviour of the signal processes.

While several kinematic variables were defined and studied, none showed discrimination in the same region for all signal points. A representative example of this is shown in figures 17.1 and 17.2 where the \( m_{bb\Delta R_{min}} \) distributions corresponding to all of the signal points are shown together with the background distributions. For low masses, the signal peak is to the left of the background peak, while higher masses tend to peak slightly above the background and also include a long tail above the (normalised) background expectation.

The variables defined are also not independent on each other, so that a cut in one variable affects the distribution of others.

This problem of finding an optimal set of requirements on some or all of a set of interdependent variables in order to classify data into signal and background is ideal for a boosted decision tree (BDT), described below.

I initiated and supervised such a study in the context of the dark meson analysis. While it couldn’t be a part of the main ATLAS analysis due to constraints on time and person-power, it is thoroughly described in reference [142] by master student Eva Mayer, who performed it. The motivations as well as the main results of the study and what they mean for the analysis are described in this chapter.

### 17.1 Boosted Decision Trees

Decision tree learning is a machine learning method which uses a labelled set of training data to construct a flowchart-like algorithm for classification or regression. When used for classification, the trained model consists of a set of threshold values on the features of the data which together define the regions of the feature space belonging to each class. In practice, a decision tree thus defined is a weak learner for most types of data, i.e. has low predictive value. A way to increase the predictive value is to use an ensemble of decision trees each trained to correct the misclassifications of the previous one. This is called...
Figure 17.1. The distribution of $m_{bbbR_{min}}$ for all signal points in the original $SU(2)_L$ signal grid. The top and bottom row shows the high- and low-$\eta$ signal points, respectively, and the left and right column shows the high mass and low mass signal points, respectively. The signal distributions are showed as dashed histogram lines, and are all individually normalised to 1. The background distribution (identical in all plots) is showed as filled, stacked histograms, with the full stack normalised to 1.
Figure 17.2. The distribution of $m_{b\bar{b}R_{\text{min}}}$ for all signal points in the original $SU(2)_R$ signal grid. The top and bottom row shows the high- and low-$\eta$ signal points, respectively, and the left and right column shows the high mass and low mass signal points, respectively. The signal distributions are showed as dashed histogram lines, and are all individually normalised to 1. The background distribution (identical in all plots) is showed as filled, stacked histograms, with the full stack normalised to 1.
boosting. A boosted decision tree (BDT) is thus a linear combination of the sequentially trained decision trees. Several algorithms for doing this exist, and in this project XGBoost [86] was used.

17.2 Training data and model
An XGBoost model with parameters that were determined by a hyperparameter scan described in reference [142] and listed in section 17.1 is trained on the simulated signal and background data used in the dark meson analysis. Prior to the training, the dataset is split in half, and one half of it is kept for testing. The training objective is the binary classification of signal versus background. Due to the large imbalance of the true classes, the signal events were weighted by the inverse fraction of signal events in the training sample.

<table>
<thead>
<tr>
<th>Table 17.1. The parameters used to define the XGBoost model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Learning rate</td>
</tr>
<tr>
<td>Maximum depth</td>
</tr>
<tr>
<td>Objective</td>
</tr>
</tbody>
</table>

The pre-selection on the simulated data is shown in table 17.2.

<table>
<thead>
<tr>
<th>Table 17.2. Summary of the selection used for the BDT training data.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of leptons                 = 1</td>
</tr>
<tr>
<td>Number of additional loose leptons = 0</td>
</tr>
<tr>
<td>Number of jets                    ≥ 5</td>
</tr>
<tr>
<td>Number of b-tagged jets           ≥ 3</td>
</tr>
<tr>
<td>$H_T$                             &gt; 300 GeV</td>
</tr>
</tbody>
</table>

The features studied are the same variables that I defined and studied for the traditional analysis, described in appendix B. From these, the most promising ones are chosen according to the methodology described in reference [141], including importance ranking of the features, correlation considerations and univariate analysis. These are used as the training features of the BDT. They were:

- $N_{jets}$
- $N_{b\text{-jets}}$
- $p_Tl$
- $H_T$
- $\Delta R$ between the leading b-tagged jet and the lepton
- $m_{\text{had}} + m_{\text{lep}}$
- $m_T \ m_{bb}$

where $p_Tl$ is the transverse momentum of the lepton and all other variables are defined in chapter 11.3.
17.3 Results

Several BDTs were trained on different signal points and then evaluated on all other signal points. The BDT that performed the best in terms of efficiency and purity simultaneously was one trained on the signal point $(\eta, m_{\pi^0})=(0.25, 800 \text{ GeV})$. Figure 17.3 shows normalised plots of the discriminant for background and all $SU(2)_L$ points in the original signal grid. Here, too, the signal distribution peaks range from right at the background peak (0) to as far away as possible from the background peak (1). Rudimentary signal regions were defined by selecting events with a discriminant value above a threshold. The study estimated the significance of signal discovery at each point using only statistical uncertainties, showing a promising increase in significance across most of the parameter space.
18. Sparks in the Dark

This following text has been submitted to Computing and Software for Big Science, and has been made public as a pre-print by me and my co-authors [168]. The main body is presented here in its entirety.

18.1 Introduction

Collider experiments in high-energy physics often deal with large amounts of experimental data. The two general-purpose experiments at CERN’s Large Hadron Collider (LHC), ATLAS and CMS, record about 10 PB of data per year. These data are then analysed for, e.g., consistency with different theoretical models, which involves both isolating a small signal from large background and data-driven corrections to phenomenological background estimates. A pre-selection of data is performed based on the particles involved in the experimental signature of the signal. Subsequently, the resulting dataset is explored with the objective to create a phase-space region enriched in signal events. This enriched region allows for a statistical analysis that is sensitive to the signal. Optimising the region involves using theoretical knowledge of the signatures and kinematic behaviour of the signal and background processes to define new variables, and a tedious process of exploring the data using 1D or 2D histograms. Machine learning classifiers are also commonly used at this stage, which both hone in on the region without the same need for manual optimisation and utilise complex relationships between variables. The downside of these methods is that the interdependence of the variables is never made explicit, and the analysis becomes harder to understand than one defined in terms of intervals in each variable. This matters not only for the understanding of the individual physicist, but also matters for reinterpretations of the results. This paper is a proof-of-concept of a new method which has the potential to produce a more sensitive signal region in a shorter time than manual optimisation, while keeping the analysis and interpretability as simple as possible.

This work builds on multi-dimensional histograms with rigorously defined arithmetic using a scalable implementation with sparse binary trees representing the data, as implemented in the SparkDensityTree library [174], following [124, 158, 159].

SparkDensityTree takes arbitrarily large sample sizes in high dimensions that are assumed to be drawn from an unknown density and returns the minimum distance estimate (MDE) of the unknown density itself as a multi-dimensional histogram. Unlike most density estimation methods, MDE histogram is
the only scalable $L_1$-smoothed density estimate with so-called universal performance guarantees [158]. In particular, calculating the coverage or highest density regions of the MDE histogram of the signal and background data allows for finding the region of phase space with the largest probability density in the signal and background. The method takes measured or simulated data for signal or background processes as input and returns the highest density region of its density estimate (MDE histogram). The signal region is given as a union of intervals, rectangles, cuboids and hyper-cuboids over the domain of the input variables.

The current proof-of-concept is largely inspired by an ongoing search for dark mesons in ATLAS data for which a preliminary result is public [16]. The data and simulation used in the following sections, as well as the selections applied, closely follow the analysis. The signal point chosen in this study has already been excluded by ATLAS [16], and so the data will be used as background.

18.2 Datasets and event selection

The study uses $2.3 \text{fb}^{-1}$ of $\sqrt{s} = 13\text{TeV}$ proton–proton ($pp$) collision data collected by the CMS experiment [84] in 2015 to model the background to the dark meson signal. The analysed data correspond to the SingleElectron [93] and SingleMuon [94] datasets released on the CERN Open Data portal [82]. Only events in the list of validated runs [92] are retained for the study. A total of about 110 million single electron and 70 million single muon $pp$ events are available for analysis.

The datasets are provided in the CMS miniAOD format, which contains high-level reconstructed objects that can be used for analysis [91]. This study is based on such reconstructed electrons, muons and jets. The data is accessed and processed using the CMS analysis code provided with the CMS open data [88]. Within this framework, jets are reconstructed using the anti-$k_T$ algorithm [74] with a fixed radius parameter $R = 0.4$ and are tagged as containing a bottom hadron based on the Combined Secondary Vertex (CSV) tagging algorithm.

A dark pion signal sample is simulated using MadGraph5_aMC@NLO 3.5.1 [32] interfaced with Pythia 8.306 [165] for showering and hadronisation. Both the resonant and the Drell-Yann-type dark pion production $pp \rightarrow \pi_D \pi_D$ are considered, allowing for any decay mode of the $\pi_D$. Fast simulation of the detector is done with Delphes 3.5.0 [102] using the standard CMS detector card. Within Delphes, jets are determined with the FastJet 3.3.4 [75] software package and the anti-$k_T$ algorithm [74]. The default $b$-tagging of the CMS Delphes card is used to identify $b$-jets. The dark pion mass is set to $m_{\pi_D} = 500\text{GeV}$ and the dark rho mass to $m_{\rho_D} = 2\text{TeV}$. A total of 50k signal events are simulated.
As previously mentioned, this signal point has already been excluded by the ATLAS collaboration [16].

Events are further selected for the study based on kinematic and quality criteria imposed on the reconstructed leptons and jets. In the MC events, any electron or muon with transverse momentum $p_T > 28$ GeV is considered as a signal lepton. In data events, the signal lepton must additionally pass the Tight selection criteria [89, 90]. Only events containing exactly one signal lepton are retained for the study.

All jets are required to have a transverse momentum $p_T > 20$ GeV and to satisfy $|\eta| < 2.5$. In addition, any jet is required to have an angular distance $\Delta R > 0.4$ from the signal lepton in the event, in order to resolve any reconstruction ambiguities between the lepton and jets. If these requirements are not met, the jet is discarded. Events are eventually required to have at least four jets, out of which at least two must be $b$-tagged.

Events passing all requirements listed here are selected for analysis. A total of 120k and 7.6k events pass this baseline selection in data and signal respectively. The signal is normalized to the integrated luminosity of the data sample and corresponds to a total of 6.47 selected events.

