UNIVERSITÉ DE GENÈVE
Département de physique nucléaire et corpusculaire

FACULTÉ DES SCIENCES
Professeur Tobias Golling

A weakly supervised search for resonant new physics in the dijet final state using $139\,\mathrm{fb}^{-1}$ of pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector.

THÈSE

présentée à la Faculté des sciences de l'Université de Genève
pour obtenir le grade de Docteur ès sciences, mention physique

par

Samuel Klein

de Nouvelle-Zélande

Thèse N° 5893

GENÈVE
Atelier d'impression ReproMail
2024

UNIVERSITÉ
DE GENÈVE

FACULTÉ DES SCIENCES

DOCTORAT ÈS SCIENCES, MENTION INTERDISCIPLINAIRE

Thèse de Monsieur Samuel KLEIN

intitulée :

«A Weakly Supervised Search for Resonant New Physics in the Dijet Final State Using 139 fb^{-1} of *pp* collisions at $\sqrt{s}$ = 13 TeV with the ATLAS detector.»

La Faculté des sciences, sur le préavis de

Monsieur T. GOLLING, professeur ordinaire et directeur de thèse
Département de physique nucléaire & corpusculaire

Monsieur S. VOLOSHYNOVSKIY, professeur ordinaire et codirecteur de thèse
Département d'informatique

Monsieur G. IACOBUCCI, professeur ordinaire
Département de physique nucléaire & corpusculaire

Monsieur M. KAGAN, docteur
SLAC National Accelerator Laboratory, California, United States

autorise l'impression de la présente thèse, sans exprimer d'opinion sur les propositions qui y sont énoncées.

Genève, le 20 février 2025

Thèse - 5893 -

La Doyenne

N.B. - La thèse doit porter la déclaration précédente et remplir les conditions énumérées dans les "Informations relatives aux thèses de doctorat à l'Université de Genève".

Contents

Abstract

The Large Hadron Collider at CERN has collected a vast amount of data at the highest energies ever achieved in a laboratory setting. This data has allowed for the discovery of the Higgs boson and the precise measurements of properties of the Standard Model of particle physics. No new particles have been discovered beyond the Standard Model, and despite an impressive physics program, there remain many extensions that have not been excluded. Many extensions to the Standard Model predict the existence of new particles that decay into a pair of visible particles, which would appear as a peak in the invariant dijet mass spectrum. The standard approach to searching for such resonances is to perform a bump hunt, where the invariant mass spectrum is examined for a peak. This approach is limited in its sensitivity as it only utilizes the information contained in the invariant mass spectrum.

This thesis presents a search that extends the bump hunt by including up to six additional variables that are sensitive to the presence of physics beyond the Standard Model. Narrow width new physics models are searched for in the dijet invariant mass spectrum using 139 fb^{-1} of proton-proton collision data collected by the ATLAS detector at a center-of-mass energy of $\sqrt{s} = 13$ TeV. Events that contain at least two large radius jets at high transverse momentum are selected. The search uses two different interpolation techniques for producing background estimates over the six additional variables. This search is the first of its kind and thus requires extensive development and validation of the analysis strategy. The development of the analysis, and the challenges that were faced, are laid out in detail. Limits are set on the production of twenty different possible new physics processes, many of which have not been studied by any other analysis. Many of the challenges that need to be addressed in future iterations of such a search are also discussed in detail, as well as avenues for interpreting the results of such a search.

In addition, studies using machine learning techniques to improve different aspects of high energy physics analyses are presented.

Résumé

Le Grand collisionneur de hadrons du CERN a recueilli une grande quantité de données aux énergies les plus élevées jamais atteintes en laboratoire. Ces données ont permis de découvrir le boson de Higgs et de mesurer avec précision les propriétés du modèle standard de la physique des particules. Aucune nouvelle particule n'a été découverte au-delà du modèle standard et, malgré un programme de physique impressionnant, de nombreuses extensions n'ont pas été exclues. De nombreuses extensions du modèle standard prédisent l'existence de nouvelles particules qui se désintègrent en une paire de particules visibles, ce qui apparaîtrait comme un pic dans le spectre de masse invariant des jets. L'approche standard pour rechercher de telles résonances consiste à effectuer une chasse aux bosses, c'est-à-dire à examiner le spectre de masse invariant à la recherche d'un pic. Cette approche est limitée dans sa sensibilité car elle n'utilise que l'information contenue dans le spectre de masse invariant.

Cette thèse présente une recherche qui étend la chasse aux bosses en incluant jusqu'à six variables supplémentaires qui sont sensibles à la présence de physique au-delà du Modèle Standard. Des modèles de nouvelle physique à largeur étroite sont recherchés dans le spectre de masse invariant des jets en utilisant 139 fb^{-1} de données de collisions proton-proton collectées par le détecteur ATLAS à une énergie de centre de masse de $\sqrt{s} = 13$ TeV. Les événements qui contiennent au moins deux jets à grand rayon et à grand moment transverse sont sélectionnés. La recherche utilise deux techniques d'interpolation différentes pour produire des estimations du bruit de fond sur les six variables supplémentaires. Cette recherche est la première du genre et nécessite donc un développement et une validation approfondis de la stratégie d'analyse. Le développement de l'analyse et les défis à relever sont décrits en détail. Des limites sont fixées pour la production de vingt processus de nouvelle physique possibles, dont beaucoup n'ont été étudiés par aucune autre analyse. De nombreux défis à relever dans les futures itérations d'une telle recherche sont également discutés en détail, ainsi que des pistes pour l'interprétation des résultats d'une telle recherche.

En outre, des études utilisant des techniques d'apprentissage automatique pour améliorer différents aspects des analyses de la physique des hautes énergies sont présentées.

Acknowledgements

I would like to first thank my supervisor Tobias Golling, particularly for giving me the freedom and support to pursue my research interests over the course of my PhD. I would also like to thank Slava Voloshynovskiy for many interesting discussions throughout my PhD.

I am particularly grateful to Matthew Leigh, for always being motivated to discuss whatever we might be working on, for investing the time into figuring out how to nicely structure projects, for your attention to detail, and for enthusiastically sharing whatever you learn as you learn it. I would also like to thank Bálint Máté for lots of interesting chats over the years, and for making the one project we managed to finish together a lot of fun. I also need to thank Debajyoti Sengupta for joining me on the Curtains projects. Likewise, I need to thank Johnny Raine, Sebastian Pina-Otey and Knut Zoch for their help throughout my PhD. Also my thanks to the group at the University of Geneva that has made for a fun office environment over the years: Tomke, Malte, Antii, Mario, Vilius, Carlos, Leon, Sarah, Stephen, Jona, Franck, Kinga, Chris, Alex and Lukas.

From the analysis team, I particularly need to thank Kees Benkendorfer, for being a great collaborator and for sticking this analysis out with me. It has been a pleasure working with you. I also would like to thank Benjamin Nachman for his help and advice throughout the analysis, and also Dennis Noll and Sascha Diefenbacher for helping to bring it to a close.

I would like to thank Chris Scheulen, Alexander Froch, Knut Zoch, Kees Benkendorfer and Vilius Cepaitis for reading and providing feedback on this thesis when they really did not have to. I would also like to thank Steven Schramm for very helpful discussions on the topic of this thesis. My thanks to Michael Kagan and Lukas Heinrich for being generous with their time and advice, and for many interesting discussions.

Thank you also to Irene Balboni, for being the best, and trying to talk me out of my worst tendencies. Also Mixue Tan, Koen Kramer, Lorina Blackburne, Matteo Turchetta, Thérèse Obrist, Maëlle Zanone, Isotta and Giulio Degano for being great friends. Also Iana, Vincent, Reka and Caroline for making learning how to dance so fun. Special thanks also to Hadley Rax, Eddy Shearer, Daniel Vincent, Callum Orr, Louis Harknett, Oliver Dean, Guy Le Noel, Tim Driver, Oliver Stewart, Joe Highton, Andre Embury, Becki Enright, Hannah Wright.

Finally, I would like to thank my family Alexandra, Mum, Dad, Ben, Matt, Lill, Georgina (yes, in that order) for their support.

List of Figures

List of Tables

Chapter 1

Introduction

Modern physics has successfully built predictive models at radically different scales. At large distances, General Relativity successfully explains the motion of planets and other celestial bodies, the formation of galaxies, and even predicted the existence of gravitational waves [1]. At ultra-cold temperatures, Bose-Einstein condensates are well described by quantum mechanics [2]. At ultra-relativistic speeds massive particles are well described by the Standard Model of particle physics [3–13]. However, the dream of a single Grand Unified Theory of everything at all scales still eludes us.

To date, arguably the most successful theory is the Standard Model, which describes three of the four known fundamental forces and all forms of matter that can currently be observed directly. However, there are many observed phenomena the Standard Model does not explain. For example, the Standard Model can not describe gravity, does not account for the existence of dark matter [14], does not explain the matter-antimatter asymmetry in the universe [14] or neutrino oscillations [15]. Given these limitations, there is a large physics program searching for new physics beyond the Standard Model.

In particular, the Large Hadron Collider at CERN, the European Organization for Nuclear Research, has collected data at the highest energy scales ever achieved in a particle accelerator. These data have been used, for example, to discover the Higgs boson [16, 17], and to search for new physics beyond the Standard Model. However, no new physics has been discovered at the Large Hadron Collider, and there remains the tantalizing possibility that unknown physics processes could be hidden in datasets that have already been collected. Such physics could take the form of new particles or even new forms of matter which require fundamentally different descriptions. It is not known if such processes are accessible at the scales probed by the Large Hadron Collider, and there is no unambiguous guide for the energy scales at which they become relevant.

There are a plethora of possible new phenomena that could exist within the reach of current physics experiments and there is no clear guide for where to search for new physics. This thesis presents an attempt to search for new physics processes with only minimal assumptions about its nature. Specifically, this thesis presents a weakly supervised search for resonant new physics using data recorded with the ATLAS experiment at the Large Hadron Collider at CERN. This thesis also presents the development of multiple machine learning methods, mostly with application to the field of High Energy Physics.

This thesis is structured as follows: Chapter 2 of this thesis briefly introduces the Standard Model of particle physics and the concepts that are relevant to the work in this thesis. Chapter 3 presents the experimental setup used to collect the data analysed in this thesis, providing some brief background on the Large Hadron Collider and the ATLAS experiment. Chapter 4 provides background on the statistics used in physics searches as relevant for the analysis presented in this thesis. Chapter 5 discusses machine learning and method developments performed as a part of this thesis. Chapter 6 introduces the concept of a weakly supervised search, the type of search pursued in this thesis. Chapter 7 introduces the analysis strategy performed in this thesis, related work and the dataset that is analysed. Chapter 8 presents the design choices made for the analysis and the rationale behind them. Chapter 9 describes how the analysis was developed, and the challenges faced in this context. This is relevant as this is the first search of this type performed at the Atlas experiment. Chapter 10 presents the results of the analysis on the final validation set and the recorded dataset. Chapter 11 contains discussion around the successes and failures of this analysis, some general conclusions about weakly supervised approaches and the outlook for future work.

Part I

Theory and Experiment

Chapter 2

Standard Model of Particle Physics

The Standard Model (SM) of particle physics [3–13] is one of the most successful physical theories ever constructed. The SM describes all known fundamental particles and all forces except gravity. In the SM matter is split into half-integer spin fermions and integer spin bosons. Fermions form the fundamental building blocks of matter and the gauge bosons mediate the fundamental forces. In the SM all fermions have associated anti-particles that are identical in mass but carry opposite quantum numbers. All particles of the SM, and the hypothesized gravity mediating graviton, are shown in Figure 2.1.

The SM is the most general dimension four local quantum field theory (QFT) with symmetry groups $SU(3)_C \times SU(2)_L \times U(1)_Y$. Each of these groups describes a different interaction, where $SU(3)_C$ is the gauge group for the strong interaction (color charge), $SU(2)_L$ is the gauge group for the weak interaction (left-handed weak isospin) and the admixture of $SU(2)_L$ and $U(1)_Y$ leads to electromagnetism. The SM Lagrangian is given by,

$$\mathcal{L}_{\text{SM}} = -\frac{1}{4}F_{\mu\nu}F^{\mu\nu} + i\bar{\psi}\gamma_\mu D^\mu\psi + (y_{ij}\bar{\psi}_i\phi\psi_j + \text{h.c.}) + |D_\mu\phi|^2 - V(\phi). \tag{2.1}$$

The term $-\frac{1}{4}F_{\mu\nu}F^{\mu\nu}$ represents the kinetic term for each of the gauge bosons and is defined by,

$$F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu - ig[A_\mu, A_\nu], \tag{2.2}$$

where A_μ is the gauge field, g is the coupling constant for the interaction and $[A_\mu, A_\nu]$ is the commutator of the gauge fields. In the SM the gauge fields are the gluons $G^a_{\mu\nu}$, the weak bosons $W^i_{\mu\nu}$ and the weak hypercharge boson B_μ. The term $i\bar{\psi}\gamma_\mu D^\mu\psi$ is the Dirac term, accounting for the kinetic energy and interactions of the fermions ψ with the gauge fields through the covariant derivative D^μ. The covariant derivative combines the spacetime derivative with the gauge interactions, with each gauge field coupling to fermions according to their respective charges. The Yukawa interactions between the fermions and the Higgs field ϕ are described by $y_{ij}\bar{\psi}_i\phi\psi_j$; h.c. is the Hermitian conjugate of this term. The kinetic term for the Higgs field is given by $|D_\mu\phi|^2$, and $V(\phi)$ represents the Higgs potential. Together, this Lagrangian describes the complete SM, describing the interactions and dynamics of fundamental particles.

In the high energy limit, which is reached by the LHC, predictions can be made with this Lagrangian using perturbation theory. A useful technique for writing down the contributions to a given process, at a given order in perturbation theory, is provided by Feynman diagrams and the corresponding Feynman rules. The SM Lagrangian gives rise to the vertices shown in Figure 2.2 as well as the conjugates of these diagrams. Time flows left to right in these diagrams, and particles with arrows pointing backward in time represent anti-particles. These

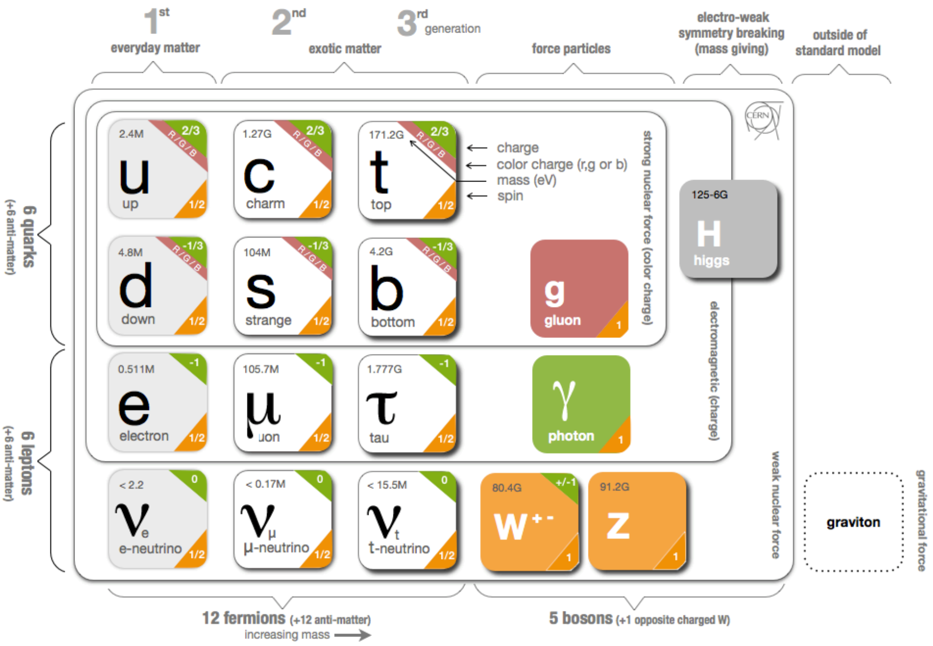


Figure 2.1: The fundamental particles described by the Standard Model of particle physics from Purcell [18].

vertices are determined by the terms in eq. (2.1) and provide a succinct visual summary of the direct particle interactions that can occur in the SM.

2.1 Electroweak theory

For this thesis, the weak and electromagnetic interactions only play a minor role and so are only described briefly. The unified electroweak theory describes weak and electromagnetic interactions in the SM. This theory was combined in the Glashow-Weinberg-Salam theory building on Fermi and quantum electrodynamics (QED) [3, 5]. The theory corresponds to the $SU(2)_L \times U(1)_Y$ symmetry group and an interesting feature of the theory is that due to gauge invariance none of the force-carrying fields can have mass. In the SM these fields (and quarks) gain mass through their interaction with the Higgs boson in the Brout-Englert-Higgs mechanism and spontaneous symmetry breaking [20–23]. In eq. (2.1) the appearance of mass is due to the potential $V(\phi)$ which is defined in such a way the vacuum expectation value for the Higgs field ϕ is not at zero. The discovery of the Higgs boson in 2012 [16, 17], first predicted in 1964, was a major achievement of the SM and CERN.

2.2 Quantum chromodynamics

The strong force describes the interactions of color charged particles using quantum chromodynamics (QCD), a non-abelian gauge theory with the $SU(3)_C$ symmetry group [24–26]. The parts of this theory that are relevant to this thesis are described in some detail in the following. There are eight different color charged gluons in QCD that mediate the strong force. The fact the force mediators in QCD are themselves color charged has interesting implications. In particular, the strong force coupling constant α_s of QCD runs in such a way that it is smaller at high energies and grows at low energies. The implication of this is that at low energies QCD is strongly interacting and at high energies it is weak. This leads to the 'confinement' principle, where color charged particles can not be observed in

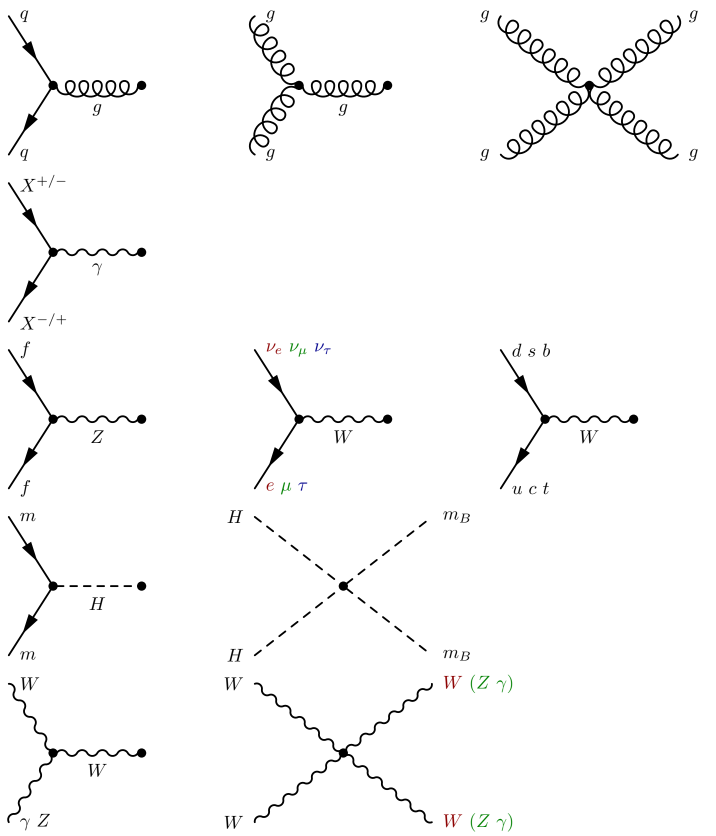


Figure 2.2: The Feynman vertices of the SM broken down by the different forces as shown in Lindon [19]. Here $X^{\pm}$ is any charged particle, m is any particle with mass and m_B is any massive boson. The q can be any quark and the f can be any fermion. All other particles are defined in Figure 2.1. Diagrams with colored labels must have two particles of the same color on the relevant legs.

isolation. It also leads to 'asymptotic freedom', where color charged particles are weakly coupled at high energies. Observable composite matter such as baryons (three quarks or three antiquarks), mesons (quark-antiquark pairs), and more exotic states such as tetraquarks (two quarks and two antiquarks) and pentaquarks (four quarks and one antiquark) are all colorless combinations of quarks and gluons referred to as hadrons.

2.2.1 Jets

QCD is a rich physical theory and its fundamental description leads to important considerations in terms of what can be experimentally observed. For example, the confinement principle means that quarks and gluons can not be observed in isolation. Instead, bound states formed from these particles and other decay products are observed, from which the initiating particles' properties can be inferred. Due to the nature of QCD, color charged particles decay into collimated sprays of particles called jets. Jets are a fundamental object in QCD and are a key object in collider experiments. The definition of a jet is ambiguous as it exists as a concept in both perturbative and non-perturbative regimes. This thesis presents an analysis of a dataset of jets and so these objects are described in some detail. A later section describes how jets are reconstructed in a detector, this chapter is concerned with their description in QCD. Jets form principally due to three properties of QCD as outlined in the following.

At high energies, the strong coupling α_s is small, which means the constituents of a proton can be treated as point-like particles. These point-like particles are the quarks and gluons that are the fundamental building blocks of the proton and are referred to as partons. Perturbative series in QCD contain divergences in the collinear limit, which means that any given parton is likely to decay into multiple parallel particles. For example, a quark can split its energy into a gluon and an anti-quark, and due to the collinear divergence, these particles are likely to be emitted in the same direction as the parent quark. The decay products can further radiate in a cascade of collimated particles. This means QCD predicts that a parton produced in a collision at the LHC decays into a stream of partons. This stream is called a jet and is described by parton showering, it occurs at scales that are accessible with perturbation theory.

Beyond the collinear divergence, it is also relevant the strong coupling α_s is small at high energies. If this coupling was large then both incident and outgoing partons would emit radiation in all directions. There would still be a preference for radiation to be emitted collinear to the parent particle, but in addition to this, high momenta emissions in any direction would be likely. A large value for α_s would lead to spherically distributed decay products.

The final reason that jets form is the gluons carry color charge. To understand this, consider a collision that produces two quarks propagating in opposite directions in the detector. Due to the color charge of gluons, the field lines connecting these two quarks form a 'color tube'. The energy density of this tube increases as the quarks become separated, and at a certain point, it becomes energetically favourable to break the tube by producing a light quark anti-quark pair. This leads to the creation of two mesons that no longer strongly interact. This process is not fully understood from first principles as it occurs at non-perturbative scales. The description provided here is referred to as the Lund string model [27].

A final important feature of QCD is the soft divergence. This divergence means that QCD partons have a high chance of emitting low momentum radiation, which can further split and even form jets. Due to the collinear divergence, this is again most likely to be parallel to the quark, but it can be emitted in any direction. Therefore, it also makes sense to think of quarks as being surrounded by a 'soft haze' of gluons. These emissions can also occur in initial state radiation.

2.2.2 Parton distribution functions

Perturbation theory in QCD is performed in powers proportional to α_s. Due to the running of α_s, perturbation theory in QCD is only valid at high energies. At the energy scales probed by the LHC, the coupling is small enough to treat quarks and gluons as point like particles, which are referred to as partons [28]. A proton at the LHC is composed of a down quark and two up quarks, all three of which are referred to as valence quarks, which are bound together by gluons. The gluons can spontaneously decay to two quarks which recombine into a gluon, these are referred to as 'sea quarks'. Each of the partons in a proton carries a certain fraction of the total momentum of the proton referred to as the Bjorken x_i. The probability of having a particle with a certain x_i is given by parton distribution functions (PDFs). These distributions can not be calculated and instead are extracted from dedicated measurements. An example PDF set is shown in Figure 2.3 as a function of x at a momentum transfer of $Q^2 = 10$ GeV^2 and $Q^2 = 10^4$ GeV^2 [29].

The PDFs depend on the energy scale of the interaction and are extracted from data at a given scale. The PDFs are then evolved to different scales using the Dokshitzer-Gribov-Lipatov-Altarelli-Parisi (DGLAP) equations [30–32].

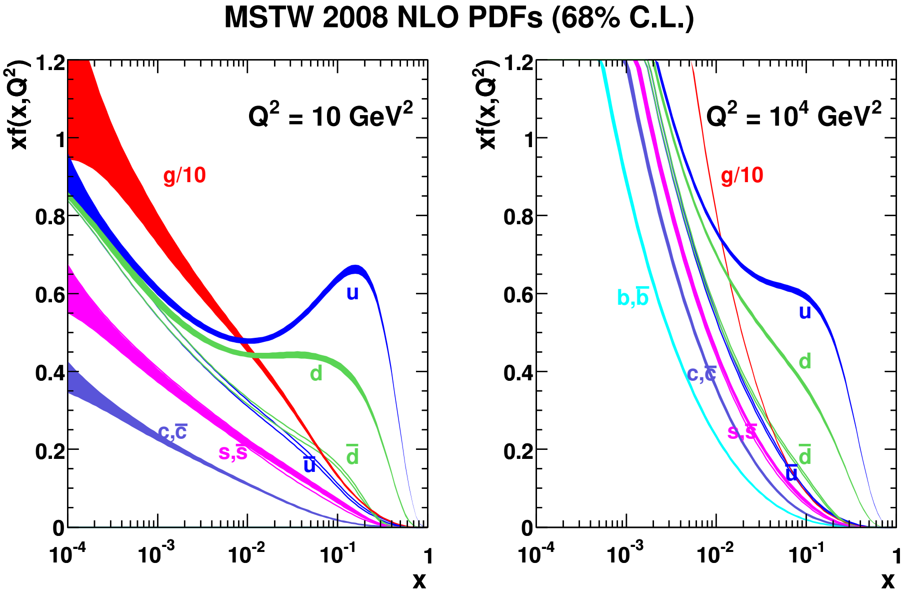


Figure 2.3: The parton distribution functions of the proton as a function of x at a scale of (left) $Q^2 = 10$ GeV^2 and (right) $Q^2 = 10^4$ GeV^2 as shown in Ref. [29].

While the PDFs are inherently non-perturbative, they can be understood in terms of the general principles of QCD. Low energy probes of protons mostly resolve the valence quarks. This is why at low Q^2 sharp peaks can be seen in the PDFs for the valence quarks and a broad distribution in the sea quarks. As Q^2 increases the peaks in the valence quarks broaden and the sea quarks become more prominent. There is an intrinsic order in the PDFs, where amongst the quarks the up quark is the most likely as the lightest and most common in a proton. Then amongst the other partons, the PDFs follow the mass hierarchy from the SM, where heavier particles are less likely to be produced and therefore are less likely to interact in a collision. Other than the masses of the quarks, there is no fundamental physical scale in QCD. In the PDFs this is why the virtual sea quarks can have any x_i, but the likelihood of having a given x_i falls according to a power law distribution. As discussed later in this thesis, the fact the PDFs have this power law distribution also means the mass of the combined system in a hard scatter event, an important quantity for this analysis, also smoothly falls as a function of x_i. While it is understood that PDFs are smooth functions, the exact form of the PDFs is not known and must be extracted from data. The NNPDF collaboration uses neural networks to fit the PDFs [33].

2.3 s-channel resonances

One channel through which particles can interact in a QFT is by annihilating and forming an intermediate virtual particle that then decays, this is referred to as an s-channel decay. The mediating virtual particle can be anything the incident particles couple to, and at higher orders in perturbation theory there are contributions from vacuum fluctuations of the mediators, called loops. If the total energy of the incident particles is the same as the mass of particle X then the contribution of X to this process increases. Therefore, at the mass of a particle, there is a peak in the cross section for the production of that particle, this is referred to as an s-channel resonance. Two particles can alternatively interact when one particle emits a mediator that then interacts with the other particle. These processes are referred to as t- and u-channel processes. A t-channel process is more likely to occur in QCD due to the gluon haze that surrounds and binds the valence quarks. It is also interesting to note that s-channel resonances at the LHC are more likely to occur through quark-gluon and gluon-gluon fusion as protons do not contain valence antiquarks and quark-antiquark fusion requires a sea antiquark to be produced. In QCD these are the only first order s-channel processes that can occur at the LHC as shown in Figure 2.2. A sea antiquark is less likely to carry a large fraction of the proton's momentum than a gluon or valence quark as shown in Figure 2.3.

2.4 Theory to experiment

In connecting theory and experiment it is important to identify what can be measured and practically predicted. One spectacular success of both theory and experiment is the match between predicted and observed cross sections for a wide array of different processes as shown in Figure 2.4. These processes can decay into multiple different final states. These channels may be characterized by the presence of electrons, muons, taus, photons or neutrinos (missing transverse energy). Experimentally it is important to be able to distinguish between these different components. Theoretical predictions for these processes are made using the SM Lagrangian and the Feynman rules. The cross sections for these processes are calculated using perturbation theory. A later section describes how these cross sections are measured at the LHC.

2.5 Beyond the Standard Model

Searches for new physics Beyond the SM (BSM) are well motivated by astrophysical observations, along with the inability of the SM to account for neutrino masses. There are also general considerations, such as the notion of the naturalness of the Higgs mass, that point to the existence of new physics within the reach of the LHC [35]. The overwhelming process at the LHC is QCD and any new physics processes are likely to be buried in this background. The abundance of the background causes issues for detecting new physics effects as discussed in detail later in this thesis. Theoretical model building of possible BSM physics,

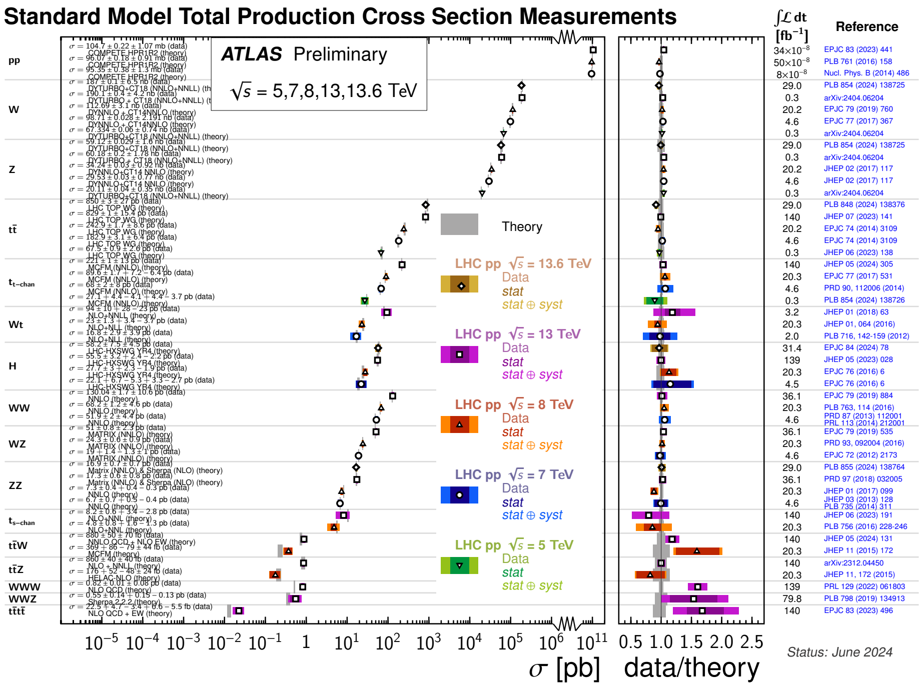


Figure 2.4: Various total production cross section measurements in the SM performed by ATLAS and compared to the associated theoretical predictions as shown in Ref. [34].

and characterizing decays within these models, is essential for discovering new physics. However, there are many possible BSM models, and it is also possible that unforeseen new physics scenarios exist. This is one of the motivations for the style of search presented in this thesis. The analysis in this thesis focuses on a dataset of jets. There are many BSM scenarios that produce jets and these objects are an excellent probe of our understanding of QCD.

Most new physics models have many free parameters, and it is not possible to test all of these parameters. Instead, simplified models are used, which contain only a few parameters but provide a rich phenomenology. For example, the Heavy Vector Triplet model [36] is a simplified model that contains a heavy vector boson that decays into two bosons. This model can represent a wide range of BSM physics scenarios and is used in this thesis to test the sensitivity of the analysis to new physics effects. These models predict the appearance of a resonance in the invariant mass spectrum of the decay products, which smoothly falls in the predominantly QCD background. This property can be leveraged to search for new physics effects, as discussed later in this thesis.

Chapter 3

Experimental set up

The data analysed in this thesis was collected by the Atlas detector [37] at the Large Hadron Collider (LHC) [38]. The following provides a broad overview of these machines and the data they produce.

3.1 The Large Hadron Collider

The LHC is the largest and most powerful particle accelerator ever built. As a proton-proton collider, it is designed to reach very high energies and instantaneous luminosities to probe the fundamental structure of matter. It is located at CERN in Geneva, Switzerland. The LHC is a circular collider with a circumference of 26.7 km. Protons are accelerated from an injection energy of 450 GeV up to energies of 7 TeV in opposite directions in two separate beam pipes. At maximum design center of mass (CoM) energy, the protons would collide with a CoM energy of $\sqrt{s} = 14$ TeV. Particles are guided around the LHC ring by superconducting dipole magnets. The magnets are operated below their superconducting transition temperature of 1.9 K, which requires a large cryogenic system of superfluid helium. Higher order multipole magnets are used to focus the beams and correct for imperfections in the magnetic field. The maximum energy of the LHC is limited by strain on the superconducting magnets, but operating CoM energy of $\sqrt{s} = 13.6$ TeV has been achieved. Sixteen radio frequency cavities are used to accelerate the protons to these energies.

The LHC operates in data taking periods called Runs, which are followed by long shutdowns to upgrade and replace the accelerator and detectors. Also, heavy ion runs are performed, where lead ions are collided at the LHC, the data collected in these runs is not relevant for this thesis. The dataset used in this thesis was collected during Run 2 of the LHC. This took place from 2015 to 2018 while the LHC was operating at a CoM energy of $\sqrt{s} = 13$ TeV, and had an estimated integrated luminosity of 139 fb^{-1} with an uncertainty of 1.7% [39, 40].

The LHC was built at CERN to leverage previous state-of-the-art detectors, and particles are accelerated through an extended accelerator complex before being injected into the LHC. The full complex as it existed in 2018 is shown in Figure 3.1. In the first stage, hydrogen atoms are stripped of their electrons to produce protons. The protons are then accelerated in the Linac2 linear accelerator to an energy of 50 MeV. Next, the protons are injected into the Proton Synchrotron Booster (PSB) where they are accelerated to 1.4 GeV. The protons are then injected into the Proton Synchrotron (PS) where they are accelerated to 26 GeV. As a final stage before the LHC, the protons are injected in sequential batches into the Super Proton Synchrotron (SPS) where they are accelerated to 450 GeV. After Run2 the Linac2 was replaced by the Linac4, which accelerates negatively charged hydrogen ions which are then stripped of their electrons before being injected into the PSB.

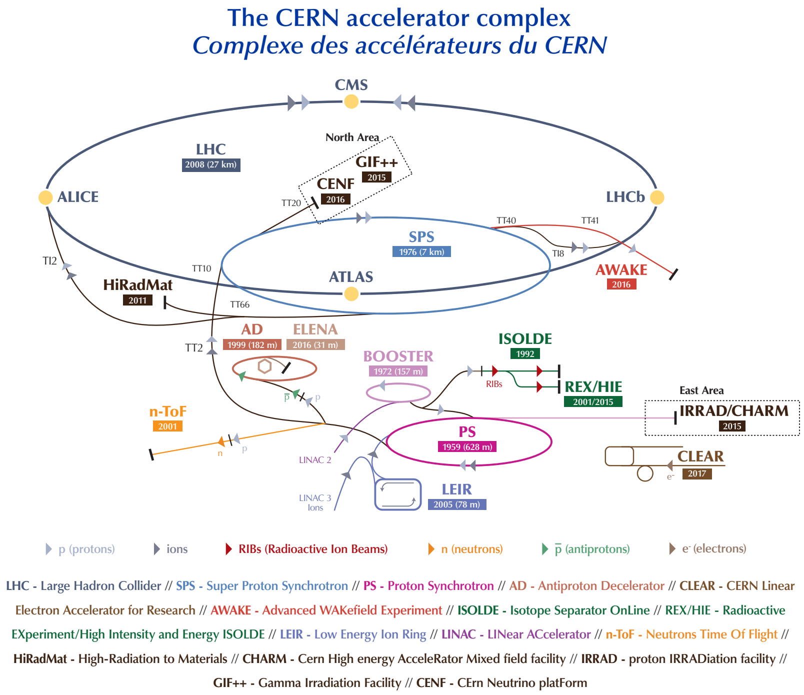

The CERN accelerator complex
Complexe des accélérateurs du CERN
CMS
North Area
GIF++
2015
CENF
2016
LHC
2008 (27 km)
ALICE
LHCb
TT20
TT40
TT41
SPS
1976 (7 km)
TI8
TI2
TT10
ATLAS
AWAKE
2016
HiRadMat
2011
TT66
AD
1999 (182 m)
ELENA
2016 (31 m)
TT2
BOOSTER
1972 (157 m)
ISOLDE
1992
RIBs
REX/HIE
2001/2015
East Area
IRRAD/CHARM
2015
n-ToF
2001
p̄
p
PS
1959 (628 m)
n
p
LINAC 2
CLEAR
2017
e⁻
LINAC 3
Ions
LEIR
2005 (78 m)
p (protons)
ions
RIBs (Radioactive Ion Beams)
n (neutrons)
p̄ (antiprotons)
e⁻ (electrons)
LHC - Large Hadron Collider // SPS - Super Proton Synchrotron // PS - Proton Synchrotron // AD - Antiproton Decelerator // CLEAR - CERN Linear Electron Accelerator for Research // AWAKE - Advanced WAKefield Experiment // ISOLDE - Isotope Separator OnLine // REX/HIE - Radioactive EXperiment/High Intensity and Energy ISOLDE // LEIR - Low Energy Ion Ring // LINAC - LINear ACcelerator // n-ToF - Neutrons Time Of Flight // HiRadMat - High-Radiation to Materials // CHARM - Cern High energy AcceleRator Mixed field facility // IRRAD - proton IRRADiation facility // GIF++ - Gamma Irradiation Facility // CENF - CErn Neutrino platForm

Figure 3.1: The LHC accelerator complex at CERN as laid out in 2018, from Lopienska [41].

To increase the luminosity of the LHC, protons are accelerated in bunches of $\sim 1.15 \times 10^{11}$ protons. Bunches of particles from each beam are crossed every 25 ns at the interaction points, corresponding to a frequency of 40 MHz. The beam intensity is high enough that multiple proton-proton collisions can occur in the same bunch crossing. Many of these collisions are from elastic scattering and are not of interest to the physics goals of the Atlas detector. These additional interactions are referred to as pileup. Pileup can either be in-time, where it occurs in the same bunch crossing as the hard scattering event, or out-of-time, where it occurs in a different bunch crossing. The observed distribution of the number of interactions per bunch crossing is shown in Figure 3.2. This makes for a complex environment in which to perform physics measurements and analyses.

One physical quantity that is measured at the LHC is the cross section of a given process. The cross section is a measure of the probability of a given process occurring. It is interesting to measure the cross section of a process to compare with theoretical predictions. However, the cross section is not directly measurable, the detectors at the LHC can measure the total number of events N. The luminosity of the LHC can also be measured, this quantifies the number of collisions per unit area per unit time. The luminosity is related to the cross section by the equation,

$$\frac{dN}{dt} = \sigma \mathcal{L}(t), \tag{3.1}$$

where dN/dt is the number of events per unit time, σ is the cross section, and $\mathcal{L}(t)$ is the instantaneous luminosity. This equation can be integrated to find the total number of events,

$$\begin{aligned} N &= \sigma \int \mathcal{L}(t) dt, \qquad (3.2) \\ &= \sigma \mathcal{L}_{int}, \qquad (3.3) \end{aligned}$$

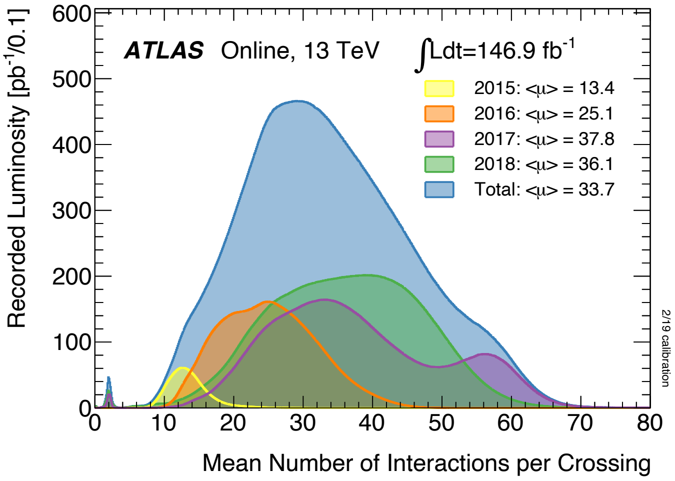


Figure 3.2: The mean number of interactions per bunch crossing in the ATLAS detector during Run 2, from Ref. [42].

where $\mathcal{L}_{int}$ is the integrated luminosity. Together with a measurement of the integrated luminosity, the cross section of the process can be extracted.

3.2 The ATLAS detector

The LHC provides an excellent source of high energy proton-proton collisions and some heavy ion collisions. To do some physically interesting inference, the particles produced in these collisions are measured. These measurements are performed by different detectors at the LHC, each of which has a different design to study different aspects of the collisions. The ATLAS detector [37], which is the focus of this thesis, is one of two multi-purpose particle detectors at the LHC, the other being the CMS detector [43]. Being general purpose, the ATLAS detector is intended to serve a wide range of physics goals.

Due to the confinement principle, and the fact that most BSM high mass particles have short lifetimes, many particles of interest can not be directly observed, only their decay products. By measuring these decay products, properties of the particles that produced them can be inferred. The decay products have different properties that dictate their interaction with ordinary matter. To measure these properties, the ATLAS detector is composed of multiple sub-detectors that are optimized to measure different properties of the particles. The sub-detectors are arranged in layers around the interaction point as shown in Figure 3.3. Measurements from each of these detectors are composed to make more precise inferences on the nature of the particles produced in the collisions. The detector covers nearly the full 4π solid angle around the interaction point and is 25m in diameter and 44 m long.

The particles the detector can measure directly are primarily muons, electrons, photons, pions, kaons, protons and neutrons. In general, the different properties of these particles are leveraged to effectively measure them. Charged particles, for example, can be bent by a magnetic field, and leave a clear trail in tracking detectors. These particles are also stopped in calorimeters such that their energy can be measured. Neutral particles, on the other hand, are not bent and do not leave a track, they do however interact strongly with high Z material and produce showers of particles that can be measured. Particles like muons do not interact strongly with the detector, and therefore interact with material in the outer layers of the

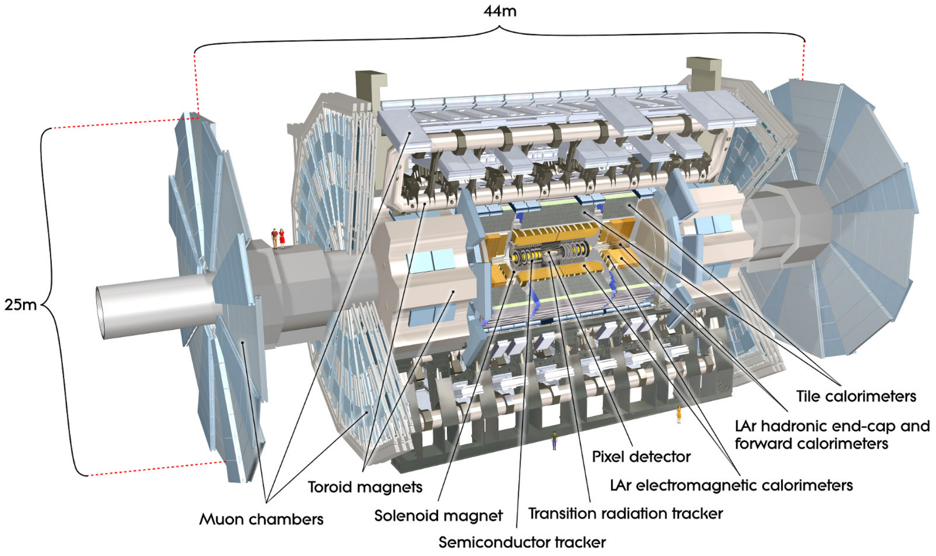

44m
25m
Tile calorimeters
LAr hadronic end-cap and forward calorimeters
Pixel detector
LAr electromagnetic calorimeters
Toroid magnets
Muon chambers
Solenoid magnet
Transition radiation tracker
Semiconductor tracker

Figure 3.3: An overview of the ATLAS detector at the LHC, from Ref. [37].

detector. The presence of particles like neutrinos can be inferred by the presence of missing transverse momentum.

The ATLAS detector uses a coordinate system with its origin at the interaction point in the center of the detector and the z-axis aligned with the beam axis, the x-axis pointing towards the center of the LHC ring, and the y-axis pointing upwards. The azimuthal angle ϕ is measured from the x-axis, and the polar angle θ from the z-axis. Particle momenta are typically described using the transverse momentum p_T, which is the momentum component perpendicular to the beam axis. This is because the transverse momentum is invariant to boosts in the beam direction. This is relevant because the colliding partons do not necessarily have equal and opposite momentum, so the center of mass frame might be boosted along the beam axis. The polar angle θ is also often replaced by the rapidity y, which is defined as,

$$y = \frac{1}{2} \ln \left(\frac{E + p_z}{E - p_z} \right), \tag{3.4}$$

where E is the energy of the particle and p_z is the momentum component along the beam axis. This is the rapidity of the boost along the beam axis which moves the lab frame to the center of mass frame. The rapidity is useful because rapidity differences are invariant under boosts along the beam axis. The pseudorapidity $\eta = -\ln\left(\tan\frac{\theta}{2}\right)$ is also used as it is also invariant under boosts along the beam axis, and is equivalent to the rapidity for massless particles, and a good approximation for massive particles at high energies. The pseudorapidity is often preferred because it is easier to measure than the rapidity as it only depends on the angle of the particle with respect to the beam axis. In contrast, to measure the rapidity, one needs to know both the energy and momentum of the particle.

The following sections describe the different sub-detectors of the ATLAS detector and their purpose, starting with the innermost sub-detector and working outwards.

3.2.1 The Inner Detector

The closest sub-detector to the interaction point is the Inner Detector (ID). This is designed to track charged particles produced in the collisions. It is important to have tracking close

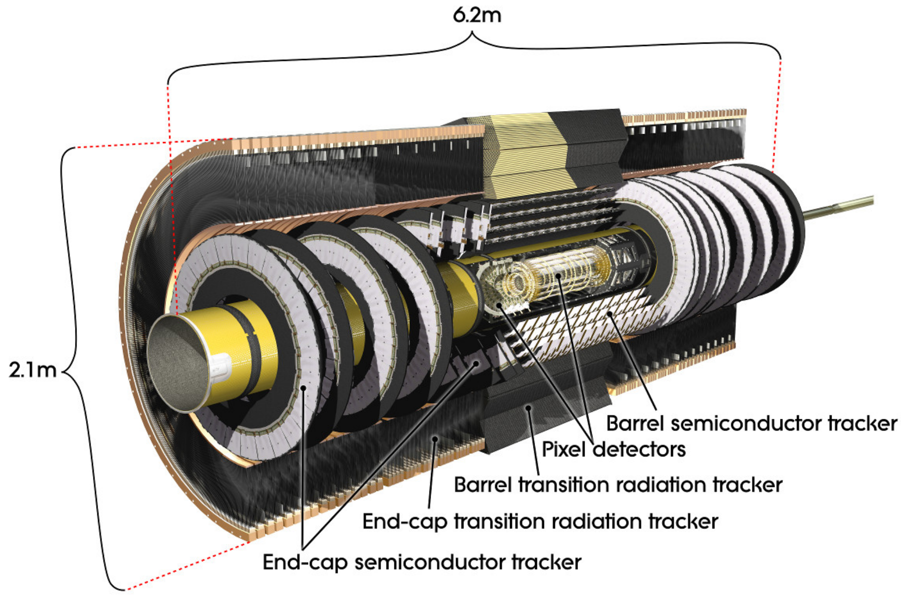

6.2m
2.1m
Barrel semiconductor tracker
Pixel detectors
Barrel transition radiation tracker
End-cap transition radiation tracker
End-cap semiconductor tracker

Figure 3.4: The layout of the Inner Detector of the ATLAS detector, from Ref. [37].

to the interaction point to accurately predict the primary and secondary vertices. The ID is composed of three sub-detectors which in order of proximity to the beam pipe are the Pixel Detector, the Semi-Conductor Tracker (SCT), and the Transition Radiation Tracker (TRT). The full layout of the ID is shown in Figure 3.4. These detectors are submerged in a $\sim$ 2 T magnetic field to bend the charged particles and measure their momentum. Low momentum particles have larger curvature and therefore their momentum can be measured more accurately. Particles with momenta below $\sim$ 1 GeV suffer from multiple scattering and are not measured accurately by the ID.

The three sub detectors of the ID are arranged in barrel like layers around the beam pipe, with end caps on either side. The Pixel detector is constructed from silicon semiconductor pixels and has more than 80 million read out channels. Per unit area, this is the most expensive part of the detector. It is cooled to -6° C to reduce the spontaneous generation of electron-hole pairs. A high granularity is needed to accurately reconstruct the primary and secondary vertices, particularly in the Run 2 conditions where the number of interactions per bunch crossing was high. For Run 2, the innermost layer of the Pixel detector in the barrel region was replaced with the Insertable B-Layer (IBL) to improve the tracking performance [44, 45]. In the barrel region, the Pixel detector has three additional layers, placed at a radial distance of 50.5 mm, 88.5 mm, and 122.5 mm from the beam pipe. There are an additional three layers of flat disks in the end-cap region.

Additional information on the track of a particle is provided by the SCT. This detector is composed of silicon microstrips and covers the region $|\eta| < 2.5$. The SCT has eight layers in the barrel region and two flat end-cap disks. This detector provides high granularity ϕ measurements as half of the barrel strips are aligned in this direction. The other half of the strips are rotated by 40 mrad to provide a measurement in the z direction.

The last sub-detector in the ID is the TRT. This detector is composed of straw tubes that are filled with a gas mixture. In the barrel region, the tubes are placed parallel to the beam pipe, while in the end-cap region, they are placed radially. This detector provides additional tracking information and is particularly useful for separating electrons from charged hadrons.

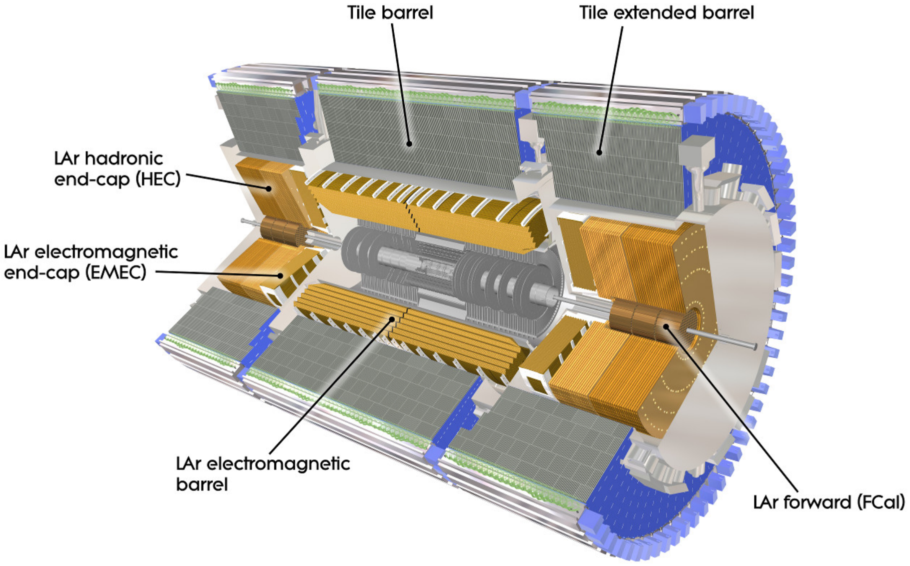

Tile barrel
Tile extended barrel
LAr hadronic end-cap (HEC)
LAr electromagnetic end-cap (EMEC)
LAr electromagnetic barrel
LAr forward (FCal)

Figure 3.5: General overview of the Atlas calorimeters, from Ref. [37].

3.2.2 The Calorimeters

The next sub-detector is the calorimeter system. These are designed to measure the energy of particles produced in the collisions. This is performed by stopping the particles in the detector by inducing a shower of particles. The energy of the particles is inferred from the energy of the particles in the shower. Two different types of calorimeters are used in the Atlas detector, the electromagnetic calorimeter (ECAL) and the hadronic calorimeter (HCAL). The former is designed to measure the energy of electrons and photons, while the latter is designed to measure the energy of hadrons. The calorimeters are placed outside the ID and the combined system is designed to cover up to $|\eta| < 4.9$. The ECAL has about 180 thousand readout channels, while the HCAL has about 14 thousand. The calorimeter system is shown in Figure 3.5.

The ECAL is divided into a barrel structure covering $|\eta| < 1.475$ and two end-cap structures covering $1.375 < |\eta| < 3.2$. The ECAL has a relatively high granularity, compared to the HCAL, to accurately measure the energy of electrons and photons. Incident electrons (positrons) emit photons due to Bremsstrahlung in the ECAL. The average length to emit a photon is characterized by the radiation length X_0. Photons are converted into electron-positron pairs in the material of the detector, on an average distance of $\sim 1.3X_0$. This defines a cascade of processes such that as a particle passes through the ECAL, a shower of particles is produced and measured. The length of the shower scales logarithmically with the energy of the particle that initiated the shower. The shower stops when the energy of the particles is below the threshold for ionization. The radiation length is dictated by the detector material, and in total the ECAL is around 20 radiation lengths thick, to ensure that most of the energy of the particles is absorbed.

The ECAL is a sampling calorimeter, meaning that it is composed of alternating layers of active and passive material. The active material is a scintillating material that emits light when particles pass through it and is used to measure the energy of the particles. The material used for this in Atlas is liquid argon. The passive material is ionized lead that has a small interaction length to induce the shower when particles pass through. With this design, only a fraction of the energy of the particles is directly measured. The total energy can be inferred from $f_{\text{sampling}} = E_{\text{visible}}/E_{\text{deposited}}$. While sampling calorimeters are less precise than homogeneous calorimeters, they are more cost effective and can be used for large detectors.

The HCAL is also a sampling calorimeter and is designed to measure the energy of hadrons. The development of showers in the HCAL follows a similar pattern to the ECAL, except the initiating particles are hadrons. Showers in the HCAL are characterized by the nuclear interaction length λ, which is the average distance a hadron travels before interacting with a nucleus. In general, the interaction length is 5-10 times larger than the radiation length, and showers in the HCAL therefore take longer to develop. In total, the HCAL is around 10 interaction lengths thick. The HCAL is intended to minimize the amount of radiation that punches through into the muon spectrometer. The HCAL has a tile calorimeter in the barrel region and a liquid argon calorimeter in the end-cap region. The tile calorimeter is composed of steel plates and scintillating tiles while the liquid argon calorimeter is composed of copper plates and liquid argon. The tile calorimeter is used in the barrel region because it is more radiation hard than the liquid argon calorimeter. The liquid argon calorimeter is used in the end-cap region because it is more compact and has a higher granularity.

3.2.3 The Muon Spectrometer

The outermost sub-detector of the Atlas detector is the Muon Spectrometer. Muons are minimum ionizing particles and do not interact strongly with the calorimeters and therefore can be measured in the outermost layers of the detector. This last layer has a toroidal magnetic field to bend the muons and measure their momentum through the curvature of their tracks. The muon spectrometer is composed of four sub-detectors, the Monitored Drift Tubes (MDT), the Cathode Strip Chambers (CSC), the Resistive Plate Chambers (RPC), and the Thin Gap Chambers (TGC). The muon spectrometer is a gas detector and operates on similar principles to the TRT. For $|\eta| < 2.7$ the magnetic field is provided by the barrel toroid and is approximately orthogonal to the muon tracks, while for $2.0 < |\eta| < 2.7$ the end-cap toroids are used. In the transition region between the barrel and end-cap regions, both toroids are used. The muon spectrometer is shown in Figure 3.6. In the barrel region, the muon spectrometer is oriented cylindrically around the beam pipe, while in the end-cap region, the detectors are placed perpendicularly. The MDT is the main precision tracking detector in the muon spectrometer and is used in the barrel region. For pseudorapidities $2.0 < |\eta| < 2.7$ the CSC is used, this detector has a higher granularity than the MDT. Together the MDT and CSC add up to 40 thousand readout channels. The RPC and TGC are used for triggering and have a lower granularity than the MDT and CSC.

3.2.4 Trigger and Data Acquisition

With the high bunch crossing rate of the LHC, it is not feasible to record all complete events that are produced. Instead, a trigger system is used to select events that are interesting for further study. The Trigger and Data Acquisition (TDAQ) system is the center of the trigger system in the Atlas detector. In Run 2 the trigger system in Atlas is composed of a hardware level trigger (L1) and a software level trigger (HLT) [46]. The L1 trigger is designed to reduce the event rate from 40 MHz to 100 kHz, which is further reduced to 1 kHz by the HLT.

The L1 trigger receives information from the low granularity components of the calorimeters and the muon trigger system. On average the L1 trigger has $\sim 1\ \mu$s to make a decision. This trigger is implemented in custom hardware designed to be fast. The L1 trigger selects events based on a predefined set of criteria, and these events are passed to the HLT.