18.3 Discriminating variables

The method is demonstrated on four event-level quantities that are suitable as discriminating variables. The first three are: $\Delta R(l, b_2)$, defined as the angle between the highest-$p_T$ lepton in the event and the second closest $b$-jet; $m_{bb\Delta R_{\text{min}}}$, defined as the invariant mass of the two $b$-jets in the event that are closest to each other; and $H_T$, defined as the scalar sum of the $p_T$ of the jets in the event. The final variable is based on $R = 1.2$ jets reclustered from the $R = 0.4$ jets using the anti-$k_T$ algorithm with a fixed radius parameter of $R = 1.2$ [149]. All leptons in the event are added to the $R = 0.4$ jet collection before the reclustering and the highest-$p_T$ large-$R$ jet containing the lepton is referred to as $J_{\text{lep}}$ while the highest-$p_T$ fully hadronic large-$R$ jet is referred to as $J_{\text{had}}$. The sum of the masses of these two jets is used as a discriminating variable and is denoted by $m_{\text{had}} + m_{\text{lep}}$. Distributions of the discriminating variables in data and signal are shown in Fig. 18.1 for events passing the baseline selection described in the previous section.

18.4 Method

The SparkDensityTree library is a library of statistical methods, with the base class being a multi-dimensional density estimator that for any sample generated from an unknown density returns an optimally smoothed histogram. The optimally smoothed histogram is taken to be the one that, per estimation, min-
imizes the $L_1$ distance to the true underlying distribution, using the minimum distance estimate (MDE) method. The statistical methods on these MDE histograms include arithmetic operations, conditional densities, coverage regions, and marginal densities.

The histogram object is represented as a binary tree in which each node represents a bisection of the phase space, and the leaves contain the event count in the finest resolution boxes thus obtained. The histogram construction begins with the definition of the root box, ideally the smallest hypercube containing all data points. From the root box, $x_p$, the support is iteratively bisected until a stopping criterion is reached, as visualized in Fig. 18.2. The Mapped Regular Pavings [124] underlying the tree structure allows for giving each box in the splitting a unique address. The combination of the leaf address and the counts is defined as the label of the box, $(\rho_v, #x_p)$. The MDE histogram is described in [121], and is taken as the optimal density estimate in this work. It is found by an adaptive search in sequentially coarser histograms, starting at the one obtained by the splitting.

Figure 18.1. Distributions of the discriminating variables in data and signal for selected events, normalized to 1. (Top, Left): $\Delta R(l,b_2)$; (Top, Right): $m_{bb_2}\Delta R_{min}$; (Bottom, Left): $H_T$; and (Bottom, Right): $m_{j\text{had}} + m_{j\text{lep}}$. 
The splitting is an inherently sequential process, but a distributed solution was developed in [159, 121]. This requires an initial splitting of the root box down to the finest resolution that might be needed instantaneously – possibly to the point that each leaf only has a count of one – and then merged again. This is accomplished by only representing the leaves with at least one data point using sparse binary trees.

In the distributed method, therefore, an additional step is added between the splitting and the MDE, which consists of merging the cells to a stopping criterion on the counts in each box, effectively representing the initial histogram for finding the MDE.

For a more in-depth explanation of the steps, the reader is referred to [124, 158, 159, 121, 160]. The procedure is sketched below:

**Stage 1:** Find the root box containing all the data points.

**Stage 2:** Define a stopping criterion for the splitting, such as a maximum box size. The root box is split until this criterion is reached, giving the finest resolution histogram. In this work, the finest splitting is determined by the stopping criterion that no leaf-box has any side length longer than the parameter finestResSideLength.

**Stage 3:** Merge leaves such that the counts are maximized, while not going higher than some limit minimumCountLimit and keeping the leaf depth as small as possible.

**Stage 4:** Starting from the histogram obtained in stage 3, find the optimally smoothed histogram using MDE as described in [121, 160].

Additionally, two user-defined parameters concerning the distributed aspect of the method are available: numTrainingPartitions and sampleSizeHint. Respectively, they correspond to how many times the training data is partitioned, related to distribution of work among computing nodes, and an initial guess of points connected to the size of the node batches [161].

The value of this method for data exploration in high-energy physics lies in the next step. When the MDE histogram is obtained, the highest density regions can be extracted by calculating the pdf coverage regions; and accordingly the highest and lowest density regions.

For simplicity, marginal densities are considered in this work, but the method can be extended to take the full density into account simultaneously.

The marginal densities for all unique pairs of the variables can be obtained from the 4-dimensional MDE histogram. In this paper, \( \binom{4}{2} = 6 \) unique pairs of variables are chosen and these six combinations are what the highest density regions are computed from. This is done separately for signal and background. The signal and background highest density regions can be defined independently of each other, and can, crucially, be flipped around to allow for finding the least dense region in the background density. From here, the user has to consider the best ways to use these marginal densities, and an example is given below.
18.5 Results

The results presented in this work are documented in a Github repository [169]. All computations for the upcoming results have been performed on Virtual Machines (VMs) hosted by Google as a part of a dataproc cluster. The cluster contains three VM instances, all of which run four Intel Skylake vCPUs and has 15 GB of RAM; all in order to utilize the distributed aspect of the method.

Figure 18.3 shows a comparison between a 2D frequency or count histogram of the data over a uniform grid and that over the optimally smoothed nonuniform partition corresponding to the MDE histogram of this method. All distributions considered in this work have been verified by eye in this way to ensure sensible density estimates are returned by the method.

The density estimate is presented at three different highest density regions for background in Fig. 18.4, and for signal in Fig. 18.5 for the $m_{\text{had}} + m_{\text{lep}}$ vs. $\Delta R(l, b_2)$ combination.

Figure 18.3. Comparison between a regular 2D histogram representation (Left) and the distribution obtained in this method (Right) of $m_{\text{had}} + m_{\text{lep}}$ vs. $\Delta R(l, b_2)$ in background from data.
Comparisons between signal and background distributions can also be made at different levels. Figure 18.6 shows the 3D and 2D combinations, together with the highest 50% density regions for $m_{\text{had}} + m_{\text{lep}}$ and $H_T$.

The density estimates for signal and background are combined to form $X_{\text{sig}} \otimes \overline{Y}_{\text{bkg}}$ density regions, where $X_{\text{sig}}$ indicates the $X\%$ highest signal density region and $\overline{Y}_{\text{bkg}}$ indicates the complement of the $Y\%$ highest background density region. These combinations are used to design kinematic regions corresponding to the most dense signal and the least dense background. The regions are achieved from the $X\%$ highest signal density region and the $Y\%$ highest background density region using a bounding box around the density region in each pair of variables. From the bounding box, the sensitive interval of each variable is taken as the projection of the box onto that axis. The intersection of the signal interval and the complement of the background interval forms the final interval of interest for each variable pair. Each variable is associated with exactly three intervals from its participation in three variable pairs. In this work, the final region is defined by the union of these intervals in each variable. Three combinations are presented: 50% $\otimes$ 50%, 90% $\otimes$ 20% and 90% $\otimes$ 10%. As an example, the obtained intervals for the 90% $\otimes$ 10%
Figure 18.5. Full signal density estimate of the $m_{\text{had}} + m_{\text{lep}}$ vs. $\Delta R(l, b_2)$ combination in 3D (Top Left) and 2D (Top Right) together with the highest 75% density region (Bottom Left) and the highest 50% density region (Bottom Right).

The combination are:

$$\Delta R(l, b_2) : [0.6, 1.1] \quad H_T : [625, 2172] \text{ GeV}$$

$$m_{bb\Delta R_{\text{min}}} : [312, 634] \text{ GeV} \quad m_{\text{had}} + m_{\text{lep}} : [552, 996] \text{ GeV}$$

When compared to the one-dimensional distributions in Fig. 18.1 it is clear that these correspond to regions with discrimination power between signal and background. The event selection corresponding to the intervals is applied to signal and data and the number of events passing the requirements are presented and compared in Table 18.1.

The method results on less than one signal event on all tested scenarios and no background events pass the selections in the most aggressive selection. Dark meson signals are usually very small, and unlikely to be accessible in 2.3 fb$^{-1}$ of data. It is possible however to naively scale the 0.57 expected events in the 50% $\otimes$ 50% scenario to, e.g. the full Run 2 data set collected by ATLAS, containing 140 fb$^{-1}$, to more than 30 events, a reasonable signal for a new physics search.
Figure 18.6. Full density estimate of the $m_{\text{had}} + m_{\text{lep}}$ vs. $H_T$ combination in 3D (Top), 2D (Middle) and the highest 50% density region (Bottom) for signal (Left) and background (Right).

The method could further be developed to identify the highest density region directly in the 4D histogram, and then project this onto the four axes. The SparkDensityTree library allows for defining arithmetic on the histograms, and it might be possible to combine the signal and background histograms and find the densest region in, e.g., number of signal events divided by number of background events, or the difference between the histograms.

Finally, scalability is a very powerful aspect of this approach. This study did not delve into it, but as mentioned in [159, 121, 160], the original method
Table 18.1. Number of signal and background events passing the baseline analysis event selection and the selections derived from the density regions applied on top of the baseline. Relative numbers of events with respect to the baseline analysis selection are given within brackets. The signal numbers are normalized to the integrated luminosity of the dataset.

<table>
<thead>
<tr>
<th>Selection</th>
<th>Signal</th>
<th>Background</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>6.47 (100.00%)</td>
<td>123951 (100.00%)</td>
</tr>
<tr>
<td>50% ⊗ 50%</td>
<td>0.57 (8.74%)</td>
<td>364 (0.29%)</td>
</tr>
<tr>
<td>90% ⊗ 20%</td>
<td>0.30 (4.57%)</td>
<td>16 (0.01%)</td>
</tr>
<tr>
<td>90% ⊗ 10%</td>
<td>0.07 (1.11%)</td>
<td>0 (0.00%)</td>
</tr>
</tbody>
</table>

has been tested on several terabytes of simulations, and great decreases in computational time can be seen with the increase of cores. This is something of interest for the field of high-energy physics, as it would be straightforward to run directly on the full collision datasets from the LHC.