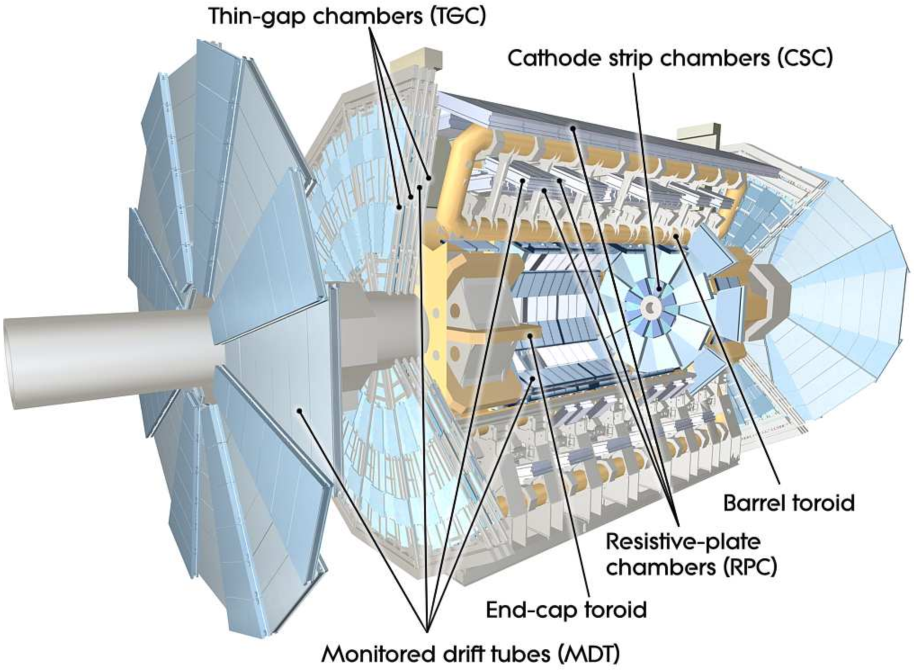

Thin-gap chambers (TGC)
Cathode strip chambers (CSC)
Barrel toroid
Resistive-plate chambers (RPC)
End-cap toroid
Monitored drift tubes (MDT)

Figure 3.6: The layout of the Muon Spectrometer of the Atlas detector, from Ref. [37].

For many regions of phase space, the event selection is not stringent enough to reduce the event rate enough so some triggers are prescaled. This means that only a fraction of the events that pass the trigger are recorded. The main bottleneck here is that the full event must be read out and stored, which is a time consuming process. Some analyses get around this limitation by only recording a fraction of an event [47].

The HLT is a software trigger that runs on a farm of computers, it runs algorithms that are more complex than the L1 trigger. The HLT has access to the full event information and can therefore make more informed decisions. It also applies some similar reconstruction as is performed offline. Trigger decisions in the HLT are made based on the candidate physics objects produced by these algorithms. The offline reconstruction algorithms are the focus of the next chapter.

3.3 Reconstruction

This chapter describes how the measurements from the different sub-detectors are combined in software to infer the properties responsible for the given measurements. The Atlas collaboration uses a software suite for simulation, reconstruction and triggering [48]. In total, the Atlas detector has around 100 million readout channels. Events only activate a sparse set of these channels, and the data from these channels is used to infer the properties of the particles that produced them. Other than the sparse nature of the data, the instantaneous luminosity of the LHC is high enough that multiple proton-proton collisions can occur in the same bunch crossing. Many of the collisions are from elastic scattering that is not of interest to the physics goals of the Atlas detector. However, there is still a high chance of multiple hard scattering events in the same bunch crossing as already shown in Figure 3.2.

3.3.1 Jet Reconstruction

Most of the particles created in a hard scattering event have a high transverse momentum, but not many of these particles are stable enough to be detected directly by the detector. For example, as already covered in section 2.2.1, color charged particles hadronize into color neutral collimated sprays of particles called jets. At the level of the detector, a jet is a collection of energy deposits in the calorimeters that are close to each other in η and ϕ. Without additional information, streams of particles initiated by a photon or electron can also look like jets. The dataset used in this analysis does not apply any object based selection, so the jets are reconstructed from the energy deposits in the calorimeters as described in the following section.

The reconstruction of jets is a complex process that is performed in multiple steps. The goal of this process is to infer the properties of the partons that produced the jets. Any charged particle in a jet deposits energy in the ECAL and leaves a bent track in the ID, both of which can be used to measure the momentum of the particle. Neutral particles only deposit energy in the calorimeters. For the analysis presented in this thesis jets are reconstructed in ATLAS from the energy deposits in the calorimeters. Additional tracking information is used to better resolve the jet's kinematics.

In defining a jet reconstruction algorithm, one must consider the underlying physics of QCD. Specifically, quarks and gluons undergo soft and collinear radiation with a high probability. This means the jet reconstruction algorithm must be robust to the presence of additional soft radiation and random splittings of particles in the jet. These properties are referred to as infrared (IR) and collinear safety respectively. The jet reconstruction algorithm used in ATLAS is the anti-k_t algorithm [49], and it satisfies both of these properties.

A particularly useful form of jet clustering algorithms are sequential recombination algorithms. These algorithms are based on the distance between particles in the $\eta - \phi$ plane and their transverse momentum, defined as,

$$d_{ij} = \min\left(p_{T,i}^{2a}, p_{T,j}^{2a}\right)\left(\frac{\Delta R_{ij}}{R}\right)^2, \tag{3.5}$$

$$d_{iB} = p_{T,i}^{2a}, \tag{3.6}$$

where $p_{T,i}$ is the transverse momentum of particle i, $\Delta R_{ij}^2 = (\eta_i - \eta_j)^2 + (\phi_i - \phi_j)^2$ is the distance between particles i and j, and R is a parameter of the algorithm that defines the size of the jet. Pairs of particles are sequentially combined following the rules,

$$d_{ij} < d_{iB}, \tag{3.7}$$

$$d_{iB} < d_{jB}. \tag{3.8}$$

If the first condition is satisfied, particles i and j are combined into a single jet. If the second condition is satisfied, particle i is considered a jet on its own. This process is repeated until all particles are clustered into jets. The parameter a in eqs. (3.5) and (3.6) balances the contributions to the jet definition of energy and $\eta - \phi$ distance of the particles. The anti-k_t algorithm uses $a = -1$, and clustering proceeds by merging particles around the hardest (largest transverse momentum) particle first. This algorithm is preferred as it produces leading jets that are conical. Using this algorithm also means the number of jets is dynamically determined by the event content. The k_t algorithm uses $a = 1$ and merges particles starting with the softest particles first.

The sequential approach to jet clustering defines a tension between the merging and stopping criteria of the algorithm. The radius parameter R defines the size of the jet and is a free parameter of the algorithm. The total kinematics of a jet are defined by the sum of the four momenta of the particles in the jet. From general kinematic considerations, one can show the radius parameter of a jet should be chosen such that,

$$R \gtrsim \frac{2m}{p_T}, \tag{3.9}$$

where m is the mass of the jet and p_T is the transverse momentum of the jet. Jets with $R = 1$ are considered large-R jets and are used as a proxy for boosted W, Z, Higgs bosons and top quarks.

Calorimeter jets

The jets used in this analysis are reconstructed from the energy deposits in the calorimeters. This is done by constructing topological clusters of energy deposits in the calorimeters. The topological clusters are then used as input to the jet reconstruction algorithm. The topological clusters are constructed by grouping adjacent energy deposits based on the significance, defined as,

$$\xi_i = \frac{E_i}{\sigma_{\text{noise},i}}, \tag{3.10}$$

where E_i is the energy of the cell and $\sigma_{\text{noise},i}$ is the noise from the electronics and pileup. Clusters are seeded by cells with $\xi_i > 4$ and are grown by adding adjacent cells with $\xi_i > 2$. In a final step, all adjacent cells are added to the cluster. Topological clusters can also be split if there are two local maxima in the cluster.

3.3.2 Jet grooming

The presence of radiation from the underlying event and pileup can have a significant impact on the jet reconstruction. In the $\eta - \phi$ plane, the effect of pileup is expected to be uniform. The impact of pileup on the jet reconstruction can be reduced by grooming the jets. In the dataset used in this thesis jets are groomed using trimming [50]. This is done by reclustering the constituents of the jet with the k_t algorithm with $R = 0.2$ and removing the softest sub-jets with $p_T < 0.05$ times the jet p_T. The justification for this is the soft radiation is more likely to be from pileup than from the hard scattering process. This soft radiation is expected to appear as a sub-jet in the jet reconstruction and can therefore be trimmed following the above procedure.

3.3.3 Jet substructure

Different particles result in different patterns in the energy deposits within a jet. This can be used to infer the nature of the particles that produced the jet. These patterns are referred to as jet substructure, and this is useful for understanding jet physics [51]. For example, a jet initiated by a highly boosted $W \to q\bar{q}$ decay often has a two pronged substructure within a large-R jet due to the two quarks. For this thesis, the 'pronginess' of jets is used to discriminate between signal and background. Other more exotic particles might have different substructure on average. The pronginess of a jet can be quantified by the N-subjettiness variables [52]. These variables are calculated by minimizing the distance between the constituents of the jet and a set of N axes. The τ_N variable is defined as,

$$\tau_N = \frac{1}{p_T^{\text{jet}}} \sum_i p_{T,i} \min\left(\Delta R_{1,i}, \Delta R_{2,i}, \ldots, \Delta R_{N,i}\right), \tag{3.11}$$

where $\Delta R_{n,i}$ is the distance between the ith constituent of the jet and the nth axis. In a two pronged jet, the τ_2 variable is small, while in a one pronged jet, it is large. For QCD initiated jets, both the τ_1 and τ_2 variables can be either small or large depending on the jet, but these variables are typically observed to have a similar scale in QCD jets, while the same is not true for jets with genuine two-prong substructure. Therefore, the ratio $\tau_{MN} = \frac{\tau_M}{\tau_N}$ variables become powerful discriminants for signal and background [53]. The analysis presented in this thesis uses the τ_{21} and τ_{32} variables. The calculation of τ_{32} is inherently more complex than τ_{21}, as multiple axes need to be minimized. Therefore, this variable is generally less sensitive to three prong substructure than τ_{21} is to two prong substructure.

3.3.4 Trigger

During data taking the physics objects used in trigger decisions are constructed on the fly. In general the trigger objects are not as well reconstructed as offline objects. Due to differences between the trigger and offline reconstruction, there is not a sharp boundary at the online threshold, instead following a smooth turn on curve. This turn on curve flattens out at a certain value, which is referred to as the offline threshold. The region above the offline threshold is referred to as the trigger plateau.

3.4 Simulation

To compare measured data to theoretical predictions, the physics processes that occur in the detector need to be simulated. This is true for both the signal and the background processes. For the analysis used in this thesis, the QCD background is simulated as well as a set of postulated BSM signals. This thesis is largely data driven, but simulation is used to validate the analysis, provide benchmark signal model samples and help estimate the background. Simulated samples are also often used in machine learning studies that are discussed in chapter 5.

The simulation chain in HEP is based on multiple steps at different scales, from the hard scattering process to the detector response. The hard scattering process is simulated using Monte Carlo (MC) event generators that use perturbative QCD to calculate the matrix element of the targeted process. Outside the hard scattering process, partons produce sprays of particles at energies accessible with perturbation theory. These processes are modelled using a parton shower model, which is applied to both initial and final state partons. At a certain scale, perturbative QCD breaks down and partons hadronize into color neutral particles. This hadronization factorizes from the hard scattering [54]. The confinement of color charged particles is modelled by a hadronization model such as the Lund string model [27] as described in section 2.2.1. After the parton shower, only colorless particles remain, and their interactions with the detector are simulated using a detector simulation. The most accurate form of the detector simulation is a full GEANT4 simulation [55]. In this simulation, individual particles are propagated through the detector material and the response of the detector is simulated.

Part II

Analysis background

Chapter 4

New physics searches

A search for new physics in HEP is a data analysis that attempts to find evidence of new phenomena in data and to exclude the existence of such processes. This defines two different statistical questions. Both exclusion and discovery are important for the field; discovery informs us about new forms of matter, and exclusion informs us which theories do not accurately explain nature. This chapter discusses how these questions are formally addressed using the frequentist approach to statistics used in HEP [56].

4.1 Introduction

To make probabilistic statements about observed data a probabilistic model of our expectations needs to be built. This involves constructing a likelihood model of our data as in section 5.7.1 and repeated here in modified form. The likelihood,

$$\mathcal{L}(\mu, \theta) = \prod_{i=1}^{N} p(x^i|\mu, \theta), \tag{4.1}$$

is a function of the parameters $\{\mu, \theta\}$ given the observed dataset $\mathcal{X} = \{x^i\}_{i=1}^N$, where N is the number of observations. A likelihood is constructed by making background and signal predictions and accounting for all uncertainties that relate to these predictions. The parameter μ is called the signal strength and allows us to compare signal plus background likelihood models ($\mu \neq 0$) and background only models ($\mu = 0$). This likelihood allows us to make statements about the likelihood of the observed data under a given model.

Due to the stochastic nature of HEP processes background and signal predictions are always probabilistic in nature. For example, one might predict $\nu \in \mathbb{R}$ events to be observed on average in a certain region. The likelihood of observing $n \in \mathbb{Z}$ events in this region is then defined by a counting experiment and our model can be parameterized by a Poisson distribution,

$$p(n|\nu) := \text{Pois.}(n|\nu) = \frac{\nu^n e^{-\nu}}{n!}. \tag{4.2}$$

The rate ν can almost never be predicted with absolute certainty, and this value can be expected to fluctuate according to the level of uncertainty. This can be either systematic or statistical in origin. Systematic (epistemic) uncertainty arises due to a lack of knowledge about the processes involved in our prediction. For example, uncertainty in predictions about the detector response in simulation gives rise to systematic uncertainties. Statistical (aleatoric) uncertainty arises due to inherent randomness in the data.

To encode knowledge about uncertainty, additional constrained parameters are introduced into the likelihood. For example, the earlier predicted value of ν might be expected to vary by

$\pm\phi$ following a Gaussian distribution. Note the uncertainty as quantified by the parameter ϕ is itself a prediction. A likelihood model for this setting can be written as,

$$\mathcal{L}(\mu, \theta) = \text{Pois.}(x|\nu + \mu S + \theta)\mathcal{N}(0; \theta, \phi^2), \tag{4.3}$$

where μ is a free parameter, S is the expected number of signal samples, ν is fixed and θ is constrained by the Gaussian distribution.

4.2 Discovery of new physics

A search for new physics is formalized as a hypothesis test. The null hypothesis (H_0) is the data is described by the Standard Model and the alternative (H_1) hypothesis is the data contains new physics. A hypothesis test returns a test statistic, which together with the sampling distribution for the test statistic can be converted to a p-value. The p-value is the probability of observing as many, or more, events than were observed assuming the null hypothesis is true. A p-value is often converted into a significance Z, which is the number of standard deviations between the mean of the null hypothesis and the observed data. In a search, the null hypothesis is rejected if the significance is greater than some threshold, often 5σ [57].

Hypothesis tests in HEP are based on the profile likelihood ratio [58, 59],

$$\lambda(\mu) = \frac{\max_{\hat{\theta}} \mathcal{L}(\mu, \hat{\theta})}{\max_{\hat{\mu},\hat{\theta}} \mathcal{L}(\hat{\mu}, \hat{\theta})}, \tag{4.4}$$

and the corresponding test statistic $t_\mu = -2\ln(\lambda(\mu))$. This test statistic is particularly useful because in the asymptotic limit of large statistics, the sampling distribution of the test statistic is known and so the p-value is straightforward to calculate [58]. Questions of discovery are interested in $\lambda(\mu = 0)$ which defines the likelihood ratio under the background only hypothesis. In this thesis, it is assumed that physical processes are additive on top of the background and therefore the signal strength $\mu > 0$. To account for this the test statistic is set to zero if the best fit $\hat{\mu} \leq 0$. For quantifying deficits, negative values of $\hat{\mu}$ are considered. The appearance of a deficit is interpreted in the context of the assumptions made by this search as a failure to accurately model the background.

To understand the test statistic the setting of a simple counting experiment is again useful. In this setting, when μ is fixed the parameter θ shifts away from the nominal value of zero if $x \neq \nu$. A shift in θ away from zero comes at the cost of moving away from the maximum likelihood of the Gaussian constraint. In contrast, when μ and θ are fit simultaneously there is no cost to shifting μ and so $\hat{\theta}$ stays fixed at zero. This results in no cost due to the Gaussian constraint. Therefore, the numerator of eq. (4.4) is always less than the denominator and $\lambda(\mu) \in [0, 1]$. Also, note the larger the discrepancy between x and the predicted rate ν the smaller the numerator becomes while the denominator remains fixed and so as the discrepancy grows $\lambda(\mu) \to 0$.

One thing that is important to account for in HEP is the so-called 'look-elsewhere' effect. This problem arises when an analysis makes multiple comparisons between model predictions and observations. The reason multiple comparisons are problematic is the p-value is uniform under resampling of the data in the background only case. That means if the background model is exactly correct, and the same quantity was measured an infinite number of times, discovery is claimed $y\%$ of the time, where y is fixed by the discovery threshold. Conversely, if the alternate hypothesis were true then discovery would be claimed more than $y\%$ of the time when repeating a measurement an infinite number of times. The key takeaway here is that discovery can be claimed in the background only case, though it is more likely if the alternate is true. If multiple hypothesis tests are run on independent quantities, where the fluctuations in the data are uncorrelated, the uniformity of the p-value distribution plays the same role. Therefore, more comparisons increase the chance of claiming a discovery in the background only case. Individual comparisons are used to construct 'local' p-values, but the 'global' p-value of the analysis has to combine these and pay a trials factor [60].

4.3 General search considerations

For this thesis, it is useful to consider binned searches, which are common in HEP. In a binned search the observed data is defined by a histogram and the likelihood in eq. (4.1) contains the product of multiple Poisson distributions. The histograms that are used in the search are defined in variables where reliable background and uncertainty predictions can be made. These variables are only ever a subset of the variables reconstructed from the detector. The following provides some general considerations for searches in HEP that are useful for understanding the work in this thesis.

Consider a histogram with a single bin and no systematic uncertainty on the number of predicted background events B or the number of signal events S. The likelihood in this case is given by,

$$\mathcal{L}(\mu) = \text{Pois.}(n|B + \mu S), \tag{4.5}$$

where n is the number of observed events, and μ is the signal strength. In this setting, $\mu = 1$ is $s = S/\sqrt{B}$ standard deviations from the median of the null hypothesis. For high statistics ($B \gtrsim 50$) this s is similar to the significance Z of the corresponding test statistic $t_{\mu=0}$. If the background can be reduced by a factor of $\alpha_0 > 1$ at the cost of reducing the signal strength by a factor of $\alpha_1 > 1$ then this significance scales like the significance improvement characteristic $\text{SIC} = \sqrt{\alpha_0}/\alpha_1$ [61]. Therefore, the sensitivity of the search can be enhanced by reducing the background faster than the signal. Note also that random downsampling of the data by a factor of $\alpha = \alpha_0 = \alpha_1 > 1$ reduces the significance of the test statistic by $1/\sqrt{\alpha} < 1$.

In general, background can be reduced by making selections on the data by making use of variables that are not fit in the hypothesis test[1]. That is variables that are not used to form the count n. These selections need to be made carefully such that the background can still be accurately predicted after selections have been made. This simple heuristic is complicated by the fact the signal and background are not known exactly. The addition of systematic uncertainties further complicates this picture.

One generic strategy for performing searches in HEP is to partition data based on a score to define a 'signal enriched' dataset that can be used to perform a hypothesis test. The score would be assigned to data samples using criteria based on some large set of features and the hypothesis test would be run on a subset of features. This allows information from the full feature set to be used indirectly to enhance the sensitivity of the final hypothesis test. To permit a background prediction to be made, the partitioning and subset of features need to be carefully selected. This is addressed later in this thesis.

[1] This generally means they are not a part of the vector x^i in the likelihood defined in eq. (4.1)

4.4 Exclusion of new physics

Conceptually, the existence of new physics can be excluded by observing its absence with confidence. Formally, this can be done by setting upper limits on the parameters that determine the signal. Limits are most often set on the cross section of new physics processes at a certain threshold defined by a pre-defined confidence level. In HEP, limits are set at 95% using a method that is referred to as the CL_s procedure [62]. This procedure sets upper limits based on pseudo frequentist p-values and is conservative as the true exclusion probability can be greater than what is quoted. The CL_s procedure is standard across multiple experiments which is useful for allowing comparisons to be made.

For exclusion with a bounded parameter of interest, as studied here, a modified test statistic is used [58],

$$\tilde{q}_\mu = \begin{cases} -2\ln\tilde{\lambda}(\mu) & \text{if } \hat{\mu} \le \mu \\ 0 & \text{if } \hat{\mu} > \mu \end{cases} = \begin{cases} -2\ln\frac{\max_{\hat{\theta}}\mathcal{L}(\mu,\hat{\theta})}{\max_{\hat{\theta}}\mathcal{L}(0,\hat{\theta})} & \text{if } \hat{\mu} \le 0 \\ -2\ln\frac{\max_{\hat{\theta}}\mathcal{L}(\mu,\hat{\theta})}{\max_{\hat{\mu},\hat{\theta}}\mathcal{L}(\hat{\mu},\hat{\theta})} & \text{if } 0 \le \hat{\mu} \le \mu \\ 0 & \text{if } \hat{\mu} > \mu \end{cases}, \tag{4.6}$$

where $\tilde{q}_\mu = 0$ for $\hat{\mu} > \mu$, because the signal is additive on top of the background and such a fit result, does not indicate less compatibility of the data with the tested value of μ.

When performing physics analyses there is always some background contamination. This is reflected by writing $s + b$ to define the signal model. When excluding new physics the CL_s approach ensures the analysis does not exclude processes to which it has no sensitivity. In the CL_s approach limits are set using a pseudo p-value,

$$p'_\mu = \frac{p_\mu}{1 - p_b} = \frac{CL_{s+b}}{CL_b} = CL_s, \tag{4.7}$$

where p_u is the p-value for $\tilde{q}_\mu$ under the hypothesis of signal strength μ,

$$p_\mu = \int_{\tilde{q}_\mu}^{\infty} f(\tilde{q}'_\mu|\mu, \hat{\theta}(\mu))d\tilde{q}'_\mu, \tag{4.8}$$

where f is the PDF for the sampling distribution of the test statistic and $\hat{\theta}$ is the maximum likelihood estimate of θ with the signal strength fixed at μ. The p-value p_b is the p-value for the same test statistic $\tilde{q}_\mu$ under the background only hypothesis,

$$p_b = \int_{-\infty}^{\tilde{q}_\mu} f(\tilde{q}'_\mu|0, \hat{\theta}(0))d\tilde{q}'_\mu. \tag{4.9}$$

The use of the CL_s procedure has some nice features in the context of exclusion. If the data is significantly in excess of what is expected under the background only likelihood model then $1-p_b$ is small, and the CL_s value is large, effectively making it hard to exclude the signal. This is expected as the signal should not be excluded if an excess is observed. If an observation in excess of what was measured is equally probable under both the signal and background likelihood models, characterizing an analysis that has no sensitivity to the targeted signal, then the numerator and denominator are approximately equal, the ratio is close to one, and the signal can not be excluded. If the data agrees with the background only hypothesis, but not with the signal hypothesis, then the fraction is small, and the signal can be excluded. This is also expected as the data is much better described by the background model than the signal plus background model, and so it can be said with confidence the signal is not produced at this cross section. A signal at a certain cross section σ, as parameterized by the signal strength μ, is excluded at 95% if $p'_\mu < 0.05$. In practice, the observed limit is found by solving $p'_\mu = 0.05$ for μ.

Analyses also calculate expected limits to provide a comparison to the observed limit. These limits are intended to represent the limit that would be set on average if the experiment were repeated many times and the background only hypothesis was true [58]. In ATLAS, these limits are calculated by setting the expected observation to match exactly the recorded observations (Asimov dataset), with no signal injected and all parameters in θ fixed to their best fit values [63, 64]. One and two sigma variations of this are also reported by setting the expected observation to the one and two sigma variations of the background only likelihood distribution centered on the observed data. The expected limits reported in this thesis are the median limit obtained under resampling of the observations assuming the recorded observation characterizes the true background only prediction.

Chapter 5

Machine learning

Over the last decade machine learning (ML) has become a valuable scientific tool [65–68]. More recently it has gained widespread attention in broader societal contexts with the advent of high-fidelity text generation models like ChatGPT [69] and image generation models like Stable Diffusion [70]. The success of these models has led to public debate about the promise and dangers they may pose. These concerns are important, but in the context of this thesis, ML is treated as a suite of tools that can be used to enhance the efficiency and precision of many data analysis related tasks. There are challenges associated with deploying ML algorithms in scientific contexts and some discussion of this is provided throughout this chapter.

This chapter introduces the basic concepts of ML and the specific elements that are required to understand the analysis this thesis presents. Work completed as part of this thesis that does not pertain to the main analysis is described briefly. The chapter closes with some discussion and outlook.

5.1 Background

In practice, an ML task is defined by data samples from some distribution, such as the simulated interactions of particles with the ATLAS detector, and a task, such as object identification or generation. An ML engineer produces a function, or model, that can perform the specified task by minimizing the 'error' in the predictions of the model. This minimization is performed by fitting a parametric ML model to data with some internal structure (architecture) using some learning algorithm. There is a fundamental connection between this algorithm, the structure of the data and the architecture of the ML model which is illustrated in Figure 5.1. Few formal results exist for how to piece these different elements together and ML engineers most often rely on experience to make decisions on how to approach any given problem.

All the ML approaches used in this thesis fit differentiable parametric models to data using some form of gradient descent in the space of model parameters to find the minima of some objective function. The act of fitting the model to data is referred to as learning or training. Models are typically fit to data using empirical risk minimization and assume independent and identically distributed data (iid). The derivatives of the ML models are calculated using backpropagation which is a special kind of reverse mode automatic differentiation [72]. The parameters of ML models are referred to as weights. The parameters that define the ML model's structure and training algorithm are referred to as hyperparameters. It is often observed that ML models can extract more accurate models from data than can be created using human crafted algorithms.

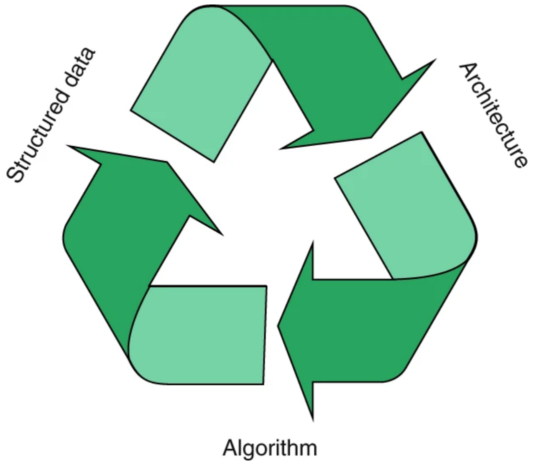


Figure 5.1: The elements of machine learning from Zdeborová [71]. All machine learning approaches involve an interaction between the structure in the data, the architecture of the neural network and the learning algorithm used to train the model.

Some ML models are known to be universal function approximators of continuous functions [73, 74]. This means they can parameterize any possible continuous function. Models that are used in practice are not universal function approximators and only have the capacity to parameterize continuous functions in a certain family. A sufficiently flexible ML model can have the capacity to 'memorize' the data on which it is trained. This quality is undesirable as then the model is unlikely to make accurate predictions on data it has never seen before.

The ability of a model to perform well on new data samples is referred to as generalization. This is an essential property for ML models as they are trained on one dataset but deployed on another. If the capacity of a model is too large, and it starts to memorize the data, its generalization performance decreases, this is referred to as overfitting. If the capacity of a model is too low then the true function may not lie within the space of functions the model can approximate, in which case a poor function is learned [75]. This is referred to as underfitting. This relationship between model capacity and generalization performance is an example of the bias variance trade off.

Overfitting can be measured by splitting the available dataset into a train and validation set. The model is trained on the training set and evaluated on the validation set, with the difference in performance on the two sets being used as a proxy for overfitting. Regularization techniques, where the model capacity is restricted, are used to improve generalization. Selecting a regularization strategy, and other model parameters, using the validation set can lead to overfitting on this dataset. This is solved by introducing a third split of the dataset, the test set.

In modern ML a phenomenon known as 'deep double descent' has been observed, where the generalization performance improves with the number of weights in a model [76]. This is paradoxical if one assumes the number of parameters in a model uniquely defines its capacity when fit to data. However, the relationships shown in Figure 5.1 are expected to play a significant role in dictating model capacity. Therefore, the number of parameters in a model is often a poor proxy for its capacity when fit to data. Both the structure of the data and the learning algorithm used to train the model can significantly impact the capacity of the model.

The modern approach to ML, particularly in the subfield of deep learning, is to overparameterize models and train them with specific learning algorithms [77–79] on large datasets. This, of course, does not work on all possible datasets and settings [80–82], particularly when the iid assumption is violated. However, in practice, this approach is successful in many settings including HEP [65].

Issues in applying ML algorithms do arise when there is a domain shift. This is defined by models being trained on one dataset and deployed on another, where the two datasets are distributed differently [83]. This is particularly relevant for HEP, where supervised algorithms are trained on simulated data and deployed on real recorded data. Due to mismodelling in simulation, these two datasets are not drawn from the same distribution.

In choosing an ML model it is important to match the architecture to the data type, this defines an important inductive bias. In general, injecting physics knowledge (inductive bias) into a model significantly increases the efficiency with which it learns [84, 85]. However, constraining models by introducing inductive bias can make them more difficult to fit to data. This may be in part why more data and bigger models – while less efficient – have consistently proven to be the highest performing [86–88]. If the goal of an ML algorithm is to be as performant as possible and resources are not limited then analysts should pursue more data and bigger models. However, if the goal is efficiency then one should try to inject as much physics knowledge as possible [89, 90].

Modern ML can be broadly categorized into two classes: supervised and unsupervised. In supervised ML a dataset of samples with targets is provided, and the objective is to learn a map from the samples to the targets. Unsupervised ML does not have such labels, and instead some inherent property of the data such as the density is estimated, or labels are defined using the data directly to try and extract some useful representation. This chapter introduces some specific examples of these types of learning. Approaches that are used in this thesis are described in detail, all other methods are only described in high level terms.

A common problem faced by ML approaches is that models operate on high dimensional spaces. This causes problems as high dimensional spaces behave in ways that are unfamiliar to us and our experience in four dimensions, leading to counterintuitive behaviour in some settings. For example, the volume of an N dimensional unit sphere V_N^S as a fraction of the volume of an N dimensional unit hypercube V_N^C goes to zero as N goes to infinity,

$$\lim_{N\to\infty} \frac{V_N^S}{V_N^C} = 0, \tag{5.1}$$

which shows that when sampling from a unit hypercube in N dimensions, for large N the corners of the cube are much more likely to be sampled than the center. This is at odds with the familiar picture in two dimensions, where samples are more likely to fall inside a circle inscribed in a unit square. This is just one example of the difficulties associated with thinking in high dimensions and the problems that result when sampling such spaces.

5.2 Data structures

Training simple neural networks is intractable if the data satisfies some worst-cast conditions [91]. This suggests that datasets where ML is successful have some special properties. It is commonly observed that almost all natural datasets have this property as ML algorithms learn almost always. This subsection is not concerned with what this property is, instead focussing on how data is represented when passed to an ML model. As is shown in the next subsection, particular representations are complemented by specific ML model architectures.

One of the simplest data representations is an N dimensional real vector $x \in \mathbb{R}^N$. In ML each entry x_i in this vector is referred to as a feature. For HEP applications this vector could represent a jet, described by its four momenta and substructure, for example. It is common for these vectors to contain features with different scales and units. To account for this the data is normally preprocessed as described in the following. Features with long tails are often log scaled to reduce their variance. ML models are assumed to be sufficiently flexible to be able to undo this transformation if it hampers learning. In practice, it is often observed in HEP that this log transform improves learning. Choosing units to represent different features is also arbitrary as ML models are expected to learn the relationship between features from data. The input features to an ML model are therefore preprocessed to a common scale, each feature is scaled to range from zero to one for example. Some features, such as particle type, are not natively represented as real numbers. These features are referred to as categorical features. Embedding these features as scalar integers is problematic as this implies an ordering to particle type that does not exist. Instead, each entry in a set of M unique categorical features is embedded as a row of an M dimensional identity matrix. This is referred to as one hot encoding. A one hot encoded label can be included in the feature vector x.

Two common data modalities in ML are text and images. Many approaches have been developed for handling data in this form. Images are represented as three-dimensional tensors of real numbers with three color channels and the spatial resolution of the image. Text can be represented as an ordered sequence of categorical features. Breaking down text into different chunks is not done on a word by word basis as this is inefficient. Instead, text is broken into more atomic blocks based on the byte-pair encoding algorithm [92, 93], which leverages common patterns to reduce redundancy.

In HEP data is often represented by physics objects like jets. A jet can be described as a collection of particles as reconstructed from the detector. Each of these particles can be assigned a feature vector x^i such that the full jet is described by the set $x = \{x^i\}$. Inherently there is no order to this set, though one can be assigned based on p_T. This introduces a bias with limited information by identifying the p_T part of the feature vectors x^i as special. Natively a jet, as represented by a collection of particles, is an unordered attributed set. In this representation of a jet the features in the vector x are treated on equal footing by an ML model. This representation is also the same as how text is represented in ML, except it is unordered, and therefore methods developed for text can be successfully translated to the HEP domain [88]. The ability of ML algorithms to bridge domains, with methods being developed for text finding significant utility in HEP, is a key strength of the field.

5.3 Model architecture

Machine learning models come in different types. The following describes some of the most important and relevant for this thesis. They are typically composed of stacked layers. The different parameters in the models are initialized randomly [94]. There are many choices in how they are designed defined by hyperparameters that are typically chosen by a grid search over a subspace. This search space is typically chosen from experience or by looking at what has been found to be successful in the literature.

5.3.1 Multi-layer perceptrons

One of the most common building blocks of ML models is the multi-layer perceptron (MLP). This model operates on vector inputs. It is constructed from interleaved linear transformations and element wise non-linear functions. The linear transformations L_i are parameterized by real $N \times M$ dimensional matrices W_i and N dimensional bias vectors b_i such that they act on some M dimensional input α as,

$$L_i(\alpha) = W_i\alpha + b_i, \tag{5.2}$$

and produce an N dimensional output vector. The non-linear functions are referred to as activations and are typically simple and differentiable almost everywhere. The most famous example is the ReLU [95] activation function,

$$\texttt{ReLU}(\alpha) = \max(0, \alpha), \tag{5.3}$$

where the $\max$ acts on each feature in α separately. There are many forms of activation functions [96] but they are always simple in form. An MLP with $n - 1$ hidden layers using activations σ_i is defined as,

$$\mathrm{MLP}(x) = \sigma_n \circ L_n \circ \sigma_{n-1} \circ L_{n-1} \cdots \sigma_0 \circ L_0(x), \tag{5.4}$$

acting on an input sample x. Normally all the internal activations are identical except for the final activation σ_n. This output activation is chosen to match the prediction task. The width of an MLP layer is defined by the size of the internal representations N, and the depth is defined by the number of stacked layers n. A finite width and depth MLP with ReLU activations is a universal function approximator [74]. Models like the MLP are often referred to as deep neural networks when they have many layers.

The parameters of an MLP are the elements of the matrices W_i and biases b_i. If the features input to a neural network take large values or have different scales then the weight matrices of an MLP can become extreme and generally makes learning more difficult. The same is true for all ML models and this is the main motivation for preprocessing the features input to a neural network.

5.3.2 Tree based algorithms

MLPs work well on data that is naturally represented by smooth, continuous, real valued data, particularly if the dataset is large. However, lots of datasets contain heterogeneous features, few samples (order 10k) and uninformative features. These datasets are referred to as tabular. On these datasets tree based classification algorithms outperform MLPs [97]. In tree based algorithms the dataset is split based on simple criteria of the features and some splitting functions. Multiple trees of different depths are ensembled, whereby their predictions are aggregated, to form performant algorithms for different tasks. These algorithms have better inductive biases for tabular data and therefore learn more efficiently on these datasets [97], where efficiency means they don't need as much data to learn. This is relevant for some ML problems in HEP [98].

5.3.3 Invertible neural networks

In various applications, it is useful to be able to parameterize flexible invertible functions. These functions allow a single map between spaces to be defined, and typically both their inverse and forward passes are required to be differentiable. An invertible function is useful because it allows the spatial deformation due to a transformation to be directly quantified. This is useful in density estimation tasks as discussed later.

A simple example of an invertible transformation is an affine transformation, $f_\theta(x) = ax + b$, which has parameters $\theta = \{a, b\}$. Parameterized in this way the transformation has little flexibility. More flexible invertible functions, such as rational quadratic splines, can be used instead where the parameters of the spline define θ [99]. However, even these transformations, defined in this way, can not represent complex functions.

A simple way to extend the flexibility of any invertible transformation is to parameterize the invertible functions f_θ in terms of the data. For example, given a two dimension input vector, an invertible transformation can be constructed by composing two invertible functions f_{θ_1} and f_{θ_2} as,

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \rightarrow \begin{bmatrix} x_1 \\ f_{\theta_1(x_1)}(x_2) \end{bmatrix} = \begin{bmatrix} x_1 \\ u_2 \end{bmatrix} \rightarrow \begin{bmatrix} f_{\theta_2(u_2)}(x_1) \\ u_2 \end{bmatrix}, \tag{5.5}$$

where the parameters of each invertible transformation are now predicted as a function of the data. This effectively means that different invertible functions are applied at every value of x. The predicted parameters $\theta_i(x_i)$ of the invertible function are typically the output of other non-invertible neural networks like MLPs.

More iterations of the steps in eq. (5.5) can be stacked together to form expressive invertible transformations. These networks are typically either constructed using autoregressive [100, 101] or coupling [102, 103] approaches. Autoregressive approaches can be either quick to encode data and slow to invert or vice versa. The choice of the fast direction is made based on the nature of the problem they are designed to solve. Coupling approaches have approximately the same speed in both directions. One of the main shortcomings of invertible neural networks (INNs) is they are dimension preserving, which limits their expressivity.

The approach outlined in eq. (5.5) defines invertible transformations by stacking transformations together. An alternative approach is to parameterize the transformation as a dynamic process through infinitesimal steps and then integrate to find the full transformation. This defines continuous time flows, which uses the ordinary differential equation,

$$\frac{du_t}{dt} = f_\theta(u_t, t), \tag{5.6}$$

with some function f_θ that satisfies some minimal conditions to ensure this equation has a unique solution. Most neural networks can be used to parameterize f_θ directly. Solving the dynamics defined by this equation for finite time results in a flexible relatively unconstrained invertible transformation.

5.3.4 Convolutional neural networks

Convolutional neural networks (CNNs) are composed of convolutional layers specifically designed to process image inputs and MLPs. The convolutional layers have an inbuilt inductive bias to handle images, and for a long time were the most successful architectures to be applied to image data [104]. Though transformers are now competitive they also use a similar inductive bias [105]. These networks are not used in this thesis and so are not detailed here.

5.3.5 Transformers

One of the most successful modern architectures is the transformer [106]. The novel layers in these architectures are based on the attention mechanism. This mechanism allows for information to be quickly propagated across long sequences and can handle variable length sequences of inputs. This is particularly relevant for jets in HEP, allowing their native embedding as unordered sets to be modelled with high performance [88]. The attention layer is more complicated than the other layers discussed so far and is not used for the work in this thesis. To form a transformer the attention mechanism is combined with MLP blocks and other layers. For an excellent introduction see Fleuret [107].

5.4 Learning algorithm

Most ML models, except tree based algorithms, are fit to data using variants of stochastic gradient descent (SGD) [108]. This is performed using derivatives with respect to the weights of the model θ of an objective (loss) function $\mathcal{L}(x, \theta)$ such that the update step for the parameters of an ML model is,

$$\theta = \theta - \eta \frac{1}{n} \sum_{i=1}^{n} \nabla \mathcal{L}(x^i, \theta), \tag{5.7}$$

where n is the number of samples used in the update and referred to as the batch size and η controls the step size taken in each update and referred to as the learning rate. The learning rate is one of the most important hyperparameters in training machine learning models. It is often found that scaling this global variable according to some schedule significantly improves model performance [109]. In true SGD the gradient is estimated using single samples $n = 1$, but it is more computationally efficient to compute this in batches with $n \gg 1$. An epoch is defined as one full pass through the dataset.

Gradient estimates calculated in this way are noisy and this can interfere with learning. Many variants of SGD exist that aim to smooth out the gradient estimate and assign per-weight learning rates to increase the importance of sparse parameters. Some use a moving average of the gradient rather than the gradient itself. Momentum, analogous to physics, can also be used to smoothen gradient estimates. The most commonly used optimizer is `Adam` [110, 111] which uses momentum with per parameter adaptive learning rates.

It has been shown that SGD finds minima that are smooth and 'simple' [112]. This means it is often assumed the function parameterized by the fit ML model looks at simple relationships in the data. Generally, this is thought of as being beneficial as simple patterns are often more robust in the spirit of Occam's razor. However, simplicity comes at the cost of ignoring complex relationships that may be as informative as, and orthogonal to, simple ones. This over reliance on simple information can cause a lack of robustness to small perturbations in the input domain, particularly when simple patterns are spurious [80]. This may explain adversarial attacks [113], where a small amount of additive noise to an input sample can radically change the output prediction.

5.5 Classification

One of the conceptually most simple tasks in ML is classification. In this setting the dataset $\mathcal{X} = \{x^i, y^i\}$ is formed from paired samples x^i and labels y^i. Typically, the labels are one hot encoded, the input features preprocessed, and then an ML model f_θ is fit to the data with a SIGMOID (SOFTMAX) activation function. This thesis focuses on binary classification, where the labels are binary, and the SIGMOID activation function is used. In HEP this is often used to discriminate signal from background, which is the focus of the discussion here. This output activation turns the prediction into a normalized probability estimate such that the cross entropy objective function can be used,

$$\mathcal{L}_{CE}(x^i, y^i, \theta) = p(y^i) \log(f_\theta(x^i)), \tag{5.8}$$

where p is the true probability of the labels and is typically a delta function.

In HEP classification models are often trained on simulated data where the generating process is known such that exact labels can be assigned. Due to the stochastic nature of processes in HEP, it is almost never the case that different classes can be perfectly separated. This is particularly true for the classification of jets in HEP. Instead, when fit to data the model learns the likelihood ratio of the two classes. This ratio can be used to select events to define data samples that are enriched in certain desirable properties.

An ML model is trained to minimize eq. (5.8) using a variant of eq. (5.7). This algorithm is then applied to data that is measured by detectors like ATLAS and is often seen to have good performance despite the mismodelling in simulation [114]. Applying ML models trained on simulated data to real data is a common practice in HEP, and requires careful calibration to ensure that simulation and data can be treated in the same way. This calibration itself is a challenging task that can also be solved using ML methods [115].

5.5.1 Classification without labels

It has also been shown that classification algorithms can be trained on mixtures of data where only the relative mixtures are known. This applies to binary classification problems such as discriminating signal from background. Consider two datasets $\mathcal{X}_1$ and $\mathcal{X}_2$ with signal to background ratios f_1 and f_2, respectively. If $f_1 > f_2$ and both signal and background are sampled from the same distribution in both datasets, then the optimal classifier for discriminating between signal and background can be found by assigning the label one to $\mathcal{X}_1$, zero to $\mathcal{X}_2$ and training a classifier. This defines the classification without labels (CWoLa) paradigm [116]. With this approach models can be trained directly on data, removing the need for simulation. This has been shown to be effective for isolating muons in data [117] for example.

Models trained in this paradigm still use the cross entropy loss function defined in eq. (5.8), and functionally the training procedure is the same. The only difference is that labels are assigned based on the relative mixtures of the datasets. There are some difficulties associated with this approach, particularly when there are only a few samples from one of the classes. In general, it is observed that CWoLa is sensitive to the absolute number of samples in each class, rather than the relative mixtures. If one of the classes does not have a sufficient number of samples then the model struggles to learn, as is typical in ML. When training a classifier in this fashion it is also observed that MLP classifiers struggle to ignore irrelevant features in the data [118]. This can limit the dimensionality of the data that can be used in the classifier. However, it has been shown that tree based algorithms can largely overcome this limitation [118]. The performance of these models is also significantly boosted by something referred to as pretraining which is discussed later in this thesis.

The CWoLa approach to training classifiers plays a significant role in the work presented in this thesis.

5.6 Decorrelation

Once a binary classifier has been trained in HEP it is often used in an event selection to create a signal enriched data sample. In an analysis, this selection is only ever one step in a chain of procedures. Some analyses rely on the classifier output having restricted correlation with certain variables m [114]. This is required because analyses often rely on assumptions about the distribution of the variables in m to estimate the background. Making a selection based on a correlated classifier can distort this distribution and violate these assumptions.

Decorrelation can be promoted by minimizing a measure of the correlation between the classifier output $f_\theta(x)$ and the protected attributes m while training a machine learning model [119–122]. In practice, this means a model is trained to minimize the expectation over a modified objective function of the form

$$\mathcal{L}(x^i, y^i, m^i, \theta) = \mathcal{L}_{\text{CE}}(x^i, y^i, \theta) + \alpha \mathcal{L}_{\text{decor}}(x^i, m^i, \theta), \tag{5.9}$$

where m^i is the protected variable for sample i and $\mathcal{L}_{\text{decor}}$ is an objective function that decreases as the correlation between m and $f_\theta(x)$ decreases. The parameter α controls the relative contribution of each of the objectives and is the minimal additional parameter that can appear in such loss functions, though more parameters can appear in $\mathcal{L}_{\text{decor}}$. Decorrelation methods that use eq. (5.9) incur additional costs when defining the classifier for calculating derivatives and tuning additional parameters like α.

5.6.1 Decorrelation with the conditional CDF

During the work on this thesis, we proposed using the conditional cumulative distribution function (CDF) to decorrelate trained classifiers [123]. This approach is compatible with all preexisting decorrelation methods and corrects the output of the classifier after it has been trained. In one dimension the CDF is defined as a map $F : \mathbb{R} \to [0, 1]$ such that $F(x) = \int_{-\infty}^{x} p(x)dx$. This means F is a monotonically increasing function, which also means that it is order preserving. An example of a probability density function (PDF) and its corresponding CDF is shown in Figure 5.2. The conditional CDF $F(x|m)$ is defined in the same way but operating on the conditional distribution $p(x|m)$.

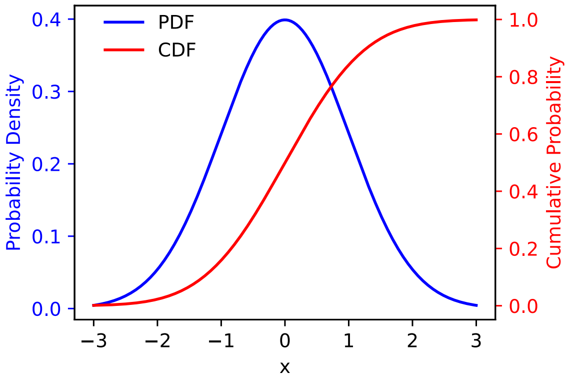


Figure 5.2: An example of a probability density function (PDF) and its corresponding cumulative distribution function (CDF).

Fundamentally, a correlation between the classifier output and the protected variable m manifests as a difference in $p(f_\theta(x)|m)$ at different values of m,

$$p(f_\theta(x)|m) \neq p(f_\theta(x)|m') \quad \forall m \neq m'. \tag{5.10}$$

If the classifier output distribution were the same at every value of m then the classifier output would be decorrelated from m. The conditional CDF maps the conditional distribution of the classifier output to a uniform distribution on the $[0, 1]$ interval. This means that at every value of the parameter m the classifier output distribution is identical.

$$F(f_\theta(x)|m) = F(f_\theta(x)|m') = U(0, 1) \quad \forall m, m'. \tag{5.11}$$

Therefore, $F(f_\theta(x)|m)$ is decorrelated from the protected variable m. This form of decorrelation also ensures the separation power of the classifier at each value of m is preserved because the CDF is order preserving. In practice, this means that in bins of m the separation power of the classifier is preserved.

When the underlying distribution of the data is unknown the CDF has to be estimated from data and this is often done using the empirical CDF $\hat{F}(x)$. This is defined as,

$$\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(x_i \leq x), \tag{5.12}$$

where $\mathbb{I}$ is the indicator function. When estimating the empirical *conditional* CDF the samples can be split into bins based on the protected variable m. However, this does not work well when the number of samples in each bin is small, and it does not leverage the continuity between bins. For this decorrelation application, we developed a novel approach to learning the conditional CDF that is referred to as a conditional flow for the remainder of this section. When this was originally presented this novel approach to CDF estimation was mixed with the idea of decorrelation when in fact they are separate. The approach to CDF estimation is described in detail in section 5.7.6.

5.6.2 Experiments

As an example of this application consider a problem where an ML model is trained to classify signal from background. For this task, a dataset of QCD jets and hadronically decaying boosted W bosons is used. Such a signal appears in many extensions of the Standard Model, and the large boost of the W boson means the decay products are mostly contained within a single large-R jet.

The samples were simulated in Refs. [121, 124] to emulate those used in an earlier ATLAS study on mass decorrelation techniques [125]. Both signal and background were generated by Pythia at $s = \sqrt{13}$ TeV with a detector simulated by Delphes [126]. Jets are reconstructed using FastJet [127, 128] and clustered using the anti-*kt* algorithm [49] with $R = 1.0$. Each jet is required to have transverse momentum $p_T \in [300, 400]$ GeV and mass $m \in [50, 300]$. For each jet ten substructure variables are calculated and used as input to the ML classifiers, these variables are the same as those used in previous studies of decorrelation [121, 122, 125].

The classifier used in our approach is an MLP (vDNN)[1] constructed from three hidden layers with 64 nodes in each hidden layer and ReLU activations [95] and a sigmoid activation on the output. The classifiers are trained for 100 epochs using the `Adam` optimizer [110] with an initial learning rate of 0.001 annealed to zero following a cosine schedule [129]. Comparisons are made to other decorrelation methods that have been developed for this task, MoDe [122] and DisCo [121]. These approaches introduce a $\mathcal{L}_{\text{decor}}$ term as in eq. (5.9) that is designed to decorrelate the classifier output from the mass. Following the prescriptions of the respective papers, models trained with MoDe decorrelation use a batch size of 16384 and 2048 for DisCo models. Both of these approaches require large batch sizes to be able to estimate $\mathcal{L}_{\text{decor}}$ accurately. The vDNN uses a batch size of 256.

The mass of the QCD background is a protected variable because the QCD background spectrum falls smoothly in the mass and the signal peaks around some resonant value. In the context of a search for boosted W bosons, the background could be estimated from data using the assumption of a smoothly falling background. Before estimating the background,

[1] MLP models are often referred to as deep neural networks (DNN).

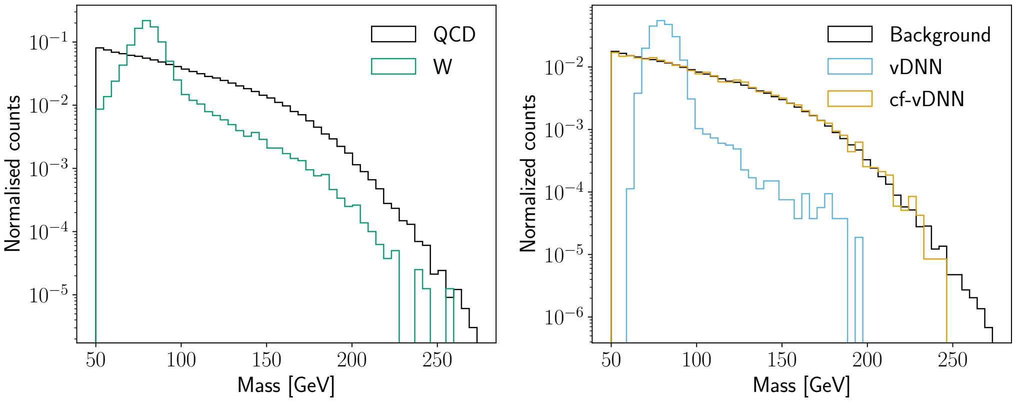


Figure 5.3: The invariant mass distribution for the background samples without applying any selection, and the mass profile after selecting a threshold that rejects 50% of the signal for a classifier (vDNN) that has no decorrelation applied, and a decorrelated discriminant that is the output of a conditional normalizing flow (cf-vDNN).

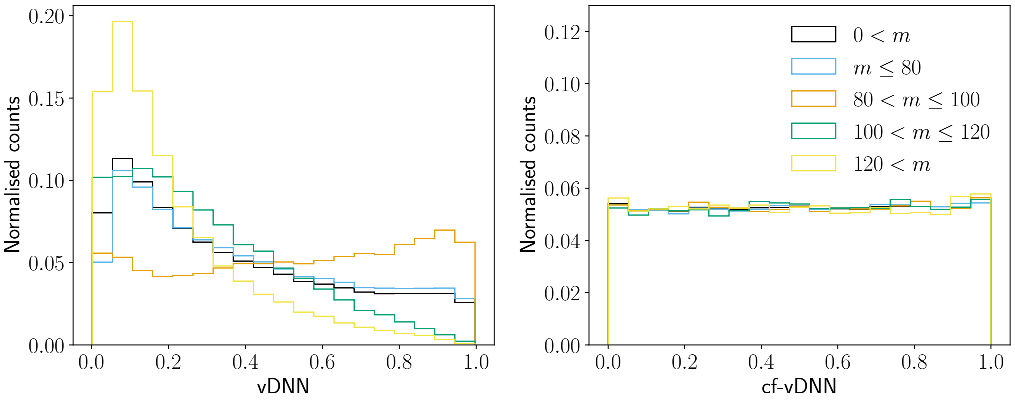


Figure 5.4: The distribution of the discriminant over QCD (background) samples for a vanilla deep neural network (vDNN) and a conditional flow trained on the vDNN output (cf-vDNN).

the signal would be enhanced by making a selection on a classifier output. If the classifier output is correlated with the mass then the background estimate becomes distorted. Note that this decorrelation is performed strictly on the QCD background, and the signal is not considered in this process. If the classifier were decorrelated on both signal and background, then no separation power would remain by definition. The mass distribution of the QCD background and signal, and the background that results after making a selection using correlated and uncorrelated classifiers are shown in Figure 5.3.

The output of a classifier that is correlated with the mass is distributed differently at different values of the mass. The conditional CDF of the classifier output $f_\theta(x)$ over the background only distribution maps this conditional distribution to a uniform distribution on the $[0, 1]$ interval. This is shown in Figure 5.4. After applying the conditional CDF (cf) the classifier output is decorrelated from the mass. This approach is applied after a classifier has been trained as a post processing step. It, therefore, can be used with classifiers trained using MoDe and DisCo approaches and significantly improves the performance of the classifier [123].

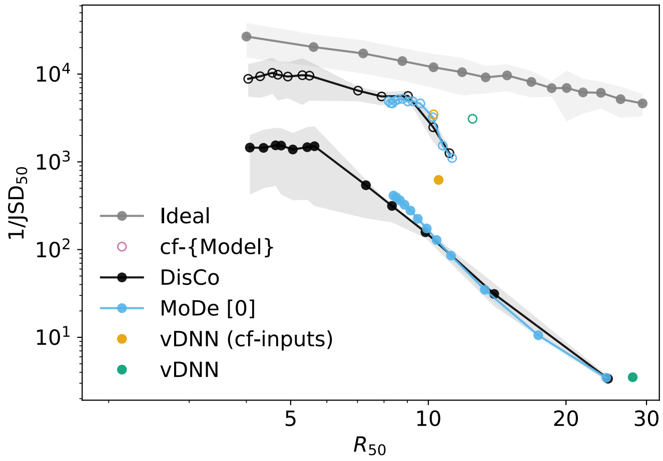


Figure 5.5: The background rejection at 50% signal efficiency (R_{50}) and the inverse of the Jensen-Shannon divergence between the mass distribution of events that do and do not pass the cut at the same threshold (1/JSD$_{50}$). The empty circles denote a classifier that uses the output of a trained conditional normalizing flow. Ten classifiers are constructed for every training paradigm and the shaded lines contain 8 of the classifiers for each paradigm. Figure we made for Klein and Golling [123].

To compare the conditional flow decorrelation we developed to existing methods the trade off between the inverse of the Jensen-Shannon divergence $1/\mathrm{JSD}_{50}$ and background rejection R_{50} at 50% signal efficiency can be studied. The JSD_{50} is computed between samples that pass and fail the selection. The $1/\mathrm{JSD}_{50}$ and R_{50} are correlated in the regime of finite statistics such that there exists an optimal trade off between these two metrics. This trade off can be estimated directly from data by applying random selections with the same proportions defined by R_{50}. This defines the optimal trade off and is referred to as ideal.

The conditional flow (cf) performs significantly boosts the performance of existing methods as shown in Figure 5.5. When applied to models trained with decorrelation objectives in the form of eq. (5.9), such as MoDe [122] and DisCo [121], the cf-decorrelation approaches the ideal limit. A conditional normalizing flow can also be used to directly decorrelate the input features to find a new representation on which a classifier (vDNN cf-inputs) can be trained. As this decorrelation is performed on all input variables directly, some correlation with the mass remains in the classifier output. This residual correlation is evidenced by the improved performance of training another conditional flow on the resulting discriminant.

5.6.3 Discussion

The same idea was recently proposed by Chakravarti et al. [130], where the motivation was the conditional CDF is the same as the optimal transport map to a uniform distribution in one dimension. An optimal transport map is a map that minimizes the cost of transporting one distribution to another. Minimal transport implies minimal changes to the distribution, and therefore minimal changes to the classifier output.

Optimal transport maps are also unique, and can therefore be used to define the CDF of a distribution. This definition is useful because it allows the CDF to be defined in higher dimensions, where the CDF is the optimal transport map to a unit hypercube [131–133]. This also means that higher dimensional CDFs can be used for decorrelation in more than one dimension. In HEP contexts this has been shown to be effective [134]. Further, in some settings, it may not be desirable to map the classifier output to a uniform distribution. The conditional optimal transport map to the targeted output distribution then replaces the conditional CDF as the more general form of decorrelation.

The issue of unwanted correlations with classifier outputs also appears in societal contexts where systematic biases appear in the datasets on which ML algorithms are trained. If these biases are not removed then the use of ML in decision-making reinforces these biases. There is a lot of work in the ML community on this problem under the umbrella of fairness [135]. Ideas from this field can be directly applied to HEP, where the same issues arise, and vice versa. This again highlights the strength of ML in being able to bridge domains.

5.7 Density estimation

Given samples $\mathcal{X}$ the task of density estimation is to estimate the true generating data distribution $p_D(x)$ from which the samples are drawn. Density estimation can be used to train generative models from which samples can be drawn, to identify rare events, or to learn maps between different distributions.

5.7.1 Maximum likelihood estimation

Given a parametric density model $p(\cdot|\theta)$ with parameters θ and observations $\mathcal{X} = \{x_i\}_{i=1}^N$ the function,

$$\mathcal{L}(\theta) = \prod_{i=1}^{N} p(x_i|\theta), \tag{5.13}$$

defines the likelihood. The maximum likelihood estimate of θ is then given by,

$$\hat{\theta} = \max_{\theta} \mathcal{L}(\theta). \tag{5.14}$$

This estimate is interesting because it defines the model that is most likely to explain the given observations given the family of densities parameterized by $p(\cdot|\theta)$. The next section looks at how parametric densities are defined in ML.

5.7.2 Normalizing flows

Normalizing flows [136] are composed of a tractable base distribution $p_b(b)$ and an INN T. A density $p_T(x)$ can be induced on the input space by applying the change of variables formula,

$$p_T(x) = p_b(T^{-1}(x))|\det J_{T^{-1}(x)}|, \tag{5.15}$$

where $J_{T^{-1}}$ is the Jacobian of the inverse INN T^{-1} [137]. This construction leverages the fact that invertible transformations deform the density in predictable ways and the density of simple distributions like a Gaussian, which is typically used for $p_b(b)$, is known exactly. Conditional densities can be learned by predicting the parameters of the INN as a function of the conditional attributes.

A normalizing flow can be fit to data using the maximum likelihood principle, where the maximization is performed over the parameters of the INN. The resulting model can be used to estimate the density directly, or it can be sampled from by first sampling from the base distribution[2] and then transforming this sample through the INN. These models have been shown to produce high fidelity samples [138]. A normalizing flow can also be fit to data by

[2] This is only possible if the base distribution is chosen such that it can be sampled from.

starting from a tractable likelihood that describes the data generating process but is hard to sample from. An example of this is the physics parameterized by a Lagrangian, as in QCD. Normalizing flows show promise in being able to generate samples in these settings [90, 139, 140].

5.7.3 Flowification

In work completed during this thesis, we showed how any network composed of MLP and CNN components that uses invertible activation functions can be converted into a normalizing flow [141]. To demonstrate this we had to overcome fundamental restrictions of invertible transformations, namely that they are dimension preserving. This can be approached by considering dimension altering operations that increase or decrease the dimension of the data.

Increasing the data dimension is straightforward, as samples u from a known distribution p_u can be produced and appended to data input to the INN, allowing arbitrary changes in dimension with the cost of replacing the exact likelihood with a lower bound. This operation is shown in Figure 5.6(a), and the idea is the data can be modelled on a larger dimensional space, this provides significant benefits and has been shown to improve the performance of normalizing flows [142]. The likelihood of the data under this operation is given by,

$$\begin{aligned}
\log p(x) &= \log \int du\, p(x,u) && (5.16)\\
&= \log \int du\, \frac{p_u(u)p(x,u)}{p_u(u)} && (5.17)\\
&\geq \int du\, p(u) \log \frac{p(x,u)}{p_u(u)} && (5.18)\\
&= \mathbb{E}_{u\sim p_u(u)}\big[\log \frac{p(x,u)}{p_u(u)}\big] && (5.19)\\
&= \mathbb{E}_{u\sim p_u(u)}\big[\log p(x,u) - \log p_u(u)\big]. && (5.20)
\end{aligned}$$

Where $p(x,u)$ is the joint distribution of the data and the augmented data, and $p_u(u)$ is the distribution of the augmented data. In modelling the joint distribution $p(x,u)$ the flow might not learn that $p(x,u) = p(x)p(u)$, in which case the lower bound is not tight.

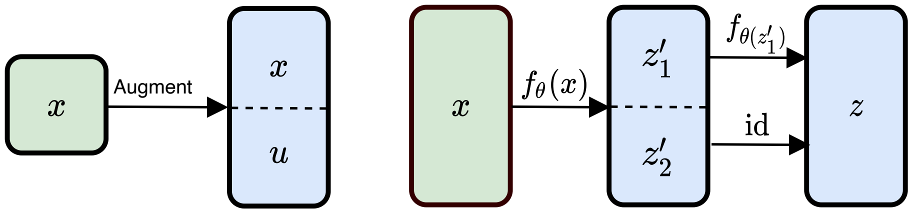
Reductions in dimension can be treated as being factorized models, where part of the input data is simply no longer processed by the INN [143] as shown in Figure 5.6(b). These two operations are enough to turn almost any neural network into a normalizing flow. For example, by decomposing the linear operations in MLPs using singular value decomposition,

$$W_i = U\Sigma V \tag{5.21}$$

every linear operation can be separated into two invertible matrices U, V, and a dimension altering and linear scaling operation Σ[3]. The two invertible matrices can be described as standard flow layers, and the dimension altering operation can be described as either a dimension preserving, increasing, or decreasing operation as described above.

The approach outlined here, while interesting, requires additional modelling components to be effective. In particular, we found it necessary to augment MLPs and CNNs with standard normalizing flows to be performant on density estimation tasks, and even then

[3] Any singular values that become zero can be treated as a dimension reducing operation followed by a dimension increasing operation.



(a) (b)

Figure 5.6: Dimension altering flow layers as shown in Máté et al. [144].

results were mixed. This work was an interesting pursuit and does have some interesting insights into how information is processed through standard ML networks. In particular, it provides a principled way of injecting noise into neural networks, which may have interesting applications [145]. It is also interesting to observe that standard neural networks are poor density estimators, especially given these models are excellent at information extraction as demonstrated by their performance on classification problems.

5.7.4 Flows for flows

A normalizing flow can be used to learn a map between distributions. In the standard training approach, a normalizing flow learns a map from the data distribution to the base distribution. In this thesis, we showed how this idea can be extended to learn a map between two data distributions. Given samples from a distribution $p_D^x(x)$ a map to another distribution $p_D^y(y)$ can be found by first fitting a normalizing flow p_{T_x} to the samples from $p_D^x(x)$ and then fitting a second flow p_{T_y} to the second dataset using the first flow as the base density,

$$p_{T_y}(y) = p_{T_x}({T_y}^{-1}(y))|\det J_{{T_y}^{-1}(y)}|. \tag{5.22}$$

In work completed as a part of this thesis, we explored the use of this idea in HEP contexts [146–148]. An example of this is shown in Figure 5.7, where a normalizing flow is trained to rotate a distribution of overlapping circles. This example is particularly interesting as the standard approach results in samples populating the space between the circles, while the flows for flows approach results in samples that are rotated versions of the original circles. The reason the standard approach fails is the base distribution is not well suited to the problem as it is topologically distinct from the target distribution. The flow for flow approach is able to learn the rotation by mapping actual data samples to the target distribution.

5.7.5 Flow matching

Recent progress in generative models has largely come from diffusion models [150] and more recently the concept of flow matching [151, 152]. The latter is a general framework that combines elements of diffusion models and continuous time normalizing flows, which are normalizing flows that use ODEs (eq. (5.6)) to define the INN. Similar approaches are the key method behind high fidelity models like Stable Diffusion [70].

The flow matching approach was designed to overcome the need to integrate in time the full ODE when using continuous time normalizing flows. The idea is to define a parametric vector field $f_\theta(u_t, t)$, as in eq. (5.6), and a conditional probability path $p_t(x|x^i)$ that interpolates between the base distribution p_b and any given sample x^i. Samples from a simple example of such a path can be drawn using the so-called linear schedule,

$$x_t^i = (1-t)x^i + tu, \quad x_t^i \sim p_t(x|x^i) \tag{5.23}$$

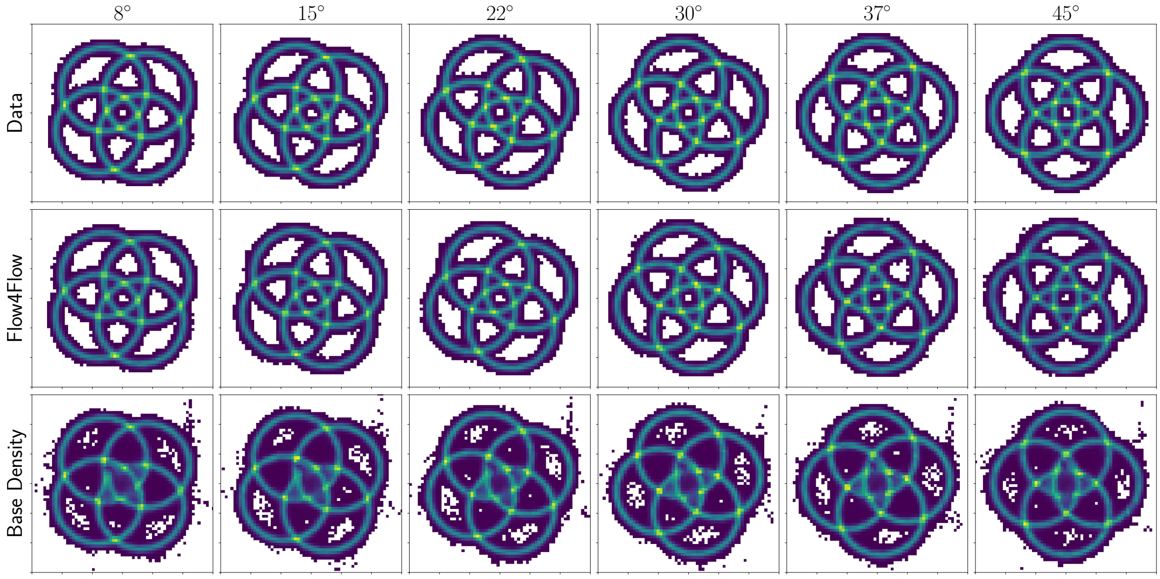


Figure 5.7: Conditional flow for flow model trained on a distribution composed of four overlapping circles conditioned on an absolute angle. The Flow4Flow model takes as input an unrotated circle and is trained to rotate the circle. The Base Density approach is the standard flow based approach to solving this problem. Figure as shown in Klein, Raine, and Golling [149].

where $u \sim p_b(u)$. Defining this probability path defines a velocity field and the objective is to regress this field. This can be done with a simple objective,

$$\mathcal{L}(x^i, t, u, \theta) = ||f_\theta(x_t^i, t) - (u - x^i)||^2, \tag{5.24}$$

which is computed over data samples $x^i \sim p_D(x)$, time samples $t \sim \mathcal{U}[0,1]$ and base distribution samples $u \sim p_b(u)$ during training. Crucially only one time sample and one base distribution sample are required for each data sample in a batch. This has turned the full integration that is required to train a standard continuous time normalizing flow, into a simple regression problem that is fast to calculate. The training dynamics are also arguably simpler, as the model does not have to learn to map from the base distribution to the data distribution *and* to find a path. Instead, the path is prescribed and the model 'simply' has to learn to reproduce this path to generate samples.

5.7.6 Estimating the conditional CDF

As discussed above, normalizing flows can be viewed as maps between different distributions, and therefore they can be used to estimate the CDF. This is because the CDF is a map from the data distribution to the uniform distribution. The only additional constraint is the CDF is monotonically increasing. This can be accounted for as all invertible functions in one dimension are monotonic, and a monotonically decreasing function can be made monotonically increasing by taking the negative of the function. Therefore, an INN that estimates the CDF can be fit using a normalizing flow with a uniform distribution as the base distribution. The same can be done for a conditional CDF, where the INN is conditioned on the protected variable m.

We developed this approach for the decorrelation of classifiers in HEP [123]. However, it is also interesting in its own right. For example, the CDF can be used to directly estimate the quantiles of a distribution. Quantile regression has many applications [153]. The approach we developed appears to estimate quantiles more robustly than other approaches [123], and future work should look at this in more detail. As a conditional quantile regressor, we observe that our approach estimates all quantiles simultaneously, and also leverages continuity in the conditional distribution. This may explain why it was observed to be more robust than other approaches [154, 155], with the estimation of additional quantiles providing additional constraints on the model, and continuity providing additional information.

5.8 Foundation models

It can be resource intensive to design and train ML models for every possible task independently. It is also inefficient because many tasks are similar and therefore require similar modelling. This is why similar reconstruction algorithms can be used for different physics analyses. A similar approach can be taken when training ML models, where an already (pre-)trained model can be 'fine-tuned' on a new dataset or task. This fine tuning works by training a model initialized with the pre-trained weights.

In ML it is becoming increasingly more common to train 'foundation models' (FMs). An FM is a large model that is trained using self-supervised learning (SSL) on a large dataset. SSL is a form of unsupervised ML where labels are derived directly from the data. It is often observed that FMs require smaller datasets and training times to be performant. This is particularly true for transformer architectures, which typically require huge volumes of data to train. In HEP it is hoped that a large FM could be trained directly on experimentally observed data and applied to an array of different tasks within different collaborations [156].

While working on this thesis we helped to develop an FM for HEP [156]. This approach is referred to as masked particle modelling (MPM). The strategy is based on some of the most successful foundation models developed for text [157] and image data [158]. In MPM a subset of the particles in a jet are removed, and a model is tasked with predicting properties of the dropped particles using the remainder of the jet. The MPM approach trained a 40 million parameter transformer model to use as an FM. Training this model on a dataset with 100 million samples, and then fine-tuning on smaller datasets from the same and different distributions, we observed the FM is more sample and training time efficient.

The MPM strategy can use data directly and is similar to but more general than Kishimoto et al. [159]. Earlier work had looked at self supervision using augmentation [160]. Several further approaches have since been developed [161–163], including an extension of the MPM work to which we contributed [164]. Notably amongst this work, Vigl, Hartman, and Heinrich [165] have looked at fine tuning the reconstruction pipeline used by experiments like ATLAS. This reflects the possibility of having end-to-end optimizable analyses in HEP. The development of FMs for HEP is an exciting area for further development.

5.9 Discussion

The use of ML tools to solve HEP problems shows great promise, and these techniques are useful throughout the rest of this thesis. While the future is bright, these models face issues in their deployment, particularly as they are further integrated into the physics workflow. In particular, the domain shift between simulation and data is a significant challenge. This occurs because ML models require labelled training data, which can only be obtained from simulation. However, the simulation is not a perfect representation of real data. There are promising techniques for this issue like invariant risk minimization [83, 161, 166], as an alternative to empirical risk minimization, and data based training [116, 117].

Another possible challenge is that ML models are black boxes in the sense that a simple formula to describe their action on data is often not available. This can mean, for example, that an ML model could identify a discrepancy between a background estimate and data, and it is not possible to characterize this discrepancy in terms of a simple formula. Interpreting such discrepancies in a physics context can be difficult and an example of this appears in this thesis.

We believe the issues we face when integrating ML tools into HEP are surmountable and these tools offer many exciting opportunities for the future.

Chapter 6

Weakly supervised searches

This chapter introduces and discusses the idea of weakly supervised searches. The search presented in this thesis is based on a weakly supervised strategy and this chapter provides the necessary background to understand the methods used in the analysis.

Definition 1. *A weakly supervised search makes generic assumptions about new physics processes.*

The discussion should be interpreted within the context of new physics searches in HEP. This discussion distinguishes between unsupervised, or weakly supervised, *searches* and unsupervised, or weakly supervised, *learning* as defined in section 5.1. As an example of the distinction, a generic assumption one can make about new physics is that it populates low density regions of phase space. It might be possible to use unsupervised learning to identify such events. This would define a weakly supervised *search* that uses unsupervised *learning*.

The need for weakly supervised searches is motivated by the large space of possible new physics models and the possibility of unforeseen new physics scenarios to which standard approaches may not be sensitive. As these searches do not make strong assumptions about new physics processes they have broad sensitivity to different processes. This comes at the cost of decreased sensitivity to any given model as compared to dedicated searches. These approaches have gained significant interest in HEP [167–170].

This chapter first presents fully supervised and unsupervised searches to provide a contrast to the weakly supervised search approach. Some approaches used to perform weakly supervised searches are then discussed. The conclusion summarizes this discussion.

6.1 Supervised searches

Definition 2. *A supervised search makes strong assumptions about new physics processes.*

If a new physics process is specified and can be simulated, then a supervised search for that process can be performed. In this setting, simulated signal samples can be drawn from $p_s(x|\theta)$, where θ defines the new physics process and x is some set of features that can be reconstructed by the detector and are useful for discriminating signal from background. Given these samples, and a background model $p_b(x)$, the likelihood ratio can be constructed. By the Neymann-Pearson lemma [171] this is known to be the most powerful discriminant for distinguishing between two distributions using the information in x. The resulting search is maximally sensitive to any θ' that satisfies $p_s(x|\theta') = p_s(x|\theta) \forall x$. Any model that deviates from this no longer uses the optimal test statistic or selection criteria, and the sensitivity of the search is necessarily reduced. These considerations justify the statement that a fully supervised search has a high sensitivity to a narrow range of models.

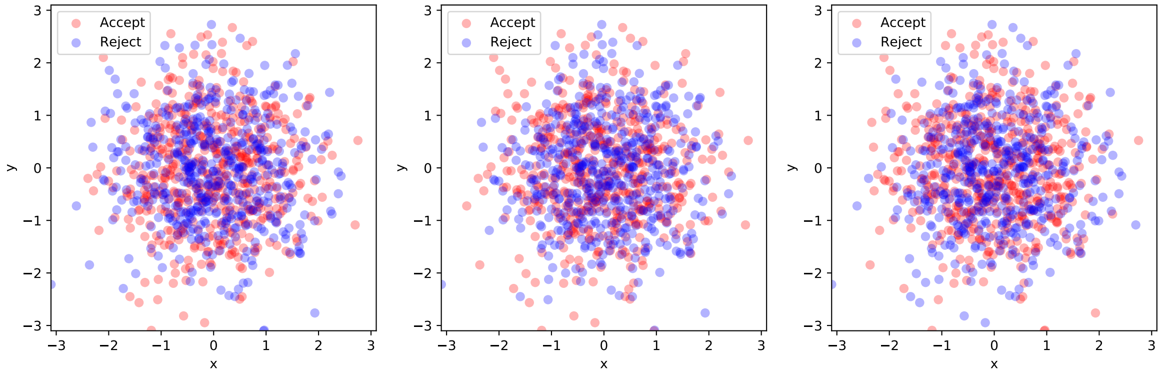


Figure 6.1: A visualisation of randomly sampling data without preference three different times. The red points are considered anomalies and accepted when applying a threshold using this score.

6.2 Unsupervised searches

Definition 3. *An unsupervised search makes no assumptions about new physics processes.*

This section discusses unsupervised search strategies. When discussing unsupervised searches there is an instinctive feeling that it should be possible. After all, humans can often spot anomalies in data. The problem with this justification is there is limited utility in using the human brain to reason by analogy in this context. The human brain can not find all types of anomalies, and it is not clear where the ability we have originated. The following attempts to elucidate why it is not possible to design an unsupervised search with meaningful sensitivity to new physics.

6.2.1 Random functions

One of the first steps to consider when designing a search is to make some selections to reduce background (section 4.3). Selections can be made by using a function to assign a score to each sample in the targeted dataset and then selecting events based on this score. To design an unsupervised search, the function that assigns scores must be chosen without any knowledge of the signal. To make the problems that arise in this context explicit one can consider the use of random functions. These functions also appear in a different context later in this thesis.

One approach to assigning scores that is unarguably unsupervised is to assign them randomly. Making a selection in this setting, shown in Figure 6.1, is equivalent to randomly downsampling the full dataset to define the signal enriched dataset. As this score has no continuity signal samples are just as likely to be sampled as background samples. Therefore, this construction results in a dataset where the signal-to-background ratio remains unchanged, and the discovery potential has strictly decreased, as discussed in section 4.3.

The main issue with assigning scores randomly is the lack of continuity in the score. Therefore, a natural minimal extension is to consider sampling random functions from a family of continuous functions. The sampled function can be used to assign scores to the data points as visualized in Figure 6.2. Each of these functions has some set of signal samples that populate the space that is accepted when making a selection and therefore each of these functions can increase the sensitivity of a search to a given set of signal processes. However, the set of signal samples to which any given function is sensitive is random, and might not be interesting for the physics goals of the experiment. For example, this set of signal

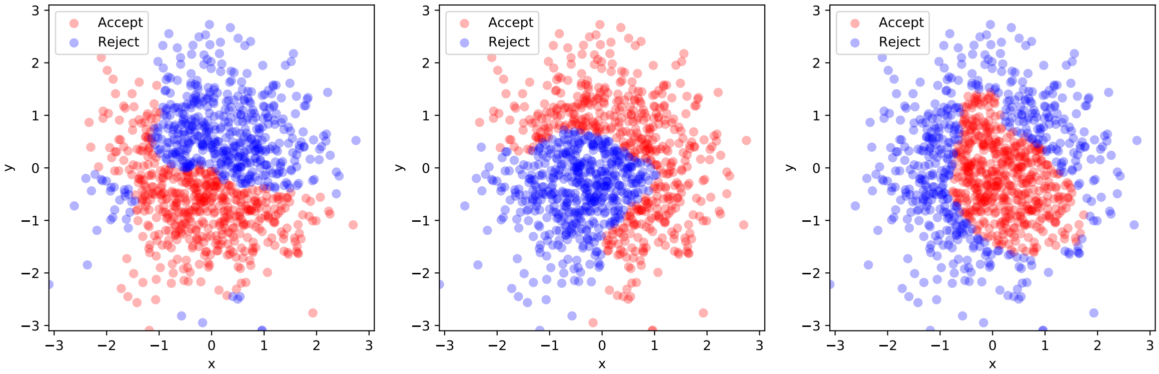


Figure 6.2: A visualization of sampling three random functions from randomly initialized neural networks and using these functions to assign a score. The red points are considered anomalies and accepted when applying a threshold using this score.

samples could be excluded by existing searches or could violate fundamental assumptions that would require significant evidence to discard [57]. These selections might also make it difficult to model the background distribution in the reduced dataset, and therefore to perform a hypothesis test.

Assuming that random functions are sampled in a way that makes any portion of the data space equally likely to be rejected, the average overall selections simplify to randomly downsampling the data. Therefore, for this approach to be useful the set of random functions needs to be biased in some way. A bias could be chosen randomly, but this is unlikely to lead to physically interesting insights. Choosing a meaningful bias requires a definition of new physics, inevitably leading us to define a weakly supervised search.

The purpose of this discussion was to demonstrate that unsupervised search strategies as defined in Definition 3 can only have random sensitivity and in trying to correct for the deficiencies of random sensitivity one arrives at a weakly supervised search. A search can only target new physics that is interesting to an experiment if it makes some assumptions about the form of this new physics. There is a trade-off between sensitivity and the number of assumptions that are made, with more assumptions resulting in higher sensitivity. Weakly supervised searches make some minimal assumptions about new physics in an attempt to develop broad sensitivity.

6.2.2 Perfect background model

An alternative approach to an unsupervised search is to assume perfect knowledge of the background model and not make any selections. New physics could in principle be found by comparing simulated samples from the background model to data. Such a comparison could be made using the new physics learning machine[1] [172] which calculates the test statistic directly by comparing two samples and can account for all systematic uncertainties [173]. This comparison requires a background only sample to be produced and a selection of features x over which to search for deviations. In the context of a supervised search, this feature set is selected for the targeted signal model, in an unsupervised setting no signal model is assumed. Choosing a large set of features increases the sensitivity to statistical fluctuations in the data and simulation, whereas a small feature set is only sensitive to some signal models. Further, a perfect background model does not exist, and simulation often

[1] This approach can be used any time a hypothesis test needs to be run.

needs to be corrected to data. This correction can only be derived by defining signal depleted regions, which again requires a definition of new physics. Given these considerations, we classify this approach as a weakly supervised search. It should also be noted that such an approach is always less sensitive to specific processes than a supervised search as a supervised search is only sensitive to fluctuations where the density $p_s(x|\theta)/p_b(x)$ is large, whereas global comparisons are sensitive to fluctuations relative to $p_b(x)$ everywhere $p_b(x)$ has support.

6.3 Rare event detection

One common assumption about interesting events is they are rare. Rarity is defined by areas of relatively low density in the data over some set of features. Such search strategies can not be sensitive to new physics that populates the bulk of the background distribution. Selecting rare events requires an estimate of the density, a task for which many methods have been developed as discussed in section 5.7.

As with all weakly supervised searches, feature selection defines an important assumption in rare event detection. However, in this setting the scaling of the features also becomes relevant. As discussed in the definition of normalizing flows in section 5.7.2, transformations of the features transform the density. Therefore, scaling the features changes the densities and therefore the definition of what is rare. This means a rare event can be transformed into a common event, or vice versa [174]. Therefore, the preprocessing of the features in this approach defines an important assumption. This can not be avoided by accounting for changes in density due to preprocessing as the original scale is defined by a choice of units, which is itself arbitrary.

Feature selection is also important in this context as it defines the dimension over which a density must be estimated. This is a particularly relevant concern in this setting as the density is small everywhere in high dimensions. In estimating small densities to identify rare events, random fluctuations in the training dataset and learning procedure become relevant. A small random fluctuation in a density estimate can easily swap the rarity of two samples. Further, in high dimensions, the typical set of a distribution can be shifted from areas of high density, which makes it difficult to estimate the density without an explicit prior [175]. In addition to all of this, rare event detection should do something more interesting than making simple selections based on high level features. For example, if the selection made by the rare event detector can be approximately reduced to selections based on the transverse momentum, then a density estimate adds little value.

Another particular problem of rare event detection is the lowest density regions contain the least statistics by definition[2]. Therefore, the largest fluctuations in the density estimate occur in exactly the regions that are the most interesting for rare event detection. Such fluctuations lead to an unreliable method for detecting rare events. These fluctuations occur due to the statistics of the dataset as well as the presumably stochastic training of the density estimator. The latter can be marginalized out through ensemble methods, but addressing the former would require low density regions to be upsampled – which requires knowledge of the density. Low density regions could be upsampled through iterative use of a density estimate, but to the best of our knowledge, this has not been explored.

Many different approaches have been developed for rare event detection based on unnormalized [176–179] and normalized density estimates [180, 181]. While difficult, these approaches

[2] Unless the dataset is weighted and these regions have been upsampled using some procedure that does not already assume knowledge of the density.

do have some successes, in particular when they are not integrated into a search. For example, CMS has integrated this approach into their L1 trigger system [182] and uses it for online data quality monitoring [183]. There are several approaches to integrating rare event detection into a search. This has been done using the ABCD method [184], a data driven background estimation technique based on partitioning the data. Another approach is to make a selection based on the score defined by the density estimate and then run a bump hunt, this places additional requirements on the selection [185]. Specifically, the score that is assigned can not be more than quadratically correlated with the feature in which the bump hunt is performed [186]. This condition is typically handled by decorrelation techniques as described in section 5.6.

6.4 Bump hunts

One broad assumption about new physics is that it is produced at resonance. Some dominant background processes are also known to yield smoothly falling spectra. The resonant production of particles can be detected as bumps in such spectra. A search for a bump is a weakly supervised search where the primary assumption is that new physics is produced at resonance. As an example, QCD multijets are the dominant contribution to SM dijet events, which results in a smoothly falling dijet mass (m_{JJ}) spectrum in which bumps can be hunted. This section describes a bump hunt using m_{JJ} as a specific example.

To perform a search for a bump it is necessary to make a background prediction. This is typically done by performing a background only fit to the m_{JJ} spectrum. Data can be directly compared with this fit in a hypothesis test, or a background plus signal fit is performed. These fits mostly use an assumed functional form [114, 187, 188]. An alternative is to use non-parametric approaches that directly incorporate knowledge of the physics [189]

A bump hunt can be performed on the full spectrum inclusively, or in sliding windows. The analysis performed in this thesis uses a sliding window bump hunt. This involves splitting the spectrum into different bands as shown in Figure 6.3. The lower end of the spectrum is referred to as sideband one (SB1), the middle region is referred to as the signal region (SR) and the final region is referred to as sideband two (SB2). Collectively, SB1 and SB2 are referred to as the sidebands (SBs). To perform a sliding window bump hunt, the data in the SBs is fit with the SR masked. The fit is then used to make predictions in the SR. These predictions are taken to be the background prediction in a hypothesis test. The fit uncertainty is calculated by considering variations in the prediction to first order using the Jacobian of the fit function. An additional uncertainty is assigned due to a lack of knowledge about the correct form for the fit [114]. Each SR is tested separately and then combined to form a global p-value as in BumpHunter [190]. The search presented in this thesis does not calculate a global p-value.

While bump hunts are an effective weakly supervised search they only leverage the information contained in the spectrum on which they are performed. To increase the sensitivity to possible new physics they can be performed on datasets that have passed additional high level criteria and some generic selections such as requiring the presence of a photon [191] or for jets to contain a b-hadron [114]. Such selections are typically motivated to enhance the sensitivity to generic classes of new physics. Bump hunts can be performed directly in higher dimensions, but this introduces new problems and there is a limit on the variables that can be considered [192, 193].

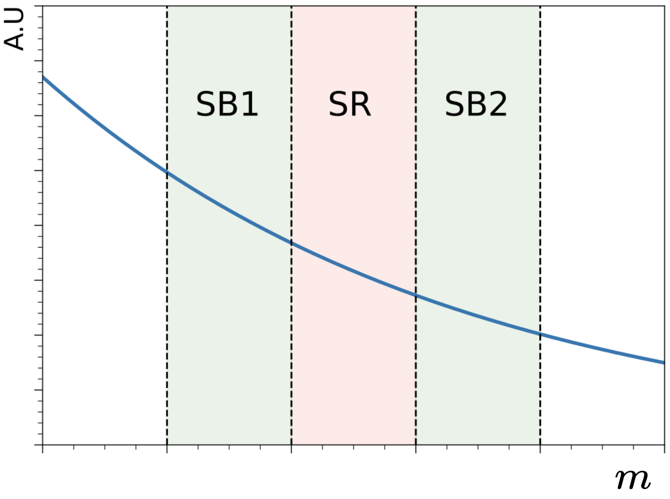


Figure 6.3: Example of a bump hunt set up on a smoothly falling spectrum m. The first sideband (SB1) is at lower values of m and the second (SB2) at higher values of m than the SR (SR).

6.5 Extending bump hunts

To increase the sensitivity of a bump hunt it is necessary to add additional information beyond that contained in the spectrum on which the hunt is performed. Additional information can be provided in the form of additional variables x. In analogy with a bump hunt, the goal is to produce a background estimate over these variables. At present no simulator can produce an accurate enough estimate of the data to be used in this context, and estimates are derived by including data in the SBs. Selecting the variables in x again introduces an important assumption about new physics. Having chosen x the goal becomes to produce an estimate of the background in these variables. Here, no functional form for x is assumed, instead, it is assumed a background estimate can be produced following some procedure. This is a non-trivial assumption that requires robust validation to be incorporated into an analysis.

The next subsections detail how background 'reference' estimates can be produced for variables in x and how they can be incorporated into an analysis. This type of search is used to perform the analysis presented in this thesis.

6.5.1 Reference generation

Following the sliding window strategy, data in the SBs can be used to produce a background estimate in the SR $p_b^R(x)$. This is typically done by interpolating conditional densities over x and then integrating over m_{JJ} such that,

$$p_b^R(x) = \int p_b^R(x|m_{\mathrm{JJ}}) p_b(m_{\mathrm{JJ}}) dm_{\mathrm{JJ}}, \tag{6.1}$$

where $p_b^R(x|m_{\mathrm{JJ}})$ is the conditional estimate of the density from the SBs and $p_b(m_{\mathrm{JJ}})$ is the estimate of the m_{JJ} distribution from a functional form fit to the SBs. Structuring the problem in this way allows knowledge of the m_{JJ} distribution to be encoded directly. Leveraging knowledge of the form of m_{JJ} minimizes the mismodelling in this feature. The other features in x are chosen to be mostly uncorrelated with, and smoothly varying as a function of, m_{JJ} such that they can be interpolated.

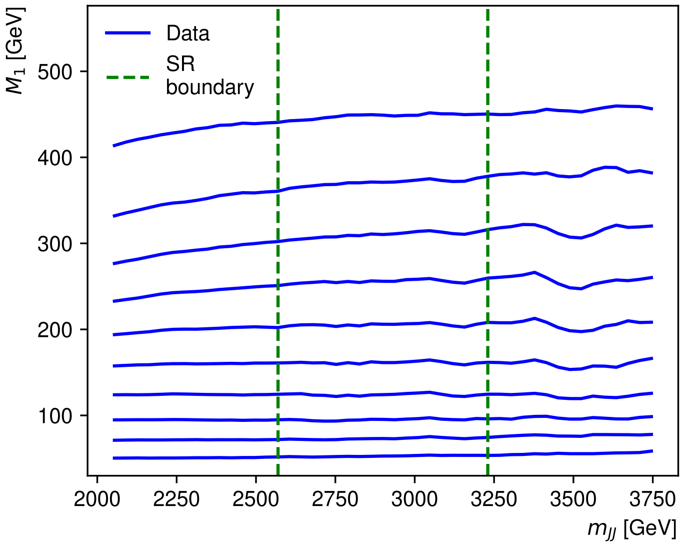


Figure 6.4: Ten different m_{JJ} conditional quantiles evenly spaced between 5% and 95% for the mass of the leading jet M_1 in a dijet system. The green lines demarcate a possible SR boundary.

With the formulation of Equation (6.1) the problem reduces to estimating the conditional distribution $p_b^R(x|m_{\mathrm{JJ}})$ using data in the SBs and interpolating it into the SR. One way to visualize this is as a conditional quantile estimation task as shown in Figure 6.4. To estimate the distribution in the SR every quantile needs to be interpolated from the SBs into the SR. Visually this can be related to the interpolation performed in the standard bump hunt of Figure 6.3. While this is a useful visualization in one dimension, the problem is more complex in higher dimensions.

Most approaches that have been developed to perform the interpolation of $p_b^R(x|m_{\mathrm{JJ}})$ either directly or indirectly estimate the conditional density. These approaches can be categorized into whether they are simulation assisted or data driven and whether they estimate densities directly, use the likelihood ratio or use optimal transport, as shown in Table 6.1. This classification is useful in that it identifies that no method has been developed for this problem that is data driven and uses the likelihood ratio. The data derived validation sets described in a later chapter (section 8.2.3) could be used in this context. The interpolation of likelihood ratios from data derived control regions (SBs) has been performed in searches [194–196], but it is unclear how to best apply it in this context.

Table 6.1: Different reference generation techniques are classified by whether they use likelihood ratios, density estimation or optimal transport and whether they correct simulation or are data driven. This table characterizes methods on the approach they use to interpolate, not how they are integrated into an analysis.

	Ratio	Density	Transport
Simulation	SALAD [197, 198]	DRAPES [199], FETA [200]	OT-KNN [196]
Data		(R-)ANODE [201, 202] (LA)CATHODE [203, 204] CURTAINS [148] DRAPES [199]	RAD-OT [205] CURTAINSOT [206]

The first approach to tackle this problem was ANODE [201] where the density is estimated in the SBs with a normalizing flow and interpolated directly to the SR. The same approach is used in Cathode [203] but with a different integration into the analysis as discussed in the next subsection. These approaches are data driven and do not apply any regularization to the density estimate. In general, when performing interpolation, it is expected that some kind of regularization would be beneficial.

In CurtainsOT [206] some implicit regularization was added by using an invertible neural network (INN) to map from $p_D(x|m^0_{\mathrm{JJ}})$ to $p_D(x|m^1_{\mathrm{JJ}})$ where p_D is the data distribution. The model is trained to minimize the difference, as quantified by an optimal transport measure, between the output of the INN and samples from the target distribution. The next iteration of Curtains improved on this by using density estimation directly to map between different m_{JJ} values using the approach outlined in section 5.7.4. This was found to give a more stable estimate of the density and resulted in better performance than the original CurtainsOT approach.

Complementary simulation assisted efforts have also been developed in this context. The Salad [197] method corrects simulation to data in the SBs using a classifier parameterized by m_{JJ}. A reference is generated with Salad by applying the classifier to the simulation in the SR. Both Salad and Curtains are used in the analysis presented in this thesis and are detailed in a later chapter.

There have also been studies into the combination of different reference generation techniques [207].

6.5.2 Analysis integration

Once an estimate of the background in the SR has been produced, the challenge is to integrate it into an analysis. A hypothesis test based on a direct comparison between the data and the reference is technically possible but presents several challenges. First, the reference is not a perfect estimate of the background, and it is unclear how to assign uncertainties. Second, even if the reference were a perfect estimate of the background, no signal model is assumed, and so a hypothesis test is sensitive to fluctuations anywhere the reference has support, as is the case in an unsupervised search.

Instead of a direct hypothesis test, the reference is used to define a score that can be used to make a selection. After the selection is made a bump hunt is performed in m_{JJ} as in the previous subsection. The selection can enhance the sensitivity of the bump hunt to new physics by leveraging the additional information in x to reduce the background. This approach has the advantage of utilizing the additional information in x to enhance the sensitivity of the search without needing to produce a perfectly accurate reference over x. Of course, the better the reference the more sensitive and less error-prone the search.

To be able to run a bump hunt after making a selection the selection must not be quadratically correlated with m_{JJ} [186] to avoid sculpting a bump in the background. The form of Equation (6.1) promotes decorrelation but does not guarantee it, as mismodelling in the reference can be correlated with m_{JJ}. Ensuring this condition is satisfied requires extensive validation of any analysis performed using these methods.

In Anode the conditional likelihood in both the SBs ($p^R_b(x|m_{\mathrm{JJ}})$) and the SR ($p_s(x|m_{\mathrm{JJ}})$) is fit and the likelihood ratio ($p_s(x|m_{\mathrm{JJ}})/p^R_b(x|m_{\mathrm{JJ}})$) of the two density estimates is used as a score. This is problematic as the two densities are fit separately, which leads to uncorrelated fluctuations in the two likelihood estimates. This makes the likelihood ratio sensitive to fluctuations in low density regions of $p^R_b(x|m_{\mathrm{JJ}})$, reducing the sensitivity of the search. This problem is resolved by r-Anode [202] by fitting the likelihood in the SR using the SB estimate as a base distribution. In R-ANODE the two likelihoods are coupled and the likelihood ratio becomes less sensitive to random fluctuations.

The other approach to assigning scores is to generate samples from the estimated density $p_b^R(x)$ and use these samples to train a classifier as in the CWoLa approach [116] (section 5.5.1). Here, the classifier learns the likelihood ratio directly and thus is smoother than the ratio of two uncorrelated density estimates. This approach was shown to be successful in CATHODE [203] and has since been the primary method for developing these approaches. The method is more powerful with more samples from the estimated density [203]. It has also been shown the type of classifier is important, with tree based algorithms performing better than neural networks [98].

While the classifier approach has been shown to be effective for signal enhancement, it leads to issues when the reference perfectly describes the data and there is no signal present. In this case, the likelihood ratio is constant everywhere, and the classifier should therefore assign a constant to all samples. However, in the setting of finite statistics and training time the classifier instead places a decision boundary randomly as shown in Figure 6.2. This can conflict with the classifier decorrelation requirement as a randomly placed boundary is correlated with the mass with a certain probability. This means that even in the absence of mismodelling in the reference the classifier can sculpt a bump in m_{JJ}.

While this approach has been shown to have the potential to increase the sensitivity of bump hunts it is unclear of its real utility. The sensitivity of the search is dependent on the amount of signal present in the data, and new physics is most likely produced at low cross sections. Further, if an excess is observed it is difficult to identify what feature of the data is responsible. Analysis failure and new physics detection are not easy to distinguish in this context. More work on procedures to follow the event of discovery is required to increase the benefits of these approaches.

6.6 Limit setting

As already discussed, searches are intended to both discover and exclude new physics. The discussion of this chapter so far has focused on the discovery aspect of searches. In searches for rare events standard procedures for setting limits can be applied directly [208, 209]. In an extended bump hunt, however, this becomes non-trivial. The sensitivity of these approaches increases with the amount of signal that is injected. If there is no signal in the data then the analysis has no sensitivity and nothing can be excluded.

To set limits in extended bump hunts it has become standard to inject simulated signals. The procedure for setting limits in this context is outlined in Figure 6.5. First, signal is injected into the data at a certain cross section. Then the analysis is run up until after the classifier selection. At this point the signal samples are removed from the data and the CL_s procedure is run with the signal strength (μ) set such that the expected signal observed after the classifier is applied corresponds to $\mu = 1$. This procedure is run for a grid of signal injections and the crossings of the relevant quantities of the 5% boundary are reported as limits.

This approach is complicated because at every level of signal injection, a valid limit can be extracted. The more signal that is injected the tighter the limit that is set. For some physics cases these limits may be interesting to report, however by injecting a simulated signal they become closer to a supervised search. As it stands the limits that are reported by extended bump hunts are consistent with how limits are normally derived in physics, where

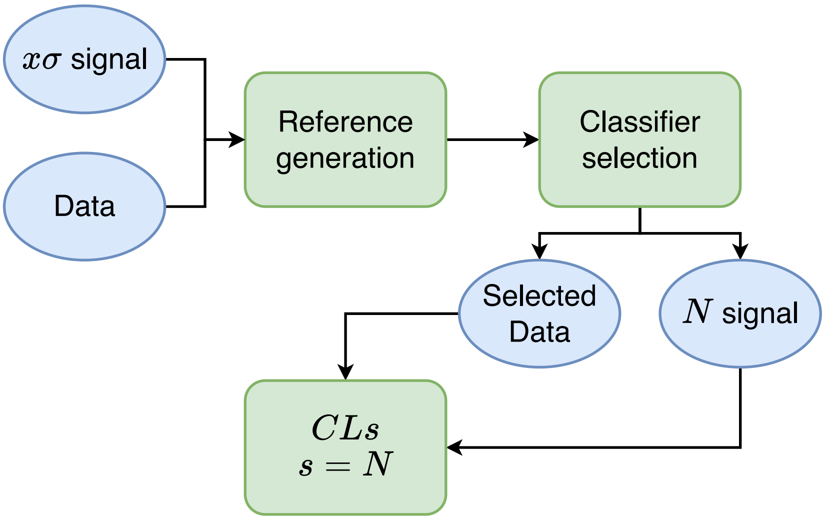


Figure 6.5: A flow chart of how limits are set in an extended bump hunt.

the signal strength (μ) is varied until the 5% boundary is hit. However, the limits are not model independent, while the analysis is intended to be 'model agnostic'.

6.7 Conclusion

This chapter introduced the concept of weakly supervised searches and methods for performing them. Searches based on other assumptions have been developed beyond what is described here [210], such as assuming the factorization of scales holds differently for signal and background [211] or that it breaks a known symmetry of the SM [212]. Enumerating the assumptions made by a weakly supervised search is useful for identifying the kinds of processes that can be detected as well as additional useful assumptions. Due to the necessity of these assumptions, we think it is important to acknowledge these approaches are not model independent. Such terms are misleading and should be avoided. To have something other than non-random sensitivity to new physics some assumptions must be made about both the form of the new physics model and the form of the background. Unfortunately, almost all methods described in this chapter have uncontrolled assumptions in the form of the hyperparameters involved in defining the ML components that parameterize them.

It is also important to be aware of the separate tasks that must be performed to leverage the weak assumptions made in these searches. In rare event detection, some form of density estimation is required, and this estimate is required to be performant in relatively low density regions. For extended bump hunts, precise high dimensional interpolation and weakly supervised classification are required. In this, it is important to acknowledge that interpolation is a separate task. For example, the main contribution that we made in developing the CURTAINS approach was the novel method for interpolating the density estimate [148]. By separating the interpolation and classification tasks the development and validation of these methods can be more focused.

In settings where limited searches have been performed or where the new physics is expected to be produced at large cross sections weakly supervised searches likely have utility. For example, extended bump hunts may prove to be effective in settings where high fidelity simulation is unavailable, discriminatory features are easily interpolatable and analyses can be rigorously validated. In developing such strategies it is particularly important to develop approaches to further increase their sensitivity and make them more interpretable.

It is unclear what impact weakly supervised approaches will have on the future of HEP. In particular, if new physics at the LHC is only accessible at small total numbers, then it is unlikely these approaches will be useful. However, as with many tools in science, the approaches developed for weakly supervised searches have found applications in other domains. For example, ANODE and CURTAINS may prove to be useful for discovering stellar streams in cosmology [213, 214]. This application is particularly interesting as robust follow ups can be performed on regions flagged by these methods. Also, weakly supervised learning can be applied to tasks like muon isolation [117] and for testing the symmetries of the standard model [212]. Developments for applications to weakly supervised searches are therefore likely to be of interest to other areas of HEP.

Part III

Analysis

Chapter 7

Analysis introduction

This chapter introduces the weakly supervised search for resonant new physics developed in this thesis using data recorded with the ATLAS experiment at the Large Hadron Collider at CERN. The analysis is data driven, meaning the background estimate used in the ultimate hypothesis test is estimated from the data itself. This chapter presents the methodology, validation, results and implications of the search.

The first section describes related work detailing some motivation for performing this search, the second section provides a high-level overview of the analysis, the next section details the strategy and assumptions the analysis uses, and the following sections detail the specific steps and ingredients used by the analysis.

7.1 Related work

The dijet topology studied in this thesis is interesting for generic searches because many proposed BSM models decay via jets. There has been an extensive set of inclusive dijet resonance searches at ATLAS. These searches are typically one of the first to be performed at each energy increase – $\sqrt{s} = 7\,\text{TeV}$ [215–217], $\sqrt{s} = 8\,\text{TeV}$ [218], and $\sqrt{s} = 13\,\text{TeV}$ [47, 114, 219, 220]. Their quick rollout after each energy increase is due to the simplicity of the approach but is also a reflection of the fact that the dijet topology is a broad probe of new physics and is useful for excluding new physics produced at large rates.

Many BSM scenarios include $A \to BC$ decays, where A is a BSM particle and the daughter (B, C) particles can be SM particles or BSM particles. This results in a large space of possible theories that are not all covered by dedicated searches [221, 222]. Furthermore, only a few searches [47, 223–237] encompass the range of possibilities where at least one of B or C is itself a BSM particle [222]. This lack of coverage is one of the main motivations for the analysis presented in this thesis.

A major benefit of using jets is that they have a rich substructure, where many BSM scenarios are distributed in a way that is different from the SM background. This allows signal enriched datasets to be produced by defining selections based on the substructure of the jets. The application of an extended bump hunt as defined in section 6.5 to a dataset of jets is therefore possible. Another benefit of jets is that the invariant mass of the dijet system (m_{JJ}) falls smoothly. This mass is defined as

$$m_{\text{JJ}}^2 = M_1^2 + M_2^2 + 2\left(E_1 E_2 - \vec{p}_1 \cdot \vec{p}_2\right), \tag{7.1}$$

with M_i the mass of each jet and $E_i^2 = |\vec{p}_i|^2 + M_i^2$ and $|\vec{p}_i| = p_T^i \cosh(\eta_i)$. At high energies, the mass of the jets is much smaller than their momentum, so the mass of the jets can be neglected. Setting the masses of the jets to zero in the calculation of m_{JJ} gives an alternate definition of the dijet invariant mass,

$$m_{\mathrm{JJ}}^2 \equiv 2\left(|\vec{p_1}||\vec{p_2}| - \vec{p}_1 \cdot \vec{p}_2\right). \tag{7.2}$$

This definition demonstrates that the m_{JJ} distribution is largely determined by the momentum of the jets, which is in turn determined by the PDFs of the partons (section 2.2.2), which are observed to be smooth. This smoothness of the background distribution allows for a bump hunt to be performed in the m_{JJ} distribution.

The analysis presented in this thesis is a follow-up to a previous ATLAS analysis that used the same dataset and a similar methodology [238]. In designing the analysis presented here, the strategy of the previous analysis was followed as closely as possible. The aim was to change as little as possible in the analysis presented here. The previous search had similarly followed well established bump hunt techniques. The dataset used in the analysis presented here is identical to the previous analysis, with the same selections applied, the same luminosity, and the same reconstruction and calibration procedures. The m_{JJ} fit framework that is used in the bump hunt step remains unchanged. The m_{JJ} SB and SRs are defined differently in this analysis than in the previous analysis. This is discussed in more detail in the following chapter.

Several improvements were made in this analysis relative to the previous iteration. In particular, the previous analysis used a single feature set with only two variables and these variables had to be decorrelated from the mass of the dijet system (m_{JJ}) to avoid sculpting an excess in the m_{JJ} distribution. The previous round trained a CWoLa classifier to distinguish between the m_{JJ} signal and SB regions. If the features used in the CWoLa step are correlated with the mass then the CWoLa classifier learns to separate the different regions based purely on m_{JJ}. When making a selection with such a classifier, the m_{JJ} distribution is sculpted to look like a signal. This approach is therefore not robust to correlations between the features used in training the CWoLa classifier and m_{JJ}.

The analysis presented here uses multiple feature sets with up to six variables, all of which have some degree of correlation with m_{JJ}. Assuming these additional features can be well modelled, they are expected to increase the sensitivity of the analysis to new physics. These additional features can be added to this analysis thanks to the use of the Curtains and Salad techniques. One important difference between the previous iteration and the analysis presented here is that the previous iteration trained a CWoLa classifier in both the m_{JJ} SR and SB on features that were decorrelated from m_{JJ}. The analysis presented here trains a classifier in the m_{JJ} SR, on features correlated with m_{JJ}, and applies this classifier to both the m_{JJ} SR and SB. This is important because it means that in the analysis presented here, there is a domain shift when applying the classifier to the m_{JJ} SB that was not present in the previous iteration.

Another improvement in this analysis is the use of the full CL_s procedure to set limits on new physics [62]. The previous analysis used a simplified version of the CL_s procedure that prevents fair direct comparisons to other analyses. To be able to benchmark the analysis presented here against the previous iteration, the same feature set is included. This feature set is used as a proxy for the sensitivity of the previous analysis, though it is expected the analysis presented here is more sensitive on the same feature set due to the use of the Salad and Curtains techniques.

The CMS collaboration has made a note public that describes a similar analysis to the one presented here [239]. The CMS work compares multiple different weakly supervised searches for new physics. They found that weakly supervised approaches had broader sensitivity to new physics than standard approaches. The note from CMS also shows results with significant deviations between the expected and observed limits for multiple signals, and the reason for this is not clear.

Other searches for new physics in the ATLAS and CMS experiments have probed hadronic final states. Inclusive dijet searches that look for narrow resonances in the dijet mass spectrum have been performed by both collaborations. The first of these searches at $\sqrt{s} = 13$ TeV was on a relatively small dataset, ≤ 3.6 fb^{-1} [219, 240]. These searches were interesting to perform as they were among the first at this new energy scale and there was still the possibility that resonant new physics was produced at large total production rates in dijet final states. Such searches are broadly sensitive to new physics but only if the total production rate is large, where the main limiting factors are the fully inclusive nature of the searches and the limited amount of information contained in m_{JJ}.

Other searches in dijets have been performed with more targeted BSM models in mind. As such searches are more targeted, they are expected to have higher sensitivity to the signal models they are designed to probe. For example, searches for diboson resonances have been performed by both collaborations as summarized in Refs. [241, 242]. A diboson search looks for particles that decay into two bosons, such as ZZ, WW, WZ, WH, ZH or $\gamma\gamma$. These resonances are predicted to appear in a variety of new physics models, such as composite Higgs models [243] or models with extra dimensions [244]. Such extensions provide solutions to the naturalness problem in the SM (Section 2.5). The CMS collaboration has also performed two three-dimensional bump hunts in the dijet and jet mass spectra to target diboson like decays [192, 193]. So far these searches have been tailored to signal models where the parent particle is a new BSM particle that decays into two SM bosons.

The ATLAS collaboration has used unsupervised machine learning, based on the assumption that new physics is rare, to search for new physics in the dijet mass spectrum [208, 209]. On top of different high level selections, these searches apply an autoencoder to estimate the density of the events in the input data sample. A selection based on the density estimate of the autoencoder is used to define a dataset on which a bump hunt is performed in m_{JJ}.

Comparisons to the ATLAS dijet search [114] and the ATLAS all-hadronic diboson search [245] are made in this thesis. The same comparisons were made in the previous round of this analysis [238]. These comparisons allow us to understand the sensitivity of the analysis relative to more standard approaches.

The diboson search in particular is expected to be an upper bound on the possible sensitivity of the analysis for certain signal models. This search targets narrow resonances in the dijet mass spectrum that decay hadronically into boosted topologies. This search used jet substructure variables to enhance the presence of signal in the data. To optimize these selections three reference models were used, one of which was from the Heavy Vector Triplet model [36] with SM daughter particles. The diboson search is expected to have high sensitivity to signals with masses close to the SM W, Z and H masses. Away from these masses, the sensitivity of the search is expected to decrease. From the benchmark signals used in this paper, only the $W'_{80,80}$ signal was generated close to the SM W mass such that the diboson search is sensitive to it. This diboson search is therefore expected to set strict limits on the $W'_{80,80}$ signal, and not be sensitive to the other signals.

The ATLAS dijet search [114] is a search for narrow resonances in the dijet mass spectrum. The search is fully inclusive but includes an analysis of a dataset that uses information about the presence of a b-hadron (b-tagging) to increase the sensitivity of the search to new physics models that have sizeable couplings to b-quarks as discussed in section 6.4. Only the fully inclusive version of this search is considered in this analysis. This search uses small radius jets with $R = 0.4$, and they also make a selection based on $|\Delta Y|/2 < 1.2$. The dijet

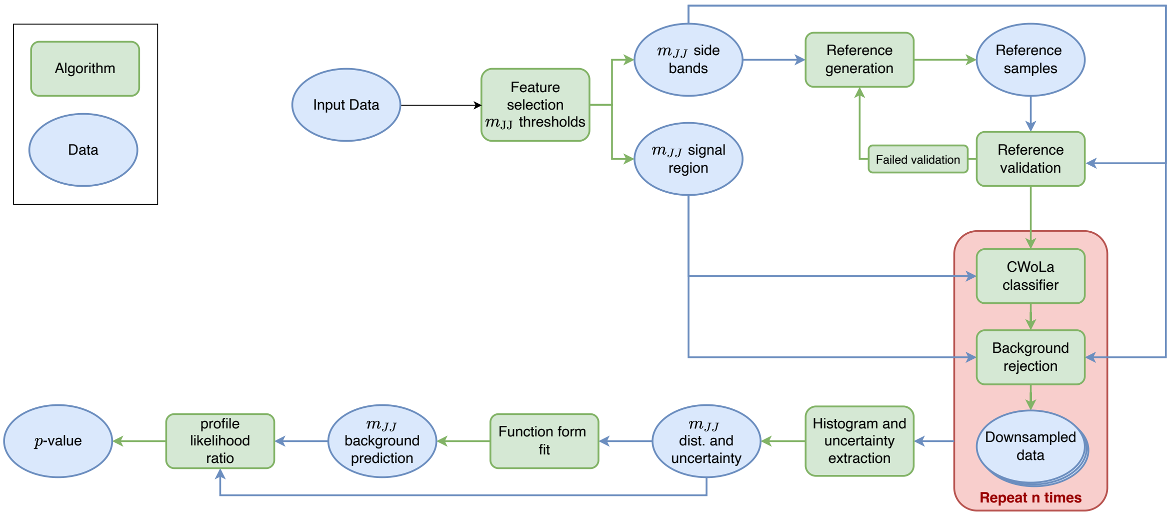


Figure 7.1: The full analysis workflow from an input dataset to the final p-value used in the hypothesis test for a single set of features and fixed m_{JJ} window.

analysis [114] is therefore likely to be more sensitive to signal models that are contained within a single small radius jet. The $W'_{80,80}$, $W'_{80,200}$ and $W'_{200,200}$ signal model decays are expected to be largely contained in small radius jets. For large radius jets, the sensitivity of the dijet search is expected to decrease. Further, as the search presented in this thesis uses additional substructure information it is expected the search presented here could be more than the dijet analysis.

7.2 Analysis overview

The analysis workflow is shown in Figure 7.1 where the input data was processed through the ATLAS software framework and passed a set of selection criteria. Data input to the analysis is divided into m_{JJ} bins defining m_{JJ} signal and SB regions. A reference generator is fit on the m_{JJ} SB data and produces a reference sample in the SR. Multiple classifiers (n) are trained to distinguish between m_{JJ} SR data and the reference sample (CWoLa). These classifiers are then used to perform n selections and define n histograms in m_{JJ}, which are averaged to extract a single histogram and corresponding uncertainty. A background only fit is performed on the SB of the averaged histogram to produce a background estimate in the SR which is used to calculate a p-value for the data in the SR. An m_{JJ} dependent correction to this p-value is derived on signal suppressed validation datasets to correct for issues relating to the possible misspecification of the function used in the background only fit to the m_{JJ} distribution. The analysis is repeated for multiple sets of features $\mathcal{X}$ and m_{JJ} windows. Upper limits are set on the cross section of a variety of new physics processes by injecting simulated signal events into the data, repeating the analysis and using the CLs prescription.

7.3 Analysis strategy

The analysis is weakly supervised and designed to be sensitive to new physics that may be produced at unknown m_{JJ} values, it follows the strategy broadly outlined in section 6.5. The data is split into m_{JJ} SB and SRs, and it is assumed the background in the m_{JJ} SR can be estimated from the SBs. In this analysis, the width of the m_{JJ} SR fixes the resolution of the search, and this analysis targets narrow width resonances such that the width of the m_{JJ} SR should be set by the mass resolution of the detector. To isolate any possible resonant signal in at least one m_{JJ} SR a sliding window bump hunt is performed. The analysis is expected to return a significant p-value in regions where a resonant signal populates the m_{JJ} SR.

The analysis is a follow-up to a previous analysis that directly decorrelated the input features [238]. The previous analysis used $\mathcal{X} = \{m_1, m_2\}$. One of the goals of this analysis is to increase the sensitivity to new physics by using feature sets with more variables. The analysis uses two different reference generation techniques, Salad and Curtains, to produce a background estimate over $\mathcal{X}$ in the SR. The signal sensitivity of these approaches is compared, but not combined.

Analyses need to be validated in controlled settings where their expected behaviour can be defined. This analysis is developed and validated using datasets on which resonant new physics is suppressed. The focus of these validations is twofold. Firstly, the analysis should behave as expected on background only data. This is characterized by the sampling distribution of the significance Z following a rectified Gaussian distribution. This distribution is expected because the significance should be Gaussian distributed, and the significance of a deficit is set to zero. Secondly, the analysis should return a significant Z when new physics is present. This is tested by injecting simulated signal events into the data at different cross sections. The sensitivity of the analysis is dependent on the amount of signal present in the data. Therefore, the entire analysis needs to be fit to data for each signal injection.