18.6 Conclusion and Outlook

This paper introduces a scalable method, originally formulated in a purely mathematical context, applied for the first time in a high-energy setting. The approach relies on optimally smoothed multi-dimensional histograms with universal performance guarantees through scalable sparse binary tree arithmetic, incorporated in the SparkDensityTree library. It enables a rigorous definition of phase space regions enriched in signal, using multiple variables at a time. This method suggests promising avenues for the exploration of new physics phenomena at the LHC.

A large number of additional options is available from the SparkDensityTree library. This library contains several arithmetic operations and statistical methods (not covered here) that can be advantageous for studies on histograms, naturally interesting in a high-energy physics context.
19. A Hardware Tracker for the Trigger

The next sub-chapter contains the main body of the ATLAS internal report that I wrote on my work in the development of the new track-trigger mentioned in chapter III. It is unchanged with the exception of formatting and editing. It was written while research and development work on the HTT was still ongoing. Since then, the HTT project has been terminated in favour of software-based track triggering.

19.1 Introduction

As the Large Hadron Collider (LHC) prepares for the upgrade to High Luminosity LHC (HL-LHC), an upgrade to the ATLAS detector to make it capable of dealing with the increase in luminosity is also being prepared. New hardware and software are being developed and prospects for detecting new physics are being studied. This is true in particular in the Trigger and Data Acquisition system. The readout rate and storage capacity will be increased, but it is still not sufficient to save the same fraction of events as in Run 2 or Run 3. Without a better trigger system, therefore, the trigger thresholds would need to be increased, which would bring the risk of interesting events being lost.

One proposal to make the trigger system more efficient in saving interesting events is to make use of hardware tracking to be able to trigger on tracks. This is something currently being developed in the Hardware Tracking for the Trigger (HTT) project. The HTT is a proposed custom piece of hardware that would do tracking on demand for the trigger system. It is currently in a research and development phase, with prospect studies as well as design of both hardware, software and simulations being undertaken.

The Extrapolation Engine (EE) is a tool that could be of interest in this endeavour. It is a fast simulation of muon tracks extrapolated through the geometry of the Inner Tracker (ITk) without taking material interactions into account. This could prove a valuable tool when a fast simulation of the HTT is needed. It could allow for quick studies of parameters such as regions or sectors and provide a tool for prospect studies of different physics cases, such as for Long Lived Particles. However, the suitability of the EE for this type of tasks needs to be studied before it can be used. In my qualification task, I have used it to study the acceptance of tracks in the ITk and compared that
method to refitting tracks from the full simulation. I have also written a stand-alone root script that generates the sectors in a sample created with the EE and compared the number of sectors in different EE samples with the number of sectors obtained with the full simulation. This report will start with an introduction to the ITk, the HTT and sectors, needed to understand these studies, followed by a description of the EE. Then the acceptance of the ITk as obtained by using the EE will be presented and compared with a method using the full simulation. Finally, a method to generate the sectors in the output of the EE will be described, the number of sectors obtained with the EE will be compared to the number obtained with the HTT simulation, and the possibilities and limitations of the EE in this context will be discussed.

19.2 Background
19.2.1 The ITk
The Inner Tracker (ITk) is the tracker currently being developed as part of the upgrade of the ATLAS detector to meet the requirements of the HL-LHC [13, 12]. Two types of sensors will be used: closest to the beam pipe will be pixel layers, layers of pixel sensors arranged in modules, and outside of them will be silicon strip layers. In the central (barrel) region of the detector, both the pixel and strip modules are arranged in cylinders around the beam pipe. There are 5 pixel layers and 4 strip layers in the barrel, and the strip modules have sensors on both sides (inner/outer). The geometry can be seen in figure 19.1.

19.2.2 The HTT
The HTT is custom hardware that will do tracking on demand for the trigger. When tracking is demanded, the first step is pattern matching. Eight of the layers will be used, and the pattern that the hits in each of them make will be matched against a pattern bank. If a match is found, the hits in the 8 layers (the outermost pixel layer and all but the innermost side of the innermost pixel layer) are fitted with a track. This is the regional HTT. The next step is the global HTT, which fits a track to hits in up to 14 layers. This necessitates storing patterns and track parameters, and since storage capacity is limited, simulations of different physics cases are made to determine which it should be. A central concept in these studies are sectors, defined as a unique combination of one module in each of the layers mentioned above. To ”generate sectors” means here to take a sample of simulated tracks and find all sectors that correspond to at least one track in the sample. The bigger the number of sectors, the greater the storage space needed.
19.3 The Extrapolation Engine

The Extrapolation Engine (EE) is a fast simulation tool which generates muon tracks and then propagates them through the geometry of the ITk, taking no material interactions into account. The ITk geometry is built by defining each point in space as either sensitive, passive or as the boundary between them. The sensitive points represent the sensors and the passive all other material in the detector. The EE is fast – it can be run for 3,000,000 tracks in 30 minutes.

19.3.1 Setup

The Extrapolation Engine[76] is set up by compiling it with Athena. This can be done by giving the following commands in the lxplus terminal[8]:

```sh
mkdir Athena
cd Athena/
git atlas init-workdir https://:@gitlab.cern.ch:8443/atlas/athena.git
cd athena/
git atlas addpkg TrkExUnitTests
mkdir ../build
cd ../build/
asetup 21.9.4,Athena
```

Figure 19.1. The layers in the ITk. $z$ is the axis along the beam pipe and $R$ is the distance from it.
cmake -DATLAS_PACKAGE_FILTER_FILE=../package_filters.txt
../athena/Projects/WorkDir
make
source x86_64-centos7-gcc62-opt/setup.sh

The file ../package_filters.txt can be obtained by

cpy ../athena/Projects/WorkDir/package_filters_example.txt ../package_filters.txt

To compile only the Extrapolation Engine the last lines in this file should be edited to read

+ Control/AthenaBaseComps
+ TrkExUnitTests
- .*

19.3.2 Job options

The job option file used for this project can be found in the appendix. It was originally written as part of the EE project and contains some explanation of the different settings. Some functionalities to keep in mind are

- Geometry: Different proposed ITk geometries have been implemented and given a tag. In this project the ATLAS-P2-ITK-22-00-00 geometry is always used.
- Kinematic variables: To set the $\eta$, $\phi$ and $p_T$ of the tracks.
- Smearing: Smearing can be applied to the transverse and longitudinal impact parameters $d_0$ and $z_0$. This refers to drawing these parameters from distributions. The options for distributions to use is gaussian and flat.
- Simulation: Other options include how many events to simulate, how many tracks in each event, what output to collect (sensitive, passive, boundary, material), and whether to use charged or neutral particles. In this project, all hits are collected but only the ones in the sensitive parts of the detector are used. Charged particles are always used.

19.3.3 Running the EE

The Extrapolation Engine is then run with the Athena command:

`athena job-options.py`

This will set up the ITk geometry specified in the job options file. It will then simulate muons with selected kinetic properties and impact parameters and extrapolate their tracks through the ITk using only information about the magnetic fields and the geometry, and taking no material interactions into account.
When a muon is propagated across a sensitive/passive/boundary volume a hit is recorded in a root flat ntuple, with leaves containing spatial and kinetic information about the hit and what type of volume it was recorded in (sensitive/passive/boundary). Each event in the root file will contain all hits corresponding to exactly one muon track. Figure 19.2 shows all hits in a sample consisting of tracks in the whole $\eta$ region of the detector and a slice corresponding to a symmetry unit of the detector in $\phi$. With a big enough number of tracks, these hits map out all the sensitive regions in the detector, and we can recognise the 5 pixel layers and 4 strip layers of the barrel region and the perpendicular layers of the endcap regions in the plots. In the $\phi$-$r$ plane (which has been plotted using a slice in $\eta$ to make the pattern visible), note that there is an overlap between modules in the same layer. This is important to keep in mind, since it can lead to one track having two hits in the same layer.

19.3.4 Output of the EE

The output from the EE is a root file containing two trees: PositionMomentumWriter contains information about position and momentum of muons and ExtrapolationEngineTest contains information about the hits. In this project, only the latter has been used. The variables starting with "Sensitive" contain information about hits in the sensitive material. This includes position of the hit and which module the hit was recorded in. SensitivePhi, SensitiveEta and the SensitivePos variables are the ones used to plot 19.2. To uniquely determine a module, six variables are needed: SensitiveBarrelEndcap, SensitiveLayerDisc, SensitiveIsPixel, SensitiveSide, SensitiveEtaModule and SensitivePhiModule. The first has value 0 if the module is in the barrel region and 2 or -2 if it is in one of the endcaps. The next three are needed to determine the layer: SensitiveLayerDisc is an integer layer index, SensitiveIsPixel is 1 for pixel layers and 0 for strip layers, SensitiveSide is always 0 for pixel layers and 0 or 1 for the innermost or outermost side of strip layers, respectively. This ITk geometry has 13 layers in the barrel and they are thus numbered as shown in table 19.1 show: First come the five pixel layers, then each side of the strip layers. SensitiveEtaModule and SensitivePhiModule are unique integer indices of the modules inside each layer.

19.4 Acceptance of the ITk using the Extrapolation Engine

The acceptance of tracks in the ITk, i.e. the region of the detector in which tracks can be measured, is an important quantity as tracks from different physics processes have different angular distributions. The HTT will require 7 or 8 hits to reconstruct a hit, so to explore the acceptance in $\eta$ a plot of hits-on-track vs. $\eta$ can be used. In 2019, the readout schema of the pixel cells in
(a) Hits shown in the $r$-$z$ plane. The five pixel layers and 4 strip layers of the barrel region can be identified, as well as the perpendicular strip and pixel layers of the endcaps.

(b) Hits shown in the $r$-$\phi$ plane, using hits only in the region $|\eta| < 1$. Here the $\phi$-overlap of the modules can be seen.

Figure 19.2. The hits in a sample made with $\text{StartingEta} < 4$ and $0.1 < \text{StartingPhi} < 0.3$ shown in two different planes. Note that the detector has a cylindrical symmetry, which is why all modules can be seen in the left-hand plot even though only a subinterval of $\phi$-values is used.
Table 19.1. The combination of variables designating each layer in the barrel, from innermost to outermost. The logical layers are the eight layers used in the first step of the HTT.

<table>
<thead>
<tr>
<th>Logical layer</th>
<th>SensitiveLayerDisc</th>
<th>SensitiveIsPixel</th>
<th>SensitiveSide</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

the ITk changed and it was important to quickly get an idea of the implications of this for the HTT project. Two studies of the acceptance using the new readout schema were performed: The one described in this section and one using refitting of full simulation tracks to the new readout schema done by Philippe Calfayan[77]. The two methods thus served to validate each other, and this also served as an initial study of the differences between the simulations.