Different overlapping feature sets $\mathcal{X}$ were used in the analysis, these were selected based on having broad sensitivity to new physics. Using simulated signal samples to select the feature sets used in the analysis is another source of signal model dependent bias introduced into the analysis. This is mitigated somewhat by using different signal models in the feature selection. Using multiple combinations of features, in the final analysis, reduces the statistical power due to the look-elsewhere effect and also increases the computational cost. Only local p-values are reported by the analysis.

Limits are set on a variety of simulated signals using the procedure outlined in Sec. 6.6. The capacity of the analysis to produce an 'interesting' result, defined as producing $Z > 2\sigma$, is explored by injecting signals at different cross sections and running the analysis. This latter test is referred to as significance enhancement and is the primary means of probing the sensitivity of the analysis during its development. The statistical tests used in hypothesis testing and signal enhancement tests are similar, but the significance enhancement does not use the modified p-value of the CLs procedure and in a sense exposes the direct discovery potential of the analysis.

7.4 Data

The data for this analysis was produced by pp collisions provided by the LHC with $\sqrt{s} = 13$ TeV, as recorded by the ATLAS detector between 2015 and 2018. The data has a total integrated luminosity of 139 fb^{-1} after requiring that all detector systems were functional and providing high-quality data. The uncertainty on the integrated luminosity is 1.7% as measured using the LUCID-2 detector for luminosity measurements. The general approach to reconstruction in the Atlas detector was discussed in section 3.3.1.

The lowest available unprescaled large-radius jet trigger was used to collect the data [46, 246]. In 2015-16 the trigger used untrimmed jet p_T, and from 2017 onward the trigger used trimmed jets and applied a jet energy scale calibration. At trigger level, events are required to have at least two large-radius jets with $p_T > 100$ GeV and $|\eta| < 3.2$ and uncalibrated jet mass $m > 30$ GeV if the jet p_T is less than 1000 GeV.

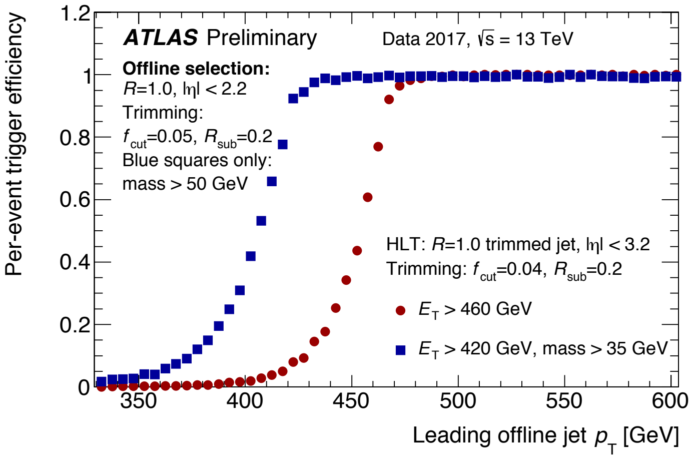


Figure 7.2: The trigger efficiency as a function of the leading jet p_T as shown in Ref. [250].

Offline, jets are reconstructed using the anti-k_t algorithm with a radius parameter of $R = 1.0$. The jets are reconstructed from topological clusters with the local cluster weighting scheme [247] and an MC based particle-level calibration is applied to jets to correct on average the reconstructed mass and p_T to their true values [248]. Jet masses are calculated using both tracking and calorimeter information, as this has been shown to have better mass resolution [249]. Jets are trimmed [50] by reclustering the jet with the k_t algorithm using $R = 0.2$ and removing subjets with $p_T < 0.05$ times the jet p_T. The trimmed jets are calibrated as detailed in Ref. [248].

In the offline selections, each event is required to have at least two jets with calibrated $p_T > 200$ GeV and $|\eta| < 2.0$, and at least one of these jets must have calibrated $p_T > 500$ GeV. The p_T selections were made to increase the boost of the jets, and the η selection is made to guarantee there is good overlap with the tracking acceptance. This selection was shown to be fully efficient with respect to the trigger as shown in Figure 7.2. No lepton overlap removal or veto is applied in this analysis. All offline selections were chosen to be maximally inclusive with respect to prospective signal models while remaining on the trigger plateau.

The two jets with the highest p_T are used in the analysis. The two jets in this analysis are ordered by their transverse momentum such that $p_T^1 > p_T^2$. The final analysis is run on data with a rapidity difference $|\Delta Y| = |y_1 - y_2| < 1.2$, where y_i is the rapidity of the ith jet. Events that pass this selection are referred to as the $|\Delta Y|$ SR. This selection is made to suppress t-channel dijet production while enhancing s-channel dijet production.

To clarify the basis of the $|\Delta Y|$ selection, consider the physical differences between s- and t-channel processes in a hard scattering event. As the colliding partons both have large transverse momenta $|p_z|$, to produce particles with small $|p_z|$ in a t-channel process, a large momentum transfer in the z direction is required. Such transfers are suppressed in QCD, so t-channel processes are more likely to produce particles with large $|p_z|$. This corresponds to large values of the rapidity, with one daughter particle with large positive rapidity and the other with large negative rapidity, and therefore large values of $|\Delta Y|$ on average. In contrast, an s-channel process produces a particle at rest in the center of mass frame of the colliding partons. Such particles can decay into two particles that propagate in any direction, uniformly populating the detector in the $\hat{\theta} - \phi$ plane, where $\hat{\theta}$ is the angle of the daughter particles from the beam axis in the center of mass frame of the colliding partons. The rapidity scales logarithmically with the z component of the momentum, and as the angle is uniformly distributed, s-channel processes populate small values of the rapidity in the center of mass frame. This corresponds to small rapidity differences between the two daughter particles in the center of mass frame. Rapidity differences are boost invariant in the z direction, so this is also true in the lab frame. Therefore, it is expected that s-channel processes on average have smaller values of $|\Delta Y|$ than t-channel processes.

The jets are required to have mass $M_1, M_2 > 30$ GeV so that jets are within the range of what has been calibrated. An upper bound of $M_1, M_2 < 500$ GeV is also applied to the jet masses, to ensure the reference generation operates on a bounded space. This selection was inherited from the previous round of this analysis [238] where it was necessary to limit the correlations between the jet masses and m_{JJ}.

While the dijet invariant mass is typically defined as the sum of the four momenta of the two jets, in this analysis it is calculated with the jet masses set to zero. This is the definition of m_{JJ} shown in eq. (7.2). This is done to lessen the correlation between the jet mass and the dijet invariant mass and was required by the previous round of this analysis [238]. Though it is not needed here, the m_{JJ} definition was left unchanged from the previous round. Updating this is expected to leave the results largely unchanged as the energy of the jets in this analysis are much larger than their masses.

The full set of selections are summarized in Table 7.1.

Table 7.1: Jet selection criteria for this analysis. The jets are ordered by transverse momentum p_T.

Observable	Selection
p_T (leading)	> 500 GeV
p_T (subleading)	> 200 GeV
$\|\eta\|$	< 2.0
$\|y_1 - y_2\|$	< 1.2
M_i	> 30 GeV, < 500 GeV

7.5 Simulation

The analysis is data driven, but simulation is used to validate the analysis, to produce reference samples in the Salad approach, and to simulate signals such that limits can be set. The simulation used in the Salad approach could be sampled from any distribution in principle, but the closer the reweighted distribution is to the true background distribution the smaller the correction and the better the analysis. As the analysis is data driven little emphasis is placed on the production of the simulation. Simulated samples are generated using Pythia8.2 [251, 252] using the A14 [253] tune and the NNPDF23LO [254] PDF set. The after-burner EvtGen [255] is used to model the decay of heavy flavor hadrons.

7.5.1 Background

The same selections as applied to data are also applied to simulated samples. Simulated samples from QCD dijet and multijet production are used to model the background in the analysis. The jet cross section is several orders of magnitude larger than that of electroweak production as shown in Figure 2.4, so samples from these processes are not considered necessary for the analysis. Simulated samples are reconstructed using a full detector simulation with simulated minimum-bias interactions superimposed to represent pile-up. The simulated background is generated in slices of the parton level $R = 0.6$ jet p_T to ensure that a

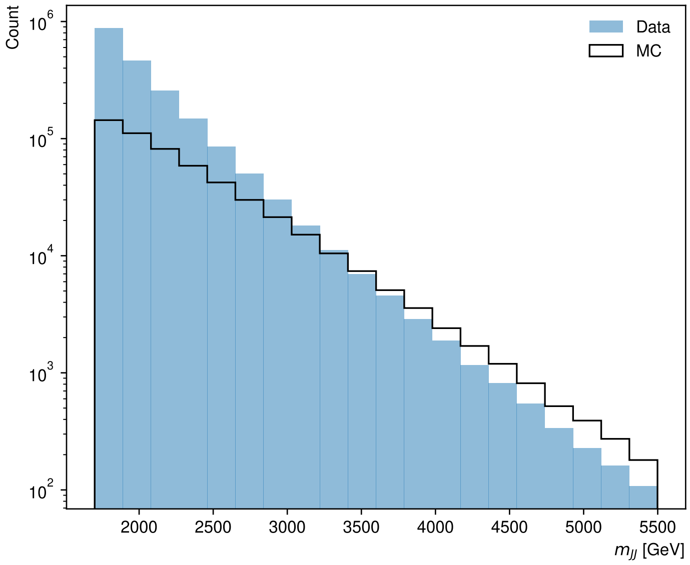


Figure 7.3: The number of events in data and simulation as a function of m_{JJ}.

wide range of detector level jet p_T is covered. However, there is much less simulation than $|\Delta Y|$ SR data overall, particularly at low m_{JJ}, as shown in Figure 7.3. The opposite is true at high m_{JJ}. At low m_{JJ} a Salad reference sample generated with this simulation has less statistical power. This simulated set is also less useful for validation at low m_{JJ}.

7.5.2 Signals

Simulated signals are used to ensure the analysis is sensitive to new physics during development and to set limits on these processes. Using simulated signals to test the sensitivity of the analysis introduces a signal model dependent bias. The simulated signals are intended to be representative of a variety of new physics processes. Further, the analysis was not specifically tuned at any point to be sensitive to any particular signal model. The main focus during development was to ensure the analysis did not report a significant p-value when no new physics was present. Samples from three different families of signal model were generated for this analysis.

The first family of signal models are $W'' \to W'Z'$ processes, where the W'' boson is a heavy resonance with a mass of 3 TeV. The W' and Z' are modelled as modified W and Z bosons with masses sampled from $\{80, 200, 400\}$ GeV. The altered bosons are required to decay hadronically without top quark decays. The same signal models were used in the previous round of this analysis [238]. The $W'' \to W'Z'$ sample, with $m_{W''} = 3\,\mathrm{TeV}$, $m_{W'} = 200\,\mathrm{GeV}$, and $m_{Z'} = 400\,\mathrm{GeV}$, was generated twice using different parton shower settings. These samples allow the impact of variations in the m_{JJ} spectrum on signal sensitivity to be studied. The signal characterized by a narrower (wider) m_{JJ} spectrum is labeled $W'_{3000}(4q)$ ($W'_{200,400}$).

The second family of signal models are $A_0 \to H'Z'$ processes, where the A_0 is a pseudoscalar resonance from the 2-Higgs doublet model [256] with a mass of 3 or 4.5 TeV. The H' and Z' are modelled as modified Higgs and Z bosons with the masses of the daughter particles set to 200 GeV and 400 GeV respectively. This family was modelled with different final states to the W'' family to test the sensitivity of the analysis to different decays. Specifically, samples were generated with final state photons in the jets, b-quarks in the final state, and 6-quark final states instead of 4-quarks.

The final family is a $V' \to VV$ process with V' a heavy vector boson from the Heavy Vector Triplet model [36] and V the SM W and Z bosons. These particles are set to always decay hadronically. The mass of the parent particle was set to $\{2.6, 2.8, 3\}$ TeV.

The widths of all signal models are set to 0.1 GeV, and the detector resolution is expected to dominate the width of the signal. All the signal models presented here are highly boosted with the decay products largely contained in a single large radius jet. The full set of processes and the names they have been assigned are shown in Table 7.2. The distribution over $|\Delta Y|$ for the $W'' \to W'Z'$ signal models is shown in Figure 7.4, all of these models are pure s-channel processes and as expected have distributions that are peaked at low $|\Delta Y|$. All signal models have the same distribution in $|\Delta Y|$.

Table 7.2: The signal models used in this analysis and the name they are assigned in plots. The subscript x can be either 3000 GeV of 4500 GeV.

Name	Process
$W'_{80,200}$	$W'_{3000} \to W_{80}Z_{200}$
$W'_{80,80}$	$W'_{3000} \to W_{80}Z_{80}$
$W'_{200,400}$	$W'_{3000} \to W_{200}Z_{400}$
$W'_{80,400}$	$W'_{3000} \to W_{80}Z_{400}$
$W'_{200,200}$	$W'_{3000} \to W_{200}Z_{200}$
$W'_{400,400}$	$W'_{3000} \to W_{400}Z_{400}$
$W'_x(6q)$	$W'_x \to W_{80}Z_{200} \to 6q$
$W'_x(4q)$	$W'_x \to W_{80}Z_{200} \to 4q$
$W'_{4500}(4q)$	$W'_{4500} \to W_{80}Z_{200} \to 2q2b$
$A_{0,x}(2\gamma 2b)$	$A_{0,x} \to H_{200}Z_{400} \to 2\gamma 2b$
$A_{0,x}(4b)$	$A_{0,x} \to H_{200}Z_{400} \to 4b$
$A_{0,x}(2q2b)$	$A_{0,x} \to W_{200}Z_{400} \to 2q2b$
VV_{2600}	$V'_{2600} \to VV \to 4q$
VV_{2800}	$V'_{2800} \to VV \to 4q$
VV_{3000}	$V'_{3000} \to VV \to 4q$

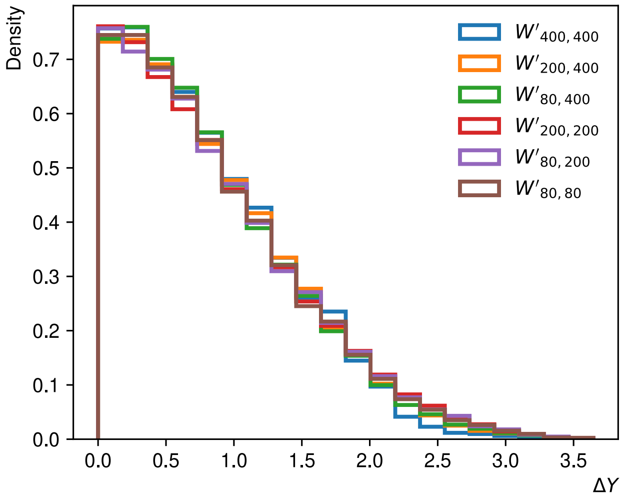


Figure 7.4: Distribution of $|\Delta Y|$ over one signal model family with a parent particle mass of 3 TeV.

Chapter 8

Analysis design

This chapter describes the design of the strategy for searching for new physics presented in this thesis. This involves defining SRs, strategies for validating the analysis, implementing the full analysis and developing the methods at every stage to work as expected. The analysis strategy was developed on validation datasets, which are signal suppressed datasets where the analysis is expected to be consistent with the background only hypothesis. Simulated samples from the signal models from the set defined section 7.5.2 are injected into the data to test the sensitivity of the analysis.

8.1 Signal regions

In this analysis, the data is divided into m_{JJ} SRs and the analysis is run on different feature sets and different classifier selections. Each of these combinations defines a different SR. The choices made in defining the SRs are detailed in the following subsections.

8.1.1 Classifier selections

To run the analysis a threshold ϵ must be set on the classifier output. For $\epsilon = x$ the samples in the largest $(x \times 100)\%$ of the classifier output are selected. In a weakly supervised search, there is no signal model to optimize the classifier selection against. The classifier selection is instead chosen based on signal injection tests, where multiple different signals are injected into the data. This means the analysis selection is optimized for a range of different signals, which is in keeping with the idea of broad sensitivity but does introduce a bias towards the kinds of signal models used for these tests. The selection is also chosen to ensure the analysis does not return a significant excess when no signal is injected. Another consideration in the selection is the amount of statistics available, overly tight cuts might select too few events for the background to be estimated effectively.

Following the previous round of the analysis, two thresholds are used, a loose selection and a tight selection. Initially, the thresholds were set to 10% and 1% as in the previous round, where the choices were made using an educated guess that was confirmed by signal injection tests. Early in the development of this analysis the tighter cut 1% was changed to 2% in an attempt to stop the analysis from reporting a large significance when no signal was injected. Changing the threshold did not fix this problem, but the threshold was not changed. Therefore, this analysis uses a loose selection of $\epsilon = 0.1$ and a tight selection of $\epsilon = 0.02$.

8.1.2 m_{JJ} band definitions

The analysis uses a sliding window bump hunt strategy to search for new physics. To define the SRs the m_{JJ} distribution is divided into bands. In this analysis all bands were defined to have a width of 600 GeV and the bands were shifted in 300 GeV increments. Shifting the bins in this way means the SRs overlap by 300 GeV. The overlap ensures any resonant process should be largely contained in at least one m_{JJ} SR.

The width of 600 GeV was chosen to match the mass resolution of the detector in the lowest m_{JJ} bins. The m_{JJ} resolution of the detector increases with m_{JJ} and the band definitions should increase to reflect this. This was not done in this analysis, but it is an improvement that should be made in future iterations. The bins were defined with fixed widths so the reference generation interpolates the same 'distance' in m_{JJ} for every region. No studies were performed to determine the optimal binning or to test if the interpolation distance has any impact on the quality of the generated reference. Using fixed bin widths might reduce the sensitivity at high m_{JJ} as the SRs are narrower than the mass resolution of the detector. This means any signal can contaminate the SB and bias both the reference interpolation from the SBs and the fit to the m_{JJ} spectrum. Another concern with fixed bin widths is there can be insufficient statistics after making a selection at high m_{JJ}.

The analysis is sensitive to injected signals at both low and high m_{JJ} SRs. Given this sensitivity, the choice of fixed width bins was not revisited. The highest m_{JJ} bin was chosen to have more than ten events in the SR after applying the tightest selection. This ensures the m_{JJ} fit is stable, and the asymptotic approximation is valid for extracting a p-value for the test statistic. The lowest m_{JJ} bin was chosen to avoid the turn on in the data sample with an inverted rapidity cut of $|\Delta Y| > 1.2$[1]. The bins on which the analysis is unblinded are dictated by the final validation results shown in the next chapter. The full set of bins that were considered in the analysis is shown in Table 8.1.

Table 8.1: The left (L) and right (R) m_{JJ} bin edges used in the analysis in GeV for the first sideband (SB1), the signal region (SR) and the second sideband (SB2).

SB1 L	SR L	SR R	SB2 R
1400	1700	2300	2900
1700	2000	2600	3200
2000	2300	2900	3500
2300	2600	3200	3800
2600	2900	3500	4100
2900	3200	3800	4400
3200	3500	4100	4700
3500	3800	4400	5000
3800	4100	4700	5300
4100	4400	5000	5600

8.1.3 Feature selection

The analysis is run on different sets of features $\mathcal{X}$. Sets comprised of different per jet features are considered. For jet i the jet mass M_i, the 2 to 1 jettiness ratios τ_{21}^{i} and the 3 to 2 jettiness

[1] The rapidity Y is correlated with m_{JJ} such that the turn on is higher in m_{JJ} for the inverted $|\Delta Y| > 1.2$ selection. The choice of the lowest m_{JJ} bin was made before the resampling described in section 8.2.1 was discovered to be necessary.

ratios τ^i_{32} are considered. The jet masses are chosen to match the previous round of this analysis [238], and as then, they are expected to be sensitive to the presence of new physics. The jettiness ratios are calculated using the N-subjettiness algorithm [53] and are expected to be sensitive to the presence of new physics as discussed in section 3.3.3.

The features under consideration have all been explored in the development of the reference generation techniques [148, 203, 204, 206] and were found to work well with the Salad and Curtains algorithms. However, these development studies only considered a single signal model and were not run on data with challenging correlations between the feature set and m_{JJ}. In the inclusive datasets considered here all of these features appear to vary smoothly as a function of m_{JJ}, and therefore there is a high chance a useful reference sample can be produced in all m_{JJ} SRs using the data in the m_{JJ} SBs. This assumption is thoroughly tested during the validation of the analysis.

Exploring multiple feature sets would require running the full analysis multiple times and therefore the number of feature sets that could be explored is limited by computational resources. Also, more feature sets increase the impact of the look-elsewhere effect. The following sets of features are considered:

1. $M = \{M_1, M_2\}$.
2. $M, \tau_{21} = \{M_1, M_2, \tau^1_{21}, \tau^2_{21}\}$.
3. $M, \tau_{32} = \{M_1, M_2, \tau^1_{21}, \tau^2_{21}, \tau^1_{32}, \tau^2_{32}\}$.

The first feature set is chosen to match what was used in the previous round of the analysis [238]. This permits comparisons between the performance of this analysis and the previous round. The previous round did not include the reference generation techniques that are used in this analysis, which is expected to enhance the performance. Unfortunately, the additional benefits of the new techniques with respect to the previous analysis are not quantified here. This is because the previous iteration did not use the full CL_s procedure to set limits. The first feature set was chosen to be used as a proxy for the previous round of the analysis as described in section 7.1.

Building feature sets by progressively adding features allows us to determine the performance of the analysis as a function of the number of features used. A priori, it is not clear whether additional features improve the sensitivity of the analysis. Additional features increase the difficulty of generating a reference sample, which increases the amount of mismodelling. The classifier performance can also degrade as the number of features increases, especially if these features are irrelevant for a certain signal [118]. The second feature set was chosen to contain the masses of the two jets and the 2 to 1 jettiness ratios. This choice was not made based on any prior knowledge, and a priori there is no good reason to prioritize a two prong scenario in the absence of a signal model.

The distributions of data and one representative signal in the $|\Delta Y|$ SR are shown in Figure 8.1. In this figure, the signal is the $W'_{200,200}$ process, and as expected a peak in the signal distribution around 200 GeV in both M_1 and M_2 is visible. The data distribution is much flatter and mostly smoothly falling, as expected of predominantly QCD processes. In the subjettiness ratios, the signal is more peaked at low values of τ_{21} than the data. This is expected as the daughter particles in this signal process are modified SM vector bosons that are expected to be highly boosted and decay to two quarks which results in a single large-R jet with a significant two prong structure. The differences in these distributions are expected to be used by the classifiers to distinguish signal from background. In Figure 8.1(g) the signal peaks at around 3 TeV in m_{JJ}, while the data distribution falls smoothly. Therefore, it is expected the bump hunting procedure will be sensitive to the presence of this signal if it is present in sufficient quantities.

An important feature of the signal injection tests is to cover a wide range of different final state topologies. Having a range of different distributions allows us to test the utility of the different feature sets for discovering new physics. In practice, this would mean a variety of different distributions in the features under consideration. The full set of features across all the signal models simulated at the 3 TeV mass point is shown in Figure 8.2. As a reminder, the jets are ordered by p_T and therefore the distribution over the subleading jet mass can be double peaked.

For the $W'' \to W'Z'$ signal models (first column) there are mass peaks around the respective masses of the W' and Z' bosons as expected. All of these signals are biased towards small values of τ_{21}. The distribution in τ_{21} is shifted to the right as the masses of the daughter particles decrease, this is expected as the boost of the jets decreases with the daughter mass and therefore the two prongs of heavier daughter particles (lower boost) are more likely to be resolved. For the VV final state, the distributions in all features are identical across the different parent masses and similar to the $W'_{80,80}$ signal model.

In the second column, it can be seen that the $4q$ final state has a peak in the τ_{21} distribution at low values, while the $6q$ final state peaks at higher values and is more similar to the QCD background. This is expected as the $4q$ final state on average has two prongs in the jet due to the presence of two quarks in each jet. In contrast, the $6q$ final state is expected to result in three prong substructure. The $4q$ final state has on average large values of τ_{32} as expected for a two prong jet. The $6q$ final state has a slight shift downwards in the τ_{32} distribution as expected for a three prong jet. This shift is moderate as higher order prongs are less well resolved in the jet substructure and the τ_{32} distribution is less sensitive to the number of prongs than the τ_{21} distribution [52].

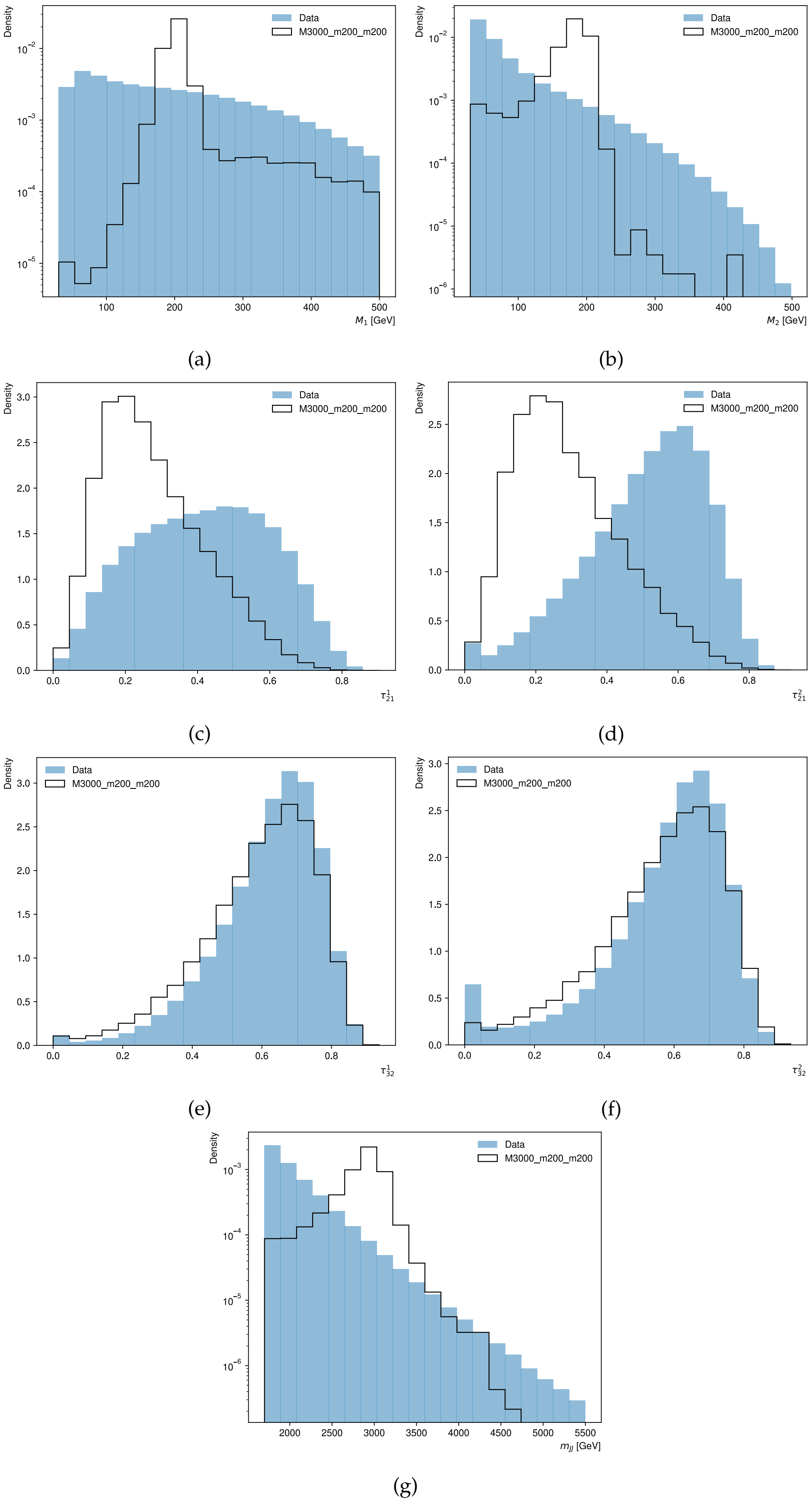
For the $A_0 \to HZ$ (third column) signal model family the mass peaks of the daughter particles at 200 and 400 GeV are pronounced. The τ_{21} distribution is peaked at low values as expected for these daughter particles. The final state containing a photon has a particularly sharp peak in the τ_{21} and τ_{32} distributions at zero. This is expected as the photon is reconstructed as a single jet with no substructure and in this case, the substructure is set to zero. As the photon is almost always the subleading jet, the τ_{21} and τ_{32} distributions for the subleading jet are more sharply peaked at zero.

The distribution of the different signals shown in Figure 8.2 are expected to be used by the classifiers to distinguish between signal and background. The more different the signal distributions are from the background the more sensitive the analysis becomes to the presence of new physics. For example, the final state containing a photon should be relatively easy to distinguish from the background as the τ_{21} and τ_{32} distributions are sharply peaked at zero.

8.2 Validation strategies

To validate the analysis datasets on which the expected behavior of the analysis is known are required. Ideally, this would be a dataset of pure background in the $|\Delta Y|$ SR. The validation sample would also ideally contain many more samples than the $|\Delta Y|$ SR. Such a dataset allows us to gather statistics on the behaviour of the analysis, specifically to test whether the reported significance follows the correct distribution[2]. The simulated samples described in

[2] When evaluating the analysis on background only data the significance is expected to follow a rectified Gaussian distribution.



(a)

(b)

(c)

(d)

(e)

(f)

(g)

Figure 8.1: Comparison of the distributions of the features used in the analysis in the $|\Delta Y|$ SR for signal and the inclusive dataset. The signal is the $W'_{200,200}$ process.

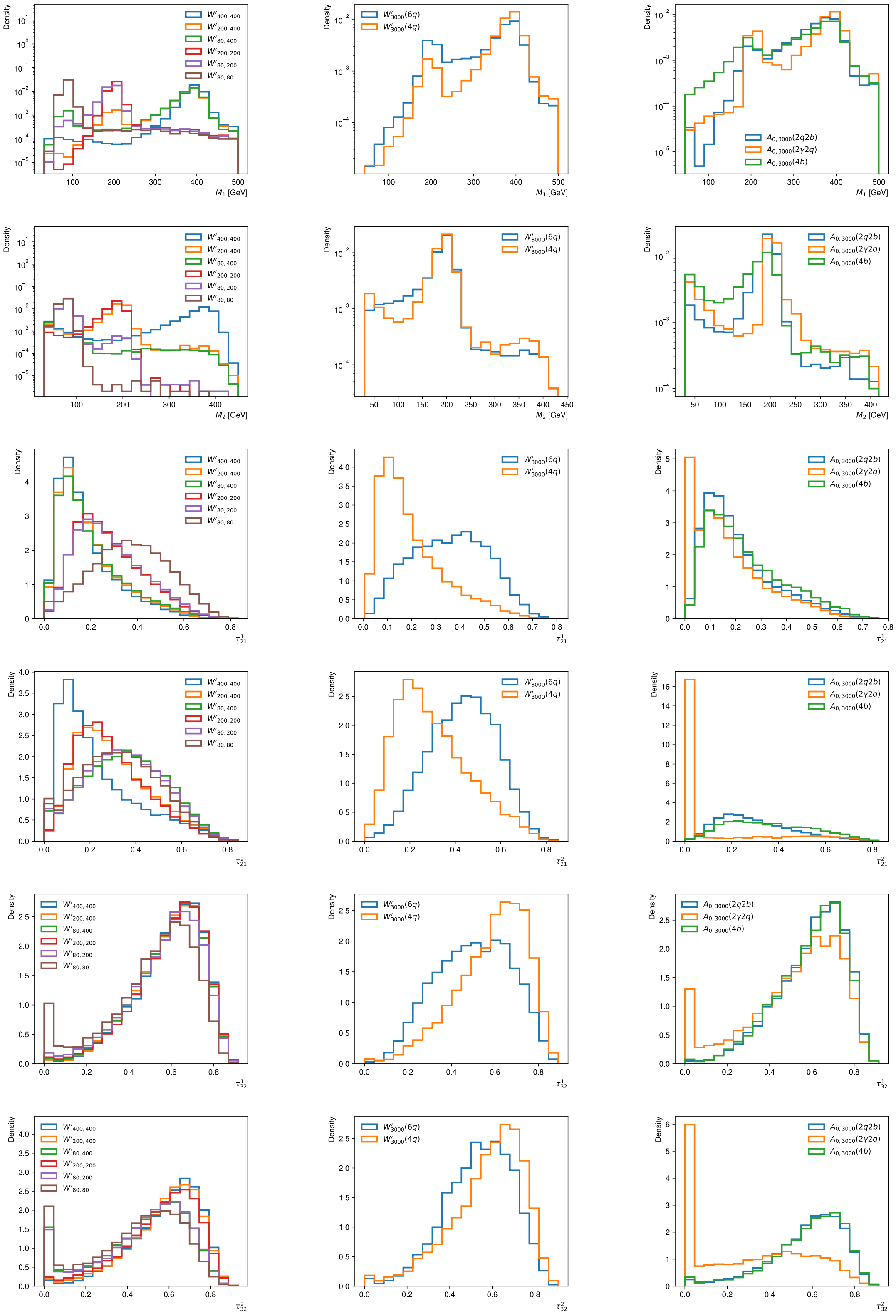


Figure 8.2: Comparisons of the distributions of the features used in the analysis across all signal models grouped by process.

section 7.5.1 are pure background, but this dataset lacks statistics and is not a perfect model of data. For this analysis, a simulated validation set is complimented by orthogonal data derived validation sets. In total, three different validation datasets are used to validate the analysis.

Data derived validation sets are required to satisfy three criteria:

- Signal suppressed.
- The density $p(\mathcal{X}|m_{\mathrm{JJ}})$ is similar in both validation and the $|\Delta Y|$ SR for all m_{JJ}.
- The density $p(m_{\mathrm{JJ}})$ matches in both validation and the $|\Delta Y|$ SR.

A perfect validation set would be background only, but in deriving validation sets from data strongly suppressing signal is sufficient. The requirements of similar $p(\mathcal{X}|m_{\mathrm{JJ}})$ and $p(m_{\mathrm{JJ}})$ are important because the analysis is fit to both of these distributions. It is essential to validate that the analysis can fit data with similar correlations to the $|\Delta Y|$ SR data. For example, it must be verified that the reference generation and classifier selection do not bias the m_{JJ} fit.

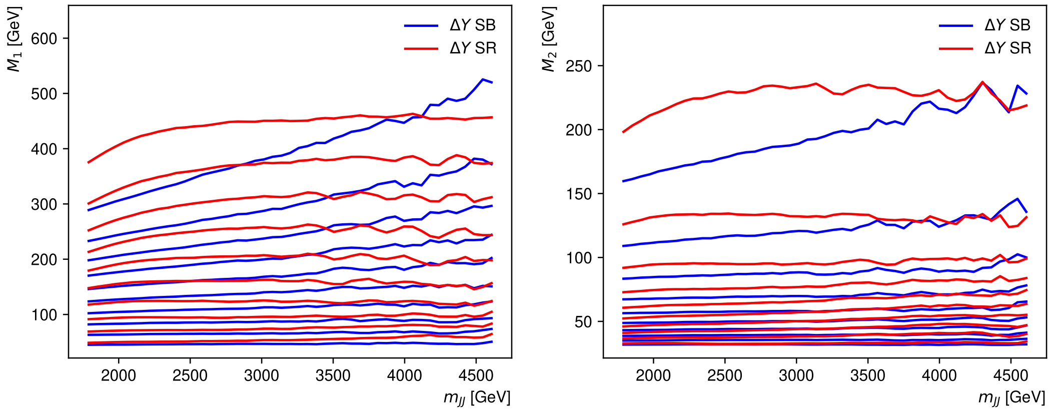
Other than validating the analysis on known datasets, there are additional constructions where the expected behaviour is known. This makes it possible to isolate the analysis stages that introduce any biases that do appear. These constructions are referred to as 'idealized' strategies and were first introduced in Hallin et al. [203]. The following sections detail the construction of the validation datasets and idealized strategies used in the development of this analysis.

8.2.1 $|\Delta Y|$ SB

The analysis was unblinded on a dataset that makes a $|\Delta Y| < 1.2$ selection to enhance the targeted s-channel resonances. This selection can be inverted to define a signal suppressed dataset. This dataset is referred to as the $|\Delta Y|$ SB. The $|\Delta Y| > 1.2$ selection has a low signal efficiency for the simulated signals used as benchmarks as shown in Figure 7.4. While this dataset is signal suppressed there are correlations between $|\Delta Y|$, m_{JJ} and $\mathcal{X}$ that limit its utility. Due to these correlations both $p(\mathcal{X}|m_{\mathrm{JJ}})$ and $p(m_{\mathrm{JJ}})$ are different in the $|\Delta Y|$ SB than in the $|\Delta Y|$ SR as shown in Figure 8.3 and Figure 8.4(b). Only a subset of the features are shown in the body for brevity, see Appendix A for the full set of features.

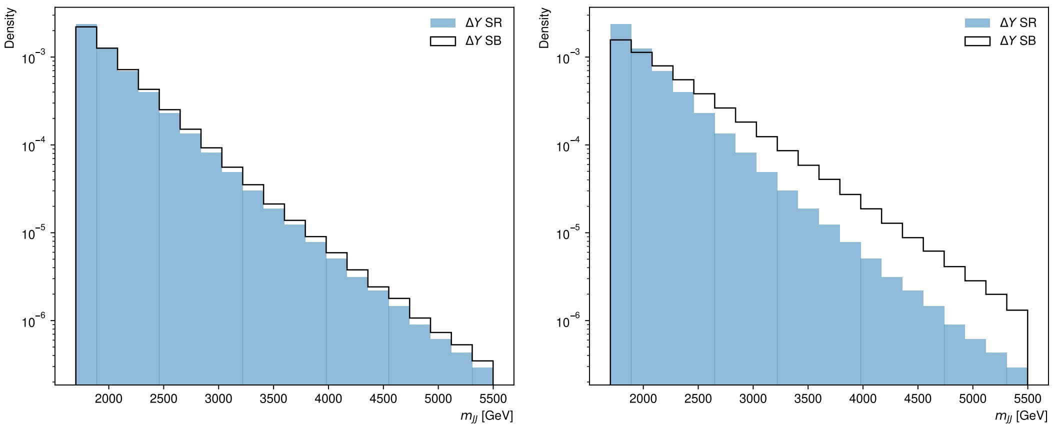
The first is composed of a random partition of the $|\Delta Y|$ SB into ten distinct datasets. Each of these datasets has fewer statistics than the $|\Delta Y|$ SR data, and the m_{JJ} distribution is different. The second set is constructed by resampling the full $|\Delta Y|$ SB dataset to match the m_{JJ} distribution in the $|\Delta Y|$ SR. The $p(m_{\mathrm{JJ}})$ distributions in the two datasets can be matched by downsampling the $|\Delta Y|$ SB as shown in Figure 8.4(a). This is possible as there are more total samples in the $|\Delta Y|$ SB. The resampling matches the m_{JJ} distribution in the $|\Delta Y|$ SB to the $|\Delta Y|$ SR. In principle, if there is signal in the $|\Delta Y|$ SR that can be detected in the fully inclusive m_{JJ} distribution, this resampling introduces a bias in the resampled $|\Delta Y|$ SB. It is assumed the fully inclusive m_{JJ} distribution in the $|\Delta Y|$ SR is not sensitive to new physics without additional information. This is justified as this distribution has been rigorously studied previously by Ref. [114]. This makes the $|\Delta Y|$ SB useful for testing the m_{JJ} fit, up to distortions introduced by the classifier.

Differences in $p(\mathcal{X}|m_{\mathrm{JJ}})$ can not be so easily mitigated as this would require distribution morphing. However, to validate the analysis it is still useful to study datasets that have correlations similar to those observed in the $|\Delta Y|$ SR. All quantiles above $m_{\mathrm{JJ}} = 2600$ GeV vary approximately linearly in both the $|\Delta Y|$ SR and the $|\Delta Y|$ SB. Therefore, testing the



(a) (b)

Figure 8.3: Distribution of the leading M_1 and subleading M_2 jet mass as a function of m_{JJ} for $|\Delta Y|$ SB and $|\Delta Y|$ SR data. Ten equally spaced quantiles between 5% and 95% are shown.



(a) (b)

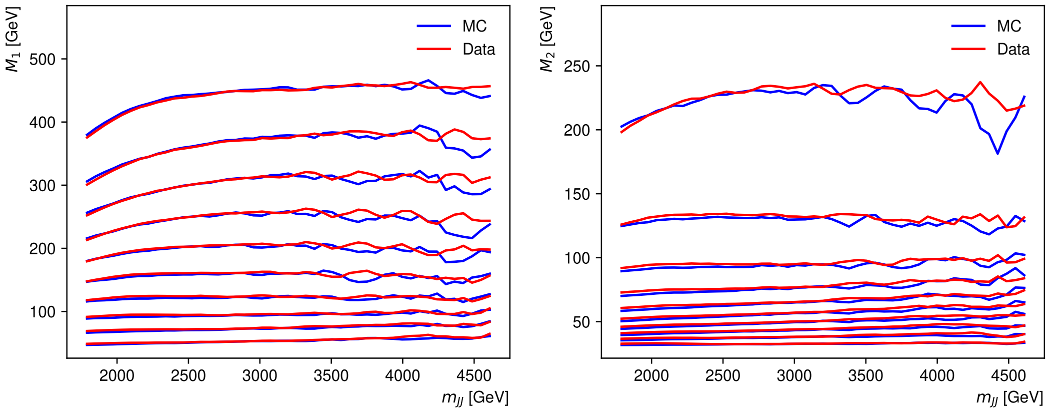
Figure 8.4: Distribution of the dijet mass for $|\Delta Y|$ SB (a) with and (b) without resampling, compared with $|\Delta Y|$ SR data.

analysis on the $|\Delta Y|$ SB is expected to provide some probe of the performance of the analysis in some SRs in the $|\Delta Y|$ SR. The character of the correlations between $\mathcal{X}$ and m_{JJ} is also simpler in the $|\Delta Y|$ SB. There are non-linear correlations between the features and m_{JJ} in the $|\Delta Y|$ SR that are not present in the $|\Delta Y|$ SB as shown in Figure 8.3. Therefore, if the analysis does not work on the $|\Delta Y|$ SB then it is unreasonable to expect it to work on the $|\Delta Y|$ SR.

The inverted $|\Delta Y|$ selection described here was used to construct the primary validation dataset used in the previous round of this analysis [238]. This validation dataset was resampled as described above. When running SALAD on the $|\Delta Y|$ SB the $|\Delta Y|$ SR simulation sample was reweighted to match data and generate the reference. As there are significant differences between these two datasets, the performance of SALAD in the $|\Delta Y|$ SB is expected to be degraded.

8.2.2 Monte Carlo validation

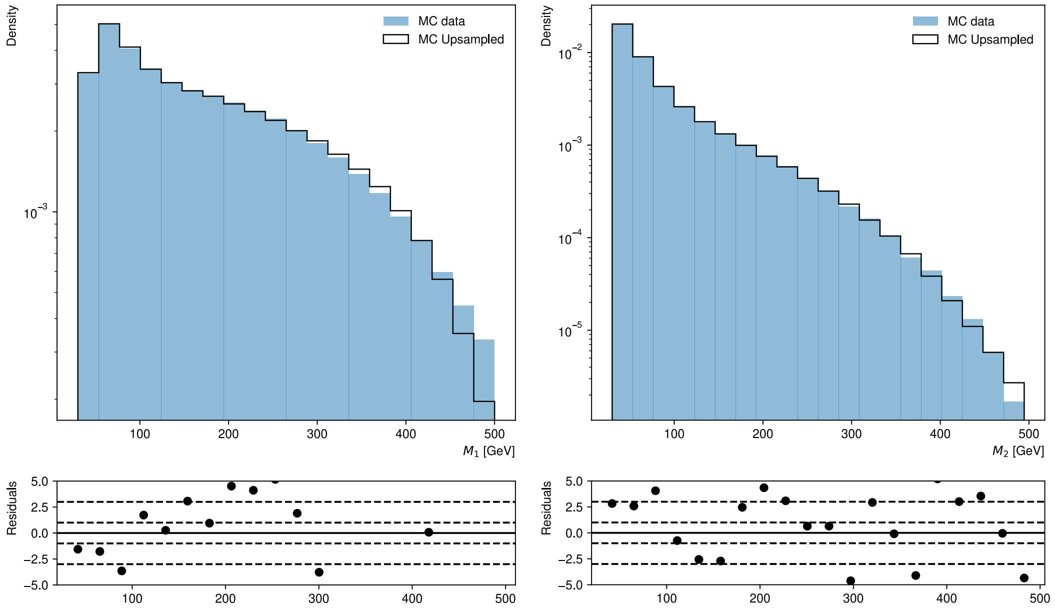
The simulated samples in the $|\Delta Y|$ SR are by definition pure background. However, as already shown in Figure 7.3 the $p(m_{\mathrm{JJ}})$ distribution does not match the $|\Delta Y|$ SR data. Fortunately, $p(\mathcal{X}|m_{\mathrm{JJ}})$ in simulation does closely match the $|\Delta Y|$ SR as shown in Figure 8.5. This dataset was used for validation in the previous round of this analysis. The m_{JJ} distributions of the two datasets could be matched by sampling from simulation with replacement, but this does not increase the effective statistics of the dataset. Instead, a generative model $p_\theta(\mathcal{X}|m_{\mathrm{JJ}})$ is fit to the simulated dataset. This distribution can be used with the m_{JJ} samples from the $|\Delta Y|$ SR to define a new validation dataset. The use of the mass distribution to upsample is justified using the same logic as the previous subsection.



(a) (b)

Figure 8.5: Distribution of the leading M_1 and subleading M_2 jet mass as a function of m_{JJ} for $|\Delta Y|$ SR MC and $|\Delta Y|$ SR data. Ten equally spaced quantiles between 5% and 95% are shown.

The distribution p_θ is of course only an estimate for the true distribution from which the simulation is sampled. However, this distribution is fit on only six features and is observed to produce high fidelity samples as shown in Figure 8.6. Note that comparing the marginals of the samples from p_θ can not be compared directly to the original sample as the two datasets have different m_{JJ} distributions. More importantly, the correlations in the dataset generated with p_θ are observed to match the original dataset as shown in Figure 8.7. Again, only a subset of the features are shown in the body for brevity. See Appendix A for the full set of features. The resulting validation set has the same effective statistics as the $|\Delta Y|$ SR data, with the correlations of the $|\Delta Y|$ SR MC as modelled by p_θ. The generative model is trained using a flow matching objective as described in section 5.7.5. Specifically, continuous flow matching [257] with logitnorm time sampling [258] was used, the vector field was parameterized by a three layer MLP with 256 hidden units and `SiLU` activations [259]. The learning rate was linearly ramped up from 10^{-5} to 10^{-4} over the first epoch and then held constant. Time was embedded into 8 dimensions using a cosine embedding and an exponential moving average of the model weights was used to stabilize training. The model was trained for 100 epochs with a batch size of 1024. In total ten different generative models were trained on the $|\Delta Y|$ SR MC dataset to produce ten different validation sets. This was done to account for any variations in the training of the generative model that could impact the performance of the analysis.



(a) (b)

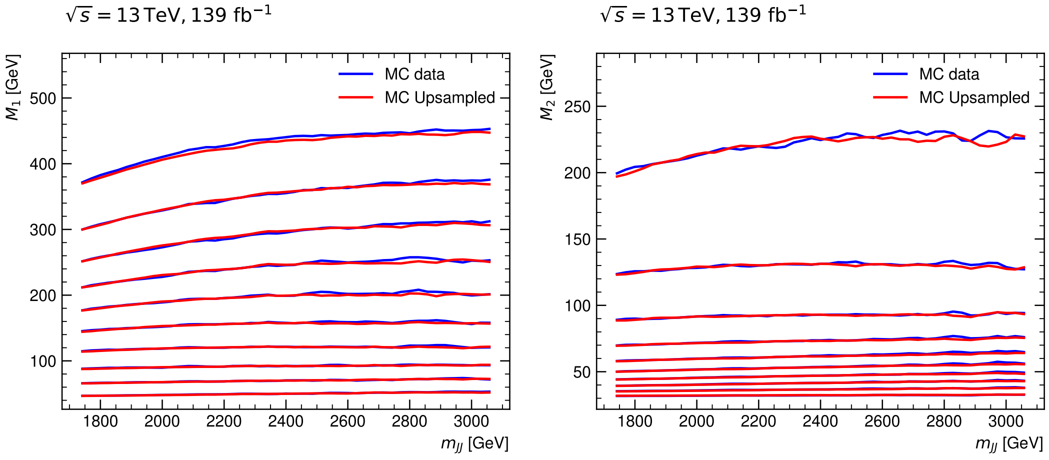
Figure 8.6: Distribution of the leading M_1 and subleading M_2 jet mass as a function of m_{JJ} for the upsampled MC dataset and the original MC dataset.

8.2.3 Down sampling validation

The final validation set is derived directly from data. This approach is based on randomly downsampling the $|\Delta Y|$ SR data. Random downsampling is expected to kill signal sensitivity and is a commonly used strategy for blinding data analyses in HEP. The signal suppression here comes from the reduction in significance discussed in section 4.3. In this case, if the dataset is randomly downsampled by a factor of $1/\alpha$ for $\alpha \in \mathbb{R}^+$ then the amount of signal and background are equally scaled on average by $1/\alpha$. Therefore, random downsampling leads to an average drop in sensitivity of $1/\sqrt{\alpha}$.

The resulting downsampled dataset allows for a controllable amount of signal suppression while maintaining samples drawn from the correct $p(\mathcal{X}|m_{\mathrm{JJ}})$ and $p(m_{\mathrm{JJ}})$ distributions. The downsampled dataset can be upsampled again by fitting a generative model p_θ as for simulated data. This generative model is trained on the downsampled dataset and can not perfectly model the original dataset. The upsampled dataset has the same, or more, signal suppression as the down sampled dataset. Therefore, the upsampled dataset is signal suppressed. This procedure can be repeated, randomly downsampling multiple times and then upsampling. Repeating the procedure results in different validation sets as the generative model is fit to different datasets in each iteration. A schematic of this procedure is shown in Figure 8.8. This procedure is referred to as the (down)upsampling.

The parameter α plays an important role here. Large α results in strong signal suppression, but also limits the available statistics for fitting the generative model p_θ used for upsampling. Also, the larger α becomes the more variability in the estimates of p_θ when repeating the down-up sampling procedure. For this analysis, $\alpha = 30$ was chosen and found to produce good generative models. This corresponds to an average signal suppression of $1/\sqrt{30} \approx 0.18$, which is expected to be large enough as no significant resonance (≥ 5) exists in the fully inclusive dataset [114]. The ability to generate multiple validation sets allows us to better probe the distribution of the significance reported by the analysis. Being able to generate



(a) (b)

Figure 8.7: Distribution of the leading M_1 and subleading M_2 jet mass for the upsampled MC dataset and the original MC dataset. Ten equally spaced quantiles between 5% and 95% are shown.

multiple copies from the same distribution p_θ permits the same test[3], but is also useful in the idealized procedures.

For this analysis, the (down)upsampling procedure was repeated ten times to produce ten validation sets. The performance of one example of such a model is shown in Figure 8.9. The upsampled distributions do not exactly match the original data, especially in the tails. A mismatch, especially in the tails, is expected as there are by definition fewer statistics in the tails of a distribution. Mismodelling in these regions is exacerbated by the fact the generative model is fit on a downsampled dataset. The mmismatch is also desirable as the validation sets should not match exactly the original $|\Delta Y|$ SR dataset. The correlations in the validation set again match the character of those in data as shown in Figure 8.10. The same flow matching generative model training was used as upsampled the MC dataset.

8.2.4 Comparison of validation datasets

Each of the validation strategies has deficiencies and strengths that probe different aspects of the analysis. The $|\Delta Y|$ SB has simple but incorrect, correlations between $\mathcal{X}$ and m_{JJ}. Any signal is strongly suppressed in this dataset, and it can be resampled to have the correct $p(m_{\mathrm{JJ}})$ distribution. Therefore, it is sensible to use this dataset for developing the analysis, but it is not reliable as a means of proper validation. The other validation sets more closely match the correlations in the $|\Delta Y|$ SR dataset. The (down)upsampled datasets in particular are seen to have correlations that match the data. These datasets also have much more variability than MC because the downsampling can be performed multiple times. The (down)upsampled datasets are expected to be the most useful for validation. A summary of the different datasets is shown in Table 8.2. Unfortunately, there is a downside risk to the generative model based upsampling procedures.

In upsampling it is possible for the generative model $p_\theta(\mathcal{X}|m_{\mathrm{JJ}})$ to introduce localized, quadratic mismodelling in m_{JJ} such that the analysis reports an excess. This corresponds to constructing a (down)upsampled validation set with an artifact consistent with signal. The other downside is the down-up sampling procedure might not sufficiently suppress any

[3] The utility of repeated upsamplings from the same p_θ is limited because each of these samples has identical m_{JJ} values.

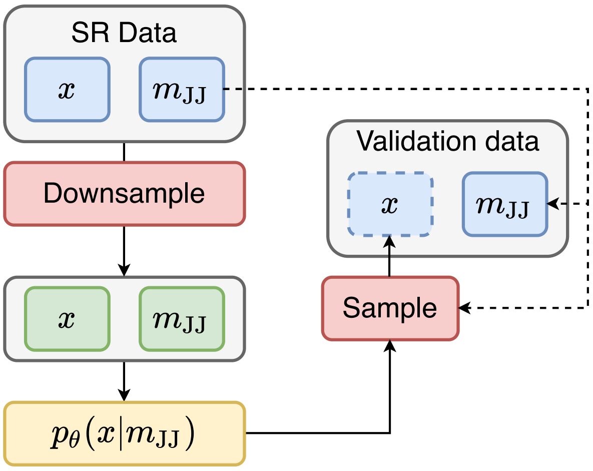


Figure 8.8: Schematic of the (down)upsampling procedure. The full $|\Delta Y|$ SR dataset is first downsampled and a generative model p_θ is fit on this downsampled dataset. A new dataset is generated using p_θ and the m_{JJ} samples from the $|\Delta Y|$ SR dataset.

Table 8.2: Summary of datasets and their properties. Each dataset is classified based on whether it has: the same statistics as the $|\Delta Y|$ signal region data; matching $p(x|m_{JJ})$ and $p(m_{JJ})$ distributions; and the ability to generate validation sets with adjustable correlations between $\mathcal{X}$ and m_{JJ}.

Dataset	Statistics	$p(m_{JJ})$	$p(x|m_{JJ})$	Variability
$|\Delta Y|$ SB	■			
$|\Delta Y|$ SB resampled	■	■		
MC raw			■	
MC upsampled	■	■	■	
(down)upsampled $|\Delta Y|$ SR	■	■	■	■

signal that is present in the $|\Delta Y|$ SR data. In both of these settings, it would be erroneously concluded that the analysis was failing the validation. To account for this the upsampling procedures themselves require validation. A dedicated study was performed to validate the generative model based upsampling procedures and is presented in section 9.8.

8.2.5 Idealized constructions

There are several idealized constructions that are useful for validating parts of the analysis and diagnosing failures. The two such constructions that were used in this analysis were idealized classifiers and idealized references. These idealized constructions were invaluable in developing this analysis.

In the absence of signal, an ideal classifier is equivalent to randomly downsampling the data. This can be integrated into the analysis by replacing the classifier step with random downsampling, which is equivalent to randomly assigning scores to samples. Using this procedure makes the selection independent of the reference generation and the training of an ML classifier. This isolates all steps downstream of the classifier selection and allows issues with these steps to be identified. For example, if the fit to the m_{JJ} distribution was biased this would be identified with this test. The m_{JJ} fit is expected to work on the fully inclusive dataset as it was used in Ref. [114] for this purpose, however it is not known how

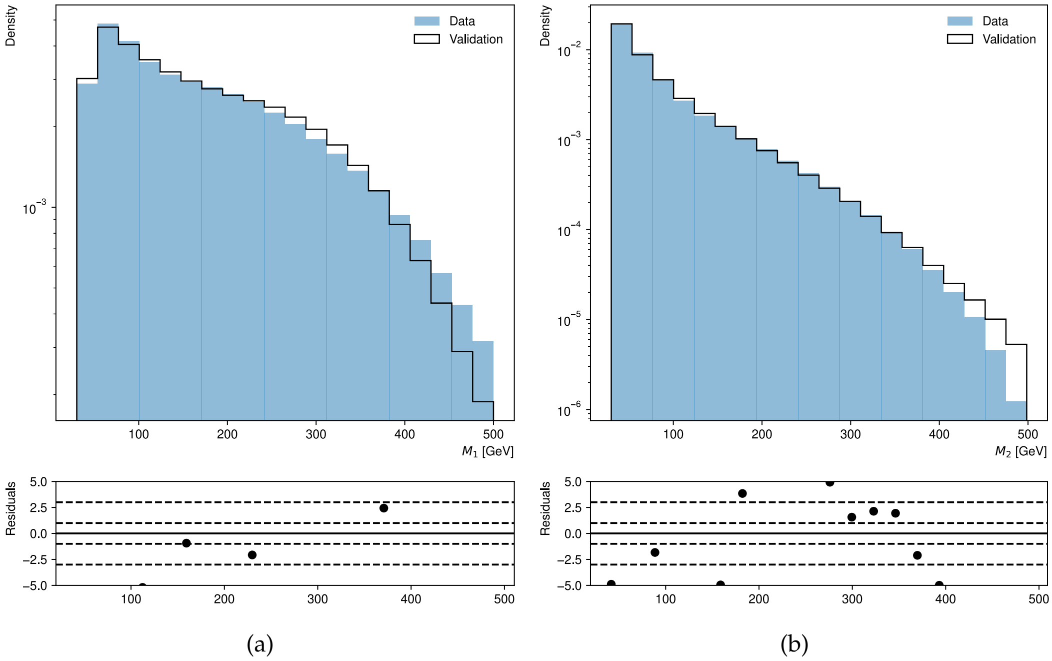
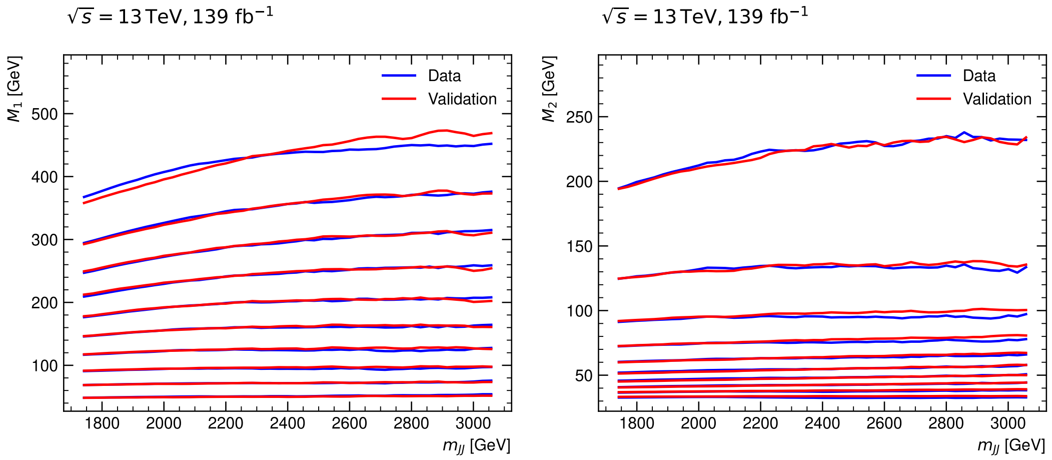


Figure 8.9: Distribution of the leading M_1 and subleading M_2 for one (down)upsampled dataset and the original $|\Delta Y|$ SR data.

the fit performs when downsampling to $\epsilon = 0.1$ and $\epsilon = 0.02$. If the fit does not work on randomly downsampled data then it can not be expected to work when a classifier is used to select the samples. The random downsampling can be performed multiple times to generate a distribution of significance. This is the expected distribution of significances when no signal is present and the analysis performs perfectly with no bias from the classifier. Due to finite statistics, this distribution is not a rectified Gaussian.

An idealized reference is a sample that is drawn from the same distribution as the data. This reference sample contains, by definition, no mismodelling but includes the effects of finite sample sizes. One way to generate an idealized reference sample is to sample from the same generative model p_θ twice and treat one sample as data and the other as a perfect reference. Training a classifier on an idealized reference isolates issues that arise from the reference generation step using CURTAINS and SALAD. If the analysis does not work as expected on an idealized reference but does work with an idealized classifier, then it can be concluded the classifier training is the source of the issue. An analysis that fails on an idealized reference is not expected to work when integrating references generated by CURTAINS or SALAD.

In addition to these idealized constructions, one could run idealized tests where the reference generation methods are trained on both m_{JJ} SB and SR data, rather than just on m_{JJ} SB data. This means the reference generation methods would do no interpolation. If the reference sample generated in this way causes the analysis to behave unexpectedly, then the reference generation methods are not working as expected. Further, all idealized constructions could be used on $|\Delta Y|$ SR data directly. This is because when running the analysis in an idealized mode the signal sensitivity can only be reduced, and it is assumed the fully inclusive dataset is insensitive to new physics. Neither of these idealized approaches was used in this analysis, but they may prove useful for later iterations.



(a)

(b)

Figure 8.10: Distribution of the leading M_1 and subleading M_2 jet mass as a function of m_{JJ} for one (down)upsampled dataset and the original $|\Delta Y|$ SR data. Ten equally spaced quantiles between 5% and 95% are shown.

8.3 Analysis implementation

This section provides details of the implementation of the analysis. The methods outlined here were chosen to be as close as possible to public studies of weakly supervised methods as they existed when this analysis was being developed [148, 197, 198, 203, 206], and to follow the general fit strategy successfully used in the previous round of this analysis [238]. The previous round of this analysis used several strategies that were taken from more standard analyses in HEP and thus expected to be well behaved in this context.

8.3.1 Reference sample generation

The reference sample was generated using either the Salad or Curtains approach. These two methods are described in detail in the following. A reference sample must be generated because the features in $\mathcal{X}$ are correlated with m_{JJ}. The generated reference is used to train a CWoLa style classifier. Two different reference generators were explored such that one could be used as a cross check on the other. In this case, a simulation based approach and a fully data driven approach were used. As these two use different assumptions to build a reference, they are expected to be useful as cross checks on each other.

Salad

The Salad approach reweights simulated background B to match the target dataset D [197, 198]. The weighted simulation is used as the reference sample. The Salad approach uses the following recipe to generate a reference sample:

1. Rescale all features in $\mathcal{X}$ according to a global scaling, such that, for each feature, the largest value in the dataset is mapped to 1 and the smallest is mapped to 0.
2. In the m_{JJ} SB assign a label of 0 to all simulated samples and 1 to all data samples.
3. Train a neural network f_θ to predict the labels.

4. Assign a weight to each simulated sample $\{x^i, m^i_{\mathrm{JJ}}\}$ in the m_{JJ} SR using,

$$w^i = \frac{f_\theta(x^i, m^i_{\mathrm{JJ}})}{1 - f_\theta(x^i, m^i_{\mathrm{JJ}})},$$

   where i indexes the samples.

5. Invert the preprocessing applied in Step 1 to the simulated data to obtain the final reference sample.

The last step is important to ensure the reference sample is generated with the same feature scaling as the data. For this analysis, the neural network f_θ is an MLP with a sigmoid activation function and is trained using the binary cross entropy loss. If many of the weights produced by the Salad procedure are significantly different from one then the effective statistics of the reference sample is reduced.

The classifiers f_θ are MLPs with 4 layers of size 64, 64, 64, 1 and activation functions `ReLu`, `ReLu`, `ReLu`, sigmoid, respectively [95]. A dropout of 5% was used between each layer to reduce overfitting. Dropout randomly masks some weights in the network during training [260]. These models were implemented in `Keras` [261] using the `Tensorflow` [262] backend. The classifiers use the binary cross-entropy loss, and the `Adam` [110] optimizer with a learning rate of 0.001 and no scheduling. Half of the data was used for training, and the other half for validation. Training proceeded for 200 epochs with early stopping, with 25 epoch patience, on the validation loss and a batch size of 512. This means the training was stopped if the validation loss did not improve for 25 epochs.

The classifier f_θ is trained on the m_{JJ} SB and applied to the m_{JJ} SR. This means that when the reference is generated the classifier is applied to m_{JJ} values which are not seen during training. No strategy was used to mitigate issues that might arise from distribution shifts between the m_{JJ} SBs and the m_{JJ} SR. This is justified as the features in $\mathcal{X}$ are assumed to change smoothly and slowly as a function of m_{JJ}. The implicit assumption is the distribution in the SR is constrained relative to the SBs. No observe issues were observed when deploying Salad in this way. It is possible that methods to account for the shift as a function of m_{JJ} could improve the performance of the analysis when using Salad. The loss updates during training were weighted such that both m_{JJ} SBs contributed equally.

Curtains

The Curtains method uses a feature morphing approach to generate a reference sample [148]. This approach is fully data driven and therefore complementary to the Salad approach. The Curtains approach uses the following recipe to generate a reference sample:

1. Rescale all features in $\mathcal{X}$ according to a global scaling, such that, for each feature, the largest value in the dataset is mapped to 3 and the smallest is mapped to -3. Any features that have long tails, such as the masses of each of the jets, are first logit scaled.

2. Train a normalizing flow $p_\phi(\mathcal{X}|m_{\mathrm{JJ}})$ to estimate the conditional density of the features in the SBs.

3. Train a second normalizing flow with INN $f_{\theta(m^1_{\mathrm{JJ}}, m^2_{\mathrm{JJ}})}$ to map features from any mass point $m^1_{\mathrm{JJ}} \in \{\mathrm{SB1}, \mathrm{SB2}\}$ to any other mass point $m^2_{\mathrm{JJ}} \in \{\mathrm{SB1}, \mathrm{SB2}\}$ such that

$$p_\phi(\mathcal{X}|m^1_{\mathrm{JJ}}) \xrightarrow{f_{\theta(m^1_{\mathrm{JJ}}, m^2_{\mathrm{JJ}})}} p_\phi(\mathcal{X}|m^2_{\mathrm{JJ}}). \tag{8.1}$$

This means that for $x \sim p(x|m_{\mathrm{JJ}} = m^1_{\mathrm{JJ}})$ the INN $f_{\theta(m^1_{\mathrm{JJ}}, m^2_{\mathrm{JJ}})}$ maps x to x' such that $x' \sim p(x|m_{\mathrm{JJ}} = m^2_{\mathrm{JJ}})$.

4. Fit the mass distribution in the SBs using $p_1(1-z)^{p_2} z^{p_3}$ where z is m_{JJ} divided by the center of mass energy and p_i are the fit parameters. Interpolate into the SR and sample [263].
5. Use the INN from Step 3 to map each SB sample to a randomly selected mass value in the SR sampled using the fit from Step 4.
6. Repeat Step 5 m times, for some choice of integer m.
7. Invert the preprocessing applied in Step 1 to obtain the final reference sample.

The scaling of the features in $\mathcal{X}$ was chosen to work with the rational quadratic spline [99] layers that are used to construct the Curtains models. The logit transform is used as it has been shown to improve the modelling of long tailed features [203]. It is again important to invert this preprocessing to ensure the reference sample is generated in the correct feature space. Note that if there are n samples in the combined SBs then Curtains generates $m \times n$ samples in the SR. This is in contrast to Salad which only generates as many samples as there are simulated SR samples. Oversampling the reference sample has been shown to dramatically increase the signal sensitivity [148, 203, 206]. For this analysis, $m = 4$ was chosen as this was found to be a good balance between signal sensitivity and computational cost. When mapping the left SB and right SB into the SR no accounting for the difference in statistics is made and the two samples contribute equally to the SR reference. During training, by default, Curtains applies no weights to the loss updates. This means that due to the statistics of the training data, the lower m_{JJ} SB contributes more to the training of the INN than the higher m_{JJ} SB.

The INN f_θ is trained using the maximum likelihood objective. It is also parameterized such that the dependence on mass is restricted. Specifically, the function is defined as $f_{\theta(g(m^1_{\mathrm{JJ}}, m^2_{\mathrm{JJ}}))}(x)$ where $g(m^1_{\mathrm{JJ}}, m^2_{\mathrm{JJ}})$ is a function that controls the dependence on mass. Control over the function g allows some control over the behaviour of the INN when interpolating from the SBs into the SR. For this thesis, two options for this function were considered.

1. $g(m^1_{\mathrm{JJ}}, m^2_{\mathrm{JJ}}) = |m^1_{\mathrm{JJ}} - m^2_{\mathrm{JJ}}|$.
2. $g(m^1_{\mathrm{JJ}}, m^2_{\mathrm{JJ}}) = \{m^1_{\mathrm{JJ}}, m^2_{\mathrm{JJ}}\}$.

The default option is the first, as it was found to work well in previous studies [148, 206]. These studies were performed on datasets where the correlations between the features and the mass were relatively weak. When the correlations are non-linear, the second option is expected to perform better.

All Curtains models used the same settings as those used in Raine et al. [206]. All normalizing flows were implemented using the `nflows` package [264] in `pytorch` [265]. All flows use rational-quadratic spline layers [99]. Both the base flow INN and the top flow INN consist of eight coupling RQ spline layers. The parameters of the spline in each layer have 8 bins and 3 layers of hidden units with 256 units each. Each spline function uses 4 bins and the parameters are predicted by an MLP with 2 hidden layers with 32 units each. The INN is trained for 100 epochs using the `Adam` optimizer with a learning rate of 0.001 that is annealed to zero following a cosine schedule.

8.3.2 CWoLa classifier

The classifier used to discriminate between the reference and data is referred to as the CWoLa classifier. This is because the classifier training relies on the CWoLa paradigm as defined in section 5.5.1, with the reference considered to be a signal depleted dataset. The classifier used in this analysis is an MLP, and the architecture mirrors that of the Salad classifier. No hyperparameter tuning was performed on this classifier. The classifiers were trained for 100 epochs with a learning rate of 0.001, no learning rate schedule, the `Adam` optimizer, batch sizes of 512, and early stopping with a 10 epoch patience.

The classifier training is based on the k-fold cross validation procedure [266]. This is done to ensure the classifier is always evaluated on data that it has not seen during training. For this procedure, the reference and m_{JJ} SR datasets are split into 5 folds. A classifier is trained on 3 of these folds, with one of the remaining folds used for validation and the other for testing. The classifiers are trained to distinguish the reference sample from the SR data. The folds are cycled through such that each fold is used as the test set once. The scores used in the analysis selection are those assigned to the test set. One of the 5 classifiers is randomly selected to assign a score to each data sample in the SBs. This makes the assignment of scores to samples in the SBs consistent with the SR.

The classifiers are never trained on SB data, and this could be a problem as the distribution over the features is not the same as in the SBs and the SR. As already discussed, the features are expected to change smoothly as a function of m_{JJ}. The distribution of the features in the m_{JJ} SBs is expected to cover the distribution of features in the m_{JJ} SR as the SBs are adjacent to the SR. However, the reverse is not expected to be true. This is a potential source of bias in the classifier. An alternative approach would be to generate a reference for both the SBs and the SR and train the CWoLa classifier on the full m_{JJ} range. This was not explored in this analysis but would be a useful avenue for future studies.

8.3.3 m_{JJ} fit

The m_{JJ} fit is performed to the data after performing a classifier selection. The fit is performed using modified least squares to data in the m_{JJ} SBs. A sequential approach, identical to the previous round, is pursued as described in Algorithm 1. This fitting procedure is based on the smoothly falling nature of the m_{JJ} distribution in QCD. The procedure uses functions that have empirically been observed to fit the m_{JJ} distribution well. No form for the signal is assumed in the fit, so there is no signal plus background fit performed and therefore no spurious signal systematic uncertainties. The choice to not include an assumed signal shape in the fit is motivated to reduce the number of signal specific assumptions in the analysis. Other approaches to performing the m_{JJ} fit would be useful to explore in future iterations of this analysis. In particular, general assumptions about the form of the signal would not introduce a strong signal bias in the analysis, because detector effects can be expected to dominate the signal shape.

8.3.4 Systematic uncertainties

Systematic uncertainties in an HEP analysis are generally either related to background or signal modelling. Signal systematic uncertainties typically only have an impact on the limit setting procedure. The motivation for this search is to discover indications of new physics, rather than to constrain specific new physics models. Therefore, in this search, consistent with past weakly supervised analyses of jet substructure [208, 238], signal systematic uncertainties are not included. This impacts the limit setting procedure, but not the discovery potential of the analysis. Ignoring signal systematics is expected to make the limits reported

Algorithm 1 Background Prediction Algorithm

```
A classifier selection is made.
The $m_{\mathrm{JJ}}$ distribution of the samples that pass the selection are used to define a histogram
  with 30 equally spaced bins. The mass of each bin is divided by the center of mass energy
  for the fit $x = m_{\mathrm{JJ}}/\sqrt{s}$.
Define fit functions $f_i$ with parameters $\{p_j\}_{j=1}^{4}$:
    $f_1 = p_1(1-x)^{p_2} x^{-p_3}$
    $f_2 = p_1(1-x)^{p_2} x^{-p_3+p_4\log(x)}$ [114]
    $f_3 = p_1 x^{p_2} e^{-p_3 x + p_4 x^2}$ [267]
$i \leftarrow 1$
while $\chi^2$ of the fit on SB data is greater than 5% do
    if $i = 4$ then
        Drop the furthest SB bins from the SR.
        $i \leftarrow 1$
    end if
    Fit SB data with $f_i$.
    if SB contains less than six bins then
        Fit failed.
        break
    end if
    $i \leftarrow i + 1$
end while
Evaluate the fit $f_i$ in the $m_{\mathrm{JJ}}$ SR to produce a background prediction.
Extract the uncertainty on the background prediction using error propagation with the
  Jacobian of the fit.
```

by this analysis slightly more strict than they would be otherwise. However, the search probes statistically limited regions of phase-space and signal systematics are not expected to have a large impact on the results.

There is an additional systematic uncertainty associated with a lack of knowledge about the correct fit function to use in the m_{JJ} fit and bias introduced by the reference. Both of these are accounted for using the m_{JJ} dependent correction that is described in section 8.3.6.

8.3.5 Likelihood fit

The likelihood fit to extract the test statistic and significance Z of an observation uses the procedure described in section 4.2. The background in the m_{JJ} SR is estimated using the fit function from section 8.3.3. The observation x in the SR is the sum of the bin counts after making selections. The background prediction b is the sum of the background prediction in each bin. The statistical uncertainty on the prediction b is the Poisson variance $\sqrt{b}$. The fit uncertainty ψ_f on this count defines a Gaussian in the likelihood fit. The likelihood function is then,

$$\mathcal{L}(\mu, \psi) = \mathrm{Pois.}(x|b + \mu + \psi)\mathcal{N}(0|\psi, \psi_f), \tag{8.2}$$

where the parameter μ is the signal strength and ψ is fit (profiled) to data. To extract a significance on the observed SR counts, a profile likelihood fit is performed [58, 59] as defined in section 4.2. All likelihood fits are performed using `pyhf` [63, 64].

8.3.6 m_{JJ} dependent correction

There is an additional systematic uncertainty associated with our lack of knowledge about the correct fit function to use in the m_{JJ} fit. This is not accounted for in the likelihood fit, instead, a correction to the significance Z returned by the likelihood fit is applied. This correction is extracted using the procedure defined in Algorithm 2. A separate correction is derived for each feature set, method of reference generation and classifier selection threshold. The moments μ and σ are extracted by assuming the significances are normally distributed but truncated at zero. This procedure was inherited from the previous round of this analysis [238].

Algorithm 2 m_{JJ} dependent correction for a fixed feature set, reference generation method and classifier selection threshold.

```
for each signal region $i$ do
    for $j = 1$ to $10$ do
        Calculate $Z_j$ on the $j$-th (down)upsampled validation dataset.
    end for
    Extract the mean ($\mu_i$) and standard deviation ($\sigma_i$) of the set $\{Z_j\}_{j=1}^{10}$.
    Calculate the SR center $m^i_{\mathrm{JJ}}$.
end for
Fit a linear function $\mu^c$ to $\{\mu_i\}$ as a function of $\{m^i_{\mathrm{JJ}}\}$.
Fit a linear function $\sigma^c$ to $\{\sigma_i\}$ as a function of $\{m^i_{\mathrm{JJ}}\}$.
```

The correction is assumed to be a function of m_{JJ} as every quantity in the analysis is assumed to vary smoothly with m_{JJ}. Fitting a function to the moments of the significances as a function of m_{JJ} is expected to capture a more accurate estimate of the systematic uncertainty. This correction is also expected to correct for any mismodelling in the analysis, where any systematic bias is expected to be caught by the validation datasets and accounted for with this correction. Such a systematic bias could arise from mismodelling in the reference generation resulting in an excess in the SR.

For an observed significance Z calculated in an SR with center m_{JJ} the corrected significance Z' is,

$$Z' = \frac{Z - \max\{\mu^c(m_{\mathrm{JJ}}), 0\}}{\max\{\sigma^c(m_{\mathrm{JJ}}), 1\}}, \tag{8.3}$$

where μ^c and σ^c are the functions derived from Algorithm 2. The $\max$ functions are used to ensure the correction is only ever applied to reduce the significance. This guarantees the correction is conservative and does not introduce any bias. The final reported significance is $\max\{Z', 0\}$ to be consistent with the assumption that new physics can not produce a deficit in the data, as described in Section 4.2.

Setting the significance to be at minimum zero is a reflection of the assumption that new physics can not produce a deficit in the data. This excludes new physics models that predict destructive interference between the signal and background, for example. Excluding signal models of this kind is another assumption of this analysis. The appearance of a deficit in this analysis is therefore assumed to be due to mismodelling in the analysis. To quantify the amount of mismodelling in the analysis, negative significances are reported. In the context of the assumption about new physics made by this search, a deficit represents a failure to properly model the background.

8.3.7 Limit setting

Upper limits on the production of various signal models are set at 95% confidence level using the CL_s prescriptions as outlined in section 6.6. Adaptive grid sizes are used to calculate the cross sections at which the signal samples are injected to ensure the limits are calculated with sufficient precision. When limits are presented an explicit example of how this procedure is performed is given. The CL_s values are calculated using `pyhf` and corrected using the m_{JJ} dependent correction. In all signal models, the cross section is a free parameter, only the shape of the signal distribution is relevant. Therefore, when model-dependent limits are drawn only the acceptance and luminosity are needed to calculate the cross section. Both the classifier and high level cuts are used to calculate the acceptance for each signal.

8.3.8 Analysis workflow

This analysis presents a technical challenge in that for every validation set, m_{JJ} SR, feature set $\mathcal{X}$ and reference generation method the full analysis pipeline must be run. This includes the training of the CURTAINS (SALAD) model, the generation of the reference sample, the training of the classifier(s), the m_{JJ} and likelihood fits. In the final analysis, there are seven different m_{JJ} SRs, three different feature sets and two different reference generation methods, so the analysis must be run 42 times to be unblinded. During validation, this is repeated ten times on the (down)upsampled validation set alone. Further, to set exclusion limits the analysis is run in full for at least 10 different signal injections for 20 different signal models across 3 different feature sets. To set limits the analysis is therefore run a total of 600 times per reference generation method. While the analysis was being developed the full analysis pipeline was run on the order of 10,000 times.

Orchestrating all the steps in the analysis, including the fitting of all ML components, is made possible by using a workflow language. This analysis was implemented in `snakemake` [268] and this made both analysis development and deployment significantly easier. The `apptainer` [269] containerization system was also used to manage environments and ensure the analysis could be run on any system with minimal setup. All methods used `numpy` [270], `pandas` [271], `scipy` [272], `matplotlib` [273], `scikit-learn` [274], `scikit-hep` [275], and pipeline runs were configured using `hydra` [276] and `omegaconf` [277].

Chapter 9

Analysis development

This type of analysis has never been performed before and therefore requires significant development. This chapter describes how the development was done and some of the challenges that were faced in this process. Some procedures defined in the previous section are modified here. This section is important for understanding how to structure the development of this kind of analysis, and the reasoning behind the modifications made to the procedures defined in the previous section.

The analysis was primarily developed using the ten partitions of the $|\Delta Y| > 1.2$ SB dataset. All results in this section use these samples, and the performance of the analysis is tested with and without signal injected. Signal injection tests are performed by injecting signal samples into the data at a certain significance s as defined in section 4.3 and then running the full analysis pipeline. The signal injection tests are used to ensure the analysis is sensitive to new physics but also introduce a signal dependent bias. When injecting signal samples into the $|\Delta Y|$ SB data the signal from the $|\Delta Y|$ SR is used. Due to the differences in the feature distributions between the $|\Delta Y|$ SB and SR data, the analysis is expected to be more sensitive to signal in the $|\Delta Y|$ SB data.

The distribution of significances returned by the analysis is expected to be consistent with the background only hypothesis when no signal is injected. As there are ten versions of this dataset the significance distribution is only an estimate of the true significance distribution. A better estimate of the mean and standard deviation of the significance distribution can be found by looking across m_{JJ} where the significance distribution in non-overlapping m_{JJ} SRs should be independent but correlated as the analysis should vary smoothly as a function of m_{JJ}.

As already stated the $|\Delta Y|$ SB data should be easier to fit than the $|\Delta Y|$ SR data. Therefore, if the analysis fails here it most likely fails in the SR. However, an important reason for using the $|\Delta Y|$ SB for analysis development is historical. As the previous round of this analysis used the inverted $|\Delta Y|$ selection to construct a validation dataset it was assumed the same procedure could be used for this analysis. This assumption proved to be flawed as described in section 8.2. For this reason, the $|\Delta Y|$ SB data was used for analysis development, but the other validation sets are the final arbiter for where the analysis can be unblinded. In an ideal scenario the analysis would be developed using validation sets that better match the behavior of the $|\Delta Y|$ SR data. The rest of this chapter is devoted to the development of the data, which involves modifying the procedures outlined so far.

9.1 Single classifier results

Using the analysis implementation described in section 8.3 large significances were reported when no signal was injected. This means the analysis does not behave as expected in the background only case. The cause of this issue is explored in detail in the following sections and the solution is presented in section 9.2.

Large significances occurred much more often than would be expected from a rectified Gaussian distribution. This was tested by running the full analysis pipeline multiple times with different random seeds on the same input data. These random seeds change the initialization and training of the CURTAINS and SALAD models, the classifier initialization and the splitting of the data into the different k-folds. This causes the classifier used in the event selection to be different in each run of the analysis. The significance reported by the analysis was found to vary significantly between runs, such that not all runs would report a significant excess. Solving this problem was the biggest challenge in developing this analysis. An example of one of these excesses is shown in Figure 9.1, where it appears the reported excess is due to a signal like artifact in the m_{JJ} SR.

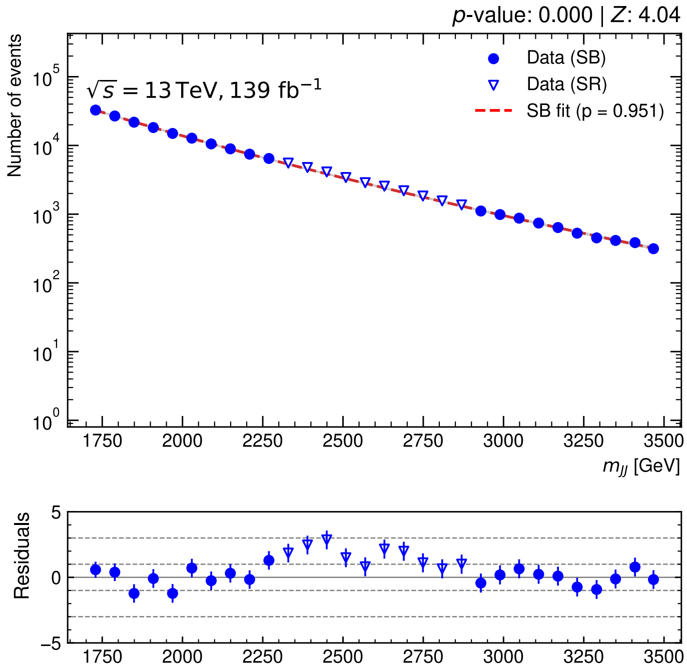
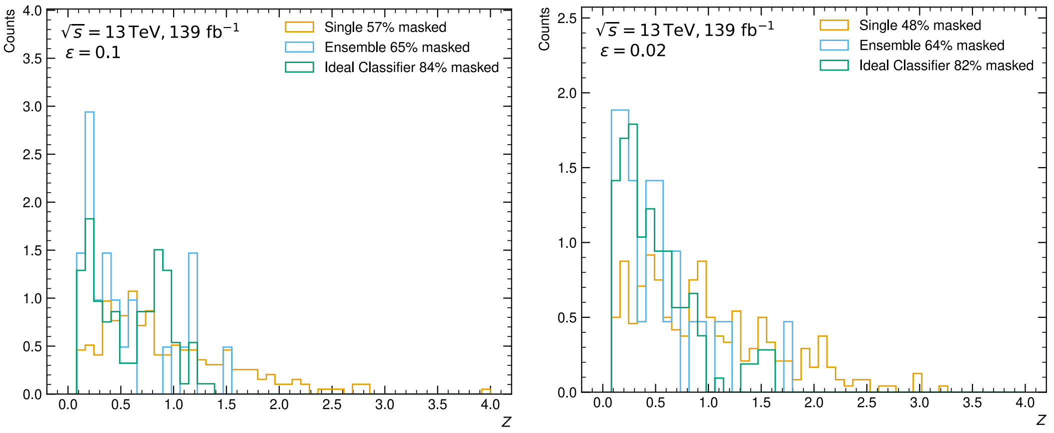


Figure 9.1: An m_{JJ} fit plot showing an excess in the m_{JJ} SR. This fit is performed after making a selection on the output of a classifier trained using a CURTAINS reference model on the M, τ_{21} feature set at a threshold of $\epsilon = 0.1$.

To isolate the source of this issue the analysis was run in an idealized mode as described in section 8.2.5. This allows different parts of the analysis to be tested with all steps upstream assumed to be performed perfectly. Specifically, the idealized classifier and idealized reference constructions are described in section 8.2.5.

Idealized classifier

The idealized classifier was used to test the m_{JJ} fit. For this, a fraction ϵ of the samples is randomly selected to match a certain classifier selection. This random downsampling is the ideal behaviour of the classifier in the absence of signal. On this randomly downsampled data, the m_{JJ} fit was performed. When randomly downsampling the underlying m_{JJ} distribution remains unchanged, and if the fit produces excesses more frequently than expected there is an issue with the fit procedure. No significant excesses are observed in the m_{JJ} fit when using the idealized classifier, as shown in Figure 9.2.



(a)

(b)

Figure 9.2: The distribution of significances for different classifiers defined in idealized settings. To show the distributions in the tails the significance (Z) has been masked at 0.1 and the fraction of events below this value is shown in the legend. The Idealized Classifier randomly assigns scores to input samples, the Single classifier is trained on an idealized classifier using an idealized reference and the Ensemble classifier is 10 classifiers trained on the same idealized reference sample and ensembled to produce a prediction.

In these tests, the fit often fails when the left SB edge goes above 5600 GeV in m_{JJ}. This is due to a lack of statistics at high m_{JJ}. As this is not a systemic failure of the analysis, but a feature of the data, the analysis is restricted to look at bins below 5600 GeV. This is how the upper m_{JJ} SR in Table 8.1 was chosen. Choosing bins according to the mass resolution of the detector would have allowed higher m_{JJ} values to be probed as the m_{JJ} SR would increase with m_{JJ}, and therefore there would be more statistics in the m_{JJ} SR and SB. This was not explored in this analysis but would be a useful avenue for future studies.

Idealized reference

Having identified the m_{JJ} fit performs as expected in the absence of a biased classifier, we move up the analysis chain to the classifier itself. To isolate the classifier training from the reference generation, the idealized reference was used. In this case two of the partitioned $|\Delta Y|$ SB datasets are compared. These two sets are drawn from the same distribution, and therefore one serves as a perfect reference dataset. The classifier should not be able to distinguish between the two datasets, up to statistical fluctuations. The analysis is run by treating one of the datasets as the reference and the other as the data and then running the rest of the analysis on the partition identified as the data. In this idealized mode, significant excesses were sometimes reported. This is shown in Figure 9.1. Therefore, there is an issue with the way classifiers are being deployed in the analysis.

While investigating the idealized reference outputs it was found the classifiers were assigning identical scores to many input samples. This is a separate problem from what has been reported so far, and it prevents a selection with exactly fraction ϵ from being made as there are not enough unique predictions to exactly match the targeted fraction. To resolve this a random tie-break was introduced. This was done by sampling random noise on the order of the machine precision and adding it to the classifier output. The issue of the tie-break did not resolve the problem of the excesses. However, the tie-break is always required and is applied to all classifiers in this analysis.

Discussion

The hypothesized cause of the issue with the classifier training is the classifier randomly placing a decision boundary in the feature space that is consistent with signal. When placing a decision boundary randomly in a feature space that is correlated with m_{JJ}, there is a probability an artificial excess is observed. This is what is assumed to be happening in the idealized reference test.

The problem of randomly assigning decision boundaries is compounded by the fact the classifiers are trained on the m_{JJ} SR and applied to the m_{JJ} SB. This makes the SB data out of distribution for the classifier. The same is true when training the SALAD classifier on the m_{JJ} SB and applying it to the m_{JJ} SR. However, as the features are assumed to change smoothly with m_{JJ}, the support of the feature distributions in the m_{JJ} SB is expected to cover the support of the distributions in the m_{JJ} SR. This is seen to be the case in all datasets under consideration. Some attempts were made to train the classifier on both the SB and SR data. However, there was evidence that this significantly reduced the signal sensitivity of the analysis and was not pursued systematically. This approach should be explored in future work using classifiers that are better suited to this weakly supervised task [118].

Another compounding feature of the classifier training is the m_{JJ} distribution is skewed such that there are more events at low m_{JJ} than at high m_{JJ}. This means the variance in the classifier output is expected to be higher at high m_{JJ} than low m_{JJ} after training. In general, this is a feature of training ML models on data with a skewed distribution and is accounted for in the training of the SALAD model by reweighting the input data [1]. The skew towards low m_{JJ} values is problematic because the variance of the classifier output is also expected to increase on data it was not trained on. Together, this represents a recipe for sculpting a bump in the m_{JJ} distribution. Accounting for this imbalance in the training of the CWoLA classifier is something that should be explored in future work, but was not investigated in this analysis.

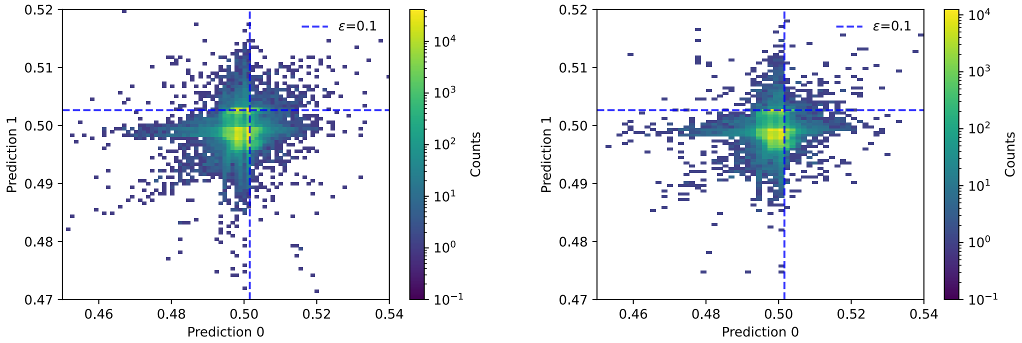
In solving the problem of false excess the analysis faces three challenges. First, the random seed should not play a large role in the final result reported by the analysis. Second, large excesses should not be reported when no signal is injected. Third, the analysis should be sensitive to signal when injected.

9.2 Ensembled classifiers

The main approach that was pursued to mitigate the issue with the classifier training was to ensemble the classifiers as described in the following. Given our hypothesis the issue arises from the random placement of the decision boundary, it is expected the decision boundaries placed by different classifiers are uncorrelated up to statistical fluctuations and mismodelling in the reference. Therefore, the issue can be resolved by training multiple classifiers $\{f_\theta^i\}_{i=1}^n$ and assigning scores to a sample x by performing some ensemble operation $G(\{f_\theta^i(x)\}_{i=1}^n)$. Varying the architecture of the classifier is not expected to resolve the issue.

[1] It also becomes necessary for CURTAINS on datasets that more closely match the $|\Delta Y|$ SR.

To explore the hypothesis that classifiers are placing random boundaries in feature space, the correlation between the outputs of two classifiers was studied. In Figure 9.3 the distribution of two different randomly selected classifiers trained on the same idealized data is shown. There is a relatively small amount of overlap between the events selected by the two classifiers. This is consistent with the classifiers placing decision boundaries in different places in the feature space. There is a difference in the classifier distribution in the m_{JJ} SB and SR, which is consistent with distribution shift playing a role in this problem. Therefore, the issue with the classifier score assignment can likely be resolved by ensembling the classifiers.



(a) (b)

Figure 9.3: The correlation between the outputs of two randomly selected classifiers trained on the same idealized reference data. The classifier outputs are shown when applied to (a) the m_{JJ} SB and (b) the m_{JJ} signal region. The $\epsilon = 0.02$ decision boundary for both classifiers is also shown. The area that is selected by both classifiers is the rectangle in the top right of each plot. The classifier is trained on the 2600-3200 GeV SR on the M, τ_{21} feature set.

In defining the ensemble function G some general properties are expected to hold. For example, ensembling the output of the classifiers using a continuous function for G should not be expected to fix the issue. Such an ensemble still results in the placement of a random decision boundary in the feature space. This random boundary is sampled from a different set of boundaries than the input classifiers, but can still be consistent with signal. Randomly selecting which classifiers or non-continuous ensemble functions might resolve the issue. However, no ensemble strategy that directly aggregated the classifier outputs was found to work for this analysis.

The ensembling procedure that was found to work is outlined in Algorithm 3. This procedure leverages the fact that each classifier produces a histogram of the m_{JJ} distribution using different data samples. In the absence of signal the classifiers are expected to randomly select events. However, when signal is injected the classifiers are expected to select events that are consistent with the signal and a study of this is presented later in this chapter.

Each classifier j produces a histogram with bins that each have counts n_i^j. The correlation in bin i between the histograms produced by classifiers j and k is given by the number of events ρ_i^{jk} that appear in both n_i^j and n_i^k. The number ρ_i^{jk} is calculated by counting the number of events that are selected by both classifiers j and k. The uncertainty on the final histogram count n_i if there are N classifiers is then,

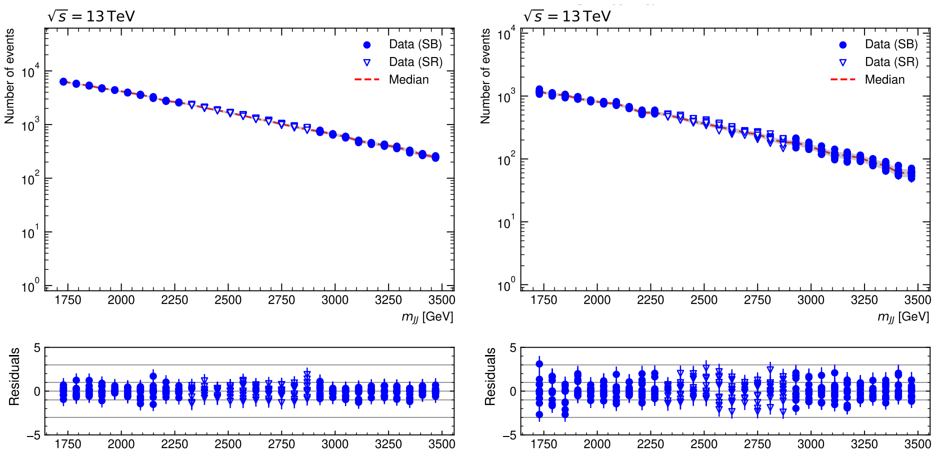
$$\sigma_i^2 = \frac{1}{N^2} \sum_{j=1}^{N} \sum_{k=1}^{N} \rho_i^{jk}, \tag{9.1}$$

Algorithm 3 Ensembled classifier histogram algorithm

```
for $i = 1$ to $N$ do
    Train a classifiers on the $m_{\mathrm{JJ}}$ SR data and reference.
    Make a selection on the dataset using classifier.
    Define an $m_{\mathrm{JJ}}$ histogram $n^i$ for the selected dataset.
end for
Calculate the median of all $m_{\mathrm{JJ}}$ histograms $\{n^i\}_{i=1}^{1}0$ to create the final observed histogram.
Account for the correlation between the histograms to determine an additional uncertainty.
Use the averaged histogram to perform the $m_{\mathrm{JJ}}$ fit.
Use the averaged histogram as the observation in the $m_{\mathrm{JJ}}$ SR.
```

where $\rho_i^{jj} = n_i^j$. This additional uncertainty is profiled in the likelihood fit where it is added in quadrature to the fit uncertainty ψ_f in eq. (8.2).

An example of the histograms produced in this approach is shown in Figure 9.4. The different histograms are seen to be consistent with each other, which is observed to be true across both CURTAINS and SALAD in all SRs. This approach was found to resolve the issue with the classifier training as shown in Figure 9.1. It was also found that this approach reduced significantly the variance in the significance reported by the analysis when changing the random seed. The analysis chose to use $N = 10$ classifiers, which was found to be well above the threshold where the variance in the significance reported by the analysis was reduced. This approach also resulted in reasonable signal sensitivity across all signal models, as is explored in the context of the full analysis pipeline in the next section. Therefore, the problem with the classifier implementation was considered solved and this ensembling approach was employed for the rest of the analysis. The next step in developing the analysis was to validate the reference generation could be integrated without biasing the classifier towards producing large excesses on the validation datasets.



(a) (b)

Figure 9.4: The histograms from ten different classifiers trained on the same idealized reference data. Each histogram is shown separately as well as its error with respect to the median count with the error calculated using Poisson statistics. The classifier is trained on the 2600-3200 GeV SR on the M, τ_{21} feature set.

9.3 Reference sample generation

The analysis pipeline has been developed to the point where the classifier is expected to be unbiased in the absence of signal. The next step is to integrate the reference generation into the analysis. It is possible the reference generation could be wrong in such a way that it biases the classifier towards producing large excesses. In this analysis, the reference generation was deemed acceptable if it did not produce an excess when no signal was injected into the data. No uncertainties on the reference generation were considered, but this should be systematically studied in future work. This section presents a visual inspection of the generated references, the performance of the classifier when trained on the reference and the full integration of the reference generation into the analysis.

An example of the reference generated by Curtains is shown in Figure 9.5 and by Salad in Figure 9.6. For both of these methods, it is observed the reference sample is a good, but imperfect, estimate of the data in the m_{JJ} SR. The better the reference sample, the less likely it is to bias the classifier and the more likely it is to amplify signal when it is injected. This piece of the analysis could be developed by optimizing measures of the quality of the reference sample to analysis independent metrics like those defined in Krause et al. [278]. The approach taken by this analysis was to test the quality of the reference generation directly in the context of the analysis by running the full analysis pipeline. This only tests the performance of the reference at specific selections and is inherently less robust than testing the reference generation using stand-alone metrics. Integrating the reference generation into the analysis is always a necessary test, but the performance of the reference generation should be quantified in a more robust way in future work.

A first concern when integrating the reference generation into the analysis is that mismodelling in the reference makes the selections made by the classifier more correlated and break the ensemble approach. An example of the correlation between two different classifiers trained on the same reference can be seen in Figure 9.7. The classifiers are observed to be more correlated than in the idealized case but still show a significant fraction of samples that are only selected by one of the classifiers. This is true for both reference generation methods in all SRs. Therefore, the classifier ensemble can still be effective in the presence of mismodelling in the reference. The correlations between two different classifiers could be quantified explicitly, but the utility of this is not clear. Correlations between different classifiers are expected, and the final test of the pipeline is the significance reported by the analysis.

The analysis was validated using both Salad and Curtains by running the full analysis pipeline on the ten partitions of the $|\Delta Y|$ SB dataset in all SRs as shown in Figure 9.8. This demonstrates the analysis behaves as expected in the absence of signal in all SRs. The analysis strategy as defined so far is therefore considered valid on the partitioned $|\Delta Y|$ SB validation dataset. Variations in the significance reported by the analysis when training on the same data with different random seeds are also reduced, observed to have an absolute spread in the reported significance of no more than 1σ in any SR. This was tested by running the full analysis pipeline on the same data with twenty different random seeds. Some variation in the reported significance is expected as the significance is a random quantity, and the downsampling of the data is expected to be random in an ideal background only setting.

In general, the performance of the analysis as summarized by Figure 9.8 is performed on all validation sets. An important consideration, that is useful for later discussion, is the number of models that need to be fit to test the analysis in this way. For each method, there are three feature sets, eight m_{JJ} SRs and ten partitions of the $|\Delta Y|$ SB dataset. This means that 240 Curtains and Salad models need to be fit to validate the analysis. On top of

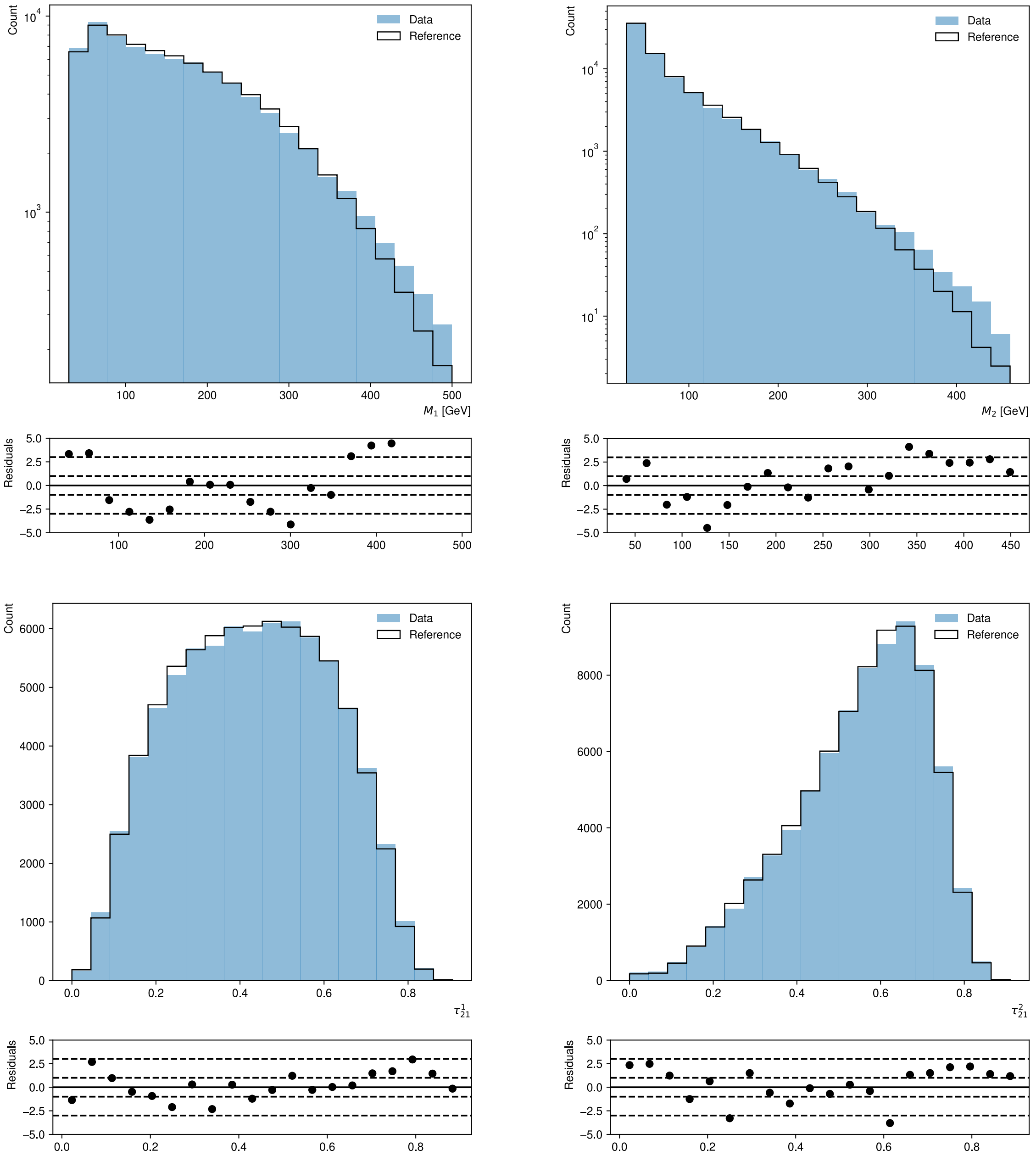


Figure 9.5: Comparison of the feature distributions in the $|\Delta Y|$ SB data and the reference sample produced by Curtains.

this, ten classifiers need to be trained for each model, and each of these classifiers is trained using 5 fold cross-validation, such that 50 classifiers need to be trained for each model. This means that 12,000 classifiers need to be trained to validate the analysis for each method on each validation set. In total, 24,000 classifiers need to be trained to produce the summary presented in Figure 9.8. This reflects a significant computation cost and, as already discussed, presents a significant technical challenge.

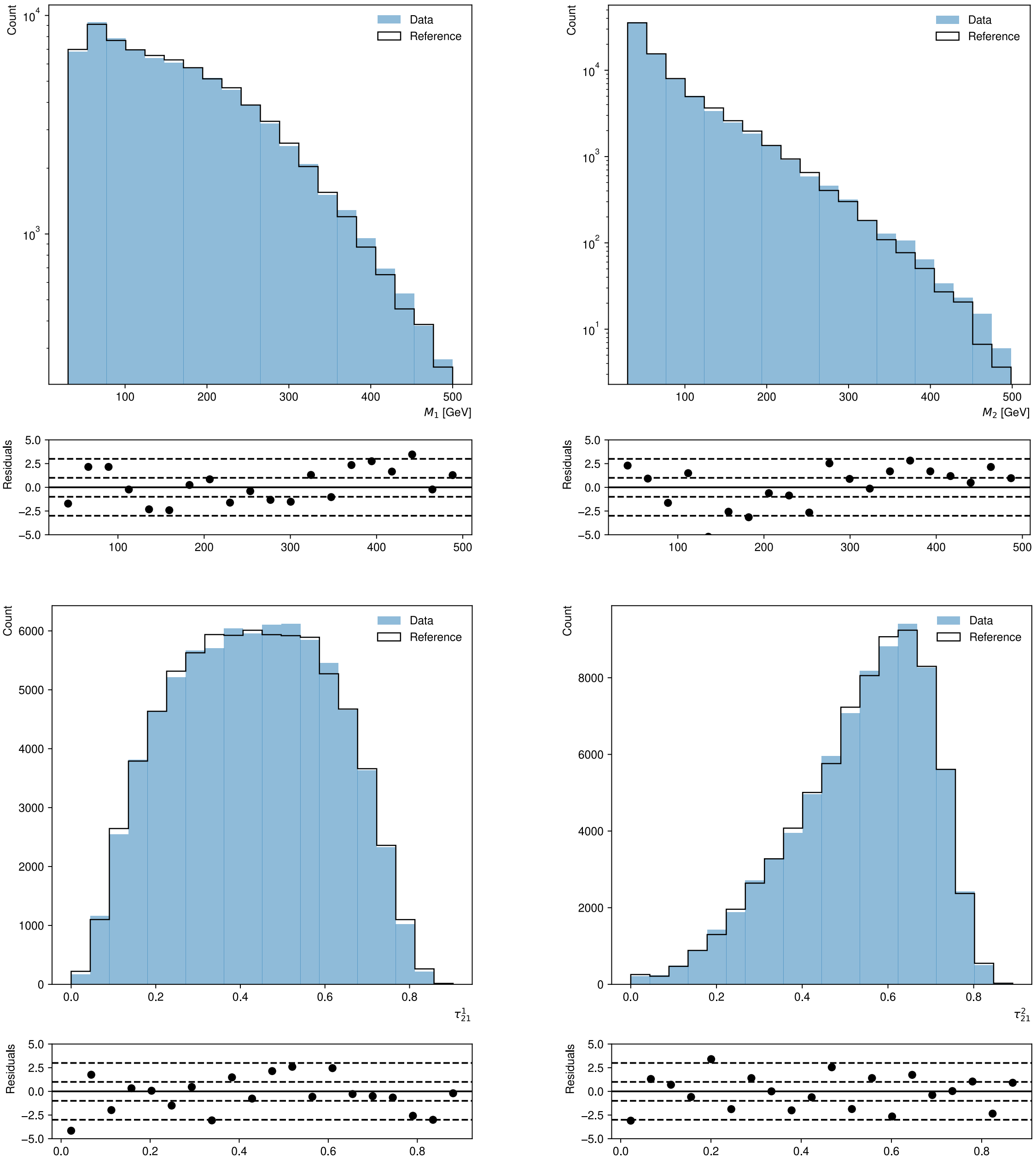
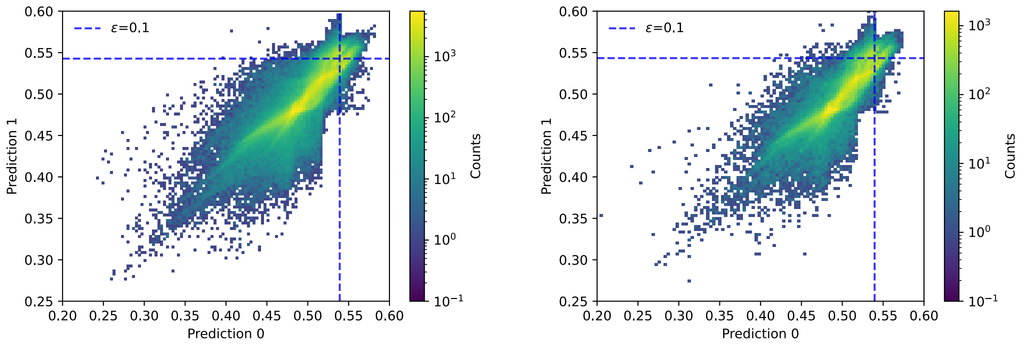


Figure 9.6: Comparison of the feature distributions in the $|\Delta Y|$ SB data and the reference sample produced by SALAD.

9.4 Signal injection

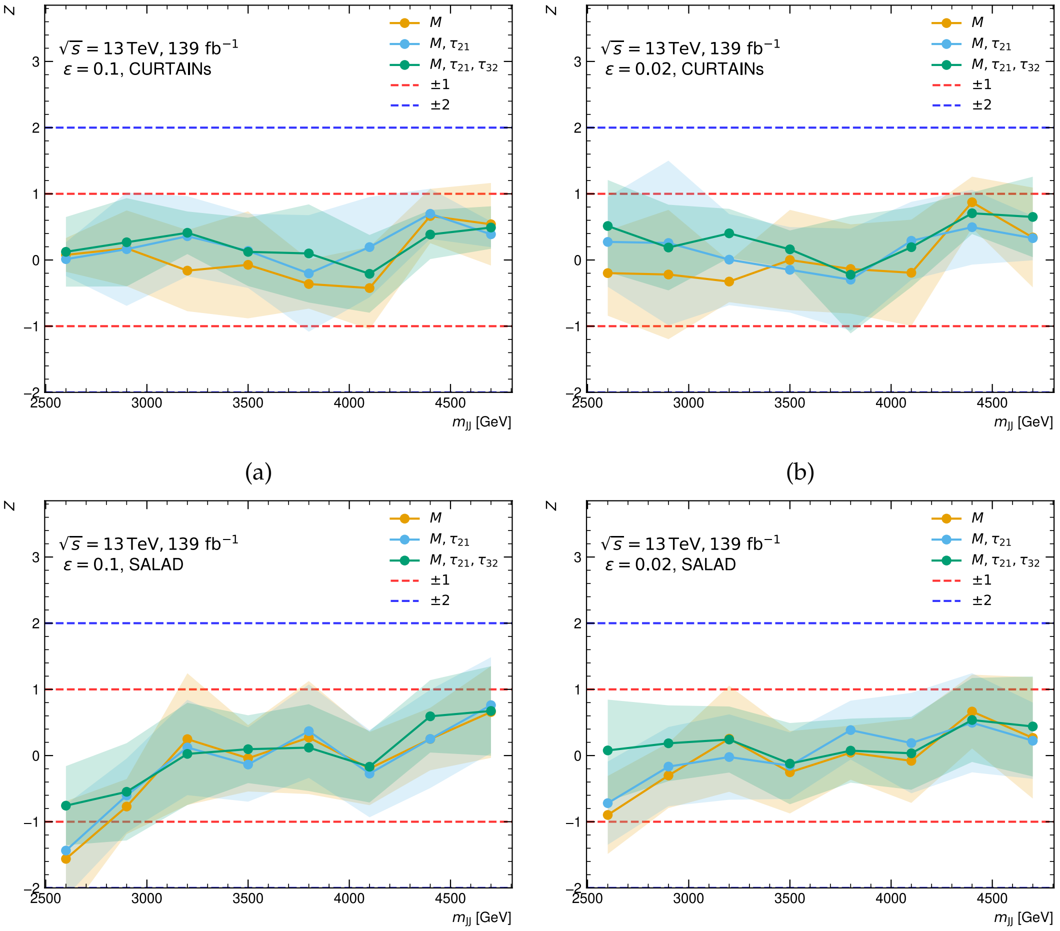
For the analysis to be considered useful it must be sensitive to signal. In the context of this analysis, sensitivity is defined as the ability to report a significant excess when signal is injected into the data. For the analysis to be considered a success, it should enhance signals injected below a 2σ threshold in the fully inclusive m_{JJ} spectrum. The threshold for 2σ is chosen because this is the threshold for identifying a region of interest in the m_{JJ} spectrum. It has also been reported in previous bump hunts that 2σ of signal can be recovered just using the fully inclusive m_{JJ} spectrum [114]. The proxy for measuring the significance of the injected signal is the local significance in the m_{JJ} SR, defined as $s = S/\sqrt{B}$ as described in



(a)

(b)

Figure 9.7: The correlation between the outputs of two randomly selected classifiers trained on a reference sample generated with CURTAINS. Same style as Figure 9.3.



(a)

(b)

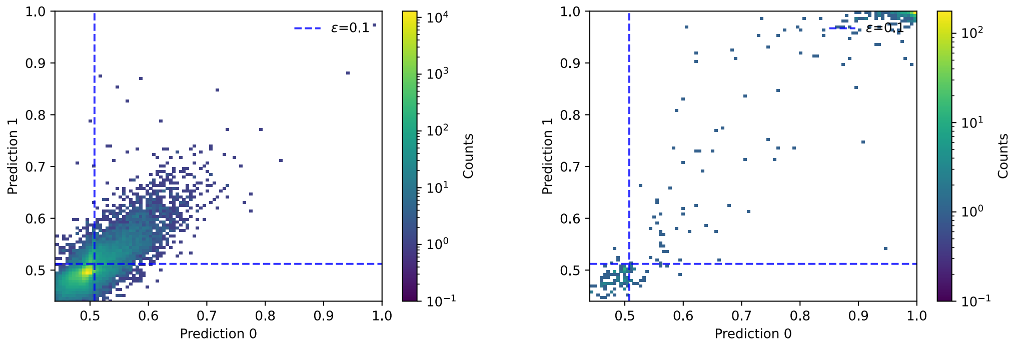
(c)

(d)

Figure 9.8: Spread of significances for (a, b) CURTAINS and (c, d) SALAD at the two different selections, (a, c) $\epsilon = 0.1$ and (b, d) $\epsilon = 0.02$. Significances are shown for all feature sets and m_{JJ} signal region centers. The spread is taken over the ten different partitions of the $|\Delta Y|$ sideband dataset. There is 3σ, measured as a local significance in the 2600-3200 GeV signal region, of signal injected into the data. Samples from the $A_0(2\gamma 2b)$ are injected into the data.

section 4.3. The amount of signal and background is calculated in a fixed m_{JJ} SR. This proxy is expected to be less than the true significance of the injected signal as it does not account for systematic uncertainties. Therefore, injecting signals at $s = 2\sigma$ is expected to be less than 2σ in the true significance.