19.4.1 Acceptance plot using the EE

Making the hits-on-track plot using the EE output can be done in the following way:

1. Loop over the hits in the track and remove the $\phi$-overlap: remove the outermost hit if a track had two hits with same $\phi$, different $\eta$ in the same layer.
2. Count the number of hits left after overlap removal.
3. Make a 2D histogram is with this number on the y-axis and the starting eta of the track (StartingEta) on the x-axis.

19.4.2 Results

The result can be seen in figure 19.3, in which it is compared to the results from the full simulation method. The job-options for this sample that differ from the generally used in this project are summarized in table 19.2. Note that the phi region used is a symmetry unit of the detector so it should give the same result as using the full range, but with better statistics. A plot of the
average number of hits-on-track for the EE and full simulation can be seen in figure 19.4. In both representations, it is clear that there are on average more tracks recorded in the EE as compared to the full simulation. This can be explained by the fact that there are no material interactions in the EE. The spread of the number of hits-on-track is also slightly smaller in the EE for the same reason. Using tracks with $p_T$ of 4-400 GeV, the plot is similar as can be seen in figure 19.5. It shows some peaks around $|\eta|=2$ that are not as pronounced with 5 GeV tracks, however.

19.5 Sector generation with the Extrapolation Engine

If the number of sectors used by a certain physics case can be calculated using the EE, it will allow for fast prospect studies of the HTT for that case. One example is studying the prospect of detecting long lived particles, i.e., particles that do not decay promptly in the beam pipe and therefore have displaced tracks, by studying whether the sectors of that signal are contained in the pattern banks of the HTT. This can also be used to determine the increase in memory need for the pattern banks, if displaced track sectors were to be added. This section describes an algorithm to generate the sectors in EE output as well as the results of applying it to different EE samples and a comparison to full HTT simulation sector generation.
19.5.1 Sector generation algorithm

To find the number of sectors in an output file of the EE, I implemented the following algorithm in a stand-alone root script. (the full code can be found in GitLab[122]): For each track:

1. the hits in the 8 logical layers (see table 19.1) are kept and the rest are discarded.
2. For each of the hits kept, a hitword is created by shifting the eta module ID (SensitiveEtaModule) bitwise by 16 bits and then taking a bitwise OR with the phi module ID (SensitiveModuleID).
3. A string is created by concatenating all the hitwords for the eight layers in order (-1 is used for layers without any hits). This is a representation of all modules hit by the track.
4. The string is inserted in a unsorted hash table: If the string is not already in the table, it is inserted with a count of 1. If it is already in the table its count is increased by 1.
Figure 19.5. The number of hits per track as a function of starting $\eta$ using EE with muons with $p_T$ between 4 and 400 GeV. Showing only tracks with more than 7 hits.

Table 19.2. The parameters used in simulating the sample used to check the acceptance.

<table>
<thead>
<tr>
<th>Number of muons</th>
<th>200 000</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p_T$</td>
<td>5 GeV</td>
</tr>
<tr>
<td>$</td>
<td>\eta</td>
</tr>
<tr>
<td>$\phi$</td>
<td>&gt; 0.1, &lt; 0.3</td>
</tr>
<tr>
<td>Smearing</td>
<td>Flat in in D0 $-2$ mm to 2 mm, $z_0=150$ mm</td>
</tr>
</tbody>
</table>

After each track has been processed, the hash table contains all the unique combinations of modules that a track in the sample has hit, i.e. the sectors. This is then sorted in descending order with respect to the count. Finally, the sorted sectors are written as a text file, with each row containing the sector and how many times it was used by the tracks in the sample.

19.5.2 Results

To study the sector generation in the EE, three samples were produced, and are summarised in table 19.4. This table also shows the result of applying the algorithm described above to these three samples. All of them were produced in an $\eta,\phi$ slice with the same smearing (see 19.3).

The first two were produced following reference [148] which deals with the use of tracker information at trigger level for detecting long lived particles at
collider experiments. Note that although the sample size increases by a factor three between sample 1 and 2, the number of sectors found increases only by 26%. Even in the larger sample 2, the number of sectors is smaller than might be expected given experience from the HTT. To quantify this, a large sector file produce in the HTT simulation framework was used to compare with. It had been generated using 100,000,000 tracks with \( p_T > 1 \) GeV, and found 90,870 sectors (also shown in table 19.4). Sample 3 was simulated to be as similar to it as possible, using the same \( p_T \). Due to computing resource limitations, however, the same number of tracks couldn’t be generated. A big increase in the number of sectors can be seen when the \( p_T \) threshold is lowered between samples 2 and 3, which can be attributed to the fact that lower \( p_T \)-tracks are more curved. In this sample 4084 sectors were found, less than 5% of the ones found with the HTT simulation. However, the HTT simulation sectors contained many wildcards, i.e. layers without any hits in them. The sector generation algorithm for the EE does take wildcards into account, but finding a layer without a hit in it is very uncommon. Figure 19.6 shows the number of hits in each layer, which are 13 in total. As can be seen in the figure, very few tracks have less than 13 hits, and further studies revealed that there are no tracks that do not have hits in all of the eight logical layers used for the sectors. This can be attributed to the fact that there are no material interactions in the EE. The comparison is made therefore between the sectors in sample 3 and the sectors in the HTT sample excluding the ones containing wildcards. As shown in table 19.4, however, the difference is still large.

Figure 19.7 shows the distribution of the counts of the different sectors in sample 3 and the HTT sample. Both have very few sectors with a high count and a lot of sectors with a very low count. The HTT sample, however, has a slope between these two extremes while sample 3 does not.

**Table 19.3. The parameters used in simulating the sector generation samples**

| \( \eta \) | \( >0.1, <0.3 \) |
| \( \phi \) | \( >0.3, <0.5 \) |
| \( p_T \) | \( < 400 \text{ GeV}, \text{ flat} \) |
| Smearing | Flat in D0 -2 mm to 2 mm |
| | Flat in Z0 -150 mm to 150 mm |

**Table 19.4. The samples simulated for the EE sector studies (1-3) and the sample used to produce the sector file with the HTT simulation.**

<table>
<thead>
<tr>
<th>Sample</th>
<th>Number of muons</th>
<th>( p_T ) [GeV]</th>
<th>Number of sectors</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1,000,000</td>
<td>&gt; 4</td>
<td>2035</td>
</tr>
<tr>
<td>2</td>
<td>3,043,908</td>
<td>&gt; 4</td>
<td>2563</td>
</tr>
<tr>
<td>3</td>
<td>3,714,480</td>
<td>&gt; 1</td>
<td>4084</td>
</tr>
<tr>
<td>HTT simulation</td>
<td>100,000,000</td>
<td>&gt; 1</td>
<td>20,533 (without wildcards)</td>
</tr>
</tbody>
</table>
19.5.3 Discussion

In light of the fact that the number of sectors generated from EE tracks is off by about 80% as compared to the HTT simulation sector generation, it is clear that the EE is – at the best – not without limitations. Due to the lack of material interactions in the EE, there is also no multiple scattering (MS), i.e. trajectories being bent by the surrounding material. The MS affected trajectories might behave in a way that is different enough to allow for big discrepancies in the sector generation. Since MS happens primarily at low track $p_T$, it is still possible that the sector generation would be comparable above a certain $p_T$ threshold. The EE could then still be used for sector generation above the threshold.

Whether or not this is the case, studying where the discrepancies occur would be interesting. This would require sector-to-sector comparison, which is something that the stand-alone root script does not allow for since it uses a different hashing than HTT offline simulation. There is currently work ongoing to facilitate this (see next section) and once that is finished, sectors could be compared to see if they only differ in the outer layers (could again be an effect of MS) or if there is any system to which sectors are overlooked by the EE method. With sector-to-sector comparison, figure 19.7 could also be studied further.

19.5.4 Modifying the Extrapolation Engine to write out the Athena hashID

In the strive toward unifying the EE sector generation with the HTT simulation sector generation, a new branch was added to the EE output containing a hash ID obtained with the getHashID function in Athena. Some effort was put into validating that the two tools do return the same value for the same module, and into fixing a bug that was discovered. The hash ID in Athena is an ID unique to each module in the detector. It is thus a one-to-one function of the six variables needed to unambiguously define a module: endcap/barrel, layer number, pixel/strip, side of strip module, eta index and phi index. Using the code HTTRawHitsWrapperAlg, which runs over a sample file of tracks from the HTT simulation and can write out any information from the HTTRawHit class, the hash ID of several hits was written out together with the six variables. This was then compared to the hash ID obtained for the same six variables in the EE. The tools now agree and this hash ID can be used in place of the hashing function described above. This would make comparisons between the tools easier and integration into the HTT offline simulation framework more straightforward.
19.6 Conclusion

In this project, the Extrapolation Engine fast simulation tool has been studied for feasibility in the context of HTT offline software. It proved to be a useful tool for studying acceptance of tracks in the ITk. The study of sector generation using the Extrapolation Engine is as of yet inconclusive. It’s been established that it cannot be used in a straightforward way in the entire $p_T$ range, but more study is needed to conclude on whether or not it is useful subject to some limitations.
Figure 19.7. Normalised count of each sector in sample 3 and the HTT simulation sample, with sectors sorted in descending order. Note that there is no correspondence between sectors of the same number in the two samples.
Part IV: Conclusions
I have presented a search for dark mesons in 140 fb$^{-1}$ of proton-proton collision data recorded with the ATLAS detector. The analysis explores complicated final states with multiple top and bottom quarks in boosted regimes, and is sensitive to a large parameter space of the model. The results presented in this thesis substantially improve what was presented in a preliminary result made public as a conference note. Unfortunately, no significant excess of events over the SM expectation were observed.

As the underlying theory used to guide the analysis is based on a generic class of strongly coupled dark sectors, these results can be used to inform new theoretical work, as well as to perform re-interpretations of different models with similar final states.

This analysis was the first of its kind at the LHC and future iterations, which would benefit from the work presented here, can be envisioned. The signal kinematics, while similar in many ways to the dominant tbb background, can be very different in different areas of the parameter space, and a future analysis could optimise for the harder to separate, non-excluded signals. Further, the analysis was only sensitive to the $SU(2)_L$ model. Future work could focus on extracting the small signals expected in the $SU(2)_R$ model, too.