A first check of the signal injection performance is to look at the correlation between the outputs of two classifiers trained on the same reference and m_{JJ} SR data. This is shown in Figure 9.9. It is observed the correlation between the classifiers increases when signal is injected into the data. However, the classifiers are much more correlated for the signal samples than the background, and while not fully efficient, the classifiers select a significant fraction of the signal. This indicates that both classifiers have learned to identify the same signal events, and the analysis is sensitive to signal. It is important that both classifiers select the same signal events as this is also a method for identifying signal in the data as studied in the next section.

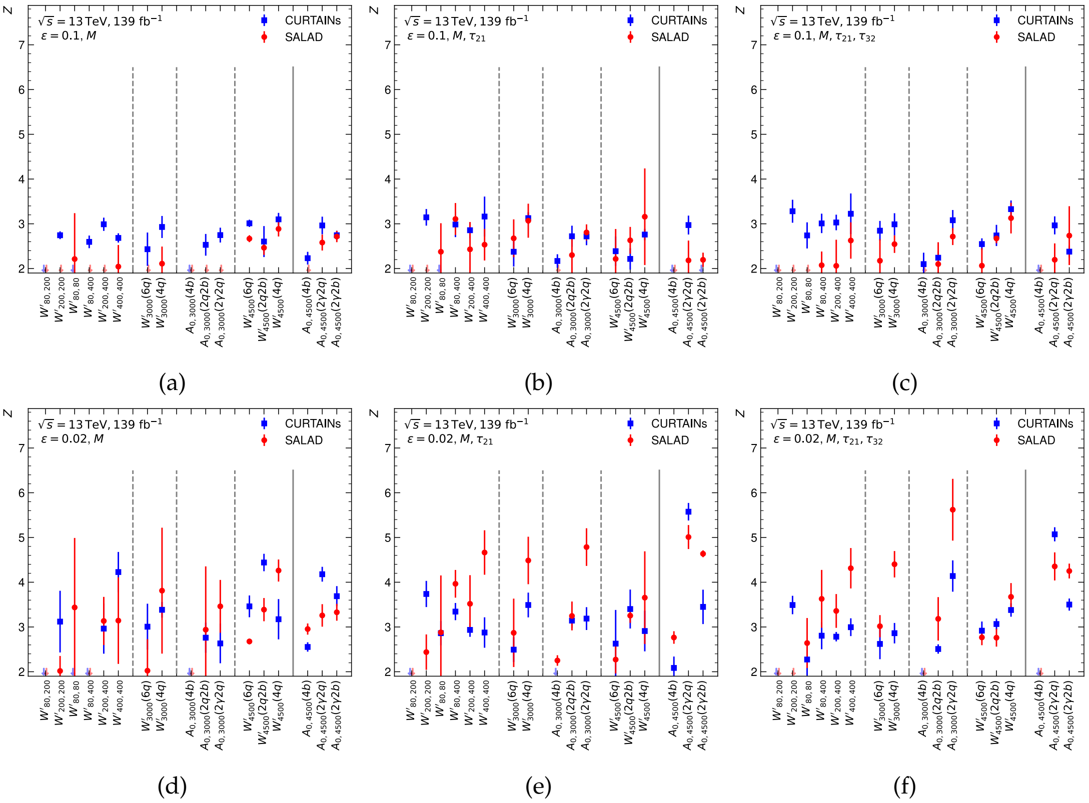


(a) (b)

Figure 9.9: The correlation between the outputs of two randomly selected classifiers trained on a reference sample generated with CURTAINS with signal samples from the $A_0(2\gamma 2b)$ model injected at $s = 2\sigma$ and the analysis run on the 2600-3200 GeV SR on the M, τ_{21} feature set. In (a) the background in the m_{JJ} SR is shown, in (b) the signal in the m_{JJ} SR is shown.

To test the sensitivity of the analysis, samples from all signal models were injected into the data at a significance of 2σ. Five different signal injections into the same data sample were considered to account for fluctuations due to the signal samples that are injected into the data. The significance reported by the analysis is summarized in Figure 9.10. It is observed the analysis enhances the presence of almost all signal models above a significance of 2σ, with a maximum average significance of $\sim 5.8\sigma$. This reflects the possibility of the analysis detecting the presence of signals in the data below the threshold achieved by standard approaches on the fully inclusive m_{JJ} spectrum. For some models the significance is actually reduced below 2σ, this is because the local significance of the injected signal is only a proxy for the true significance, but also because if the classifier does not learn to discriminate the signal from the background the sensitivity of the search is reduced. The ability of the analysis to enhance signals below 2σ suggests the analysis is sensitive to new physics on data that is distributed similarly to the $|\Delta Y|$ SB data.

In the signal injection tests the larger feature sets, M, τ_{21} and M, τ_{21}, τ_{32}, are only slightly more sensitive to signal than the M feature set. The $\epsilon = 0.02$ selection is more sensitive to signal than the $\epsilon = 0.1$ selection, which is expected if the classifiers have learned to discriminate signal from background. For some signal models there is a significant difference between the significance reported by CURTAINS and SALAD. This latter observation is possibly problematic,



(a) (b) (c)

(d) (e) (f)

Figure 9.10: Signal injection tests at (a, b, c) $\epsilon = 0.1$ and (d, e, f) $\epsilon = 0.02$. For both Salad and Curtains all feature sets are used (a, d) $\mathcal{X} = M$, (b, e) $\mathcal{X} = M, \tau_{21}$ and (c, f) $\mathcal{X} = M, \tau_{21}, \tau_{32}$. The local significance (Z) reported by the analysis pipeline is calculated after running the analysis with $s = 2\sigma$ of signal injected into the data in the m_{JJ} signal region centered on the different signals. The analysis is run for five different signal injections on one partition of the $|\Delta Y|$ sideband dataset.

as the two approaches are intended to serve as a cross check for each other, and when unblinding, if one method reports an excess and the other does not it is unclear what should be done. However, there are only a few signal models where one method reports evidence of signal and the other does not. Therefore, the analysis is considered to be sensitive to signal when injected into the data.

9.5 Signal identification

Beyond being sensitive to signal when injected into the data, the analysis would ideally be able to characterize the signal in the data. This means the analysis should be able to identify the distributions of signal samples in the features $\mathcal{X}$ that are used to make the classifier selection. One approach to understanding the signal identification strategy of the analysis is to look at the histograms of the events that pass the classifier selection. This is shown in Figure 9.11, for one signal model, where it can be seen the distributions of the features after the classifier selection closely match the characteristics of the signal. This is expected as the classifier is trained to select signal events, but this behaviour is not guaranteed. A general difficulty analyses of this type face is interpreting the nature of any excesses that are reported. In particular, it is not clear how one should discriminate between analysis failure and the presence of new physics.

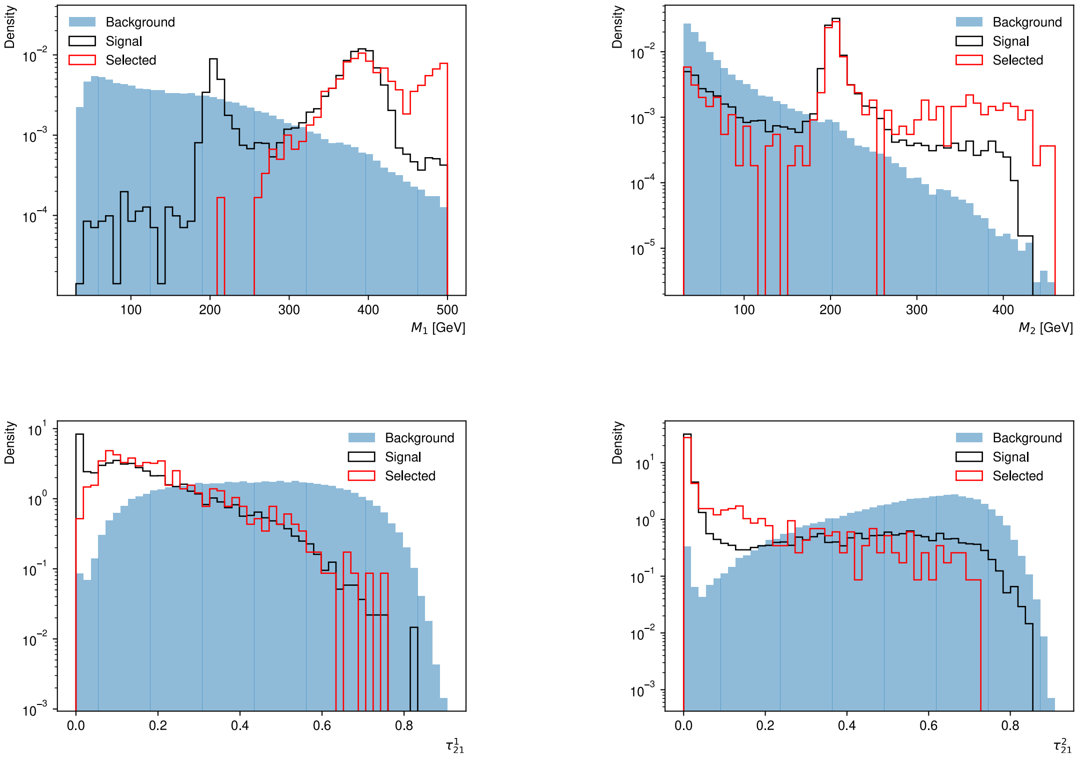


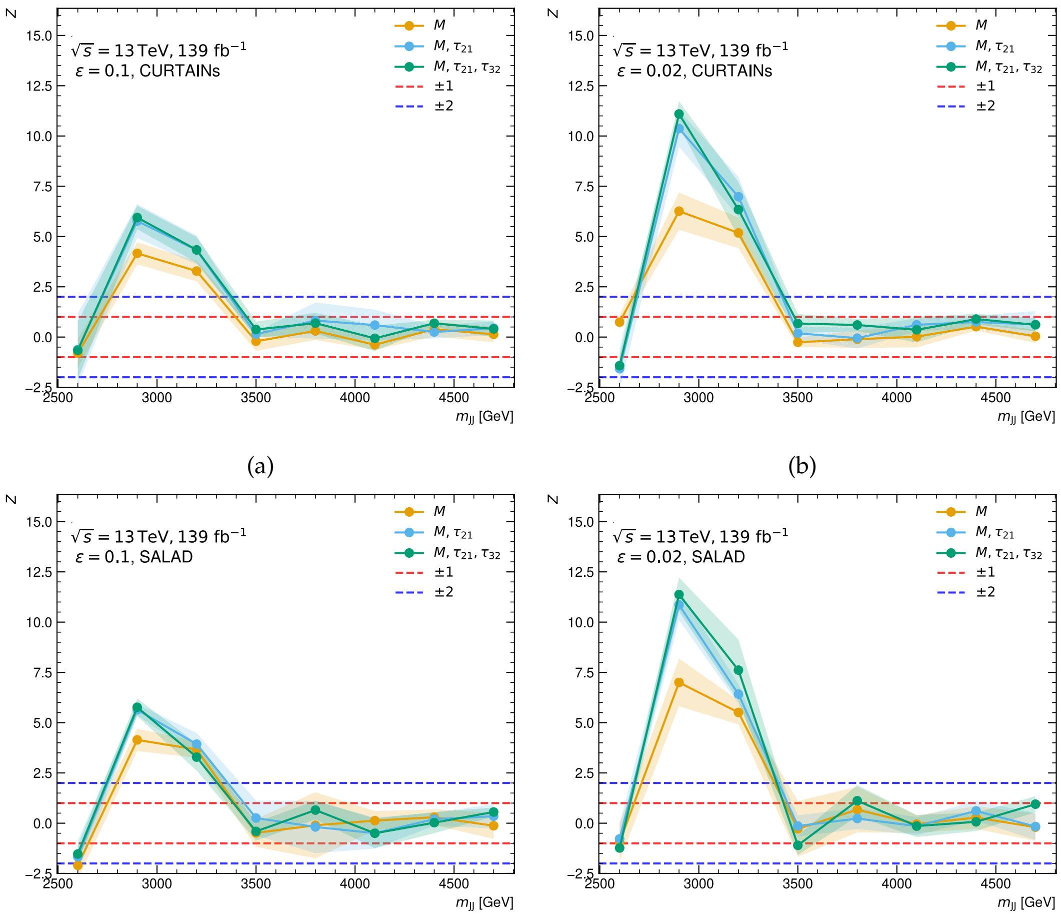
Figure 9.11: Signal identification plot where samples from the $A_0(2\gamma 2b)$ signal model were injected at $s = 2\sigma$ into the data and the analysis was run on the 2600-3200 GeV signal region on the M, τ_{21} feature set. The histograms of the events that pass all classifier selections at $\epsilon = 0.02$ are shown.

9.6 Signal scan

Beyond validating that the analysis has sensitivity to signal when injected in a single m_{JJ} bin, it is important to understand how the analysis behaves across the full m_{JJ} spectrum in the presence of signal. To test this 3σ of the $A_0(2\gamma 2b)$ signal model centered on 3000 Gev was injected, calculated as the local significance in the 2600-3200 GeV SR, into the data in all m_{JJ} SRs. This means the analysis can be run on the full m_{JJ} spectrum with a signal present. The result of running a full scan in the presence of signal is shown in Figure 9.12. The analysis reports a significant excess in two of the SRs, 2600-3200 GeV and 2900-3500 GeV. The behaviour of CURTAINS and SALAD is seen to differ in the m_{JJ} SRs that overlap with the injected signal. It is observed that SALAD has a greater average tendency to report deficits in the immediate neighbouring bins of the signal. In the rest of the m_{JJ} spectrum, the analysis behaves as expected in the absence of signal. Therefore, signal in the data is expected to have effects in multiple m_{JJ} SRs. This effect is expected to be reduced in the presence of smaller signal injections, and a large signal injection was chosen to demonstrate the effect.

9.7 Addtional validation sets

The analysis has been developed using the ten partitions of the $|\Delta Y|$ SB dataset, and it has passed all the criteria set out in the previous sections for validation success. The next step in the analysis development is to validate the analysis on a $|\Delta Y|$ SB dataset that has been resampled to match the $|\Delta Y|$ SR m_{JJ} distribution and on the $|\Delta Y|$ SR MC samples. For



(a)

(b)

(c)

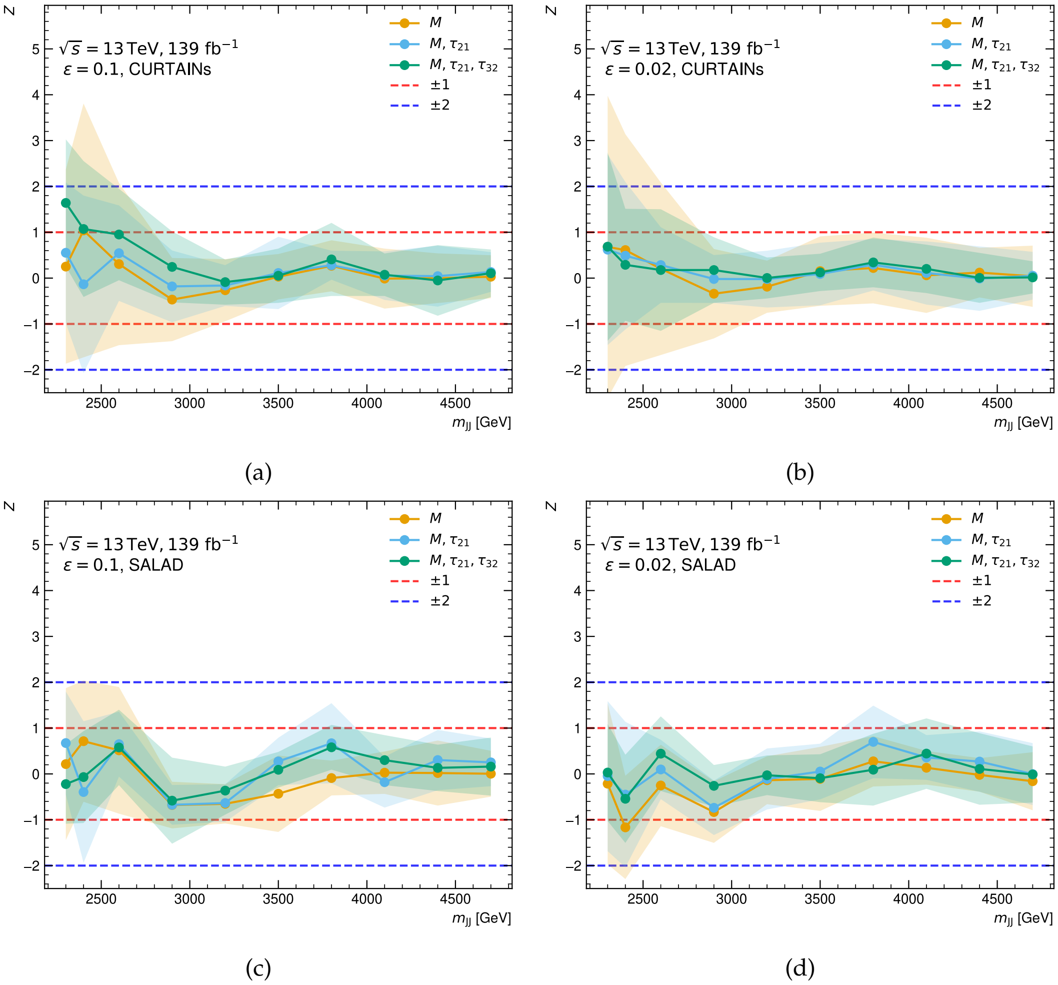
(d)

Figure 9.12: Spread of significances for (a, b) CURTAINS and (c, d) SALAD at the two different selections, (a, c) $\epsilon = 0.1$ and (b, d) $\epsilon = 0.02$. Significances are shown for all feature sets and m_{JJ} signal region centers. The spread is taken over the ten different partitions of the $|\Delta Y|$ sideband dataset. There is 3σ, measured as a local significance in the 2600-3200 GeV signal region, of signal injected into the data. Samples from the $A_0(2\gamma 2b)$ are injected into the data.

these tests, no signal injections are shown as the expected sensitivity is defined in the next chapter. The analysis validation is presented in this way as it parallels the development of the analysis and serves to demonstrate clearly why certain decisions were made.

When the analysis was applied to the resampled $|\Delta Y|$ SB dataset, the SALAD branch of the analysis behaved as expected, but CURTAINS was seen to fail in multiple SRs, this was not tested systematically as a single failure was enough to demonstrate the issue. This is the first example of the reference generation method sculpting a spurious excess in the m_{JJ} spectrum and demonstrates that this is a failure mode of the analysis. This was identified to be caused by the fact that CURTAINS did not account for the statistical imbalance in the m_{JJ} distribution. As already shown in Figure 8.4 the m_{JJ} distribution in the $|\Delta Y|$ SR is much more steeply falling than the $|\Delta Y|$ SB. Therefore, the statistical imbalance in the m_{JJ} distribution is expected to be much larger in the resampled $|\Delta Y|$ SB dataset than in the original $|\Delta Y|$ SB dataset. To account for this the CURTAINS approach was modified such that the model was fit on a dataset that was resampled to be flat in m_{JJ}. This downsampling was performed by resampling the m_{JJ} spectrum in both SBs such that it followed a uniform distribution. With this modification, the CURTAINS approach performed as expected on the resampled $|\Delta Y|$ SB dataset. The results of both Salad and Curtains on the resampled $|\Delta Y|$ SB dataset are discussed in more detail in the next section where they are also relevant and are shown in Figure 9.14.

The analysis was also run on the $|\Delta Y|$ SR MC samples. Here the Salad approach performed as expected almost everywhere, but Curtains resulted in a significant excess similar to Figure 9.1 when tested in the 2600-3200 GeV SR. The method was not tested on the full m_{JJ} spectrum as this failure was anticipated, for the reasons outlined in the following. The choice of the function g was left as the default $g(m^1_{\mathrm{JJ}}, m^2_{\mathrm{JJ}}) = |m^1_{\mathrm{JJ}} - m^2_{\mathrm{JJ}}|$ for the studies on the $|\Delta Y|$ SB data. However, the correlations in the $|\Delta Y|$ SR are more complicated, as shown in Figure 8.4. Therefore, using $g(m^1_{\mathrm{JJ}}, m^2_{\mathrm{JJ}}) = \{m^1_{\mathrm{JJ}}, m^2_{\mathrm{JJ}}\}$ was expected to be a more suitable choice for the $|\Delta Y|$ SR. When changing to this conditioning function the Curtains approach performed as expected on the $|\Delta Y|$ SR MC samples. The full validation of both Curtains and Salad is shown in Figure 9.13. The change in Curtains was again validated on the resampled $|\Delta Y|$ SB dataset. This demonstrates that both Salad and Curtains are capable of producing valid reference samples on data samples with similar correlations to the $|\Delta Y|$ SR data almost everywhere.



(a)

(b)

(c)

(d)

Figure 9.13: Spread of significances for (a, b) Curtains and (c, d) Salad at the two different selections, (a, c) $\epsilon = 0.1$ and (b, d) $\epsilon = 0.02$. Significances are shown for all feature sets and m_{JJ} signal region centers. The spread is taken over ten different upsamplings of the MC dataset in the $|\Delta Y|$ signal region.

At low m_{JJ} both Curtains and Salad can be seen to be biased towards producing an excess slightly more than is expected in the background only case. To explore the nature of this, the analysis was run in lower m_{JJ} SRs. The turn on in m_{JJ} is at ~ 1200 GeV in the $|\Delta Y|$ SR, and so the m_{JJ} fit function is expected to work in this region. The likelihood of producing an excess increases significantly at low m_{JJ} when using Curtains, but not when using Salad. This could be due to a random fluctuation produced when upsampling the MC samples, such fluctuations are particularly likely in low m_{JJ} as there are fewer MC events in this region. However, it could also be due to the more complex correlations observed in this region as shown in Figure 8.4. In either case it is possible there are issues at low m_{JJ} that neither method is capable of addressing properly. This motivates reconsidering the choice of m_{JJ} bins in the analysis. This choice is made using the final validation set, the (down)upsampled $|\Delta Y|$ SR data.

In summary, the Curtains approach to reference generation was updated to account for the statistical imbalance in the m_{JJ} distribution and the function g was updated to account for the more complex correlations in the $|\Delta Y|$ SR. All results are now produced with these updates in place. The only consideration made on the final validation set is the choice of SRs to unblind.

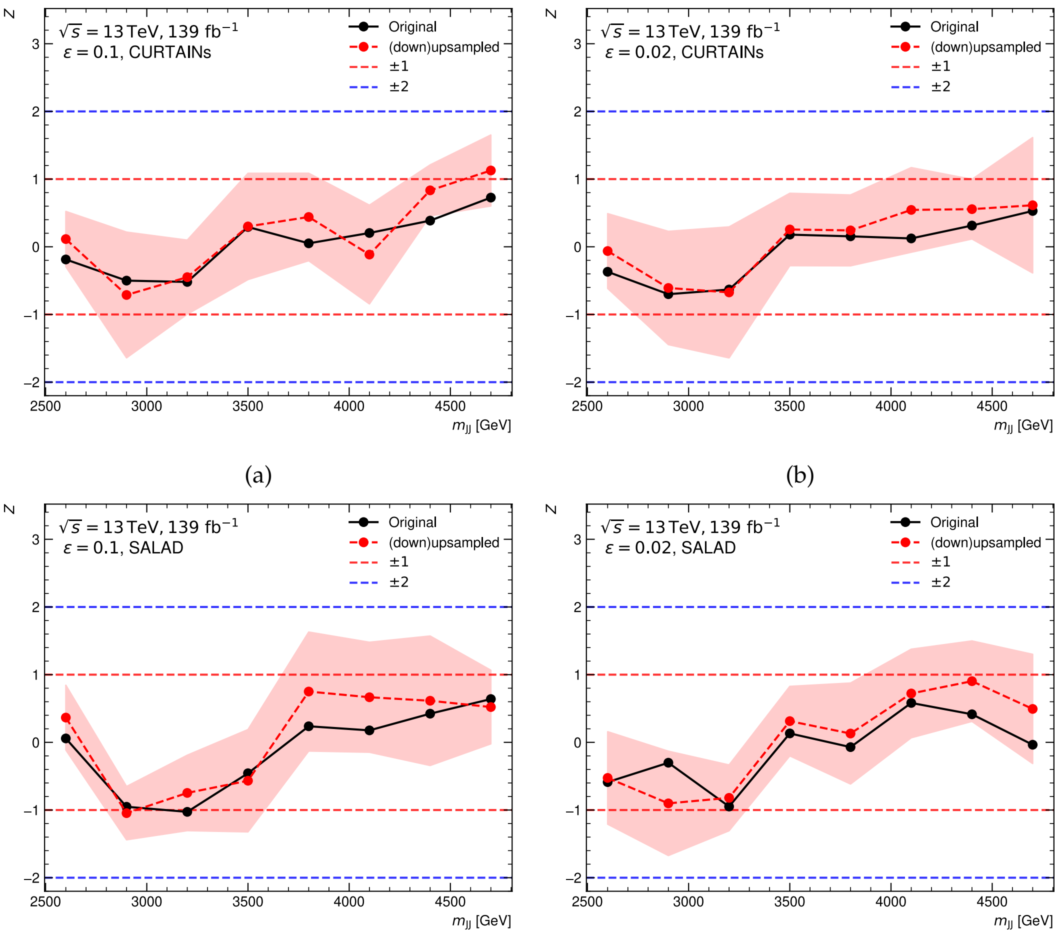
9.8 Validating validation strategies

The (down)upsampled validation set may be susceptible to random signal like fluctuations, as discussed in section 8.2.4[2]. There is also the potential that through some unknown mechanism the (down)upsampled validation sets are not sufficiently signal suppressed. The analysis pipeline developed in the previous section does not produce any excesses on the $|\Delta Y|$ SB validation set, and it behaves as expected almost everywhere on the $|\Delta Y|$ SR MC samples. Therefore, this pipeline is used to test the validity of the procedures used to generate the (down)upsampled validation sets.

To test this validation approach signal samples from the $A_0(2\gamma 2b)$ model are injected into the resampled $|\Delta Y|$ SB data such that the analysis reports an excess of $> 5\sigma$ for both Salad and Curtains in the 2600-3200 GeV SR. The (down)upsampling procedure on the combined simulated signal and data samples is repeated 10 times and the full analysis is run on all feature sets and m_{JJ} windows. Similar behaviour to the original resampled $|\Delta Y|$ SB samples from which these datasets were constructed was observed as shown in Figure 9.14. This confirms the (down)upsampling procedure suppresses signal and does not produce any spurious excesses on $|\Delta Y|$ SB data. It also demonstrates the upsampling procedure produces datasets that are similar to those from which they were created, as demonstrated by the correlation between the significances reported by the original resampled $|\Delta Y|$ SB data and the (down)upsampled data. A full set of these studies on all feature sets and methods is provided in Appendix B. No bias due to the signal injection is observed.

It is still possible the (down)upsampling approach could break down in the $|\Delta Y|$ SR where the correlations between the features and m_{JJ} are more complex than those of the $|\Delta Y|$ SB. To fully validate the (down)upsampling procedure it should be tested on a dataset with similar correlations to the $|\Delta Y|$ SR data. Validating the upsampling of the MC dataset is a partial validation of this, but a stronger test would be to repeat the (down)upsampling test on the upsampled MC dataset. This would also be an incomplete test as the upsampled MC dataset still mismodels the $|\Delta Y|$ SR correlations. However, this would be a sensible test for future iterations, especially if an equal statistics MC sample were produced. This does expose a gap in the validation strategy of this procedure, where it is possible the (down)upsampling

[2] As discussed this applies to the upsampled MC as well, but the analysis has already been seen to pass this validation everywhere in m_{JJ} that is relevant.



(a)

(b)

(c)

(d)

Figure 9.14: Spread of significances for (a, b) CURTAINS and (c, d) SALAD at the two different selections, (a, c) $\epsilon = 0.1$ and (b, d) $\epsilon = 0.02$. Significances are shown for the M feature set and all m_{JJ} signal region centers. The spread is taken over ten different upsamplings of the downsampled resampled $|\Delta Y|$ sideband dataset.

procedure can produce a dataset that causes the analysis to fail where it would not have otherwise. To be conservative, it was decided that any failure on the (down)upsampled $|\Delta Y|$ SR data would be considered a failure of the analysis.

9.9 Summary

This chapter detailed the development of the analysis. The analysis was primarily developed using the partitioned $|\Delta Y|$ SB dataset, where the analysis was seen to behave as expected in its default configuration, except for the classifier training. The classifier training was seen to be sensitive to the random seed, where large excesses when no signal was injected were observed much more often than expected in the background only case. This was resolved by ensembling the classifiers, which was seen to reduce the variance in the significance reported by the analysis when changing the random seed and to remove the appearance of large excesses when no signal was injected. The issues around the classifier bias highlight the inherent difficulty in constructing an analysis strategy around the weakly supervised approach pursued here. Multiple studies have explored the use of reference generation techniques in the context of weakly supervised searches as discussed in Section 6.5, and none have reported the classifier issues observed here. Further development of these methods should take these approaches into account.

On the $|\Delta Y|$ SB dataset the generated references were observed to be of high quality, and the analysis was seen to be sensitive to signal when injected. The analysis was also validated on the resampled $|\Delta Y|$ SB dataset, where the Curtains approach was seen to fail. This was resolved by accounting for the statistical imbalance in the m_{JJ} distribution, which the Salad implementation had already accounted for. The analysis was then validated on the $|\Delta Y|$ SR MC samples, where the Curtains approach was again seen to fail. This was resolved by accounting for the more complex correlations in the $|\Delta Y|$ SR. The repeated failure of the Curtains approach reflects the need to tune these approaches to the specific dataset being used. A robust validation strategy is also required to ensure the analysis is not biased by the reference generation. The next chapter presents additional studies on the final validation dataset, and also the unblinded results of the analysis.

Chapter 10

Analysis results

This chapter presents the results of the analysis on the final validation datasets and the $|\Delta Y|$ SR data. The analysis procedures are considered to be fixed at this point. The only remaining consideration is the SRs – defined by classifier selection, feature set and m_{JJ} bin – to unblind. The choice to freeze the analysis procedures was made for expediency, as the analysis was already in an advanced stage when it became evident the validation presented here was necessary. Given the analysis timeline, making adjustments and fully validating them was not feasible. Ideally, an analysis would be developed on more realistic representations of the $|\Delta Y|$ SR data from the outset to avoid such situations.

10.1 Validation results

The (down)upsampling construction is now assumed to be valid and is used in this chapter to produce the primary validation datasets. This choice is well motivated because the (down)upsampled validation set does not have the low statistics issue of MC at low m_{JJ}, and has similar correlations to the $|\Delta Y|$ SR data. A full comparison of these validation sets was provided in section 9.8. The frozen analysis has already passed all validation tests on the upsampled MC dataset and the $|\Delta Y|$ SB dataset in all SRs.

10.1.1 Single run

This subsection presents a complete run of the analysis on a single dataset constructed using the (down)upsampling procedure. The purpose of this is to state in one place the complete analysis pipeline and show the main intermediate results, as well as to outline the challenges faced in the $|\Delta Y|$ SR data. This demonstration is most clearly made for a single analysis run and so only the M, τ_{21} feature set in the SR from $2600 - 3200$ GeV using the CURTAINS method is shown. In all later sections, only summaries of the final results or partial intermediate results are presented.

The first step of the analysis produces the reference distribution over the features in M, τ_{21} shown in Figure 10.1. This reference is visually worse than that produced for the $|\Delta Y|$ SB validation set where the analysis was developed, shown in Figure 9.5. A similar degradation is observed for the SALAD method. This is expected as the correlations between M, τ_{21} and m_{JJ} in the $|\Delta Y|$ SR data are much more complex than in the $|\Delta Y|$ SB. Therefore, the interpolation of the reference sample from the m_{JJ} SBs is more challenging. The same degradation is not observed in the MC validation set. This reflects the fact that the (down)upsampled validation set is a more stringent test of the analysis than the MC validation set.

Once a reference sample has been produced the SR data is split into 5 folds. A classifier is trained as defined in section 8.3.2, and a small amount of random noise on the order of

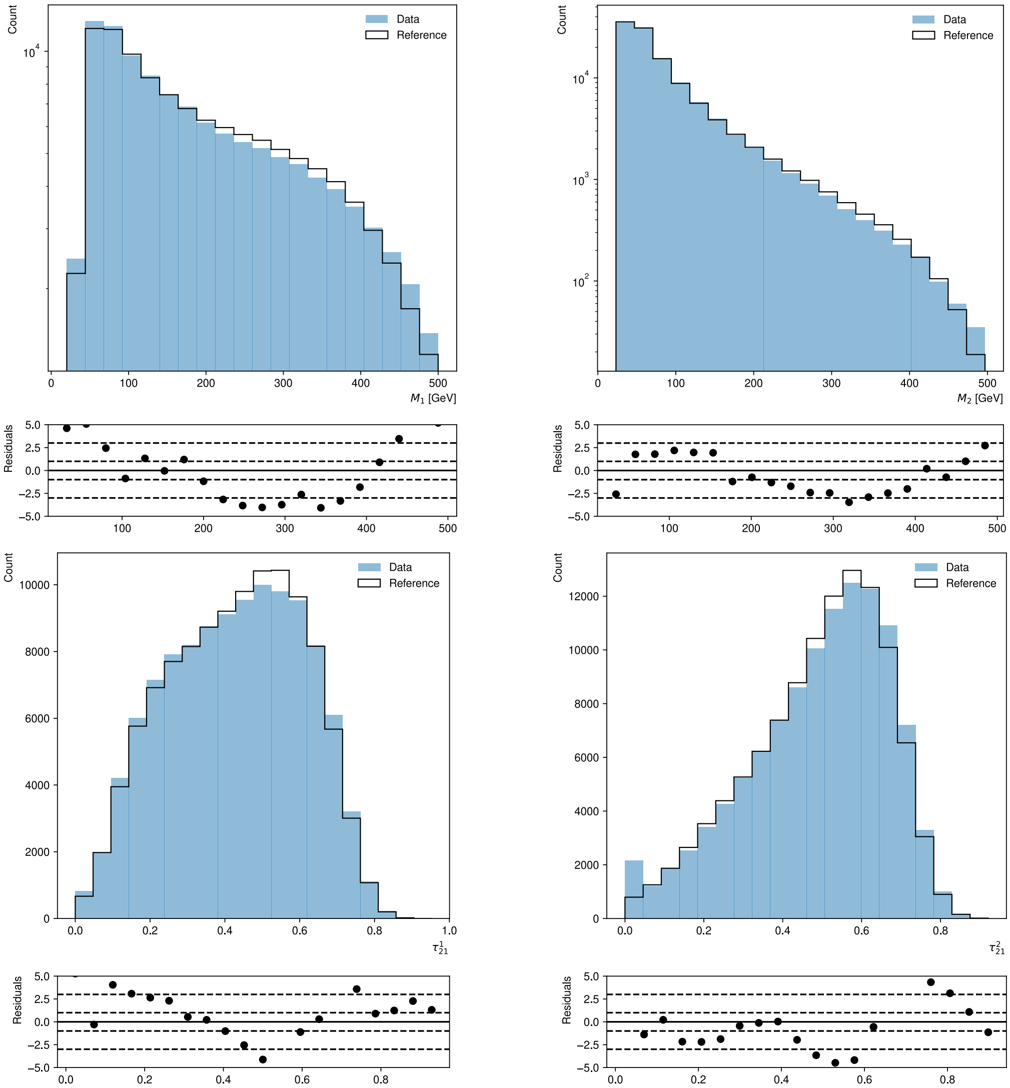
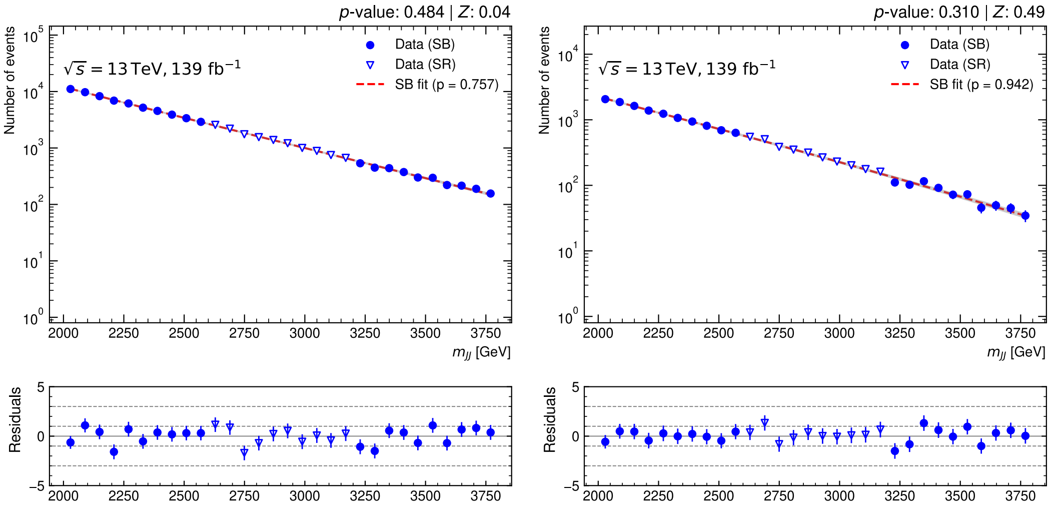


Figure 10.1: Comparison of the feature distributions in the $|\Delta Y|$ signal region data and the reference sample produced by Curtains in the $2600 - 3200$ GeV m_{JJ} signal region.

machine precision is added to the classifier scores to break ties as described in section 9.1. Ten classifiers are trained on the same reference sample and $|\Delta Y|$ SR data. In each iteration the splitting of the data into folds is different, and the classifier initializations are different. This randomization results in variations amongst the scores assigned to different data samples.

For each of the ten classifiers, all samples in the top $\epsilon = 0.02$ fraction of scores are selected. Each of the ten selected datasets is used to build a histogram with 30 bins in m_{JJ}. The histograms are then combined to produce a final histogram for the SR. Uncertainties are calculated using the error propagation of the different histograms defined in section 9.2. The histogram bins in the m_{JJ} SBs are used to estimate the background in the SR using the parametric fit described in section 8.3.3. The final histogram and the fit function are shown in Figure 10.2.

A likelihood fit is performed in the m_{JJ} SR following the procedure described in section 4.2 with $\mu = 0$. The quantities relevant to this fit are shown in Figure 10.3, where the observed data can be seen to be consistent with the predicted background. The dominant uncertainty



(a)

(b)

Figure 10.2: Representative fits for the CURTAINS method using the M, τ_{21} feature set. Fits are shown at (a) $\epsilon = 0.1$ and (b) $\epsilon = 0.02$. The m_{JJ} sideband p-value is generally high, indicating a good fit. The m_{JJ} histograms change smoothly, and the interpolated fit function agrees with the data in the signal region.

in the fit is the uncertainty from the background fit, followed by the statistical error (Poisson) and then the uncertainty from the mass histogram ensemble. The likelihood is converted to a significance using the asymptotic formula. In the final unblinded analysis this significance is corrected by a linear fit to the observed significances on ten different (down)upsampled validation sets. The final significance is corrected using eq. (8.3). All unblinded results use the corrected significance, including the limit setting procedure.

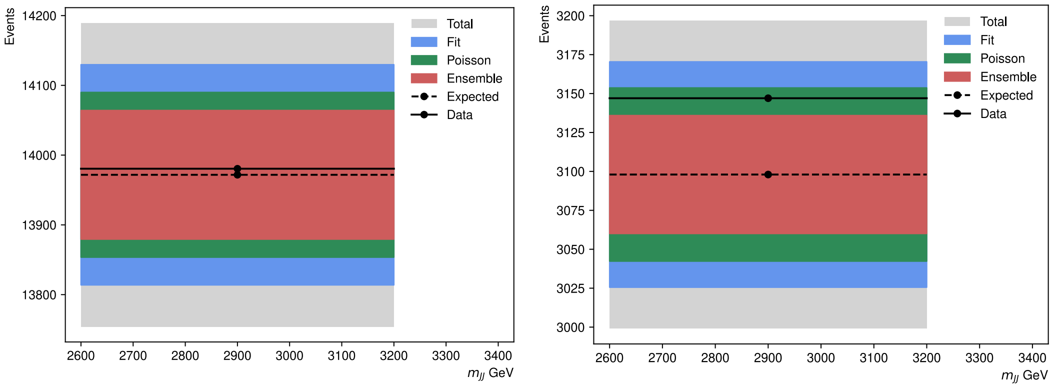
The performance of the analysis is summarized through the distribution of the output significances Z extracted from the likelihood fit. In calculating this significance a modified version of the significance is used, where negative significances are not set to zero as discussed in section 4.2.

10.1.2 Signal region scan

To identify the SRs that can be unblinded the analysis is performed multiple times on different (down)upsampled validation sets. Each of these sets is generated from a different downsampling of the $|\Delta Y|$ SR data. The analysis is considered to pass the validation if both the mean and the standard deviation of the significances are less than one. This choice is arbitrary and would in principle allow for a 2σ excess to be within one standard deviation of the mean. However, this was deemed acceptable as the linear fit to the significances is expected to correct for any such biases.

The results of the scan are shown in Figure 10.4. The SRs considered for unblinding are centered on 2600, 2900, 3200, 3500, 3800, 4100, 4400 and 4700 GeV. A finer grid is considered at low m_{JJ} to better probe the performance of the analysis in this region. The analysis is seen to perform as expected at high m_{JJ}, but there are significant excesses at low m_{JJ}. These excesses appear across multiple m_{JJ} bins and all feature sets and methods.

From Figure 10.4 it was concluded the m_{JJ} bin centered on 2600 GeV could not be used for unblinding. This is because the analysis reports on average an excess greater than one



(a)

(b)

Figure 10.3: The contributions to the likelihood fit from the different uncertainties, the observed data and the predicted background (expected) at (a) $\epsilon = 0.1$ and (b) $\epsilon = 0.02$. The uncertainties from the background fit are shown in blue, the uncertainties from the mass histogram ensemble are shown in red, the variance of the Poisson with the predicted background is shown in green and the total uncertainty is shown in grey.

for all feature sets, classifier selections and methods in this m_{JJ} bin. All other m_{JJ} bins are considered suitable for unblinding. The primary reason for performing the scan in Figure 10.4 was to identify the m_{JJ} bins that could be unblinded, but the behaviour of the analysis at low m_{JJ} is not understood and so is explored further as detailed in the next subsection.

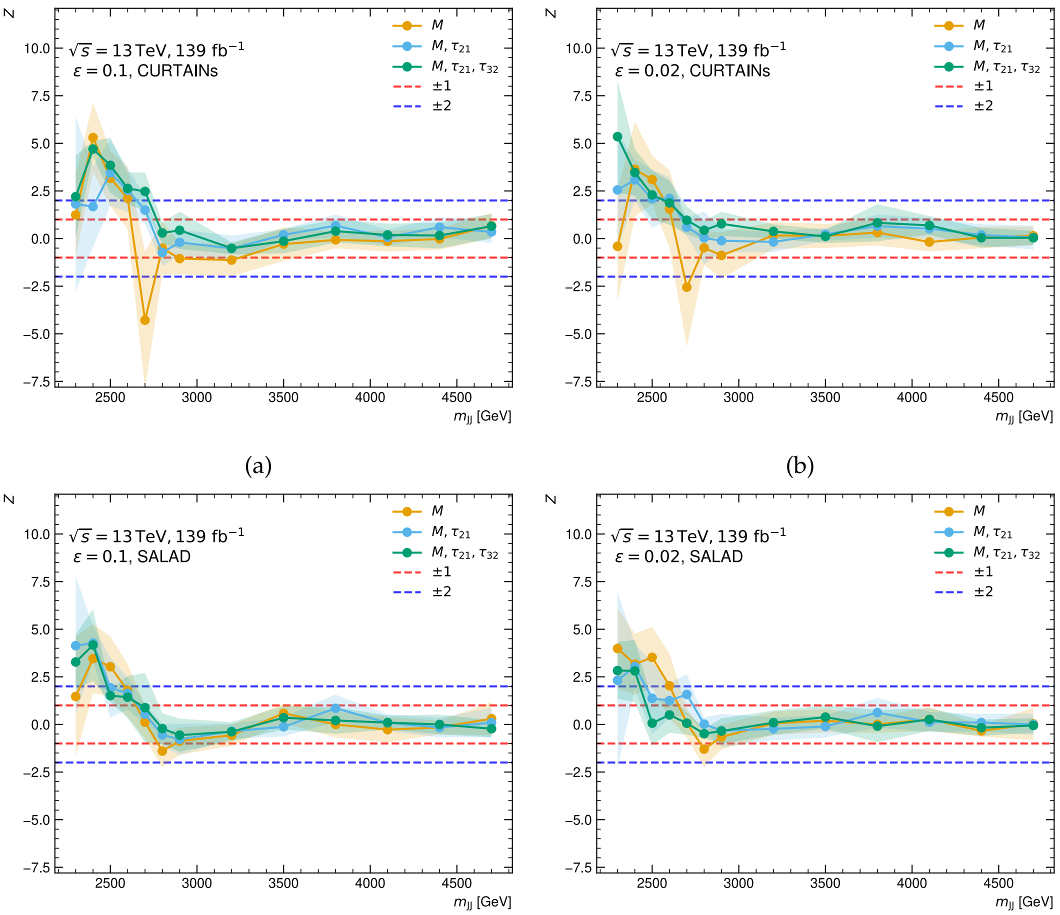
10.1.3 Issues at low m_{JJ}

The behaviour of the analysis at low m_{JJ}, shown in Figure 10.4, could be due to a number of reasons. The analysis could be failing due to a previously unseen issue, the generation of the (down)upsampled validation set could be flawed, or there could be real signal in the data. This subsection explores and discusses these possibilities. The (down)upsampling construction has been assumed to be valid, but this is still discussed in the context of the observations at low m_{JJ}.

For real signal contamination to be the cause of the observed excesses the signal must be produced at a large cross section. The (down)upsampling procedure suppresses the signal by a factor of $\sqrt{30}$ and the largest excess observed in Figure 10.4 is on average $\sim 5\sigma$. Accounting for the signal enhancement capabilities of the analysis, based on the best case signal injection shown in Figure 9.10, the signal could be on the order of 2σ in the data. Therefore, this 'signal' in the $|\Delta Y|$ SR dataset is most likely $\geq 2\sqrt{30} \sim 10\sigma$. Such a large excess is expected to be visible in the fully inclusive dataset, and so this possibility is considered unlikely.

The observed excesses could be introduced by mismodelling in the generative model used to build the (down)upsampled validation sets. This is possible because, as discussed in section 9.8, there is a gap in the validation of this procedure. However, the low m_{JJ} region is the highest statistics region, so the generative model should be the most stable here. If the generative modelling procedure were to introduce false excesses they would be expected to appear in low statistics regions[1]. In the $|\Delta Y|$ SR MC the analysis also reports excesses at low m_{JJ}, as shown in Figure 9.13. Therefore, it is unlikely the upsampling procedure has introduced the excesses at low m_{JJ}.

[1] At high m_{JJ} or below the trigger plateau.



(a)

(b)

(c)

(d)

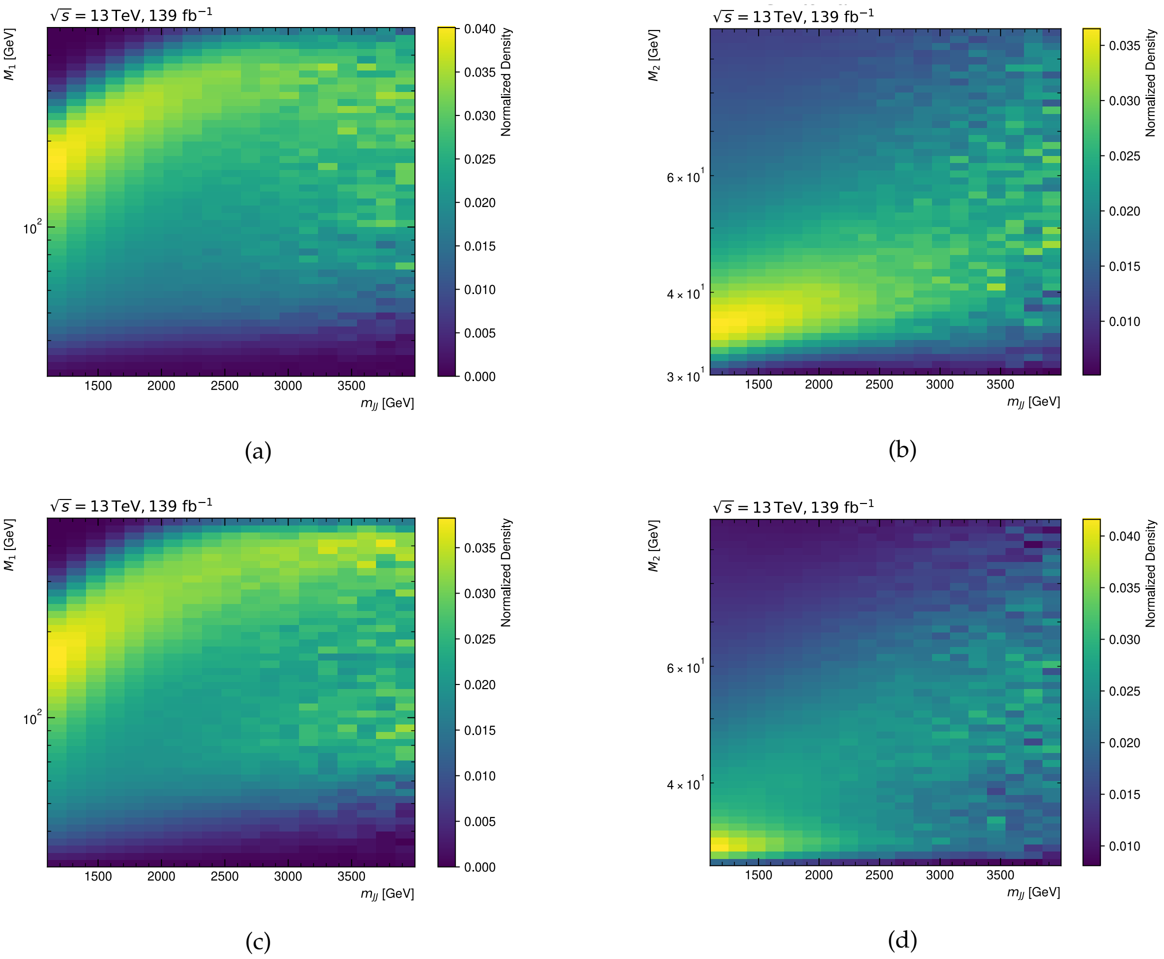
Figure 10.4: Spread of significances for (a, b) Curtains and (c, d) Salad at the two different selections, (a, c) $\epsilon = 0.1$ and (b, d) $\epsilon = 0.02$. Significances are shown for all feature sets and m_{JJ} signal region centers. The spread is taken over ten different (down)up-sampling validation sets.

This points to the most likely cause of the issue being an analysis failure. Assuming this to be true the proximate question is if the reference sample generation is flawed, or if the classifier training is again introducing a bias as in section 9.1. To try and identify the cause of the issue a number of tests were performed as detailed in the following.

The first test was to repeat the use of the idealized reference sample to isolate the effect of the reference sample generation. As the (down)upsampled validation set is generated using a generative model, a perfect reference can be sampled from this model. This test is run in all m_{JJ} SRs with centers below 2600 GeV in Figure 10.4. In these tests, the absolute value of the average, and the standard deviation, of the reported significance is less than one for all feature sets. This demonstrates that mismodelling in reference sample is the cause of the excesses at low m_{JJ}. Having isolated the cause of the excess, the next step is to identify the cause of the reference generation failure. In contrast to earlier reference generation failures, such as those with Curtains in Section 9.7, the reference generation failure at low m_{JJ} is not well understood.

Given that both Curtains and Salad produce references that bias the analysis towards reporting an excess, it is assumed that there is some property of the data that changes at low m_{JJ}. One way to understand such properties is to look at the correlations between the features and m_{JJ}. From the results of Figure 10.4 the M feature set already results in significant excesses at low m_{JJ}. Therefore, studying this feature set should provide insight into the cause of the issue. This feature set reduction is important as high dimensional feature sets are inherently difficult to interpret. As already shown in Figure 8.3 the nature of the correlations between the masses of the two jets and m_{JJ} changes in character at low m_{JJ}. This change in character is likely to be the cause of the issue.

The correlation between the masses of the two jets and m_{JJ} is shown in detail in Figure 10.5. At low m_{JJ} the leading jet mass does not reach the maximum cut off that is set at 500 GeV by the high level selections, as described in Section 7.4. Both features vary smoothly with m_{JJ} everywhere, but the nature of the correlations changes at low m_{JJ} where they appear to be non-linear. This behaviour starts at $\sim$ 2300 GeV, which coincides with the analysis breakdown. It is difficult to draw a sharp connection between the observed correlations and the analysis failure. However, the correlations changing in character at the same point as the analysis fails strongly indicates that this is the cause of the issue.



(a) (b) (c) (d)

Figure 10.5: The distribution of the masses of the two jets as a function of m_{JJ} in (a, b) one (down)upsampled validation set and (c, d) the MC validation dataset. Every column in the histogram is normalized to one.

Assuming the cause of the issue is the relationship between the jet masses and m_{JJ} at low m_{JJ} is the cause of the issue, the next step would be to understand the origin of these correlations. The MC samples exhibit the same behaviour as shown in Section 7.4 and therefore, this is not a new physics effect. No detailed investigation of the source of these correlations was performed.

Some signal identification tests as defined in section 9.5 were performed to try and identify the cause of the excesses. The results of these tests were inconclusive, likely because the cause of the excess was not due to signal but rather to reference generation failure. However, this does demonstrate that these tests do not always allow the cause of an excess to be identified. Being able to interpret the results of an analysis is crucial to its success. In this setting, it is particularly important to be able to differentiate between real signal and analysis failure. The inability to identify the cause of the excesses at low m_{JJ} is one of the most significant issues with the analysis and is a topic for future work.

While the exact cause of the issues at low m_{JJ} is unknown, the analysis appears to behave as expected on m_{JJ} SRs centered on 2900 GeV and above from the results on the (down)upsampled validation sets in these regions, and the two other validation data sets. This means seven m_{JJ} bins, out of the eight considered, are considered suitable for unblinding. The bins that succeed in the validation are all contiguous in m_{JJ}. Given our assumption the analysis performance varies smoothly as a function of m_{JJ}, this provides additional confidence the analysis is performing as expected.

10.1.4 Corrections fit

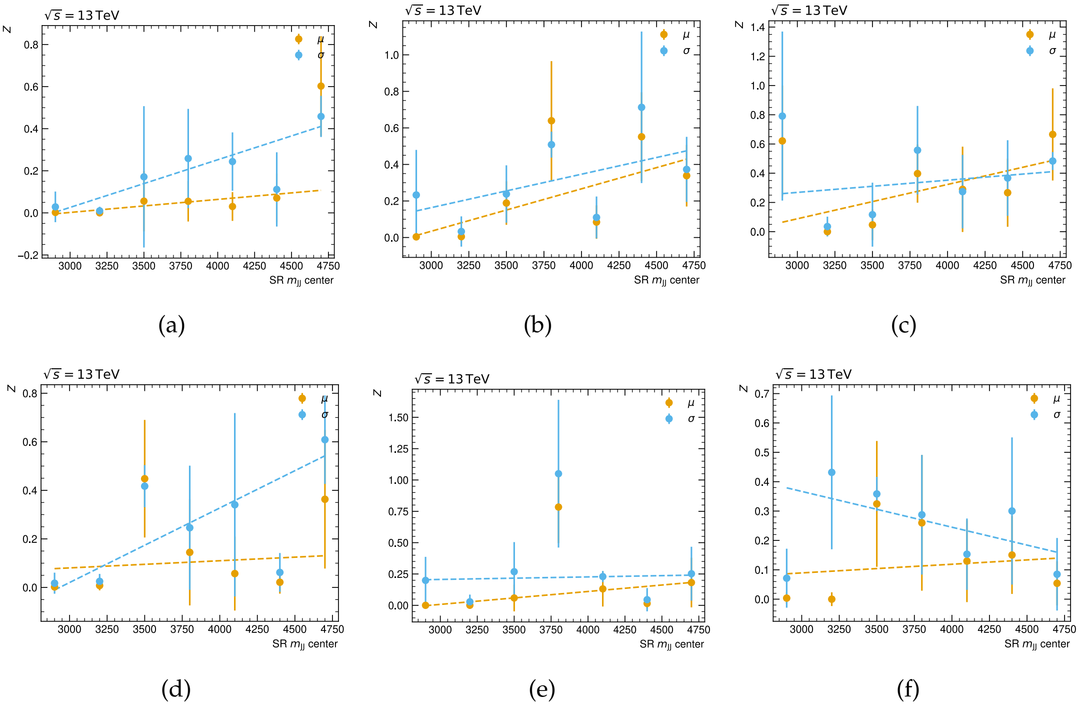
Proceeding with the unblinding of the $|\Delta Y|$ SR data for m_{JJ} SRs centered on 2900 GeV and above, the first step is to fit the corrections to the significances. This is the correction described in section 8.3.6. This correction is applied to the reported significance to account for our uncertainty about the functional form of the fit used in the analysis and as a non-closure correction. The mean and standard deviations of the significances are shown in Figure 10.6, where deficits have been set to zero following the assumption that any new physics appears as an excess, as set out in section 4.2. The standard deviation is always less than one, and so has no impact on the reported significance when used in eq. (8.3). The mean however is always positive and therefore always decreases the reported significance. This correction therefore always impacts the reported significance. All future results are reported with this correction applied.

The linear fit in Figure 10.6 does not model the measured moments well. A linear fit is used as it is the simplest model that can be used to correct the significance. More expressive functions with regularization were not explored. Future analyses that use this method should consider more expressive functions for the correction. Other approaches for estimating the correction could also be considered, like that used in Ref. [114]. Future iterations of this kind of analysis should also consider using different approaches to integrating this correction into the search, rather than just directly applying it to the significance.

10.1.5 Signal injection

This subsection reports the results of the signal injection tests on the (down)upsampled validation sets. This quantifies the expected performance of the analysis on the $|\Delta Y|$ SR data. The signal injection tests are performed on five of the ten (down)upsampled validation sets[2]. For this test, $s = 2, 3\sigma$ of different signals is injected separately into the $2600 - 3200$ GeV SR and the $4100 - 4700$ GeV SR. The signal injection tests follow the same procedure as performed in section 9.4.

[2] The full ten is not considered necessary as a precise estimate of the signal sensitivity for these comparisons is not required.



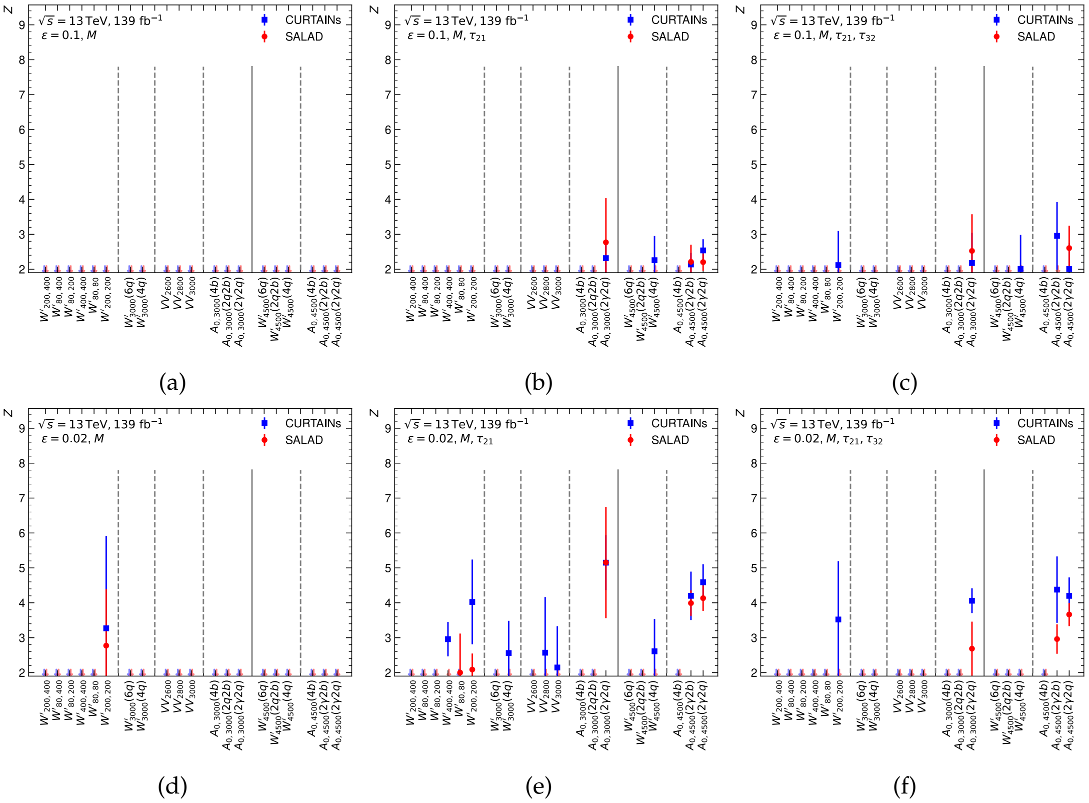
(a) (b) (c)

(d) (e) (f)

Figure 10.6: Mean (μ) and standard deviation (σ) of the significance (Z) distribution for the m_{JJ} signal regions. The averages are estimated across the ten different (down)upsampling validation sets. The correction derived using a linear fit to both the mean and the standard deviation as a function of the central m_{JJ} value of each m_{JJ} signal region is also plotted. The top row (a, b, c) shows the result for the Curtains method and the bottom row (d, e, f) shows the result for the Salad method. The first column (a, d) shows the result for the M feature set, the second column (b, e) shows the result for the M, τ_{21} feature set, and the third column (c, f) shows the result for the M, τ_{21}, τ_{32} feature set.

Both methods are mostly insensitive to the $s = 2\sigma$ signal injection as shown in Figure 10.7, with the analysis returning a significance $Z \geq 2\sigma$ for only a few signal models. This is in sharp contrast to what is observed in the $|\Delta Y|$ SB data where the analysis recovered $Z \geq 2$ for $s = 2\sigma$ signal injections as shown in Figure 9.10. This is a failure of the analysis to enhance the presence of signals in the interesting regime of $s \leq 2\sigma$ in the $|\Delta Y|$ SR data. The most likely cause for this failure is the significant degradation of the performance of the reference generation on $|\Delta Y|$ SR data as shown in Figure 10.1.

At the larger signal injection of $s = 3\sigma$ both methods can be seen to return a significance of $Z > 3$ for most of the injected signals. This is shown in Figure 10.8 and is particularly true in the M, τ_{21} feature set at $\epsilon = 0.02$. This reflects the significance enhancement ability of both Salad and Curtains in the presence of a significant amount of signal. In general Curtains is more sensitive to the signal injections than Salad. This is not a general statement about the methods but is true in this specific case where the methods have been implemented in a specific fashion. The difference in the signal sensitivity of the different methods is concerning. When unblinding, if Curtains reports an excess and Salad does not there is no way to know if this is because Curtains has identified a signal that Salad is not sensitive to, or if the reference produced by Curtains has resulted in a spurious excess. This concern is somewhat alleviated by the fact that both methods have been validated on multiple datasets



(a) (b) (c)

(d) (e) (f)

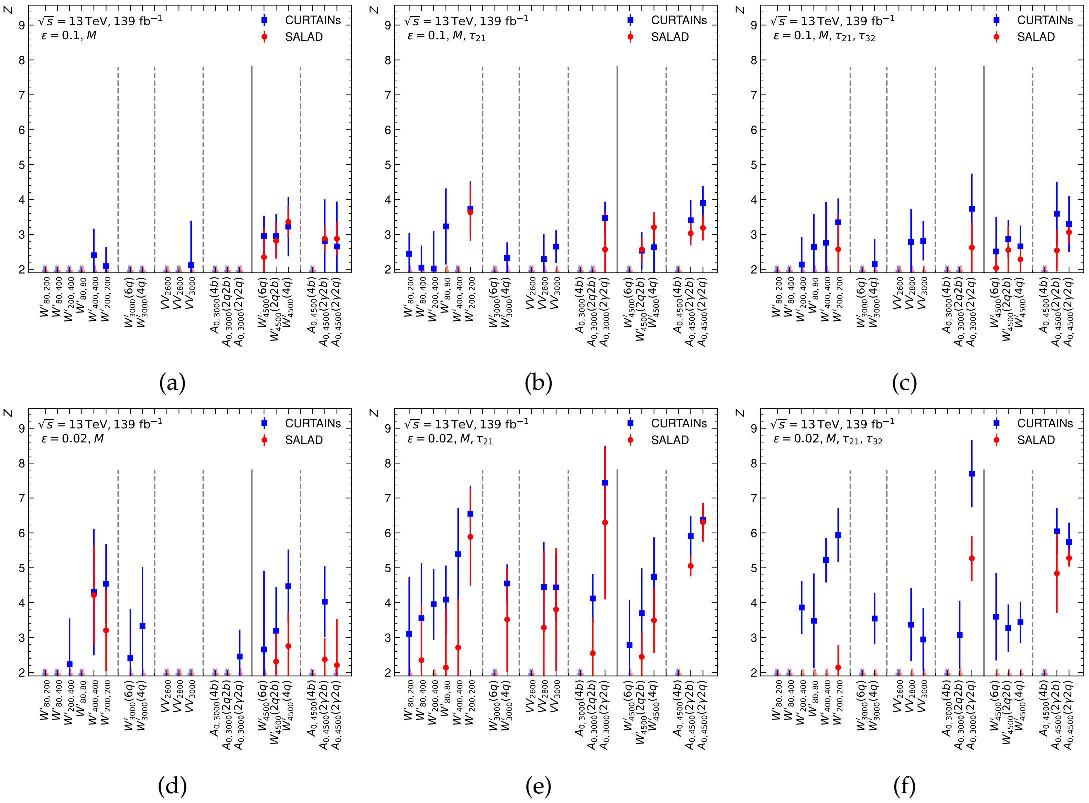
Figure 10.7: Signal injection tests at (a, b, c) $\epsilon = 0.1$ and (d, e, f) $\epsilon = 0.02$. For both Salad and Curtains all feature sets are used (a, d) $\mathcal{X} = M$, (b, e) $\mathcal{X} = M, \tau_{21}$ and (c, f) $\mathcal{X} = M, \tau_{21}, \tau_{32}$. The local significance (Z) reported by the analysis pipeline is calculated after running the analysis with 2σ of signal injected into the data in the m_{JJ} SR centered on the different signals. The errors show the variance in the reported significance as calculated across different signal injections and analysis runs.

and do not produce false excesses in any of the m_{JJ} regions where they are unblinded. This is a concern that should be addressed in future work.

In general, the tighter selection is more sensitive to a wide range of signals at both signal injections and in all feature sets for both methods. This suggests the classifier is better able to enrich signal in the tighter selection, which is expected if it has learned to discriminate signal from background. The M, τ_{21} feature set is the most sensitive to injected signals, with the M, τ_{21}, τ_{32} feature set the second most sensitive and the M feature set the least sensitive.

The failure of the M, τ_{21}, τ_{32} feature set relative to the M, τ_{21} feature set is likely due to the increased complexity of the feature set. This complexity makes the reference generation more difficult, and the classifier less able to identify signal. The additional τ_{32} feature is irrelevant for most of the signal samples considered in this analysis as shown in Figure 8.2. Further, it has been shown that MLP classifiers, like those used in this analysis, are not able to ignore irrelevant features [118]. Therefore, the drop in sensitivity is likely to be due to the classifier being unable to discriminate signal from background. To properly understand the source of the drop in sensitivity a more detailed study using idealized references would be required.

The sensitivity of the analysis to different signal models can be understood through the physics of the signal model and the features used in the analysis. The distribution of the



(a) (b) (c)

(d) (e) (f)

Figure 10.8: Plot in the style of Figure 10.7 but with 3σ of signal injected.

signals overall features was shown in Figure 8.2 and some discussion of the signals was provided in the associated section 8.1.3. The analysis is almost always sensitive to the signal models with photons in the final state $A_{0,3000}(2\gamma 2b)$ and $A_{0,4500}(2\gamma 2b)$. The sensitivity to this signal model is significantly higher on the M, τ_{21} feature set than the M feature set. This is likely due to the additional τ_{21} variable, which has a sharp peak at zero for this signal model due to the photon misidentification as a jet. The analysis is also more sensitive to the $W'(4q)$ signal models when adding τ_{21}, which is expected as this final state results in a distinct jet topology with two prong substructure. The analysis is also more sensitive to this model than the $W'(6q)$ signal models, which is again expected as the $6q$ final state should result in a three prong substructure[3]. In general, the sensitivity of the different feature sets to the signal models is consistent with the feature distributions of the signals. In conclusion, the analysis is most sensitive to signal models that are significantly different from the QCD background.

Another signal specific conclusion to draw is analysis is more sensitive to signals that are better localized in m_{JJ}. This is reflected by the analysis almost always being more sensitive to the $W'_{3000}(4q)$ model than the $W'_{200,400}$ model. These two signal models are identically distributed in all features except for m_{JJ}, where the former signal has a narrower distribution. For the VV_x signal models ($V'_x \to VV \to 4q$), all feature distributions are identical except for m_{JJ}, where the m_{JJ} distributions have different centers. The analysis is never sensitive to the VV_{2600} signal model, which is centered on the edge of the $2600 - 3200$ GeV SR into which it is injected. The analysis has similar sensitivity to the VV_{2800} and VV_{3000} signal models, which is surprising as the latter is exactly centered on the SR. The lack of a clear difference between these two models is possibly due to the lack of a precise test of the sensitivity of the analysis to the signal models.

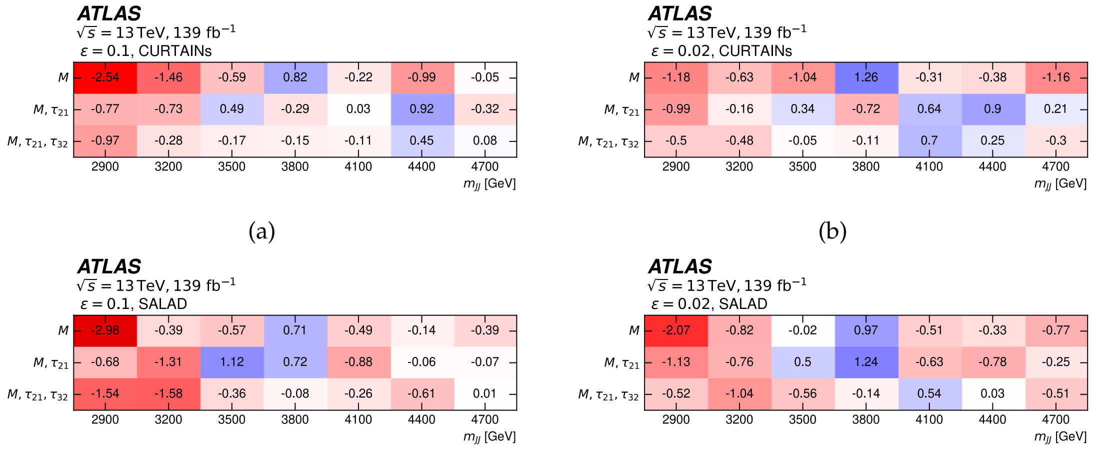
[3] This signal model does not have a significant peak in the τ_{32} distribution, which is expected as higher order τ variables are less sensitive to the substructure.

10.2 Search results

This section presents the results of the analysis on the $|\Delta Y|$ SR data. The analysis was unblinded on the m_{JJ} bins that passed the previous validation. The significance corrected by the linear fit as a function of m_{JJ} is reported, and used in the limit setting procedure.

10.2.1 Search

The largest observed excess has a local significance of $Z = 1.26\sigma$. A significant deficit is seen in the first m_{JJ} SR for both CURTAINS and SALAD when running on the M feature set at $\epsilon = 0.1$ as quantified in Figure 10.9. The nature of this deficit is shown in Figure 10.10 where the entire SR is reduced. Only the first four non-overlapping m_{JJ} SRs are shown in Figure 10.9 as this set contains the region where a deficit is observed. Between neighbouring SRs of Figure 10.10 sharp discontinuities are observed in the m_{JJ} spectrum. These appear because the classifier selection is performed in each SR independently. Therefore, there is no guarantee that adjacent SRs are continuous in m_{JJ}[4]. For all other selections and m_{JJ} SRs both CURTAINS and SALAD behave similarly to the validation region results. A full set of histograms is shown in Appendix C.



(a) (b)

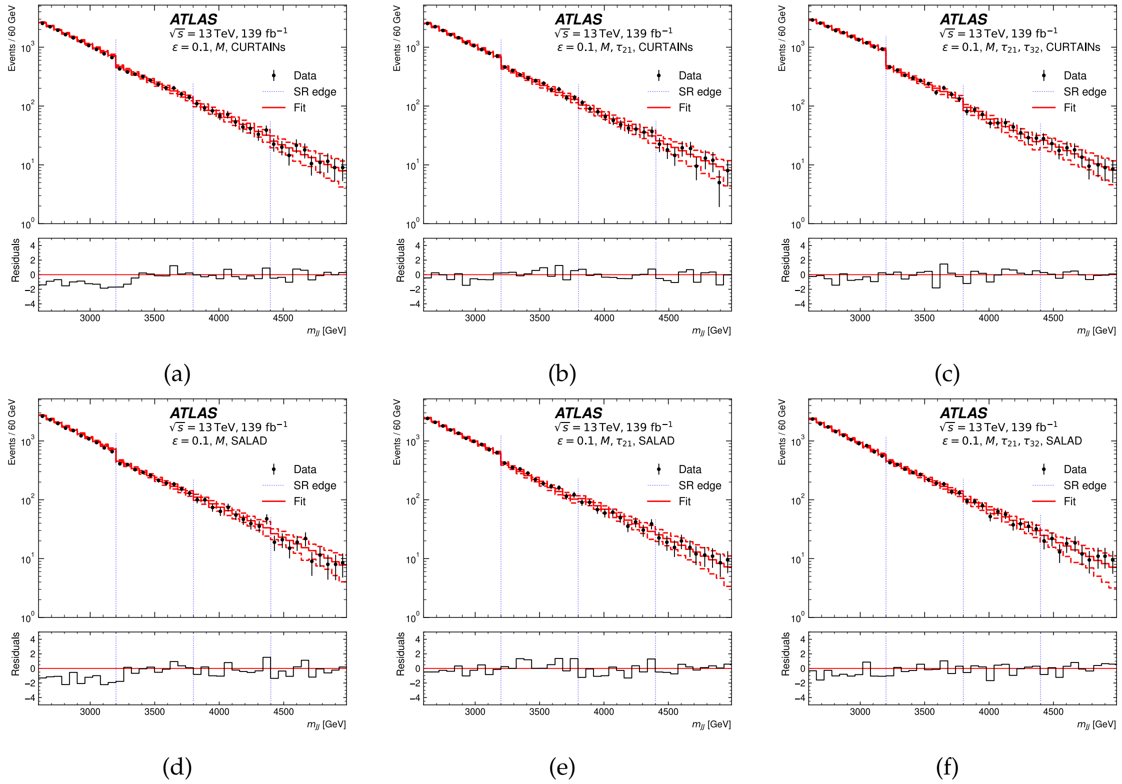
(c) (d)

Figure 10.9: Table of significances for (a, b) CURTAINS and (c, d) SALAD at the two different selections, (a, c) $\epsilon = 0.1$ and (b, d) $\epsilon = 0.02$. Significances are shown for all feature sets and m_{JJ} signal regions.

It is difficult to provide a physical interpretation of the deficit observed in the first m_{JJ} SR. To understand this effect in terms of new physics a signal model that causes a similar deficit would need to be found. However, it is more likely this is related to the effect observed in the (down)upsampled validation sets at low m_{JJ} shown in Figure 10.4. On the validation sets, there are significant deficits at low m_{JJ} that are most likely to appear in the M feature set. In validation, these deficits are more likely in m_{JJ} SRs centered on $2700 - 2800$, while the deficit in the $|\Delta Y|$ SR data is at 2900 GeV. However, it is expected the effect causing the deficit in validation would be enhanced in real data and could extend higher in m_{JJ} than what is observed in validation. This suggests the unblinding strategy should be reconsidered, and possibly SRs should only pass the validation if all neighbouring SRs also pass.

To understand the cause of the deficit in the lowest m_{JJ} bins the approach to understanding the cause of the excesses at low m_{JJ} in the validation sets should be repeated as detailed in

[4] The discontinuities were also observed in the previous ATLAS weakly supervised search [238].



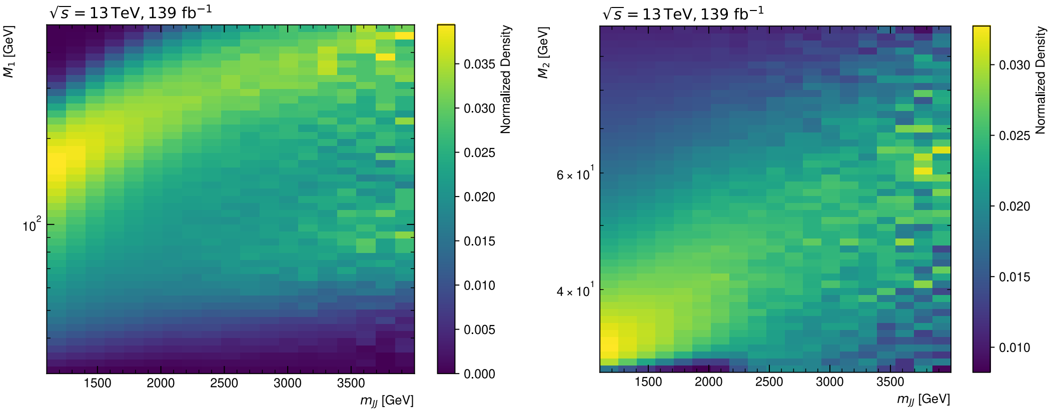
(a) (b) (c)

(d) (e) (f)

Figure 10.10: Histograms of m_{JJ} in the first set of non-overlapping m_{JJ} signal regions on all feature sets at the $\epsilon = 0.1$ classifier selection on one (down)upsampled validation set. Dashed histograms represent the fit uncertainty. The rows show different methods: (a, b, c) CURTAINS and (d, e, f) SALAD. The columns show different feature sets: (a, d) is the result of $\mathcal{T} = M$, (b, e) is the result of $\mathcal{T} = M, \tau_{21}$ and (c, f) is the result of $\mathcal{T} = M, \tau_{21}, \tau_{32}$. The fit is derived from the background-only fit interpolated from the sidebands and the uncertainty on the fit comes from the fit uncertainty and the Poisson statistical uncertainty. The uncertainty on the observed counts is the Poisson uncertainty plus the uncertainty from the mass bin ensembling procedure. The vertical dashed lines mark the edges of each signal region in m_{JJ}. The lower panel in each plot shows the Gaussian-equivalent significance of the deviation between the fit and data.

section 10.1.3. Pursuant to this, the correlations between the leading jet masses and m_{JJ} are shown in Figure 10.11. As in Figure 10.5 the correlations change in character at low m_{JJ}. However, there is unexpected behaviour in the low subleading jet mass where there is a dearth of events in the data. This effect is present in the (down)upsampled validation sets but is more pronounced in the $|\Delta Y|$ SR data. The effect is contained in the lower m_{JJ} SB of the 2600-3200 GeV SR, and so this influences the reference generation methods. It is expected that this kind of non-linear correlation can cause the reference generation to fail and is likely the cause of the deficit in the first m_{JJ} SR. The fact the deficit is suppressed in the M, τ_{21} and M, τ_{21}, τ_{32} feature sets is likely due to the classifier failing to ignore the irrelevant τ features. This is a similar effect to what is observed in signal injection tests, where the analysis is most sensitive in feature sets that fully contain the signal information.

The observed behaviour of the subleading jet mass is most likely not due to a new physics effect, but no investigation into the source of these correlations was performed. The inability to sharply diagnose the cause of the deficit is a significant issue with the analysis. This again



(a)

(b)

Figure 10.11: The distribution of the masses of the two jets as a function of m_{JJ} in the $|\Delta Y|$ SR data. Every column in the histogram is normalized to one.

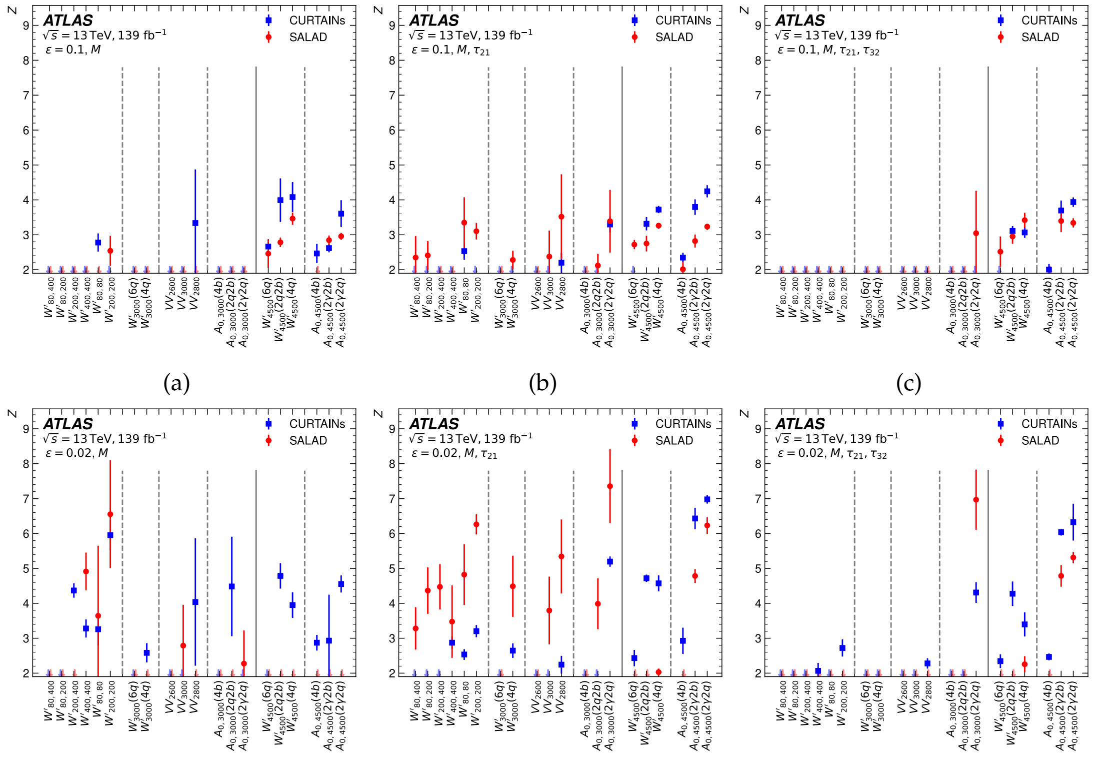
falls under the category of interpreting the results of the analysis, which is emerging as a significant issue. It should be noted that a significance correction as defined in Section 8.3.6 is not applied to negative significances. If this were to be applied, then the significance of the deficit would reduce in magnitude, as there is a $\sim -1\sigma$ deficit observed in the M feature set at $\epsilon = 0.1$ in the validation set as shown in Figure 10.4.

10.2.2 Signal enhancement

The signal enhancement tests were performed on the $|\Delta Y|$ SR data. As the expected sensitivity at $s = 2\sigma$ is small, only the $s = 3\sigma$ signal enhancement tests were performed. The results of these tests are shown in Figure 10.12. These tests consider variations of the signal injection at a fixed level. The full analysis pipeline is run ten times with different signal samples injected for each signal for every feature set and method. The different analysis runs on the same signal model are used to estimate the variance in the reported significance.

In contrast to what was observed during the signal injection tests on the validation sets, the SALAD method is more sensitive to signals than the CURTAINS method in the low m_{JJ} SR. The performance of the CURTAINS approach is seen to drop significantly, while the SALAD approach improves. A simple explanation for this might be that the MC simulation used by SALAD is a better representation of the $|\Delta Y|$ SR data than the (down)upsampled validation sets. This could be compounded by the deficit that was observed in the first m_{JJ} SR, which might reduce the sensitivity of both methods. Another difference in the signal enhancement tests is the M feature set is now more sensitive than the M, τ_{21}, τ_{32} feature set. This is unexpected as the deficit in the M feature set is expected to reduce the signal sensitivity of the analysis.

As SALAD is better than CURTAINS at low m_{JJ}, and the opposite is true at high m_{JJ}, the two methods are complementary. Complementarity is in principle a desirable property, however, as already discussed, in this setting it is undesirable. This issue is exacerbated by the fact the signal injection tests performed during validation CURTAINS was more sensitive to signals than SALAD. Therefore, an observed excess by SALAD that is not observed by CURTAINS would be concerning when unblinding. This is a significant issue that needs to be addressed in future work.



(a) (b) (c)

(d) (e) (f)

Figure 10.12: Signal injection tests at (a, b, c) $\epsilon = 0.1$ and (d, e, f) $\epsilon = 0.02$. For both Salad and Curtains all feature sets are used (a, d) $\mathcal{X} = M$, (b, e) $\mathcal{X} = M, \tau_{21}$ and (c, f) $\mathcal{X} = M, \tau_{21}, \tau_{32}$. The local significance (Z) reported by the analysis pipeline is calculated after running the analysis with 3σ of signal injected into the data in the m_{JJ} signal region centered on the different signals. The errors show the variance in the reported significance as calculated across different signal injections and analysis runs.

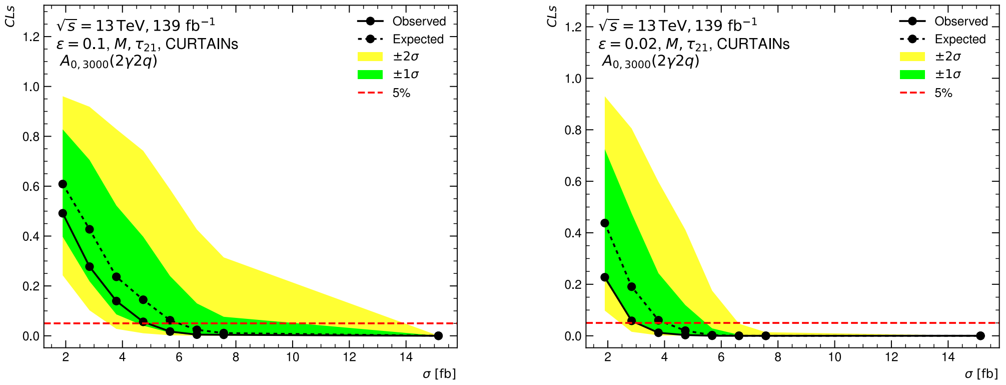
The signal specific conclusions drawn from the signal injection tests on the validation sets are confirmed in the unblinded results. In general, the higher m_{JJ} SR has a similar sensitivity in the validation injection tests and the $|\Delta Y|$ SR injection tests, the same is not true at low m_{JJ}. This suggests that something about the $|\Delta Y|$ SR data is causing the analysis to fail at low m_{JJ}, which is consistent with the deficit observed in the first m_{JJ} SR. The most sensitive SR is still the M, τ_{21} feature set at $\epsilon = 0.02$.

10.2.3 Limit setting

The limit setting procedure is performed on the $|\Delta Y|$ SR data. This procedure was defined in full in section 8.3.7. The following presents the intermediate results used to extract limits as well as the final results of the limit setting.

Limit derivation

Signals were injected using an adaptive procedure where signal is injected as necessary to resolve the relevant quantities. The inputs used to extract the values for setting limits are shown in Figure 10.13. Each point in this plot represents a different signal injection. At each injection, the expected and observed CL_s values are calculated as well as the ±1 and ±2 variations. Linear interpolation between the points for each of these quantities is used to extract the limit. The limit is set by the intersection with $CL_s = 0.05$. This procedure is repeated for all feature sets, selections, methods and signal samples. At each signal injection level, the entire analysis pipeline is run to calculate the CL_s values. All signal samples are used to estimate the classifier selection efficiency.



(a)

(b)

Figure 10.13: The extracted CL_s values for the $A_{0,3000}(2\gamma 2q)$ signal model at different signal injection levels. At every injection the expected and observed CL_s values are calculated and the ± 1 and ± 2 variations for (a) the $\epsilon = 0.1$ and (b) $\epsilon = 0.02$ selections. The CL_s values are calculated using the CURTAINS method with the $\mathcal{X} = M, \tau_{21}$ feature set. The line at 5% represents the $CL_s =$ 0.05 threshold and the intersection with this line is used to set the limit.

The curve at $\epsilon = 0.02$ is sharper than that at $\epsilon = 0.1$ in Figure 10.13. This is consistent with the large difference in the expected sensitivity between the two selections for the signal enhancement tests of the previous section, shown in Figure 10.12. It is interesting to note that while the shape of the curves are different, the crossing point for the observed CL_s is similar. This suggests that at small signal injections, the differences in sensitivity are less pronounced. The steepness of the curves shows the sensitivity of the analysis increases faster at $\epsilon = 0.02$ than at $\epsilon = 0.1$.

The points needed for finding the crossing point at $\epsilon = 0.02$ are not needed at $\epsilon = 0.1$. This is particularly clear at low signal injections. If limits were to be derived for only one selection then fewer signal injections would be needed. Running the full analysis pipeline for each signal injection level is computationally expensive. This restricts the limit setting capacity of the analysis and makes it less useful for setting limits on many signal models. In particular, this makes it difficult to reinterpret the results of the analysis on new signal models after the analysis is published [279].

Method comparison

The limits set by CURTAINS and SALAD are compared in Figure 10.14. This result is shown for $\epsilon = 0.02$ with the M, τ_{21} feature set. Both CURTAINS and SALAD set similar limits on the signal cross section at both ends of the m_{JJ} spectrum. At low m_{JJ} SALAD sets slightly stricter limits, while at high m_{JJ} CURTAINS sets slightly stricter limits. This is consistent with the expected sensitivity of the two methods as shown in Figure 10.12. The same is true for all selections, feature sets and signals as shown in Appendix D. The observed limits are always less than the expected limits at low m_{JJ}. This is because there is a deficit in the first m_{JJ} SR in all feature sets for the signal injection at which the limits are set.

Comparisons are made to the ATLAS dijet search [114] and the ATLAS diboson search [245], both of which were detailed in Section 7.1. As expected, the diboson search sets much stricter limits on the $W'_{80,80}$ signal model than both the dijet search and this analysis. This search however has extremely limited sensitivity to the other signal models considered in this analysis, due to the selections it applies to data. The dijet search sets stricter limits on signal models that produce decays that are on average contained within two small radius jets ($R > 0.4$), while this analysis appears to set stricter limits on signal models that are not. This reflects another failure of the analysis presented in this thesis. If the Atlas dijet strategy were to be reoptimized for large radius jets it is likely the resulting limits would be stricter than those set by this analysis. Further, if the analysis presented here fully leveraged the information in the M, τ_{21} feature set then it should set stricter limits on signal models that are contained within small radius jets. A concrete example of this is provided later in this chapter. It is also important to remember the limits reported by this search do not account for systematic uncertainties in the signal models, and therefore the limit comparisons are not entirely fair.

In the high m_{JJ} SR, the analysis sets stricter limits on the signal cross section than at low m_{JJ}. Tighter limits at high m_{JJ} are consistent with the previous weakly supervised search in ATLAS [238]. In the previous analysis, it was shown the two comparison searches [114, 245] set stricter limits at high m_{JJ}. Without a benchmark analysis in this region, it is difficult to interpret the relative performance of the analysis. The two comparison searches could not be reinterpreted using the signals simulated for this analysis due to technical issues. As is discussed in the next subsection, the M feature set can be used as a proxy for the sensitivity of the previous weakly supervised search [238].

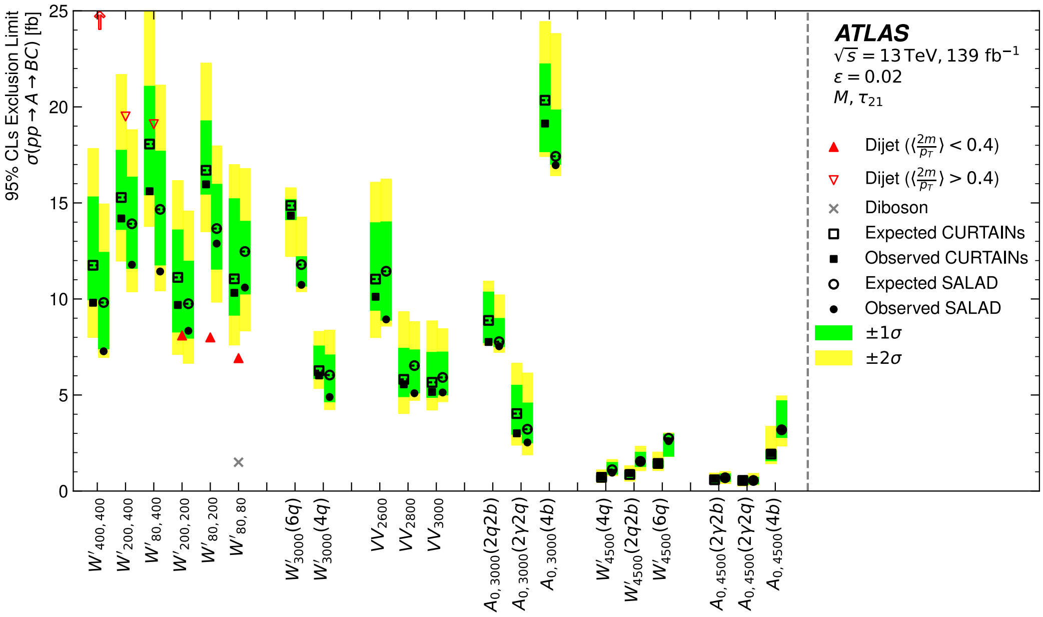


Figure 10.14: Comparison of the limits set by Curtains and Salad at $\epsilon = 0.02$ with $\mathcal{T} = M, \tau_{21}$. The one and two sigma variations on the expected limits for both Salad and Curtains are shown as the shaded regions. The signal models are described in detail in section 7.5.2. The observed limits from the ATLAS dijet search [114] and the ATLAS all-hadronic diboson search [245] as derived in the previous weakly supervised ATLAS search [238]. Limits for the inclusive dijet search are calculated using the W' signals from this paper and the analysis of Ref. [114]; the diboson search limits are computed using the Heavy Vector Triplet [36] W' signal from Ref. [245]. The acceptance for the W' in this paper, compared to the W' acceptance in Ref. [245], is 86%. Missing dijets limits are shown with red arrows.

Comparison of feature sets

The M, τ_{21} feature set results in the strictest limits on the signal cross section across almost all signals at $\epsilon = 0.02$ as shown in Figure 10.15. This is true for both SALAD and CURTAINS. In the $2600 - 3200$ GeV SR at $\epsilon = 0.1$, the M feature set results in the strictest limits on the signal cross section across almost all signals due to the deficit observed in Figure 10.10, which only appears in this region for this feature set. Otherwise, the M, τ_{21} feature set most often sets the strictest limits in all regions and selections for both methods as shown in Appendix D.

This analysis uses the full CL_s prescription to derive limits, whereas the previous weakly supervised dijet search used a different approach [238]. Therefore, direct comparisons can not be made between the two analyses. However, the previous search used the M feature set, which can therefore serve as a proxy for the sensitivity of the previous analysis. Given the M, τ_{21} feature set generally improves the sensitivity of the analysis this iteration of the analysis represents an improvement over the previous search. This is reinforced by the signal injection plots of Figure 10.12 where the M, τ_{21} feature set is more sensitive to signals than the M feature set.

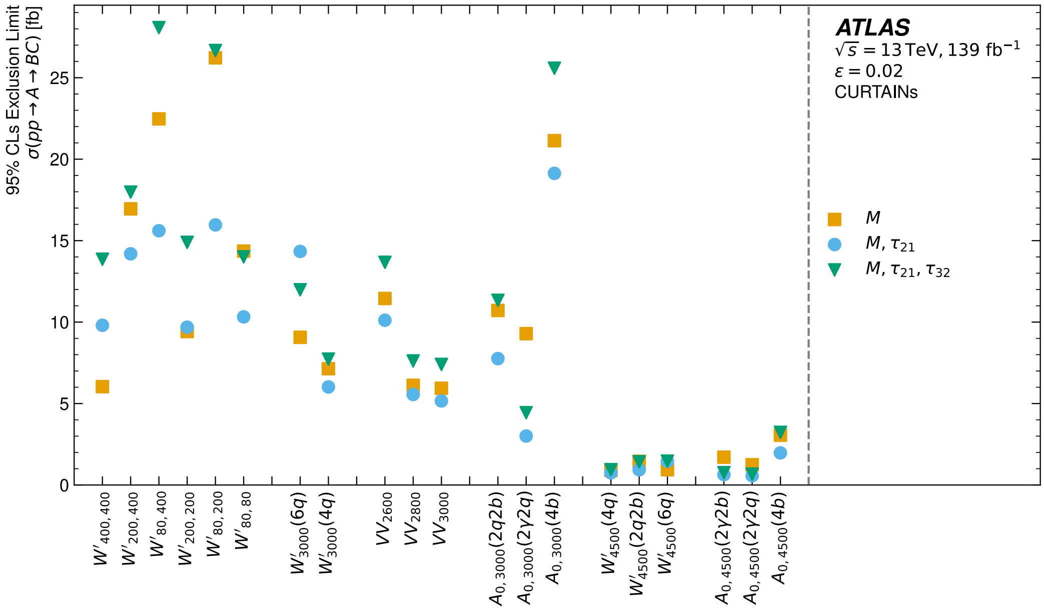


Figure 10.15: Comparison of the observed limits set with the different feature sets. All limits use the CURTAINS method with the $\epsilon = 0.02$ selection. The signal models are described in detail in section 7.5.2.

Comparison of classifier selections

The limits set at $\epsilon = 0.1$ and $\epsilon = 0.02$ are compared in Figure 10.16. Generally, the limits set by the two selections are similar. This is not consistent with what might be expected from the signal enhancement tests in Figure 10.12. However, as shown in Figure 10.13 the improved signal enhancement at $\epsilon = 0.02$ results in a steeper curve in the CL_s vs signal injection plot. This translates to a smaller spread in the ± 1 and ± 2 variations on the expected limits at $\epsilon = 0.02$. This is an issue because the sensitivity of the analysis is most important at small cross sections as new physics is expected to be rare. The similarity of the limits set by the two selections suggests the signal enhancement of $\epsilon = 0.02$ degrades at low signal injection levels. This reflects the need to have signal injection tests as a function of injected cross section rather than at a fixed signal injection.

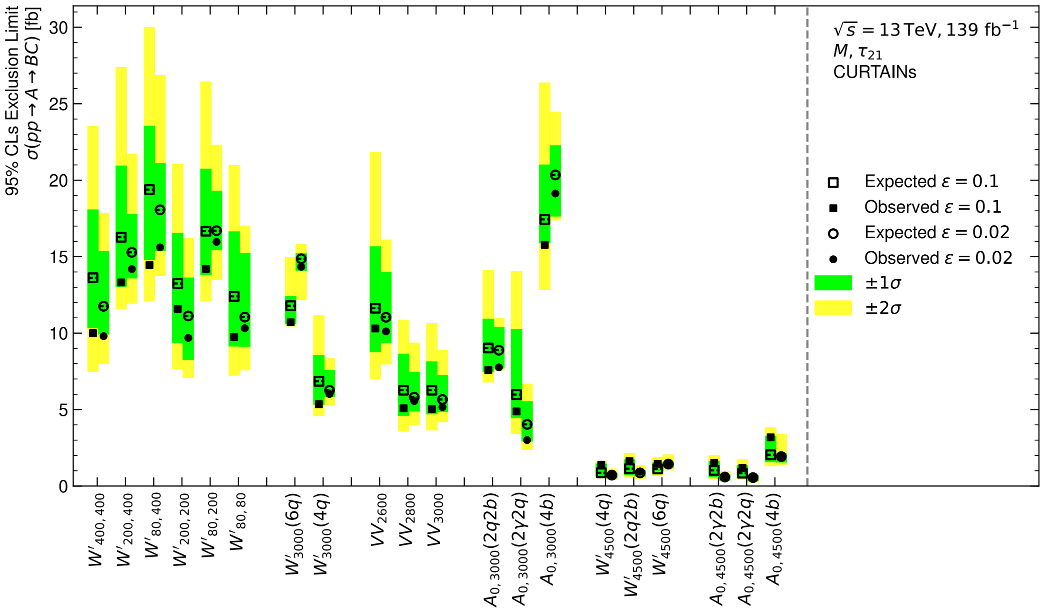


Figure 10.16: Comparison of the limits set by CURTAINS and SALAD at $\epsilon = 0.02$ with $\mathcal{T} = M, \tau_{21}$. The same style as fig. 10.14 is used.

Best possible limit

The limits presented so far are based on the approach described in section 10.2.3. However, every time the analysis is run it is possible to derive a valid limit for any signal model. When no signal is injected the analysis is expected to set weak limits. As the amount of injected signal increases the analysis becomes closer to a supervised search so the analysis should set tighter limits. Therefore, the best possible limit set by the analysis differs from the previously reported limits.

The best possible limit is derived by injecting signal at cross section σ and running the full analysis pipeline. An upper limit on the signal cross section is then set using the CL_s prescription which sets an upper bound on the parameter μ in the likelihood fit. This parameter is a scale that multiplies the signal cross section, it is set to one when all signal is injected. The upper bound on the parameter μ can be used to set an upper bound on the signal cross section.

To derive the best possible limit the signal injections of the adaptive grid are used. At every signal injection, a limit is calculated and the best limit across all points in the signal injection grid is reported as the best possible limit. The best possible limit, compared to the limit reported by the analysis is shown in Figure 10.17. This limit is seen to be significantly stricter than the limit reported by the analysis. For some signal models, there is a significant shift in the limit, while for others the shift is small. The full set of these limits across all features and selections is shown in Appendix D. The best possible limit is better than the atlas dijet search [114] for all signal models considered in this analysis. This demonstrates the capacity of the analysis to be more sensitive to signals than the dijet search when the information in M, τ_{21} is properly leveraged.

When testing analyses of this type this limit is interesting because it shows the performance the analysis could achieve. These limits also allow signal models to be excluded at lower cross sections. The limits reported using the approach defined in section 10.2.3 are a representation of how low in signal cross section the analysis can go and are useful for making fair comparisons to other searches. The best possible limit is not a fair comparison in these contexts as this limit represents something closer to a supervised search. However, both of these limits are valid and should be reported.

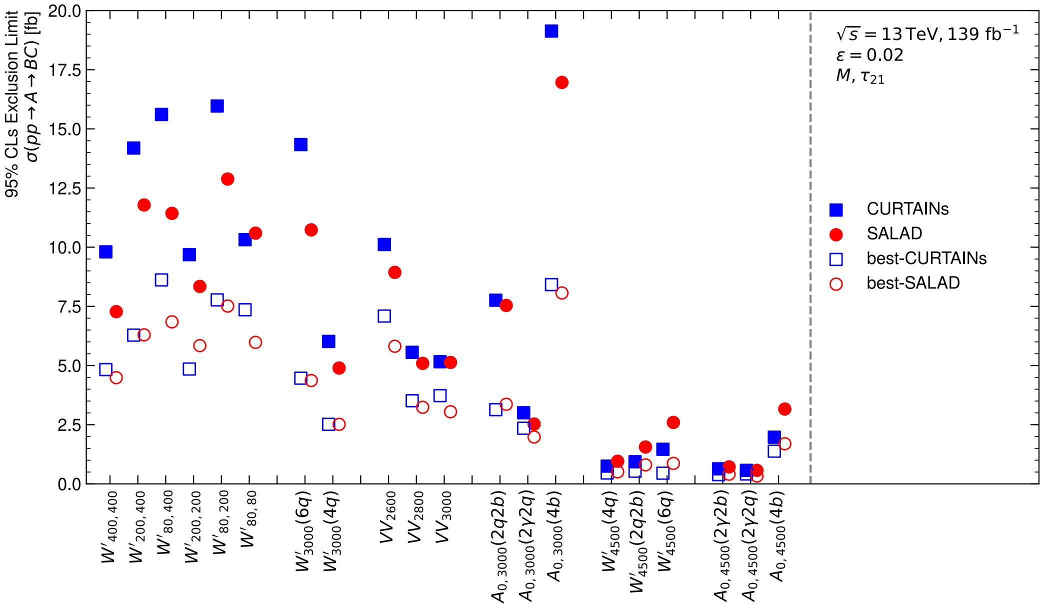


Figure 10.17: Comparison of the limits set by CURTAINS and SALAD, at $\epsilon = 0.02$ with $\mathcal{T} = M, \tau_{21}$, using the best possible derived limit and the limit the analysis reported.

Chapter 11

Discussion

The search presented in this thesis attempted to develop an analysis with broad sensitivity to possible new physics processes. This was done using substructure information to extend a bump hunt in the dijet invariant mass spectrum and search for narrow width resonances. A bump hunt is already well established as a useful data analysis tool, particularly for exploring datasets at new energy scales [47, 114, 215–220]. The results presented in this thesis make it clear that extending bump hunts by including additional information is possible but faces many challenges. In the following, the key summary points are discussed based on what was learned about nature, the method, and some outlook for the future.

11.1 Inference about nature

No significant excess was observed in the data, and the analysis was used to set limits on the production of new physics processes. Many of these signal processes have not been studied by any other analysis, and the reported limits appear to be tighter than those set by the other relevant searches. A significant deficit was observed in the 2600-3200 GeV mjj signal region, but this is assumed to be attributed to a failure to model the background correctly.

11.2 Method

In developing the extended bump hunt that was presented here, several challenges were encountered. The method can be seen to fail in some areas, but there were also some successes. These are discussed in the following sections.

11.2.1 Challenges

The analysis presented here faced many difficulties, and in particular, there were four main fundamental areas where future work is needed.

First, it was shown the CWoLa classifier can introduce a bias into the m_{JJ} spectrum that results in the analysis reporting a false excess. In this analysis, this was resolved by training an ensemble of classifiers, but this solution significantly increased the computational cost of the analysis and may not be suitable for all analyses. The source of this failure is likely due to the random placement of decision boundaries, the correlation of the features on which the classifier is trained with the dijet invariant mass, and the smoothly falling nature of the m_{JJ} spectrum. Future work should focus on developing methods to better explore and mitigate this bias.

Second, it is unclear how to robustly validate the analysis. During validation, it was observed the analysis passed the validation tests in all the proposed signal regions on the $|\Delta Y|$

sideband dataset and the simulated background dataset. The same was not true for the novel validation set generated from the $|\Delta Y|$ signal region data, where the analysis was seen to fail at small m_{JJ} values. Further, after unblinding the analysis reported a significant deficit in the 2600-3200 GeV signal region, which was much more significant than what was observed in the validation datasets. This deficit was almost captured by the novel validation strategy, and it is possible further investigations of similar approaches may allow for a more robust validation of the analysis. An alternative approach would be to follow a progressive unblinding procedure and to set stricter criteria for succeeding in the validation phase.

Third, it is unclear how to account for systematic bias in the significance reported by the analysis. In this analysis, this was done by directly correcting the reported significance, but this should be included in the profile likelihood fit in future work. Methods for accounting and measuring such biases would significantly increase the robustness of the analysis.

Fourth, the analysis was not able to characterize the properties of all excesses that were observed. The analysis failure on the validation dataset constructed from the $|\Delta Y|$ signal region data is attributed to the challenging correlations between the features used in the analysis and m_{JJ}. However, it was not identified what exact property of the data caused the analysis to fail. The same statement is true for the deficit observed after unblinding. Future work should focus on developing methods to concretely identify the cause of any excesses reported by the analysis.

11.2.2 Successes

While the analysis faced many challenges, there were also some successes, which are discussed in this section. This analysis identified idealized constructions as a useful tool for developing the analysis strategy and diagnosing failures. Future iterations should leverage these approaches, and also explore more sophisticated methods for constructing idealized datasets. Also, the use of a workflow language [268] for the analysis development was seen to be an indispensable tool for managing the complexity of the analysis. Another success of the analysis was the development of a novel validation strategy that was seen to be a more stringent test of the analysis than standard validation strategies. This approach is still in its infancy and could be further developed to provide a more robust validation. The analysis was also sensitive to small levels of signal in the $|\Delta Y|$ sideband validation dataset, which demonstrates these kinds of approaches can be successful on real measured data. This analysis also clearly identified multiple failures and subtle issues that arise when developing a weakly supervised search of this type. These lessons are valuable for the ongoing work in this area. There is significant interest in searches of this type, and this analysis serves as a useful proof of concept for the development of future searches.

11.3 Outlook

The analysis presented in this thesis is the first of its kind at the ATLAS experiment, and it is clear there is much work to be done in this area to develop a robust procedure. In the future, analyses of this type may provide useful extensions to the existing bump hunt paradigm. Such searches may be useful in ensuring there is no new physics produced at significant cross sections in the data that has been missed by existing searches. What is unclear is how much utility this will bring to a given physics program, and whether these approaches justify the significant resource investment they currently claim.

A weakly supervised search like the one presented here will only ever be sensitive to a subset of possible new physics processes. This sensitivity is always partially random, as it is dictated by the choice of those same features and the ability of the classifier to learn the signal. The choice of features dictates the space over which the signal model can be differentiated from the background, and how interpolatable the background is from the sidebands, and impacts the ability of the classifier to learn the signal. The choice of classifier and its hyperparameters dictates how well the signal can actually be learned if present in data. Such randomness may be deemed acceptable in the context of a broad search, but this begs the question of whether something more systematic could be done to increase the sensitivity of the search.

One clear benefit of developing these searches is that in interpreting them there is the option to set useful limits by reporting the best possible limit. However, it is worth noting that when setting such limits the analysis is essentially fit to a specific signal model. This means that something closer to a dedicated search is performed when setting limits, with the analysis tailored towards the injected simulated signal. If this kind of program were to be pursued, it would likely make more sense to simply perform a dedicated search for each signal model that is being reinterpreted. Such a strategy would be more sensitive than any weakly supervised search and would allow for strict limits to be set on a wide range of signal models.

Bibliography

[1] Abbott, B. P. et al. *Observation of Gravitational Waves from a Binary Black Hole Merger*. In: *Phys. Rev. Lett.* 6 (2016).

[2] Gross, E. P. *Structure of a quantized vortex in boson systems*. In: *Nuovo Cimento* (1961).

[3] Glashow, S. L. *Partial-symmetries of weak interactions*. In: *Nuclear Physics* 4 (1961).

[4] Weinberg, S. *A Model of Leptons*. In: *Phys. Rev. Lett.* (21 1967).

[5] Salam, A. *Weak and Electromagnetic Interactions*. In: *Conf. Proc. C* (1968).

[6] Glashow, S. L., Iliopoulos, J., and Maiani, L. *Weak Interactions with Lepton-Hadron Symmetry*. In: *Phys. Rev. D* (1970).

[7] 't Hooft, G. *Renormalization of Massless Yang-Mills Fields*. In: *Nucl. Phys. B* (1971).

[8] 't Hooft, G. *Renormalizable Lagrangians for Massive Yang-Mills Fields*. In: *Nucl. Phys. B* (1971). Ed. by J. C. Taylor.

[9] Georgi, H. and Glashow, S. L. *Unified Weak and Electromagnetic Interactions without Neutral Currents*. In: *Phys. Rev. Lett.* (22 1972).

[10] 't Hooft, G. and Veltman, M. J. G. *Regularization and Renormalization of Gauge Fields*. In: *Nucl. Phys. B* (1972).

[11] 't Hooft, G. and Veltman, M. J. G. *Combinatorics of gauge fields*. In: *Nucl. Phys. B* (1972).

[12] Gross, D. J. and Wilczek, F. *Ultraviolet Behavior of Non-Abelian Gauge Theories*. In: *Phys. Rev. Lett.* (26 1973).

[13] Politzer, H. D. *Reliable Perturbative Results for Strong Interactions?* In: *Phys. Rev. Lett.* (26 1973).

[14] Planck Collaboration et al. *Planck 2018 results. VI. Cosmological parameters*. In: *Astron. Astrophys.* (2020).

[15] Fukuda, Y. et al. *Evidence for oscillation of atmospheric neutrinos*. In: *Phys. Rev. Lett.* (1998).

[16] ATLAS Collaboration. *Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC*. In: *Phys. Lett. B* (2012).

[17] CMS Collaboration. *Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC*. In: *Phys. Lett. B* (2012).

[18] Purcell, A. *Go on a particle quest at the first CERN webfest. Le premier webfest du CERN se lance à la conquête des particules*. 2012.

[19] Lindon, J. *Particle Collider Probes of Dark Energy, Dark Matter and Generic Beyond Standard Model Signatures in Events With an Energetic Jet and Large Missing Transverse Momentum Using the ATLAS Detector at the LHC*. Birmingham U., 2020.

[20] Anderson, P. W. *Plasmons, Gauge Invariance, and Mass*. In: *Phys. Rev.* (1 1963).

[21] Higgs, P. W. *Broken symmetries, massless particles and gauge fields*. In: *Phys. Lett.* (1964).

[22] Englert, F. and Brout, R. *Broken Symmetry and the Mass of Gauge Vector Mesons*. In: *Phys. Rev. Lett.* (1964). Ed. by J. C. Taylor.

[23] Guralnik, G. S., Hagen, C. R., and Kibble, T. W. B. *Global Conservation Laws and Massless Particles*. In: *Phys. Rev. Lett.* (1964). Ed. by J. C. Taylor.

[24] Fritzsch, H., Gell-Mann, M., and Leutwyler, H. *Advantages of the Color Octet Gluon Picture*. In: *Phys. Lett. B* (1973).

[25] Gross, D. J. and Wilczek, F. *Asymptotically Free Gauge Theories - I*. In: *Phys. Rev. D* (1973).

[26] Weinberg, S. *Nonabelian Gauge Theories of the Strong Interactions*. In: *Phys. Rev. Lett.* (1973).

[27] Sjöstrand, T. *Jet fragmentation of multiparton configurations in a string framework*. In: *Nucl. Phys. B* (1984).

[28] Feynman, R. P. *The behavior of hadron collisions at extreme energies*. In: *Conf. Proc. C* (1969).

[29] Martin, A. D. et al. *Parton distributions for the LHC*. In: *Eur. Phys. J. C* (2009).

[30] Gribov, V. N. and Lipatov, L. N. *Deep inelastic e p scattering in perturbation theory*. In: *Sov. J. Nucl. Phys.* (1972).

[31] Dokshitzer, Y. L. *Calculation of the Structure Functions for Deep Inelastic Scattering and e+ e- Annihilation by Perturbation Theory in Quantum Chromodynamics.* In: *Sov. Phys. JETP* (1977).

[32] Altarelli, G. and Parisi, G. *Asymptotic freedom in parton language*. In: *Nucl. Phys. B* 2 (1977).

[33] Ball, R. D. et al. *Parton distributions for the LHC run II*. In: *JHEP* (2015).

[34] ATLAS Collaboration. *Standard Model Summary Plots June 2024*. 2024.

[35] Panico, G. and Wulzer, A. *The Composite Nambu-Goldstone Higgs*. Springer International Publishing, 2016.

[36] Pappadopulo, D. et al. *Heavy Vector Triplets: Bridging Theory and Data*. In: *JHEP* (2014).

[37] ATLAS Collaboration. *The ATLAS Experiment at the CERN Large Hadron Collider*. In: *JINST* (2008).

[38] Evans, L. and Bryant, P. *LHC Machine*. In: *JINST* (2008).

[39] ATLAS Collaboration. *Luminosity determination in pp collisions at $\sqrt{s} = 8$ TeV using the ATLAS detector at the LHC*. In: *Eur. Phys. J. C* (2016).

[40] Avoni, G. et al. *The new LUCID-2 detector for luminosity measurement and monitoring in ATLAS*. In: *JINST* 07 (2018).

[41] Lopienska, E. *The CERN accelerator complex, layout in 2022. Complexe des accélérateurs du CERN en janvier 2022*. 2022.