In this thesis I propose a set of additional approaches that could benefit such extensions of this analysis or similar ones. I have presented a first study of using machine learning for signal extraction yielding promising results. This can be extended to more sophisticated machine learning methods, such as deep learning, and be implemented in a conventional signal-dependent way.

I have shown that a mathematical package used for density estimation, the SparkDensityTree package, can be applied to LHC data to quickly hone in on phase space regions enriched in signal. This could significantly reduce data exploration time in future analyses, and could even be used to find the optimal signal region within a multidimensional density function, estimated from simulations of signal and background.

The UCluster method, that performs unsupervised multiclass classification, is a good fit for the dark meson analysis, as it includes boosted jets with characteristic substructure, while having a broad range of additional uses, like e.g. anomaly detection. The strength of the method lies in its versatility on the one hand, and its scalability on the other, ensuring it can be applied at any step in the analysis chain. This is also true for the SparkDensityTree method, and will be an important feature at the high-luminosity LHC, where the datasets are expected to be very large.
The high-luminosity upgrade of LHC comes with additional challenges, and a dedicated track trigger would be central to perform a similar analysis to this one in ATLAS then, considering the extreme pileup conditions expected. I have presented my part in the research and development work for the ATLAS upgrade to include such a track trigger system, and while the detector I studied was not the solution chosen for the upgrade, there will be tracker information included in the trigger chain in ATLAS for the HL-LHC.
I mer än ett decennium har elementarpartiklarnas tillkomst och interaktioner studerats vid partikelacceleratorn Large Hadron Collider (LHC) i partikelfysiklaboratoriet CERN i Genève. Den här avhandlingen är en samling projekt som alla har anknytning till ATLAS-experimentet vid LHC, vars syfte är generell elementarpartikelforskning.

Det största av dessa projekt, beskrivet i Del II (Part II), behandlar ett av partikelfysikens olösta mysterier, nämligen vad mörk materia består av. Detta görs genom dataanalys med utgångspunkt i en teori som förutspår så kallade mörka mesoner, vilka skulle lämna spår i datan om de fanns. De andra projekten syftar alla till att göra denna analys, och andra liknande, mer effektiva.

**Mörk materia**

Mörk materia är benämningen på den del av universums uppmätta massa som överstiger den synliga materians massa. Dess existens kan inte förklaras av någon av de kända elementarpartiklarna. Att ta reda på vad mörk materia består av är därför intressant inte bara för sin egen skull, utan också för att man skulle kunna upptäcka nya elementarpartiklar.

En experimentell svårighet vad gäller mörk materia är att den per definition inte är synlig för konventionella detektionsmetoder. Ett vanligt sätt att hantera det i partikelfysikexperiment är att leta efter en obalans i energi- och rörelsemängdekvationerna, som skulle kunna tyda på att en osynlig partikel har skapats.

Experiment relaterade till mörk materia delas ofta grovt upp i direkta detektionsexperiment, indirekta detektionsexperiment, och acceleratorexperiment. De ger information om den mörka materians interaktionstvärsnitt, annihileringsstvärsnitt respektive produktionstvärsnitt. Vid LHC i allmänhet och inom ATLAS i synnerhet har dataanalyser baserade på effektiva fältteorier och förenklade teorier gjorts under hela den verksamma tiden. Idén bakom dessa var att göra dem generella nog att fånga den starkaste produktionsmekanismen för de nya partiklarna. Eftersom det inte än har hänt har man i allt större utsträckning vänt sig till mer komplicerade teorier som i många fall beskriver flera nya partiklar vilka kan interagera med varandra och t.o.m. bilda sammansatta partiklar. Sådana teorier kallas för mörka sektorer, och det är just en sådan teori som ligger till grund för dataanalysen som presenteras i den här avhandlingen.
ATLAS-experimentet

Under marken i CERN, och långt in i Frankrike, ligger den 27 km långa cirkulära partikelacceleratorn LHC. I den cirkuleras två strålar av hadroner – så som protoner – åt motsatt håll och vid fyra punkter längs med ringen styrs de till kollision. Kollisionspunktarna är omgivna av partikeldetektorer: de två generella partikelfysikdetektorerna ATLAS och CMS och LHCb och ALICE, specialiserade på b-hadronfysik respektive tunga joners fysik.


Mörka mesoner

De precisionsmätningar som gjorts av Standardmodellens observabler i kombination med de analyser som gjorts baserade på teorier om mörk materia begränsar möjligheterna för nya teorier. Teorin som testats i den här avhandlingen förutspår en hel uppsättning nya partiklar, som likt de kända kvarkarna kan bilda sammansatta partiklar, baryoner. Den lättaste baryonen är stabil och skulle kunna utgöra mörk materia. Samma partiklar skulle även kunna bilda andra baryoner, så kallade mesoner, som inte är stabila utan sönderfaller direkt efter produktion till redan kända partiklar. Vi kallar de nya sammansatta partiklarna för mörka mesoner och mörka baryoner för att indikera deras koppling till mörk materia.
I den här avhandlingen presenteras en dataanalys som är känslig för tecken på den nya modellen, när de mörka mesonerna sönderfaller till t- och b-kvarkar, \(ttbb\) och \(tttb\), som sedan i sin tur sönderfaller till ett slutställande som innehåller exakt en elektron eller myon.

Även om de mörka mesonerna skulle skapas i kollisionerna så skulle det hända så pass sällan att signalen skulle vara svag. Dessutom varierar signalens attribut mycket med ändringar i modellens parametrar. Denna analys är den första som letar efter mörka mesoner vid LHC, och vi tittar därför på stora intervall för parametrarna. Varje kombination av parametervärdet kallas för en signalpunkt.


I slutändan används maximum likelihood-metoden för att skatta signalstyrkan i data och för att definiera en testvariabel. Ett hypotestest utförs därefter, där nollhypotesen är att data består endast av bakgrund och den alternativa hypotesen är att datan består av både signal och bakgrund.

Analysen har visat sig vara känslig för stora delar av modellens parameterrum. För varje signalpunkt görs ett separat hypotestest, och i inget av fallen kan vi förkasta nollhypotesen. För flera signalpunkter kan vi dock vända på testet, och kan utesluta mörka mesoners existens för stora delar av det undersökta parameterommet.

**Nya metoder för ökad känslighet**

till godtyckliga mängder data då den kan köras på obegränsat med beräkningsnoder samtidigt. I avhandlingen visas att den snabbt kan ta fram en region i vilken en statistisk analys skulle vara känslig för signalhypotesen.

Det andra är en modifiering av den oövervakade maskinlärningsmetoden UCluster. Den kan tränas för olika ändamål, t.ex. signalisolering, avvikelsedetektion och partikelidentifikation. Nyheten i den här avhandlingen är att metoden nu kan dela upp beräkningarna på många noder samtidigt, och därför kan appliceras på en godtycklig mängd data.

Det tredje har utförts av Eva Mayer under handledning av författaren, och visar hur en boosted decision tree-metod kan användas för att särskilja signal och bakgrund i analysen för mörka mesoner. Det är en preliminär studie, men ger lovande resultat för användningen av maskinlärningsmetoder i framtida upprepningar av analysen.
Author’s contributions

Part I of this thesis describes the work done by others needed to understand the rest of the thesis. As such, my only contribution to this part is the synthesis of the information.

All work in part II was done by the (small) dark meson analysis team. My own contributions were as follows:

In chapter 9 my contribution consisted of validating the event generation setup for the signal, as well as occasional comparisons between different recommended configurations. The description of the individual simulated samples are standard ATLAS internal texts. The text about $t\bar{t} + HF$ overlap removal was written by J. J. Heinrich, who implemented the procedure in the analysis framework.

In chapters 10 and 11 I was involved in the online event selection and in charge of the object definitions, definitions of discriminating variables, and the offline selection. I was also part of developing the analysis framework in which these selections were implemented, and in charge of running it over the samples described in 9. All work in chapter 12 was done by me. In chapter 13 I did all work except the implementation of the $t\bar{t} + HF$ correction in the final fit. In chapter 14 I was only in charge of producing samples and configuring our framework to produce systematic uncertainties. The text in this chapter is written by others in the analysis team and includes some standard ATLAS texts. Chapter 15 presents the work done by others in the analysis team.

All figures in chapters 8, 9, 10 and 15 are made by members of the analysis team other than me, unless otherwise specified. All others are made by me, unless otherwise specified, except figure 11.1 which was done by J.J. Heinrich.

The text in part II is heavily based off of the ATLAS internal documentation, for which I was the editor, but which contains contributions from other parts of the analysis team.

In part III, the projects have been done with different collaborators. My contributions were as follows:

In chapter 17 my contribution was identifying the possibility for this project within the analysis, and steering it through the major analysis choices. The work was done by master student Eva Mayer, whom I supervised. In chapter 16, my contributions were conceptualising the project, including finding both the method and the simulated data, as well as high energy physics domain knowledge and part of the coding. I also authored the paper, which was proof-read and edited by the co-authors. In chapter 18 I identified the use case for the method in high energy physics and coordinated the team. I supplied
the knowledge about the method including statistical and distributed computing aspects. I contributed 50% to the code and 25% to the paper. Finally, all work in chapter 19 is my own.

It should also be mentioned that any analysis done in ATLAS relies on the work of all other members of the collaboration. A lot of work is required to keep such a large experiment running and we all contribute to the running and share the results with each other. This is true for running the detector, developing trigger and object reconstruction algorithms, maintaining the software, studies of systematic uncertainties, studies of event generators and generating the events and many other common tasks. The thesis therefore includes references to different groups within ATLAS, and sometimes even snippets of text from the experts.
Acknowledgements

First, I would like to thank my supervisor, Rebeca Gonzalez Suarez, for your generosity with your time, your advice and the credit you give to others. Thank you also for the opportunities you have presented me with during my time at UU, and for always encouraging me to take ownership of my career and explore what interests me.

Next, I would like to thank Jochen Jens Heinrich, who has taught me everything I know about the technical side of ATLAS. Thank you for leading the dark meson analysis with such competency and optimism, and for creating such a good work environment in the group.

Thank you to Christina, who has walked every step of this journey beside me. I cannot imagine this time without you, your support has been invaluable and I am so proud of you.

Thank you to Giulia and Axel who joined the team late in my PhD but made it that much more enjoyable. Thank you also for your amazing effort on the Sparks in the dark paper, without you it would not exists.

Thank you to the rest of the dark meson group. To Galen Gledhill for your patience and kindness when I was a new PhD student, and for your friendship later on. To Federica Piazza, Tim Mathew, Anni Xiong and Stephanie Majewski for being amazing people to work with, and a special thank you to Federica and Jochen for working around the clock to finish the fit before my thesis deadline.

Thank you to Raazesh Saiinudiin for hours long discussions on statistics, and for always being up to something new. Thank you also for encouraging me to develop ideas into full projects.

Than you to Arnaud Ferrari for being endlessly encouraging and friendly, and for offering to extend my contract at UU to be able to finish my analysis. To Richard Brenner for encouraging me to do what I am interested in and allowing me to go to CERN even as a master student. To Elin Bergeås Kuutmann for your efforts to include everyone, and for being generous with your knowledge. To Mattias Ellert for your humor and impressive knowledge not only of computing, but everything from Russian to geography. To Max Isacson and Mikael Mårtensson for being such role-models when I started, and for adopting the two over-eager master students that were me and Christina

To Myrto Asimakopoulou for being the glue that held the group together back then, and for being so bad-ass in the lab, to Thomas Mathisen for many conversations over coffee that just might have saved my mental health, to Venu Ellajosyula for the friendship and for the support early on. Thank you to Petar Bokan and Serhat Ordek for being so funny, and so easy to talk to.
A big thank you to Jakob Beise for the TUFF project we worked on together, for your infectious enthusiasm and grit and for your friendship.

Thank you to Nora Valtonen-Mattila, Timea Vitos, Nils Heyer, An Di, Yong Sheng Koay, Michael Papenbrock, Marco Chiodaroli, Maja Olvegård, Rikard Enberg, Stefan Leupold, Ebba Ahlgren Cederlöf, Jana Rieger, Jenny Regina, Malin Bohman, Philipp Rincke, Gunnar Ingelman, Carlos Perez de los Heros, Bo Cao, Diego Turenne, Geoffrey Mullier and Jonas Steentoft for friendship and/or guidance.

Thank you also to Karl Bengtsson Bernander, Daniel Gedon and Colin Desmarais for the great collaboration on the UCluster project.

I am grateful to Shalu Solomon for being an ever supportive presence in my work even though we live far apart. Thank you also to the rest of the ECSB; Harish Potti, Ben Davis-Purcell, Dilia Portillo Quintero, Michele Faucci Giannelli, Giordon Stark, Max Swiatlowski, Laura Bruce, Henry Day-Hall, Hannah Arnold, for showing me life in the collaboration.

Thank you to all the speakers of the ATLAS Induction Day and the ATLAS Lecture Series for your valuable contributions to this thesis.

I want to thank Viktor Thorén, Timea Vitos and Johan Henriksson in particular for their valuable comments on parts of this thesis.

Thank you to Eva Mayer who was a great student and great collaborator on the BDT project, it was a pleasure working with you.

Thank you to my family; my mom for her endless patience with my patience, my dad for recognising my interest in science and computing, my siblings for inspiring me with their own academic journeys. Thank you to both of my parents for your unwavering support in everything I do and all situations I find myself in.

Thank you to my extended family on both sides, for inspiring me with your own excellence, being my safety net, and bringing me up to feel like I belong in professional spaces.

Thank you to all my friends, past and present, for always inspiring me. You are the reason I know what I want out of life.

And finally, thank you to Viktor, for your love and support, for lending your intellect and intelligence to this thesis, for your generosity in every aspect, and for giving me the life that I have. The thought of the rest of my life with you is enough to keep me going.
21. References

[3]


[29] Kim Albertsson, Piero Altoe, Dustin Anderson, John Anderson, Michael Andrews, Juan Pedro Araque Espinosa, Adam Aurisano, Laurent Basara,


[36] Nedaa Asbah, Simon Paul Berlendis, Markus Cristinziani, Frederic Deliot,


recommendations, 2011.
In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’16, pages 785–794, New York, NY, USA, 2016. ACM.


[95] Wikimedia Commons. File:standard model of elementary particles.svg — wikimedia commons, the free media repository, 2024. [Online; accessed 10-April-2024].


[111] Rikkert Frederix, Davide Pagani, and Marco Zaro. Large NLO corrections in $t\bar{t}W^\pm$ and $t\bar{t}t$ hadroproduction from supposedly subleading EW contributions. *JHEP*, 02:031, 2018.


[119] Galen Rhodes Gledhill, Olga Sunneborn Gudnadottir, Jochen Jens Heinrich, Stephanie Majewski, Rebeca Gonzalez Suarez, Giulia Ripellino, Anni Xiong, and Timothy Thankachen Mathew. Search for dark mesons decaying to top


[140] A. D. Martin, W.J. Stirling, R.S. Thorne, and G. Watt. Uncertainties on $\alpha_S$ in


[155] Michele Pinamonti, Maurizio De Santis, Salvatore Loefredo, Mohammed Faraj, Leonid Serkin, Jacopo Magro, Giancarlo Panizzo, and Lucio Cerrito. Search for $t\bar{t}$ resonances in the dilepton channel in 139 fb$^{-1}$ or $pp$ collisions at $\sqrt{s}=13$ TeV with the ATLAS experiment. ATL-COM-PHYS-2020-819, 2020.


[158] Raazesh Sainudiin and Gloria Teng. Minimum distance histograms with


Appendix A.
Full signal region cutflow table

The cutflow table for the Signal region can be found in A
<table>
<thead>
<tr>
<th>Process</th>
<th>Preselection</th>
<th>(N_{\text{jets}} \geq 5)</th>
<th>(\Delta R(l, b_2) &lt; 2.7)</th>
<th>(m_{bb\beta\gamma_{\text{miss}}} &gt; 100\text{GeV})</th>
<th>(s/b)</th>
<th>(s/\sqrt{(s+b)})</th>
</tr>
</thead>
<tbody>
<tr>
<td>ttlight</td>
<td>175081 ± 151 (100.00%)</td>
<td>129749 ± 130 (74.11%)</td>
<td>100835 ± 114 (77.72%)</td>
<td>46245 ± 77 (45.86%)</td>
<td>5.397e-02</td>
<td>1.644e+01</td>
</tr>
<tr>
<td>ttge1b</td>
<td>74674 ± 33 (100.00%)</td>
<td>63852 ± 30 (85.51%)</td>
<td>51085 ± 27 (80.01%)</td>
<td>22765 ± 18 (44.56%)</td>
<td>2.221e-02</td>
<td>6.766e+00</td>
</tr>
<tr>
<td>tt1c</td>
<td>55305 ± 84 (100.00%)</td>
<td>47831 ± 78 (86.49%)</td>
<td>38015 ± 69 (79.48%)</td>
<td>17809 ± 47 (46.85%)</td>
<td>2.234e-02</td>
<td>7.201e+00</td>
</tr>
<tr>
<td>Single top</td>
<td>12954 ± 37 (100.00%)</td>
<td>9405 ± 32 (72.60%)</td>
<td>6971 ± 27 (74.12%)</td>
<td>3582 ± 20 (51.38%)</td>
<td>1.255e-02</td>
<td>7.631e+00</td>
</tr>
<tr>
<td>ttPlusX</td>
<td>5234 ± 9 (100.00%)</td>
<td>4872 ± 9 (93.08%)</td>
<td>4096 ± 8 (84.06%)</td>
<td>1756 ± 5 (42.87%)</td>
<td>4.929e-02</td>
<td>1.093e+01</td>
</tr>
<tr>
<td>V + jets</td>
<td>5233 ± 46 (100.00%)</td>
<td>3911 ± 38 (74.74%)</td>
<td>2223 ± 28 (56.83%)</td>
<td>617 ± 13 (27.77%)</td>
<td>1.851e-02</td>
<td>4.284e+00</td>
</tr>
<tr>
<td>Multiboson</td>
<td>403 ± 7 (100.00%)</td>
<td>311 ± 7 (77.13%)</td>
<td>220 ± 6 (70.84%)</td>
<td>57 ± 3 (25.71%)</td>
<td>4.385e-03</td>
<td>1.103e+01</td>
</tr>
<tr>
<td>Total background</td>
<td>328884 ± 186 (100.00%)</td>
<td>259931 ± 162 (79.03%)</td>
<td>203445 ± 142 (78.27%)</td>
<td>92831 ± 96 (45.63%)</td>
<td>3.295e-02</td>
<td>1.083e+01</td>
</tr>
<tr>
<td>(SU(2)_L) 0.45 300</td>
<td>17482 ± 486 (100.00%)</td>
<td>15142 ± 450 (86.62%)</td>
<td>13897 ± 432 (91.77%)</td>
<td>5010 ± 262 (36.05%)</td>
<td>1.777e-02</td>
<td>1.375e+00</td>
</tr>
<tr>
<td>(SU(2)_L) 0.35 200</td>
<td>15604 ± 409 (100.00%)</td>
<td>10077 ± 328 (64.58%)</td>
<td>9313 ± 315 (92.42%)</td>
<td>2062 ± 151 (22.14%)</td>
<td>1.731e-02</td>
<td>2.936e+00</td>
</tr>
<tr>
<td>(SU(2)_L) 0.40 300</td>
<td>13406 ± 96 (100.00%)</td>
<td>11638 ± 89 (86.81%)</td>
<td>10787 ± 86 (92.69%)</td>
<td>4951 ± 58 (45.90%)</td>
<td>3.992e-02</td>
<td>1.333e+01</td>
</tr>
<tr>
<td>(SU(2)_L) 0.45 400</td>
<td>12610 ± 75 (100.00%)</td>
<td>11878 ± 72 (94.19%)</td>
<td>10967 ± 70 (93.33%)</td>
<td>5315 ± 49 (48.46%)</td>
<td>1.875e-02</td>
<td>1.247e+01</td>
</tr>
<tr>
<td>(SU(2)_L) 0.47 450</td>
<td>10195 ± 131 (100.00%)</td>
<td>9639 ± 127 (94.54%)</td>
<td>8901 ± 122 (92.35%)</td>
<td>4773 ± 90 (53.62%)</td>
<td>1.542e-02</td>
<td>1.620e+01</td>
</tr>
<tr>
<td>(SU(2)_L) 0.35 300</td>
<td>8466 ± 62 (100.00%)</td>
<td>7368 ± 58 (87.04%)</td>
<td>6750 ± 56 (91.60%)</td>
<td>3665 ± 41 (54.30%)</td>
<td>1.345e-02</td>
<td>1.841e+00</td>
</tr>
<tr>
<td>(SU(2)_L) 0.30 250</td>
<td>8390 ± 147 (100.00%)</td>
<td>6994 ± 134 (83.36%)</td>
<td>6193 ± 126 (88.55%)</td>
<td>2896 ± 86 (46.76%)</td>
<td>1.119e-02</td>
<td>1.133e+00</td>
</tr>
<tr>
<td>(SU(2)_L) 0.40 400</td>
<td>8266 ± 46 (100.00%)</td>
<td>7842 ± 45 (94.87%)</td>
<td>7279 ± 44 (92.82%)</td>
<td>4070 ± 33 (55.91%)</td>
<td>3.848e-02</td>
<td>1.336e+01</td>
</tr>
<tr>
<td>(SU(2)_L) 0.35 350</td>
<td>7535 ± 43 (100.00%)</td>
<td>7080 ± 42 (93.95%)</td>
<td>6498 ± 40 (91.78%)</td>
<td>3582 ± 30 (55.13%)</td>
<td>3.859e-02</td>
<td>1.176e+01</td>
</tr>
<tr>
<td>(SU(2)_L) 0.47 500</td>
<td>6829 ± 86 (100.00%)</td>
<td>6529 ± 84 (95.61%)</td>
<td>6051 ± 81 (92.67%)</td>
<td>3520 ± 62 (58.18%)</td>
<td>3.792e-02</td>
<td>1.151e+01</td>
</tr>
<tr>
<td>(SU(2)_L) 0.25 200</td>
<td>6555 ± 138 (100.00%)</td>
<td>4190 ± 110 (63.92%)</td>
<td>3573 ± 102 (85.29%)</td>
<td>869 ± 50 (24.32%)</td>
<td>9.361e-03</td>
<td>2.852e+00</td>
</tr>
<tr>
<td>(SU(2)_L) 0.45 500</td>
<td>5851 ± 102 (100.00%)</td>
<td>5576 ± 99 (95.30%)</td>
<td>5143 ± 95 (92.24%)</td>
<td>3055 ± 74 (59.41%)</td>
<td>3.291e-02</td>
<td>1.003e+00</td>
</tr>
<tr>
<td>(SU(2)_L) 0.30 300</td>
<td>4877 ± 35 (100.00%)</td>
<td>4251 ± 32 (87.16%)</td>
<td>3827 ± 31 (90.03%)</td>
<td>2359 ± 24 (61.63%)</td>
<td>2.541e-02</td>
<td>7.741e+00</td>
</tr>
<tr>
<td>(SU(2)_L) 0.47 550</td>
<td>4825 ± 59 (100.00%)</td>
<td>4635 ± 58 (96.06%)</td>
<td>4256 ± 55 (91.83%)</td>
<td>2711 ± 44 (63.69%)</td>
<td>2.920e-02</td>
<td>8.898e+00</td>
</tr>
<tr>
<td>(SU(2)_L) 0.35 400</td>
<td>4734 ± 26 (100.00%)</td>
<td>4520 ± 25 (95.47%)</td>
<td>4152 ± 24 (91.87%)</td>
<td>2617 ± 19 (63.02%)</td>
<td>2.819e-02</td>
<td>8.589e+00</td>
</tr>
<tr>
<td>(SU(2)_L) 0.25 250</td>
<td>4221 ± 72 (100.00%)</td>
<td>3442 ± 65 (81.54%)</td>
<td>2973 ± 60 (86.38%)</td>
<td>1570 ± 44 (52.82%)</td>
<td>1.692e-02</td>
<td>5.154e+00</td>
</tr>
<tr>
<td>(SU(2)_L) 0.30 350</td>
<td>4115 ± 52 (100.00%)</td>
<td>3855 ± 50 (93.69%)</td>
<td>3460 ± 47 (89.76%)</td>
<td>2194 ± 38 (63.40%)</td>
<td>2.364e-02</td>
<td>7.201e+00</td>
</tr>
</tbody>
</table>

Continued on next page
<table>
<thead>
<tr>
<th>SU(2)</th>
<th>0.45 550</th>
<th>4028 ± 48 (100.00%)</th>
<th>3872 ± 47 (96.13%)</th>
<th>3582 ± 46 (92.51%)</th>
<th>2396 ± 37 (66.88%)</th>
<th>2.581e-02</th>
<th>7.863e+00</th>
</tr>
</thead>
<tbody>
<tr>
<td>SU(2)</td>
<td>0.40 500</td>
<td>3611 ± 19 (100.00%)</td>
<td>3477 ± 19 (96.29%)</td>
<td>3231 ± 18 (92.93%)</td>
<td>2174 ± 15 (67.29%)</td>
<td>2.342e-02</td>
<td>7.136e+00</td>
</tr>
<tr>
<td>SU(2)</td>
<td>0.35 450</td>
<td>2963 ± 35 (100.00%)</td>
<td>2835 ± 34 (95.69%)</td>
<td>2629 ± 33 (92.72%)</td>
<td>1802 ± 27 (68.53%)</td>
<td>1.941e-02</td>
<td>5.913e+00</td>
</tr>
<tr>
<td>SU(2)</td>
<td>0.45 600</td>
<td>2899 ± 48 (100.00%)</td>
<td>2801 ± 47 (96.62%)</td>
<td>2569 ± 45 (91.72%)</td>
<td>1775 ± 37 (69.07%)</td>
<td>1.912e-02</td>
<td>5.825e+00</td>
</tr>
<tr>
<td>SU(2)</td>
<td>0.30 400</td>
<td>2471 ± 30 (100.00%)</td>
<td>2355 ± 29 (95.31%)</td>
<td>2130 ± 27 (90.44%)</td>
<td>1503 ± 23 (70.58%)</td>
<td>1.619e-02</td>
<td>4.934e+00</td>
</tr>
<tr>
<td>SU(2)</td>
<td>0.40 550</td>
<td>2397 ± 28 (100.00%)</td>
<td>2311 ± 27 (96.40%)</td>
<td>2153 ± 26 (93.17%)</td>
<td>1538 ± 22 (71.43%)</td>
<td>1.657e-02</td>
<td>5.048e+00</td>
</tr>
<tr>
<td>SU(2)</td>
<td>0.25 300</td>
<td>2229 ± 50 (100.00%)</td>
<td>1879 ± 46 (84.27%)</td>
<td>1621 ± 42 (86.31%)</td>
<td>1114 ± 35 (68.69%)</td>
<td>1.2e-02</td>
<td>3.655e+00</td>
</tr>
<tr>
<td>SU(2)</td>
<td>0.35 500</td>
<td>1921 ± 10 (100.00%)</td>
<td>1852 ± 10 (96.41%)</td>
<td>1706 ± 9 (92.08%)</td>
<td>1245 ± 8 (73.00%)</td>
<td>1.341e-02</td>
<td>4.087e+00</td>
</tr>
<tr>
<td>SU(2)</td>
<td>0.20 250</td>
<td>1622 ± 28 (100.00%)</td>
<td>1275 ± 25 (78.58%)</td>
<td>1060 ± 23 (83.14%)</td>
<td>589 ± 17 (55.55%)</td>
<td>6.343e-03</td>
<td>1.933e+00</td>
</tr>
<tr>
<td>SU(2)</td>
<td>0.45 700</td>
<td>1457 ± 24 (100.00%)</td>
<td>1419 ± 23 (97.41%)</td>
<td>1299 ± 22 (91.58%)</td>
<td>982 ± 19 (75.58%)</td>
<td>1.058e-02</td>
<td>3.223e+00</td>
</tr>
<tr>
<td>SU(2)</td>
<td>0.35 550</td>
<td>1250 ± 14 (100.00%)</td>
<td>1213 ± 14 (97.07%)</td>
<td>1128 ± 14 (92.93%)</td>
<td>854 ± 12 (75.76%)</td>
<td>9.202e-03</td>
<td>2.804e+00</td>
</tr>
<tr>
<td>SU(2)</td>
<td>0.25 400</td>
<td>1000 ± 17 (100.00%)</td>
<td>941 ± 16 (94.14%)</td>
<td>830 ± 15 (88.20%)</td>
<td>626 ± 13 (75.38%)</td>
<td>6.74e-03</td>
<td>2.054e+00</td>
</tr>
<tr>
<td>SU(2)</td>
<td>0.45 300</td>
<td>932 ± 7 (100.00%)</td>
<td>821 ± 6 (88.13%)</td>
<td>758 ± 6 (92.30%)</td>
<td>267 ± 4 (35.21%)</td>
<td>2.874e-03</td>
<td>8.757e+01</td>
</tr>
<tr>
<td>SU(2)</td>
<td>0.30 500</td>
<td>921 ± 12 (100.00%)</td>
<td>887 ± 12 (96.25%)</td>
<td>798 ± 11 (89.98%)</td>
<td>627 ± 10 (78.54%)</td>
<td>6.752e-02</td>
<td>2.057e+00</td>
</tr>
<tr>
<td>SU(2)</td>
<td>0.35 600</td>
<td>812 ± 13 (100.00%)</td>
<td>787 ± 13 (96.96%)</td>
<td>724 ± 12 (91.92%)</td>
<td>571 ± 11 (78.92%)</td>
<td>6.153e-03</td>
<td>1.875e+00</td>
</tr>
<tr>
<td>SU(2)</td>
<td>0.20 300</td>
<td>807 ± 13 (100.00%)</td>
<td>676 ± 12 (83.73%)</td>
<td>571 ± 11 (84.55%)</td>
<td>411 ± 9 (71.96%)</td>
<td>4.428e-03</td>
<td>1.349e+00</td>
</tr>
<tr>
<td>SU(2)</td>
<td>0.45 800</td>
<td>767 ± 12 (100.00%)</td>
<td>749 ± 12 (97.64%)</td>
<td>692 ± 12 (92.35%)</td>
<td>550 ± 10 (79.51%)</td>
<td>5.927e-03</td>
<td>1.806e+00</td>
</tr>
<tr>
<td>SU(2)</td>
<td>0.35 200</td>
<td>613 ± 22 (100.00%)</td>
<td>515 ± 20 (84.04%)</td>
<td>468 ± 19 (90.84%)</td>
<td>43 ± 6 (9.22%)</td>
<td>4.645e-04</td>
<td>1.415e-01</td>
</tr>
<tr>
<td>SU(2)</td>
<td>0.30 550</td>
<td>581 ± 7 (100.00%)</td>
<td>560 ± 6 (96.38%)</td>
<td>508 ± 6 (90.68%)</td>
<td>404 ± 6 (79.65%)</td>
<td>4.357e-03</td>
<td>1.327e+00</td>
</tr>
<tr>
<td>SU(2)</td>
<td>0.35 300</td>
<td>467 ± 9 (100.00%)</td>
<td>409 ± 9 (87.49%)</td>
<td>376 ± 8 (91.93%)</td>
<td>197 ± 6 (52.46%)</td>
<td>2.123e-03</td>
<td>6.47e-01</td>
</tr>
<tr>
<td>SU(2)</td>
<td>0.45 400</td>
<td>452 ± 9 (100.00%)</td>
<td>398 ± 8 (88.01%)</td>
<td>365 ± 8 (91.83%)</td>
<td>201 ± 6 (54.90%)</td>
<td>2.161e-03</td>
<td>6.585e-01</td>
</tr>
<tr>
<td>SU(2)</td>
<td>0.45 900</td>
<td>425 ± 7 (100.00%)</td>
<td>415 ± 7 (97.79%)</td>
<td>379 ± 6 (91.21%)</td>
<td>307 ± 6 (81.04%)</td>
<td>3.308e-03</td>
<td>1.008e+00</td>
</tr>
<tr>
<td>SU(2)</td>
<td>0.25 200</td>
<td>381 ± 11 (100.00%)</td>
<td>312 ± 10 (81.86%)</td>
<td>270 ± 9 (86.64%)</td>
<td>35 ± 3 (12.90%)</td>
<td>3.756e-04</td>
<td>1.144e-01</td>
</tr>
<tr>
<td>SU(2)</td>
<td>0.45 200</td>
<td>368 ± 28 (100.00%)</td>
<td>337 ± 26 (91.59%)</td>
<td>321 ± 26 (95.37%)</td>
<td>30 ± 7 (9.29%)</td>
<td>3.212e-04</td>
<td>9.785e-02</td>
</tr>
<tr>
<td>SU(2)</td>
<td>0.35 700</td>
<td>350 ± 6 (100.00%)</td>
<td>340 ± 6 (97.24%)</td>
<td>315 ± 5 (92.60%)</td>
<td>259 ± 5 (82.20%)</td>
<td>2.791e-03</td>
<td>8.503e-01</td>
</tr>
<tr>
<td>SU(2)</td>
<td>0.25 500</td>
<td>334 ± 6 (100.00%)</td>
<td>317 ± 5 (95.09%)</td>
<td>277 ± 5 (87.28%)</td>
<td>229 ± 5 (82.55%)</td>
<td>2.463e-03</td>
<td>7.505e-01</td>
</tr>
<tr>
<td>SU(2)</td>
<td>0.20 400</td>
<td>324 ± 4 (100.00%)</td>
<td>296 ± 4 (91.50%)</td>
<td>251 ± 4 (84.82%)</td>
<td>195 ± 3 (77.63%)</td>
<td>2.101e-03</td>
<td>6.402e-01</td>
</tr>
</tbody>
</table>

Continued on next page
| \(SU(2)_L\) | \(0.35\) | \(1100\) | \(17 \pm 0 (100.00\%)\) | \(17 \pm 0 (97.01\%)\) | \(16 \pm 0 (92.59\%)\) | \(14 \pm 0 (87.40\%)\) | \(1.463e-04\) | \(4.458e-02\) |
| \(SU(2)_L\) | \(0.25\) | \(800\) | \(15 \pm 0 (100.00\%)\) | \(14 \pm 0 (94.80\%)\) | \(12 \pm 0 (88.59\%)\) | \(11 \pm 0 (88.12\%)\) | \(1.157e-04\) | \(3.526e-02\) |
| \(SU(2)_R\) | \(0.15\) | \(400\) | \(12 \pm 0 (100.00\%)\) | \(11 \pm 0 (89.22\%)\) | \(10 \pm 0 (88.90\%)\) | \(8 \pm 0 (77.58\%)\) | \(8.09e-05\) | \(2.465e-02\) |
| \(SU(2)_L\) | \(0.15\) | \(700\) | \(9 \pm 0 (100.00\%)\) | \(8 \pm 0 (96.55\%)\) | \(8 \pm 0 (91.18\%)\) | \(6 \pm 0 (79.86\%)\) | \(6.501e-05\) | \(1.981e-02\) |
| \(SU(2)_L\) | \(0.25\) | \(900\) | \(5 \pm 0 (100.00\%)\) | \(5 \pm 0 (94.97\%)\) | \(5 \pm 0 (89.75\%)\) | \(4 \pm 0 (89.06\%)\) | \(4.449e-05\) | \(1.356e-02\) |
| \(SU(2)_L\) | \(0.15\) | \(800\) | \(4 \pm 0 (100.00\%)\) | \(4 \pm 0 (96.96\%)\) | \(4 \pm 0 (92.11\%)\) | \(3 \pm 0 (83.76\%)\) | \(3.467e-05\) | \(1.056e-02\) |
| \(SU(2)_R\) | \(0.15\) | \(500\) | \(4 \pm 0 (100.00\%)\) | \(4 \pm 0 (91.63\%)\) | \(3 \pm 0 (90.26\%)\) | \(3 \pm 0 (83.40\%)\) | \(3.08e-05\) | \(9.383e-03\) |
| \(SU(2)_L\) | \(0.25\) | \(1000\) | \(2 \pm 0 (100.00\%)\) | \(2 \pm 0 (94.44\%)\) | \(2 \pm 0 (88.80\%)\) | \(2 \pm 0 (88.65\%)\) | \(1.697e-05\) | \(5.169e-03\) |

Continued on next page
Appendix B.  
Variables studied in the 1-lepton channel

Section 12.1 describes how the discriminating variables used to define the signal region were chosen. The full list of variables studied at this step are listed below.

- $\mathcal{J}_{lep}$ and $\mathcal{J}_{had}$
- $m_{\mathcal{J}_{had}} + m_{\mathcal{J}_{lep}}$
- $H_T$
- $\Delta R(l, b_2)$
- $m_{bbs_{\text{min}}}$
- $\mathcal{J}_8^{\mathcal{lep}}$ and $\mathcal{J}_8^{\mathcal{had}}$: jets reclustered with $R = 0.8$ with the inclusion of a lepton as described in 10.4. $\mathcal{J}_8^{\mathcal{lep}}$ contains the lepton, and $\mathcal{J}_8^{\mathcal{had}}$ is the leading all hadronic reclustered jet
- $m_{\mathcal{J}_8^{\mathcal{had}}} + m_{\mathcal{J}_8^{\mathcal{lep}}}$
- $\Delta R$ between the leading b-tagged jet and the lepton
- Minimum invariant mass of any two b-tagged jets in the event
- Minimum $\Delta R$ between any two b-tagged jets in the event
- The invariant mass of $\mathcal{J}_{lep}$ and the closest b-tagged jet
- The invariant mass of $\mathcal{J}_8^{\mathcal{lep}}$ and the closest b-tagged jet

$$m_T = \sqrt{2p_TlE_T^{\text{miss}}(1 - \cos(\Delta \phi(l, p_T^{\text{miss}})))}$$ where $l$ indicates the property is associated with the lepton.

Figure 21.1 shows plots of the variables above for all backgrounds and benchmark signals after selecting exactly one tight lepton (vetoing events with additional loose leptons), $N_{\text{jets}} \geq 4$ and $N_{b-\text{jets}} \geq 2$. 
Figure 21.1. Distributions in 4 benchmark signal samples and the total background of the variables studied when defining the signal region. All distributions are normalised to 1.
Acta Universitatis Upsaliensis
*Uppsala Dissertations from the Faculty of Science*
Editor: The Dean of the Faculty of Science


32. **Kjell Janson**: Experimental investigations of the proton and deuteron structure functions. 1991.
34. **Jan Blomgren**: Experimental Studies of Giant Resonances in Medium-Weight Spherical Nuclei. 1991.

Prior to January 1994, the series was called *Uppsala Dissertations from the Faculty of Science*.

**Acta Universitatis Upsaliensis**

*Uppsala Dissertations from the Faculty of Science and Technology*

Editor: The Dean of the Faculty of Science

64. **Pär Samuelsson**: Control of Nitrogen Removal in Activated Sludge Processes. 2005.
69. Fredrik Robelius: Giant Oil Fields – The Highway to Oil. Giant Oil Fields and their Importance for Future Oil Production. 2007.
70. Anna Davour: Search for low mass WIMPs with the AMANDA neutrino telescope. 2007.
77. Samson Keleta: Double Pion Production in \(dd \rightarrow \alpha \pi \pi\) Reaction. 2008.
82. Arvid Pohl: Search for Subrelativistic Particles with the AMANDA Neutrino Telescope. 2009.
85. Henrik Petrén: \(\eta\) Meson Production in Proton-Proton Collisions at Excess Energies of 40 and 72 MeV. 2009.
88. Sophie Grape: Studies of PWO Crystals and Simulations of the \(\bar{p}p \rightarrow \bar{\Lambda}, \bar{\Xi}^{0}\) Reactions for the PANDA Experiment. 2009.
100. *Patrik Adlarson*. Studies of the Decay $\eta \rightarrow \pi^+\pi^-\pi^0$ with WASA-at-COSY. 2012.
118. *Li Caldeira Balkeståhl*. Measurement of the Dalitz Plot Distribution for $\eta \rightarrow \pi^+\pi^-\pi^0$ with KLOE. 2015.
137. Elisabeth Unger. The Extremes of Neutrino Astronomy. From Fermi Bubbles with IceCube to Ice Studies with ARIANNA. 2019.
149. Thomas Mathisen. Exotic Decays of Vector-Like Quarks and Development of a Test Procedure for the ITk Strip Module at the ATLAS Detector for the HL-LHC. 2022.