[42] Collaboration, A. *Luminosity Public Results Run 2*.

[43] CMS Collaboration. *The CMS Experiment at the CERN LHC*. In: *JINST* (2008).

[44] ATLAS Collaboration. *ATLAS Insertable B-Layer: Technical Design Report*. 2010. Addendum: ATLAS-TDR-19-ADD-1; CERN-LHCC-2012-009. 2012.

[45] Abbott, B. et al. *Production and integration of the ATLAS Insertable B-Layer*. In: *JINST* (2018).

[46] ATLAS Collaboration. *Performance of the ATLAS trigger system in 2015*. In: *Eur. Phys. J. C* (2017).

[47] ATLAS Collaboration. *Search for Low-Mass Dijet Resonances Using Trigger-Level Jets with the ATLAS Detector in pp Collisions at $\sqrt{s} = 13$ TeV*. In: *Phys. Rev. Lett.* (2018).

[48] ATLAS Collaboration. *Software and computing for Run 3 of the ATLAS experiment at the LHC*. 2024.

[49] Cacciari, M., Salam, G. P., and Soyez, G. *The anti-k_t jet clustering algorithm*. In: *JHEP* (2008).

[50] Krohn, D., Thaler, J., and Wang, L.-T. *Jet Trimming*. In: *JHEP* (2010).

[51] ATLAS Collaboration. *Performance of jet substructure techniques for large-R jets in proton–proton collisions at $\sqrt{s} = 7$ TeV using the ATLAS detector*. In: *JHEP* (2013).

[52] Stewart, I. W., Tackmann, F. J., and Waalewijn, W. J. *N Jettiness: An Inclusive Event Shape to Veto Jets*. In: *Phys. Rev. Lett.* (9 2010).

[53] Thaler, J. and Van Tilburg, K. *Identifying Boosted Objects with N-subjettiness*. In: *JHEP* (2011).

[54] Collins, J. C., Soper, D. E., and Sterman, G. *FACTORIZATION OF HARD PROCESSES IN QCD*. In: *Perturbative QCD*.

[55] GEANT4 Collaboration, Agostinelli, S., et al. *Geant4 – a simulation toolkit*. In: *Nucl. Instrum. Meth. A* (2003).

[56] Cranmer, K. *Practical Statistics for the LHC*. 2015.

[57] Lyons, L. *Discovering the Significance of 5 sigma*. 2013.

[58] Cowan, G. et al. *Asymptotic formulae for likelihood-based tests of new physics*. In: *Eur. Phys. J. C* 2 (2011).

[59] Collins, J. H., Howe, K., and Nachman, B. *Extending the search for new resonances with machine learning*. In: *Phys. Rev.* 1 (2019).

[60] Gross, E. and Vitells, O. *Trial factors for the look elsewhere effect in high energy physics*. In: *Eur. Phys. J. C* 1–2 (2010).

[61] Gallicchio, J. et al. *Multivariate discrimination and the Higgs+W/Z search*. In: *JHEP* 4 (2011).

[62] Read, A. L. *Presentation of search results: the CL_S technique*. In: *J. Phys. G* (2002).

[63] Heinrich, L. et al. *pyhf: pure-Python implementation of HistFactory statistical models*. In: *Journal of Open Source Software* 58 (2021).

[64] Heinrich, L., Feickert, M., and Stark, G. *pyhf: v0.7.6*. Version 0.7.6.

[65] Radovic, A. et al. *Machine learning at the energy and intensity frontiers of particle physics*. In: *Nature* 7716 (2018).

[66] Carleo, G. et al. *Machine learning and the physical sciences*. In: *Reviews of Modern Physics* 4 (2019).

[67] Hey, T. et al. *Machine learning and big scientific data*. eng. In: *Philos. Trans. A Math Phys. Eng. Sci.* 2166 (2020).

[68] Jumper, J. et al. *Highly accurate protein structure prediction with AlphaFold*. In: *Nature* 7873 (2021).

[69] OpenAI. *ChatGPT (GPT-4)*. 2023.

[70] Podell, D. et al. *SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis*. 2023.

[71] Zdeborová, L. *Understanding deep learning is also a job for physicists*. In: *Nature Physics* 6 (2020).

[72] Baydin, A. et al. *Automatic differentiation in machine learning: A survey*. In: *Journal of Machine Learning Research* (2018).

[73] Hornik, K., Stinchcombe, M., and White, H. *Multilayer feedforward networks are universal approximators*. In: *Neural Networks* 5 (1989).

[74] Maiorov, V. and Pinkus, A. *Lower bounds for approximation by MLP neural networks*. In: *Neurocomputing* 1 (1999).

[75] Wilson, A. G. and Izmailov, P. *Bayesian Deep Learning and a Probabilistic Perspective of Generalization*. 2022.

[76] Nakkiran, P. et al. *Deep Double Descent: Where Bigger Models and More Data Hurt*. 2019.

[77] Mei, S., Montanari, A., and Nguyen, P.-M. *A mean field view of the landscape of two-layer neural networks*. In: *Proceedings of the National Academy of Sciences* 33 (2018).

[78] Gu, Y. et al. *How to Characterize The Landscape of Overparameterized Convolutional Neural Networks*. In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Curran Associates, Inc., 2020.

[79] Zhang, C. et al. *Understanding deep learning (still) requires rethinking generalization*. In: *Commun. ACM* 3 (2021).

[80] Beery, S., Horn, G. van, and Perona, P. *Recognition in Terra Incognita*. 2018.

[81] Geirhos, R. et al. *Shortcut learning in deep neural networks*. In: *Nature Machine Intelligence* 11 (2020).

[82] Zhang, X. et al. *NICO++: Towards Better Benchmarking for Domain Generalization*. 2022.

[83] Arjovsky, M. et al. *Invariant Risk Minimization*. 2020.

[84] Rezende, D. J. et al. *Equivariant Hamiltonian Flows*. 2019.

[85] Bogatskiy, A. et al. *Explainable Equivariant Neural Networks for Particle Physics: PELICAN*. 2024.

[86] Radford, A. et al. *Language Models are Unsupervised Multitask Learners*. 2019.

[87] Brown, T. B. et al. *Language Models are Few-Shot Learners*. 2020.

[88] Collaboration, A. *Flavour tagging with graph neural networks with the ATLAS detector*. Geneva: CERN, 2023.

[89] ATLAS Collaboration. *Active Learning reinterpretation of an ATLAS Dark Matter search constraining a model of a dark Higgs boson decaying to two b-quarks*. Geneva: CERN, 2022.

[90] Cranmer, K. et al. *Advances in machine-learning-based sampling motivated by lattice quantum chromodynamics*. In: *Nature Reviews Physics* 9 (2023).

[91] Blum, A. L. and Rivest, R. L. *Training a 3-node neural network is NP-complete*. In: *Neural Networks* 1 (1992).

[92] Gage, P. *A new algorithm for data compression*. In: *The C Users Journal archive* (1994).

[93] Sennrich, R., Haddow, B., and Birch, A. *Neural Machine Translation of Rare Words with Subword Units*. 2016.

[94] Narkhede, M. V., Bartakke, P. P., and Sutaone, M. S. *A Review on Weight Initialization Strategies for Neural Networks*. In: *Artificial Intelligence Review* 1 (2022).

[95] Agarap, A. F. *Deep learning using rectified linear units (relu)*. In: (2018).

[96] Rasamoelina, A. D., Adjailia, F., and Sinčák, P. *A Review of Activation Function for Artificial Neural Network*. In: *2020 IEEE 18th SAMI*. 2020.

[97] Grinsztajn, L., Oyallon, E., and Varoquaux, G. *Why do tree-based models still outperform deep learning on tabular data?* 2022.

[98] Finke, T. et al. *Tree-based algorithms for weakly supervised anomaly detection*. In: *Phys. Rev. D* 3 (2024).

[99] Durkan, C. et al. *Neural Spline Flows*. 2019.

[100] Kingma, D. P. et al. *Improving Variational Inference with Inverse Autoregressive Flow*. 2017.

[101] Papamakarios, G., Pavlakou, T., and Murray, I. *Masked Autoregressive Flow for Density Estimation*. 2018.

[102] Dinh, L., Krueger, D., and Bengio, Y. *NICE: Non-linear Independent Components Estimation*. 2015.

[103] Dinh, L., Sohl-Dickstein, J., and Bengio, S. *Density estimation using Real NVP*. 2017.

[104] Liu, Z. et al. *A ConvNet for the 2020s*. 2022.

[105] Dosovitskiy, A. et al. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. 2021.

[106] Vaswani, A. et al. *Attention Is All You Need*. 2023.

[107] Fleuret, F. *The Little Book of Deep Learning*. 2023.

[108] Bottou, L. *Online Algorithms and Stochastic Approximations*. In: *Online Learning and Neural Networks*. Ed. by D. Saad. Cambridge, UK: Cambridge University Press, 1998.

[109] Defazio, A. et al. *When, Why and How Much? Adaptive Learning Rate Scheduling by Refinement*. 2023.

[110] Kingma, D. P. and Ba, J. *Adam: A Method for Stochastic Optimization*. 2017.

[111] Loshchilov, I. and Hutter, F. *Decoupled Weight Decay Regularization*. 2019.

[112] Shah, H. et al. *The Pitfalls of Simplicity Bias in Neural Networks*. In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Curran Associates, Inc., 2020.

[113] Szegedy, C. et al. *Intriguing properties of neural networks*. 2014.

[114] ATLAS Collaboration. *Search for new resonances in mass distributions of jet pairs using* $139\,fb^{-1}$ *of pp collisions at* $\sqrt{s} = 13\,TeV$ *with the ATLAS detector*. In: *JHEP* (2020).

[115] ATLAS Collaboration. *A Continuous Calibration of the ATLAS Flavor-Tagging Classifiers via Optimal Transportation Maps*. Geneva: CERN, 2024.

[116] Metodiev, E. M., Nachman, B., and Thaler, J. *Classification without labels: learning from mixed samples in high energy physics*. In: *JHEP* 10 (2017).

[117] Witkowski, E., Nachman, B., and Whiteson, D. *Learning to isolate muons in data*. In: *Phys. Rev. D* (9 2023).

[118] Finke, T. et al. *Back To The Roots: Tree-Based Algorithms for Weakly Supervised Anomaly Detection*. 2023.

[119] Louppe, G., Kagan, M., and Cranmer, K. *Learning to Pivot with Adversarial Networks*. 2016.

[120] Shimmin, C. et al. *Decorrelated jet substructure tagging using adversarial neural networks*. In: *Phys. Rev. D* 7 (2017).

[121] Kasieczka, G. and Shih, D. *Robust Jet Classifiers through Distance Correlation*. In: *Phys. Rev. Lett.* 12 (2020).

[122] Kitouni, O. et al. *Enhancing searches for resonances with machine learning and moment decomposition*. In: *JHEP* 4 (2021).

[123] Klein, S. and Golling, T. *Decorrelation with conditional normalizing flows*. In: (2022).

[124] Kasieczka, G. and Shih, D. *Datasets for Boosted W Tagging*. Version v1. Zenodo, 2020.

[125] ATLAS Collaboration. *Performance of mass-decorrelated jet substructure observables for hadronic two-body decay tagging in ATLAS*. ATL-PHYS-PUB-2018-014. 2018.

[126] Favereau, J. de et al. *DELPHES 3, A modular framework for fast simulation of a generic collider experiment*. In: *JHEP* (2014).

[127] Cacciari, M. and Salam, G. P. *Dispelling the N^3 myth for the k_t jet-finder*. In: *Phys. Lett. B* (2006).

[128] Cacciari, M., Salam, G. P., and Soyez, G. *FastJet user manual*. In: *Eur. Phys. J. C* (2012).

[129] Loshchilov, I. and Hutter, F. *Sgdr: Stochastic gradient descent with warm restarts*. 2016.

[130] Chakravarti, P. et al. *Robust semi-parametric signal detection in particle physics with classifiers decorrelated via optimal transport*. 2024.

[131] Carlier, G., Chernozhukov, V., and Galichon, A. *Vector Quantile Regression: An Optimal Transport Approach*. 2015.

[132] Rosenberg, A. A. et al. *Fast Nonlinear Vector Quantile Regression*. 2023.

[133] Vedula, S. et al. *Continuous Vector Quantile Regression*. In: *ICML Workshop on New Frontiers in Learning, Control, and Dynamical Systems*. 2023.

[134] Algren, M., Raine, J. A., and Golling, T. *Decorrelation using optimal transport*. In: *Eur. Phys. J. C* 6 (2024).

[135] Caton, S. and Haas, C. *Fairness in Machine Learning: A Survey*. 2020.

[136] Tabak, E. G. and Turner, C. V. *A Family of Nonparametric Density Estimation Algorithms*. In: *Communications on Pure and Applied Mathematics* 2 (2013).

[137] Papamakarios, G. et al. *Normalizing Flows for Probabilistic Modeling and Inference*. 2021.

[138] Krause, C. and Shih, D. *Fast and accurate simulations of calorimeter showers with normalizing flows*. In: *Phys. Rev. D* 11 (2023).

[139] Máté, B. and Fleuret, F. *Learning Interpolations between Boltzmann Densities*. 2023.

[140] Máté, B. and Fleuret, F. *Multi-Lattice Sampling of Quantum Field Theories via Neural Operator-based Flows*. 2024.

[141] Máté, B. et al. *Flowification: Everything is a normalizing flow*. In: *Advances in Neural Information Processing Systems*. Ed. by S. Koyejo et al. Curran Associates, Inc., 2022.

[142] Huang, C.-W., Dinh, L., and Courville, A. *Augmented Normalizing Flows: Bridging the Gap Between Generative Flows and Latent Variable Models*. 2020.

[143] Nielsen, D. et al. *SurVAE Flows: Surjections to Bridge the Gap between VAEs and Flows*. 2020.

[144] Máté, B. et al. *Flowification: Everything is a Normalizing Flow*. 2023.

[145] Mohan, D. and Scaife, A. M. M. *Evaluating Bayesian deep learning for radio galaxy classification*. 2024.

[146] Golling, T. et al. *Flow-enhanced transportation for anomaly detection*. In: *Phys. Rev. D* 9 (2023).

[147] Golling, T. et al. *Morphing one dataset into another with maximum likelihood estimation*. In: *Phys. Rev. D* 9 (2023).

[148] Sengupta, D. et al. *CURTAINs flows for flows: Constructing unobserved regions with maximum likelihood estimation*. In: *SciPost Phys.* 2 (2024).

[149] Klein, S., Raine, J. A., and Golling, T. *Flows for Flows: Training Normalizing Flows Between Arbitrary Distributions with Maximum Likelihood Estimation*. 2022.

[150] Ho, J., Jain, A., and Abbeel, P. *Denoising Diffusion Probabilistic Models*. 2020.

[151] Albergo, M. S. and Vanden-Eijnden, E. *Building Normalizing Flows with Stochastic Interpolants*. 2023.

[152] Chen, R. T. Q. and Lipman, Y. *Flow Matching on General Geometries*. 2024.

[153] Koenker, R. and Bassett, G. *Regression Quantiles*. In: *Econometrica* 1 (1978).

[154] Koenker, R. and Hallock, K. F. *Quantile Regression*. In: *Journal of Economic Perspectives* 4 (2001).

[155] Koenker, R. *Quantile Regression*. Econometric Society Monographs. Cambridge University Press, 2005.

[156] Golling, T. et al. *Masked Particle Modeling on Sets: Towards Self-Supervised High Energy Physics Foundation Models*. In: *Machine Learning: Science and Technology* 3 (2024).

[157] Devlin, J. et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2019.

[158] Bao, H. et al. *BEiT: BERT Pre-Training of Image Transformers*. 2022.

[159] Kishimoto, T. et al. *Pre-training strategy using real particle collision data for event classification in collider physics*. 2023.

[160] Dillon, B. M. et al. *Symmetries, safety, and self-supervision*. In: *SciPost Phys.* 6 (2022).

[161] Harris, P. et al. *Re-Simulation-based Self-Supervised Learning for Pre-Training Foundation Models*. 2024.

[162] Birk, J., Hallin, A., and Kasieczka, G. *OmniJet-α: The first cross-task foundation model for particle physics*. 2024.

[163] Mikuni, V. and Nachman, B. *OmniLearn: A Method to Simultaneously Facilitate All Jet Physics Tasks*. 2024.

[164] Leigh, M. et al. *Is Tokenization Needed for Masked Particle Modelling?* 2024.

[165] Vigl, M., Hartman, N., and Heinrich, L. *Finetuning foundation models for joint analysis optimization in High Energy Physics*. In: *Machine Learning: Science and Technology* 2 (2024).

[166] Zhou, X. et al. *Sparse Invariant Risk Minimization*. In: *Proceedings of the 39th International Conference on Machine Learning*. Ed. by K. Chaudhuri et al. Proceedings of Machine Learning Research. PMLR, 2022.

[167] Kasieczka, G. et al. *The LHC Olympics 2020 a community challenge for anomaly detection in high energy physics*. In: *Rept. Prog. Phys.* 12 (2021).

[168] Aarrestad, T. et al. *The Dark Machines Anomaly Score Challenge: Benchmark Data and Model Independent Event Classification for the Large Hadron Collider*. In: *SciPost Phys.* 1 (2022).

[169] Karagiorgi, G. et al. *Machine learning in the search for new fundamental physics*. In: *Nature Rev. Phys.* 6 (2022).

[170] Belis, V., Odagiu, P., and Aarrestad, T. K. *Machine learning for anomaly detection in particle physics*. In: *Reviews in Physics* (2024).

[171] Neyman, J. and Pearson, E. S. *On the Problem of the Most Efficient Tests of Statistical Hypotheses*. In: *Phil. Trans. Roy. Soc. Lond. A* 694-706 (1933).

[172] D'Agnolo, R. T. and Wulzer, A. *Learning new physics from a machine*. In: *Phys. Rev. D* 1 (2019).

[173] D'Agnolo, R. T. et al. *Learning New Physics from an Imperfect Machine*. 2021.

[174] Kasieczka, G. et al. *Anomaly detection under coordinate transformations*. In: *Phys. Rev. D* 1 (2023).

[175] Kirichenko, P., Izmailov, P., and Wilson, A. G. *Why Normalizing Flows Fail to Detect Out-of-Distribution Data*. 2020.

[176] Heimel, T. et al. *QCD or what?* In: *SciPost Physics* 3 (2019).

[177] Cerri, O. et al. *Variational autoencoders for new physics mining at the Large Hadron Collider*. In: *JHEP* 5 (2019).

[178] Blance, A., Spannowsky, M., and Waite, P. *Adversarially-trained autoencoders for robust unsupervised new physics searches*. In: *JHEP* 10 (2019).

[179] Roy, T. S. and Vijay, A. H. *A robust anomaly finder based on autoencoders*. 2020.

[180] Dillon, B. M. et al. *A normalized autoencoder for LHC triggers*. In: *SciPost Phys. Core* (2023).

[181] Eble, F. *Unsupervised tagging of semivisible jets with normalized autoencoders in CMS*. Geneva: CERN, 2024.

[182] Collaboration, C. *2024 Data Collected with AXOL1TL Anomaly Detection at the CMS Level-1 Trigger*. 2024.

[183] Abadjiev, D. et al. *Autoencoder-Based Anomaly Detection System for Online Data Quality Monitoring of the CMS Electromagnetic Calorimeter*. In: *Computing and Software for Big Science* 1 (2024).

[184] Mikuni, V., Nachman, B., and Shih, D. *Online-compatible unsupervised nonresonant anomaly detection*. In: *Phys. Rev. D* 5 (2022).

[185] Harris, P. et al. *Machine learning techniques for model-independent searches in dijet final states*. Geneva: CERN, 2023.

[186] Kitouni, O. et al. *Enhancing searches for resonances with machine learning and moment decomposition*. In: *JHEP* 4 (2021).

[187] Abrahamyan, S. et al. *Search for a New Gauge Boson in Electron-Nucleus Fixed-Target Scattering by the APEX Experiment*. In: *Phys. Rev. Lett.* (19 2011).

[188] ATLAS Collaboration. *Search for the Standard Model Higgs Boson in the Diphoton Decay Channel with $4.9 fb^{-1}$ of pp Collisions at $\sqrt{s} = 7\,TeV$ with ATLAS*. In: *Phys. Rev. Lett.* (2012).

[189] Frate, M. et al. *Modeling Smooth Backgrounds and Generic Localized Signals with Gaussian Processes*. 2017.

[190] Choudalakis, G. *On hypothesis testing, trials factor, hypertests and the BumpHunter*. 2011.

[191] ATLAS Collaboration. *Search for low-mass resonances decaying into two jets and produced in association with a photon using pp collisions at $\sqrt{s} = 13\,TeV$ with the ATLAS detector*. In: *Phys. Lett. B* (2019).

[192] CMS Collaboration. *A multi-dimensional search for new heavy resonances decaying to boosted WW, WZ, or ZZ boson pairs in the dijet final state at $13\,TeV$*. In: *Eur. Phys. J. C* (2020).

[193] Tumasyan, A. et al. *Search for new heavy resonances decaying to WW, WZ, ZZ, WH, or ZH boson pairs in the all-jets final state in proton-proton collisions at s=13TeV*. In: *Phys. Lett. B* (2023).

[194] CMS Collaboration. *Search for Higgs Boson Pair Production in the Four b Quark Final State in Proton–Proton Collisions at $\sqrt{s} = 13\,TeV$*. In: *Phys. Rev. Lett.* (2022).

[195] ATLAS Collaboration. *Search for non-resonant pair production of Higgs bosons in the $b\bar{b}b\bar{b}$ final state in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector*. Geneva: CERN, 2022.

[196] Manole, T. et al. *Background Modeling for Double Higgs Boson Production: Density Ratios and Optimal Transport*. 2024.

[197] Andreassen, A., Nachman, B., and Shih, D. *Simulation Assisted Likelihood-free Anomaly Detection*. In: *Phys. Rev. D* 9 (2020).

[198] Benkendorfer, K., Le Pottier, L., and Nachman, B. *Simulation-assisted decorrelation for resonant anomaly detection*. In: *Phys. Rev. D* 3 (2021).

[199] Sengupta, D. et al. *Improving new physics searches with diffusion models for event observables and jet constituents*. In: *JHEP* 4 (2024).

[200] Golling, T. et al. *FETA: Flow-Enhanced Transportation for Anomaly Detection*. In: (2022).

[201] Nachman, B. and Shih, D. *Anomaly Detection with Density Estimation*. In: *Phys. Rev. D* (2020).

[202] Das, R., Kasieczka, G., and Shih, D. *Residual ANODE*. 2023.

[203] Hallin, A. et al. *Classifying anomalies through outer density estimation*. In: *Phys. Rev. D* 5 (2022).

[204] Hallin, A. et al. *Resonant anomaly detection without background sculpting*. In: *Phys. Rev. D* 11 (2023).

[205] Leigh, M. et al. *Accelerating template generation in resonant anomaly detection searches with optimal transport*. 2024.

[206] Raine, J. A. et al. *CURTAINs for your sliding window: Constructing unobserved regions by transforming adjacent intervals*. In: *Front. Big Data* (2023).

[207] Golling, T. et al. *The interplay of machine learning-based resonant anomaly detection methods*. In: *Eur. Phys. J. C* 3 (2024).

[208] ATLAS Collaboration. *Anomaly detection search for new resonances decaying into a Higgs boson and a generic new particle X in hadronic final states using $\sqrt{s} = 13$ TeV pp collisions with the ATLAS detector*. In: (2023).

[209] ATLAS Collaboration. *Search for New Phenomena in Two-Body Invariant Mass Distributions Using Unsupervised Machine Learning for Anomaly Detection at s=13 TeV with the ATLAS Detector*. In: *Phys. Rev. Lett.* 8 (2024).

[210] Oleksiyuk, I. et al. *Cluster Scanning: a novel approach to resonance searches*. 2024.

[211] Metodiev, E. M., Thaler, J., and Wynne, R. *Anomaly detection in collider physics via factorized observables*. In: *Phys. Rev. D* (5 2024).

[212] Birman, M. et al. *Data-directed search for new physics based on symmetries of the SM*. In: *Eur. Phys. J. C* 6 (2022).

[213] Shih, D. et al. *via machinae: Searching for stellar streams using unsupervised machine learning*. In: *Mon. Not. Roy. Astron. Soc.* 4 (2021).

[214] Sengupta, D. et al. *SkyCURTAINs: Model agnostic search for Stellar Streams with Gaia data*. 2024.

[215] ATLAS Collaboration. *Search for New Particles in Two-Jet Final States in 7 TeV Proton–Proton Collisions with the ATLAS Detector at the LHC*. In: *Phys. Rev. Lett.* (2010).

[216] ATLAS Collaboration. *A search for new physics in dijet mass and angular distributions in pp collisions at $\sqrt{s} = 7$ TeV measured with the ATLAS detector*. In: *New J. Phys.* (2011).

[217] ATLAS Collaboration. *Search for new physics in the dijet mass distribution using $1\,fb^{-1}$ of pp collision data at $\sqrt{s} = 7$ TeV collected by the ATLAS detector*. In: *Phys. Lett. B* (2012).

[218] ATLAS Collaboration. *Search for new phenomena in the dijet mass distribution using pp collision data at $\sqrt{s} = 8$ TeV with the ATLAS detector*. In: *Phys. Rev. D* (2015).

[219] ATLAS Collaboration. *Search for new phenomena in dijet mass and angular distributions from pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector*. In: *Phys. Lett. B* (2016).

[220] ATLAS Collaboration. *Search for new phenomena in dijet events using $37\,fb^{-1}$ of pp collision data collected at $\sqrt{s} = 13$ TeV with the ATLAS detector*. In: *Phys. Rev. D* (2017).

[221] Craig, N. et al. *The unexplored landscape of two-body resonances*. In: *Acta Phys. Polon. B* (2019).

[222] Kim, J. H. et al. *The motivation and status of two-body resonance decays after the LHC Run 2 and beyond*. In: *JHEP* 4 (2020).

[223] CMS Collaboration. *Search for third-generation scalar leptoquarks and heavy right-handed neutrinos in final states with two tau leptons and two jets in proton–proton collisions at $\sqrt{s} = 13$ TeV*. In: *JHEP* (2017).

[224] CMS Collaboration. *Search for low mass vector resonances decaying to quark–antiquark pairs in proton–proton collisions at $\sqrt{s} = 13\,TeV$*. In: *Phys. Rev. Lett.* (2017).

[225] ATLAS Collaboration. *Search for a heavy Higgs boson decaying into a Z boson and another heavy Higgs boson in the $\ell\ell bb$ final state in pp collisions at $\sqrt{s} = 13\,TeV$ with the ATLAS detector*. In: *Phys. Lett. B* (2018).

[226] ATLAS Collaboration. *A search for resonances decaying into a Higgs boson and a new particle X in the $XH \to qqbb$ final state with the ATLAS detector*. In: *Phys. Lett. B* (2018).

[227] CMS Collaboration. *Search for low mass vector resonances decaying into quark–antiquark pairs in proton–proton collisions at $\sqrt{s} = 13\,TeV$*. In: *JHEP* (2018).

[228] ATLAS Collaboration. *Search for pairs of highly collimated photon-jets in pp collisions at $\sqrt{s} = 13\,TeV$ with the ATLAS detector*. In: *Phys. Rev. D* (2019).

[229] ATLAS Collaboration. *Search for long-lived particles produced in pp collisions at $\sqrt{s} = 13\,TeV$ that decay into displaced hadronic jets in the ATLAS muon spectrometer*. In: *Phys. Rev. D* (2019).

[230] ATLAS Collaboration. *Search for long-lived neutral particles in pp collisions at $\sqrt{s} = 13\,TeV$ that decay into displaced hadronic jets in the ATLAS calorimeter*. In: *Eur. Phys. J. C* (2019).

[231] CMS Collaboration. *Search for heavy neutrinos and third-generation leptoquarks in hadronic states of two τ leptons and two jets in proton–proton collisions at $\sqrt{s} = 13\,TeV$*. In: *JHEP* (2019).

[232] CMS Collaboration. *Search for a W' boson decaying to a vector-like quark and a top or bottom quark in the all-jets final state*. In: *JHEP* (2019).

[233] CMS Collaboration. *Search for a heavy resonance decaying to a top quark and a vector-like top quark in the lepton+jets final state in pp collisions at $\sqrt{s} = 13\,TeV$*. In: *Eur. Phys. J. C* (2019).

[234] ATLAS Collaboration. *Search for light resonances decaying to boosted quark pairs and produced in association with a photon or a jet in proton–proton collisions at $\sqrt{s} = 13\,TeV$ with the ATLAS detector*. In: *Phys. Lett. B* (2019).

[235] CMS Collaboration. *Search for low mass vector resonances decaying into quark–antiquark pairs in proton–proton collisions at $\sqrt{s} = 13\,TeV$*. In: *Phys. Rev. D* (2019).

[236] CMS Collaboration. *Search for low-mass quark–antiquark resonances produced in association with a photon at $\sqrt{s} = 13\,TeV$*. In: *Phys. Rev. Lett.* (2019).

[237] CMS Collaboration. *Search for a W' boson decaying to a vector-like quark and a top or bottom quark in the all-jets final state at $\sqrt{s} = 13\,TeV$*. In: *JHEP* (2022).

[238] ATLAS Collaboration. *Dijet Resonance Search with Weak Supervision Using $\sqrt{s} = 13\,TeV$ pp collisions in the ATLAS detector*. In: *Phys. Rev. Lett.* (2020).

[239] Collaboration, C. *Model-agnostic search for dijet resonances with anomalous jet substructure in proton-proton collisions at $\sqrt{s}$ = 13 TeV*. 2024.

[240] Khachatryan, V. et al. *Search for narrow resonances decaying to dijets in proton-proton collisions at $\sqrt{s} = 13$ TeV*. In: *Phys. Rev. Lett.* 7 (2016).

[241] ATLAS Collaboration. *Summary of Diboson Resonance Searches at the ATLAS experiment using full Run-2 data*. Geneva: CERN, 2023.

[242] Collaboration, C. *Summary of Diboson Resonance Searches in CMS*. CERN.

[243] Agashe, K., Contino, R., and Pomarol, A. *The Minimal Composite Higgs Model*. In: *Nucl. Phys. B* 1 (2005).

[244] Randall, L. and Sundrum, R. *Large Mass Hierarchy from a Small Extra Dimension*. In: *Phys. Rev. Lett.* (17 1999).

[245] ATLAS Collaboration. *Search for diboson resonances in hadronic final states in $139\,fb^{-1}$ of pp collisions at $\sqrt{s} = 13\,TeV$ with the ATLAS detector*. In: *JHEP* (2019).

[246] ATLAS Collaboration. *The performance of the jet trigger for the ATLAS detector during 2011 data taking*. In: *Eur. Phys. J. C* (2016).

[247] ATLAS Collaboration. *Topological cell clustering in the ATLAS calorimeters and its performance in LHC Run 1*. In: *Eur. Phys. J. C* (2017).

[248] ATLAS Collaboration. *In situ calibration of large-radius jet energy and mass in* 13 *TeV proton–proton collisions with the ATLAS detector*. In: *Eur. Phys. J. C* (2019).

[249] ATLAS Collaboration. *Jet mass reconstruction with the ATLAS Detector in early Run 2 data*. ATLAS-CONF-2016-035. 2016.

[250] ATLAS Collaboration. *Image from Jet Trigger Public Results*. Online.

[251] Sjöstrand, T., Mrenna, S., and Skands, P. Z. *PYTHIA 6.4 physics and manual*. In: *JHEP* (2006).

[252] Sjöstrand, T., Mrenna, S., and Skands, P. *A brief introduction to PYTHIA 8.1*. In: *Comput. Phys. Commun.* (2008).

[253] ATLAS Collaboration. *ATLAS Pythia 8 tunes to* 7 *TeV data*. ATL-PHYS-PUB-2014-021. 2014.

[254] Ball, R. D. et al. *Parton distributions with LHC data*. In: *Nucl. Phys. B* (2013).

[255] Lange, D. J. *The EvtGen particle decay simulation package*. In: *Nucl. Instrum. Meth. A* (2001).

[256] Branco, G. C. et al. *Theory and phenomenology of two-Higgs-doublet models*. In: *Phys. Rept.* (2012).

[257] Lipman, Y. et al. *Flow Matching for Generative Modeling*. 2023.

[258] Esser, P. et al. *Scaling Rectified Flow Transformers for High-Resolution Image Synthesis*. 2024.

[259] Elfwing, S., Uchibe, E., and Doya, K. *Sigmoid-Weighted Linear Units for Neural Network Function Approximation in Reinforcement Learning*. 2017.

[260] Srivastava, N. et al. *Dropout: A Simple Way to Prevent Neural Networks from Overfitting*. In: *Journal of Machine Learning Research* 56 (2014).

[261] Chollet, F. et al. *Keras*. 2015. URL: `https://keras.io`.

[262] Martín Abadi et al. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. 2015.

[263] Eschle, J. et al. *zfit: Scalable pythonic fitting*. In: *SoftwareX* (2020).

[264] Durkan, C. et al. *nflows: normalizing flows in PyTorch*. Version v0.14. 2020.

[265] Paszke, A. et al. *PyTorch: An Imperative Style, High-Performance Deep Learning Library*. 2019.

[266] Hastie, T., Tibshirani, R., and Friedman, J. *The Elements of Statistical Learning*. In: Springer Series in Statistics. New York, NY, USA: Springer New York Inc., 2001. Chap. 7.

[267] Alitti, J. et al. *A Measurement of two jet decays of the W and Z bosons at the CERN $\bar{p}p$ collider*. In: *Z. Phys. C* (1991).

[268] Mölder, F. et al. *Sustainable data analysis with Snakemake*. In: *F1000Research* (2021).

[269] Kurtzer, G. M. et al. *hpcng/singularity: Singularity 3.7.3*. 2021.

[270] Harris, C. R. et al. *Array programming with NumPy*. In: *Nature* 7825 (2020).

[271] team, T. pandas development. *pandas-dev/pandas: Pandas*. Version latest. 2020.

[272] Virtanen, P. et al. *SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python*. In: *Nature Methods* (2020).

[273] Hunter, J. D. *Matplotlib: A 2D graphics environment*. In: *Computing In Science & Engineering* 3 (2007).

[274] Pedregosa, F. et al. *Scikit-learn: Machine Learning in Python*. In: *Journal of Machine Learning Research* (2011).

[275] Rodrigues, E. et al. *The Scikit HEP Project – overview and prospects*. In: *EPJ Web Conf.* (2020). Ed. by C. Doglioni et al.

[276] Yadan, O. *Hydra - A framework for elegantly configuring complex applications*. 2019.

[277] Yadan, O., Sommer-Simpson, J., and Delalleau, O. *omegaconf*. `https://github.com/omry/omegaconf`. 2019.

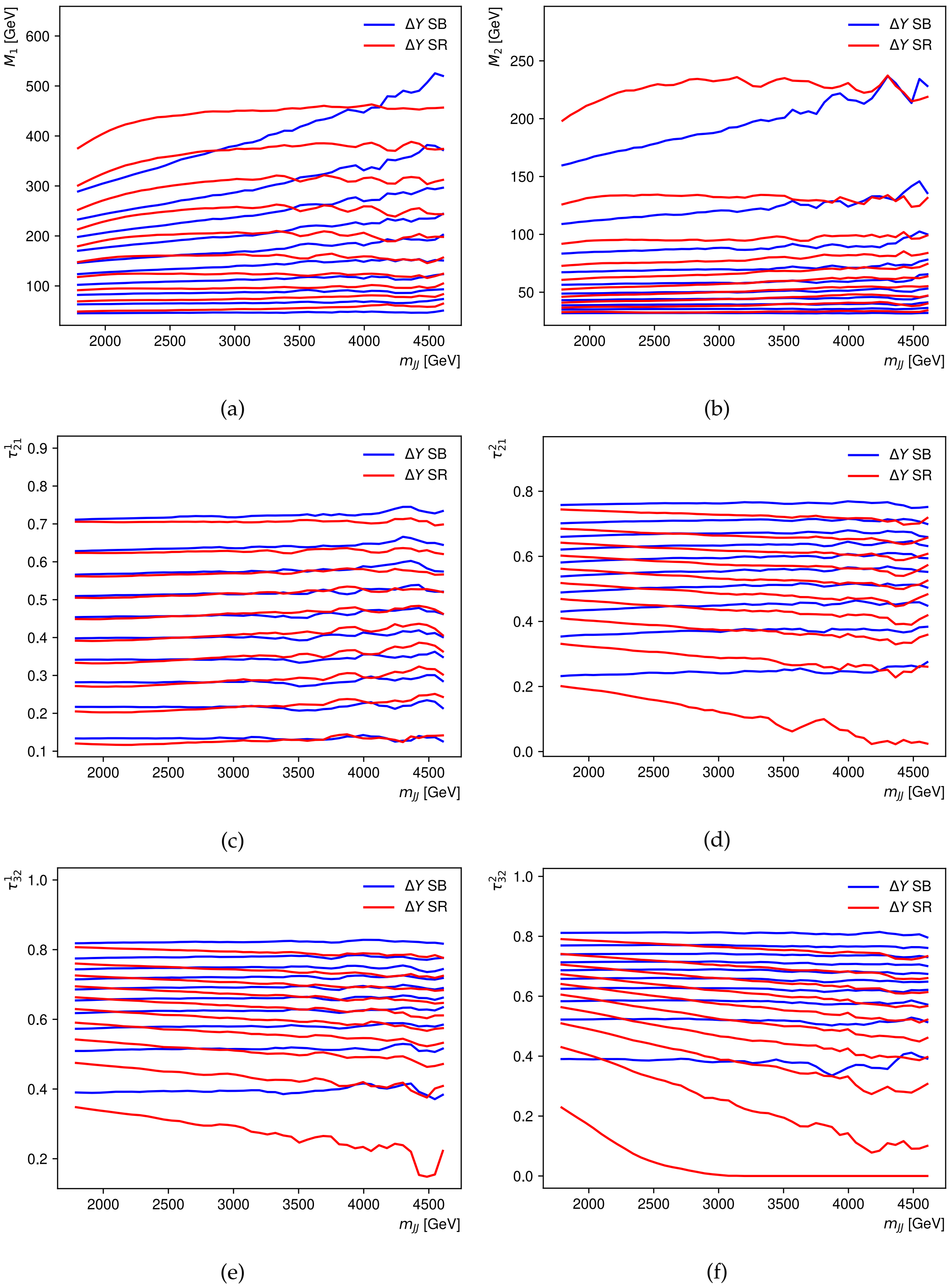
[278] Krause, C. et al. *CaloChallenge 2022: A Community Challenge for Fast Calorimeter Simulation*. 2024.

[279] ATLAS Collaboration. *RECAST framework reinterpretation of an ATLAS Dark Matter Search constraining a model of a dark Higgs boson decaying to two b-quarks*. ATL-PHYS-PUB-2019-032. 2019.

Appendix A

Feature distributions

This appendix contains the full set of features used to extend the bump hunt, M, τ_{21}, τ_{32}, for different datasets. Some of these feature distributions were not shown in the body for brevity. A comparison of the correlations between the features for the $|\Delta Y|$ SB and $|\Delta Y|$ SR datasets is shown in Figure A.1. The marginal distributions of the upsampled MC dataset and the original MC dataset are shown in Figure A.2. The comparison of the correlations between the features for the upsampled MC dataset and the original MC dataset is shown in Figure A.3.



(a)

(b)

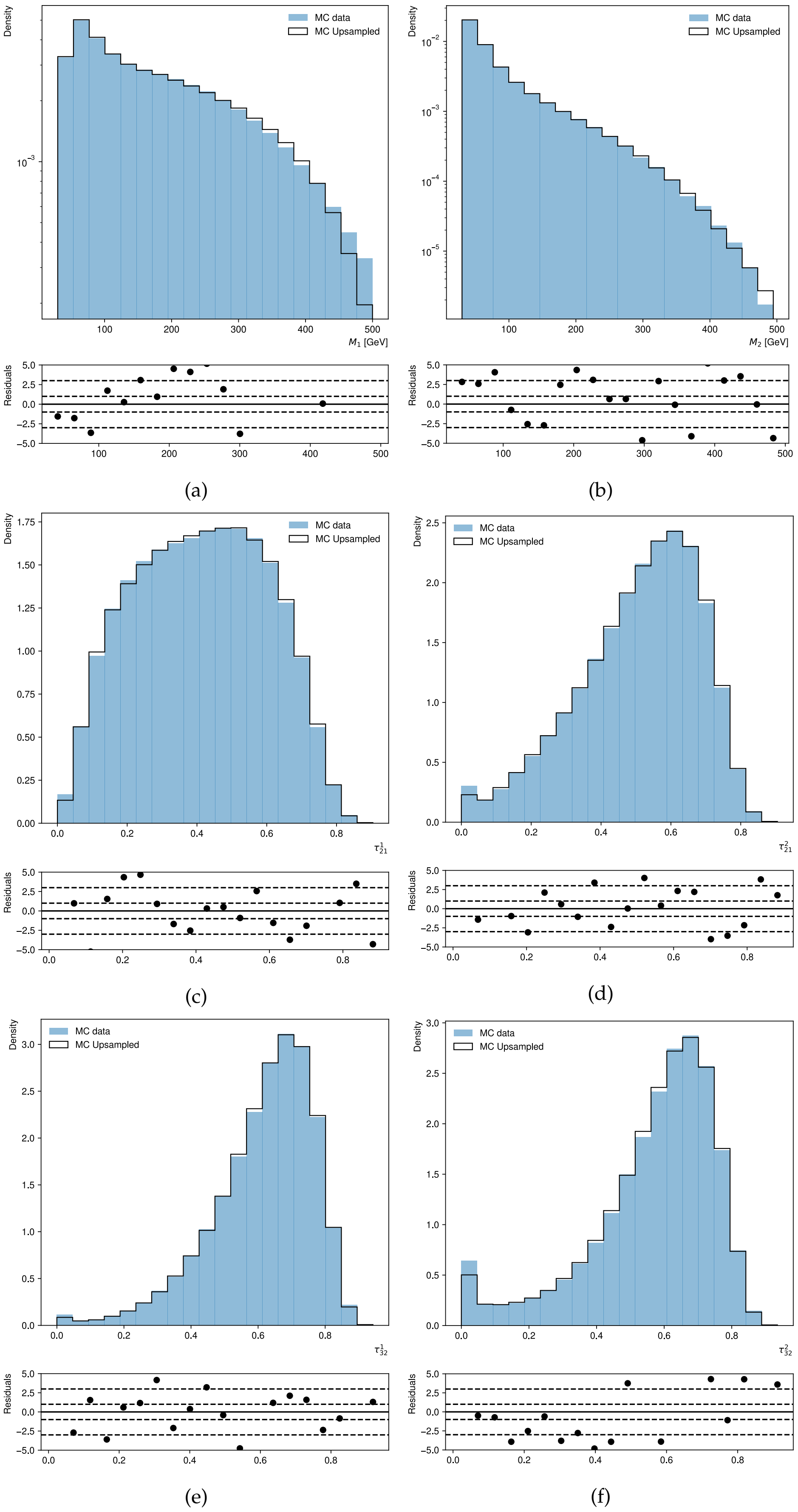
(c)

(d)

(e)

(f)

Figure A.1: Distribution of the additional features used to extend the bump hunt as a function of m_{JJ} for $|\Delta Y|$ SB and $|\Delta Y|$ SR data. Ten equally spaced quantiles between 5% and 95% are shown.



(a)

(b)

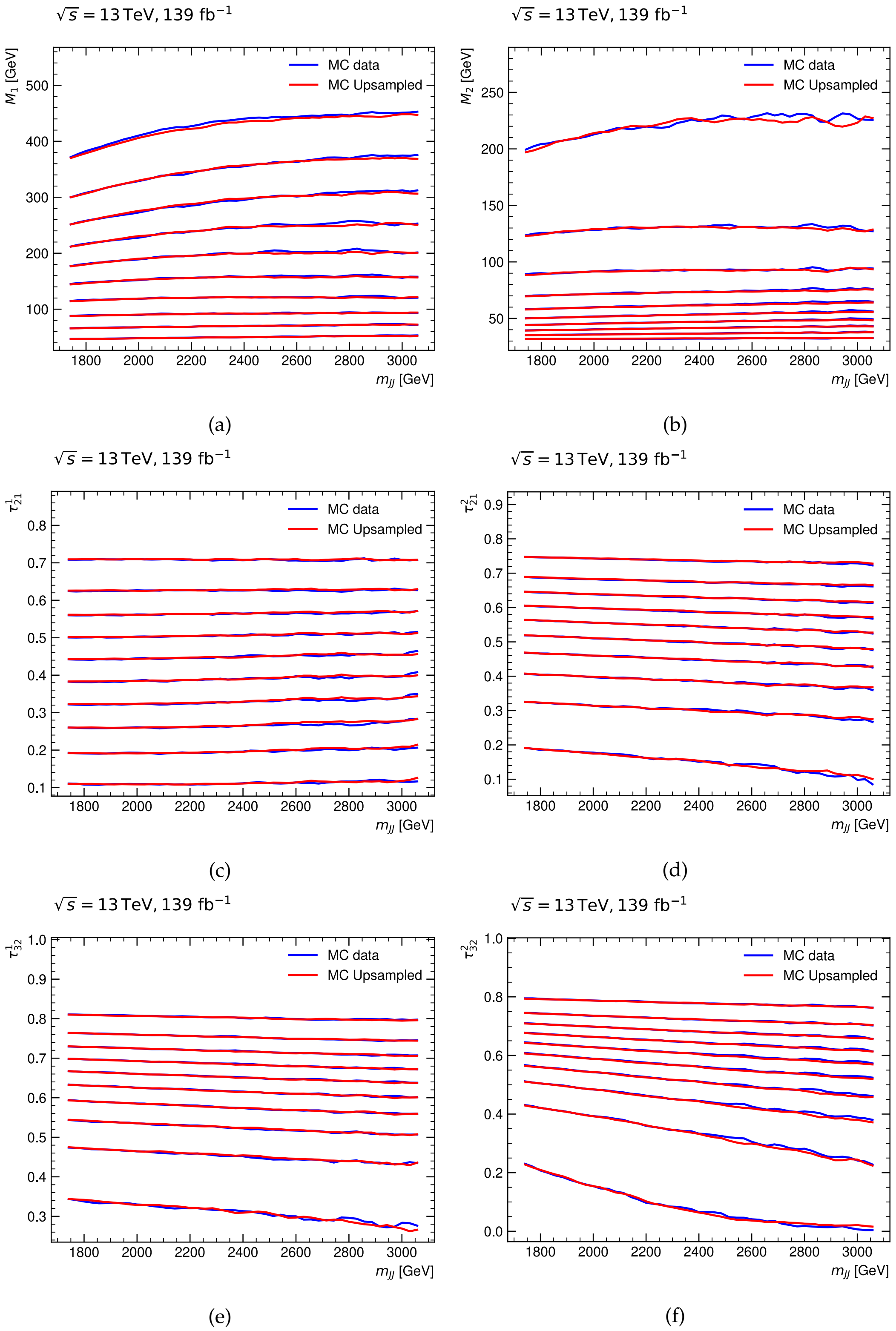
(c)

(d)

(e)

(f)

Figure A.2: Distribution of the additional features used to extend the bump hunt for the upsampled MC dataset and the original MC dataset.



(a)

(b)

(c)

(d)

(e)

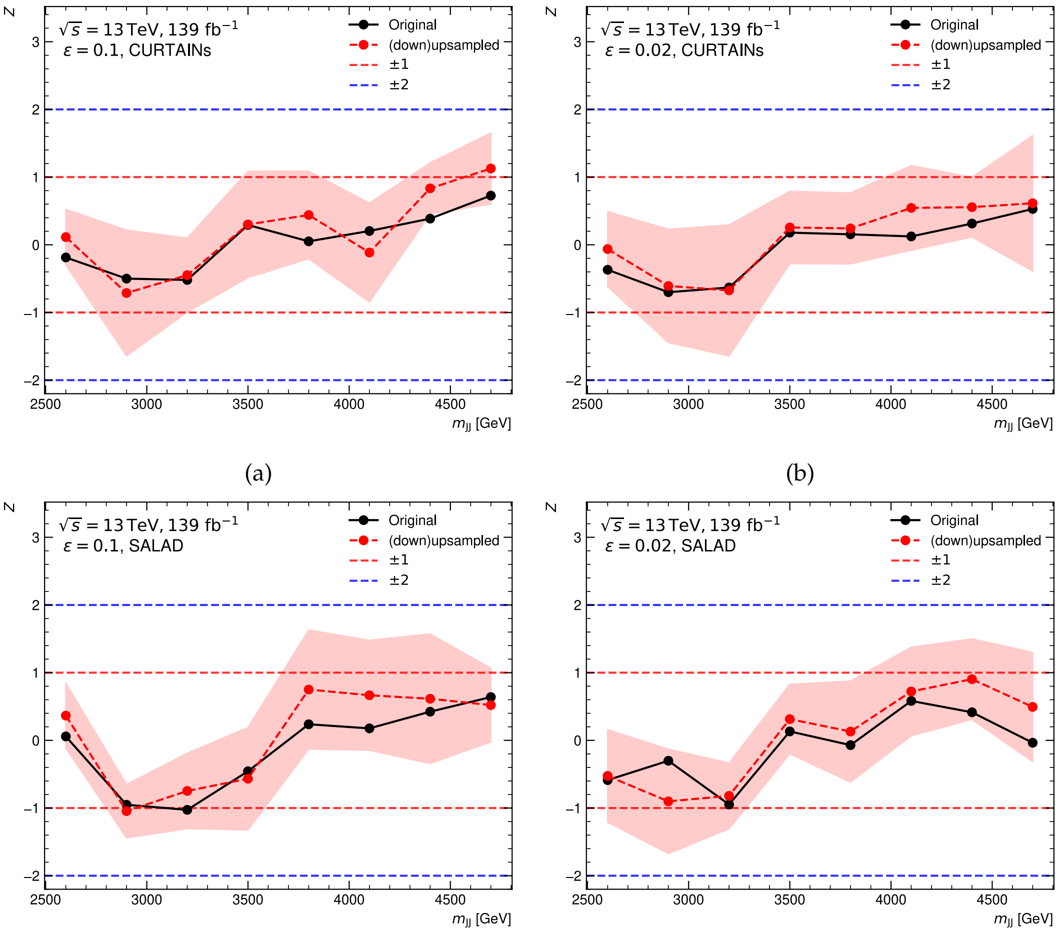
(f)

Figure A.3: Distribution of the additional features used to extend the bump hunt as a function of m_{JJ} for the upsampled MC dataset and the original MC dataset. Ten equally spaced quantiles between 5% and 95% are shown.

Appendix B

(Down)upsampling validation

The full set of validation studies for the (down)upsampling validation procedure on the $|\Delta Y|$ SB dataset. The spread of significances for CURTAINS and SALAD at the two different selections, $\epsilon = 0.1$ and $\epsilon = 0.02$, is shown in Figure B.1, Figure B.2, and Figure B.3.



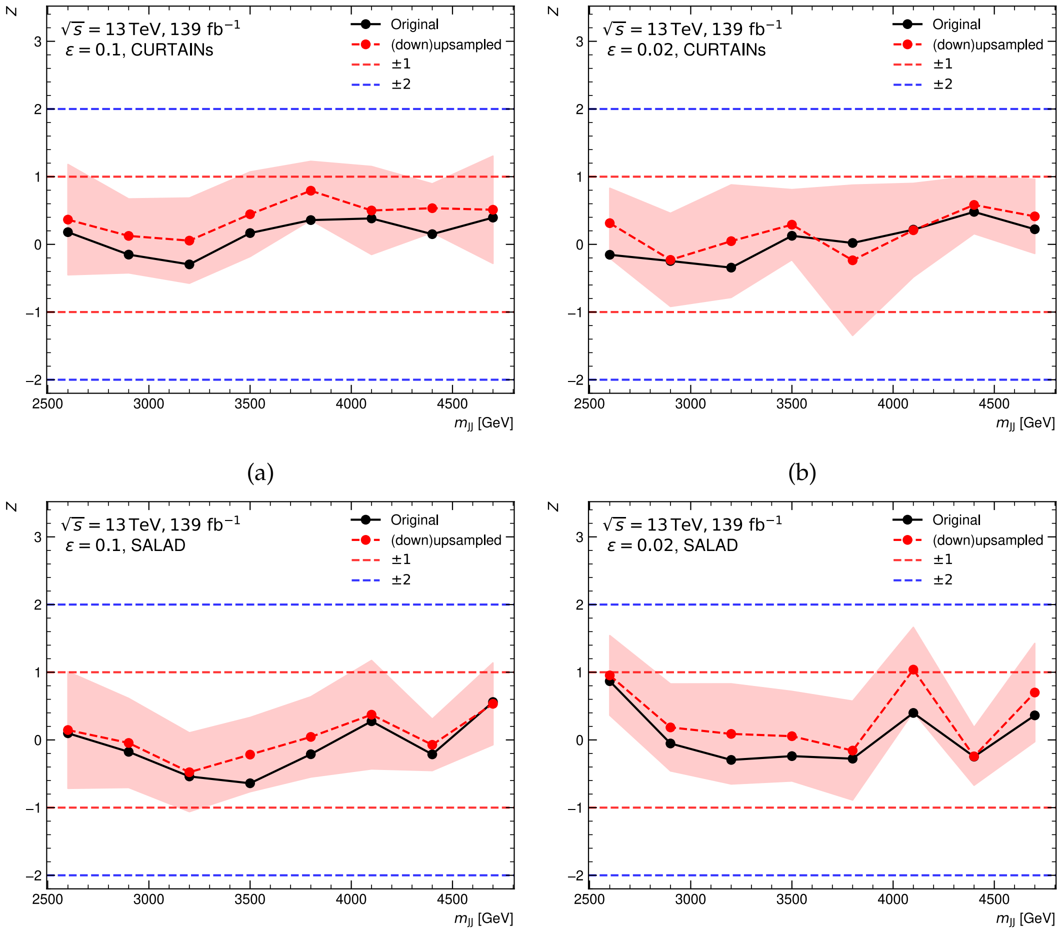
(a)

(b)

(c)

(d)

Figure B.1: Spread of significances for (a, b) CURTAINS and (c, d) SALAD at the two different selections, (a, c) $\epsilon = 0.1$ and (b, d) $\epsilon = 0.02$. Significances are shown for the M feature set and all m_{JJ} signal region centers. The spread is taken over ten different upsamplings of the downsampled resampled $|\Delta Y|$ sideband dataset.



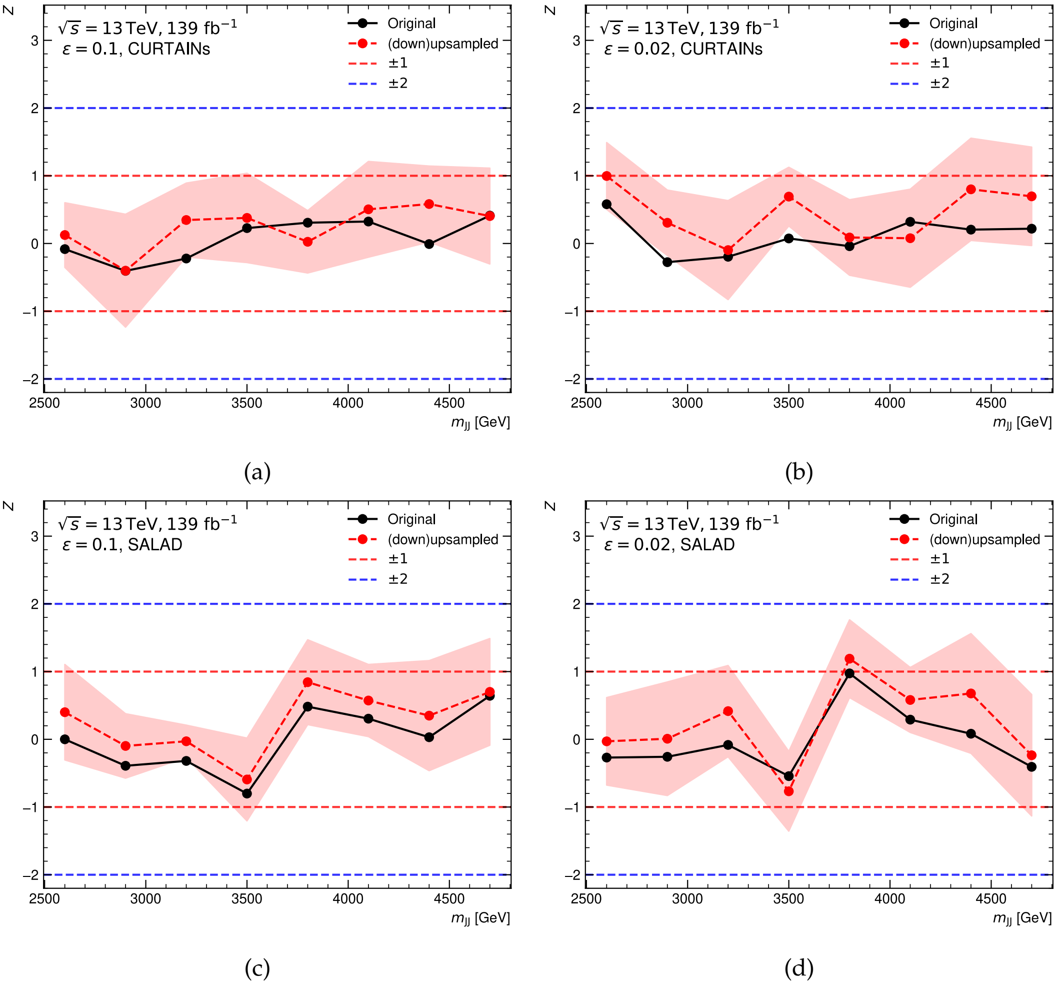
(a)

(b)

(c)

(d)

Figure B.2: Spread of significances for (a, b) CURTAINS and (c, d) SALAD at the two different selections, (a, c) $\epsilon = 0.1$ and (b, d) $\epsilon = 0.02$. Significances are shown for the M, τ_{21} feature set and all m_{JJ} signal region centers. The spread is taken over ten different upsamplings of the downsampled resampled $|\Delta Y|$ sideband dataset.



(a)

(b)

(c)

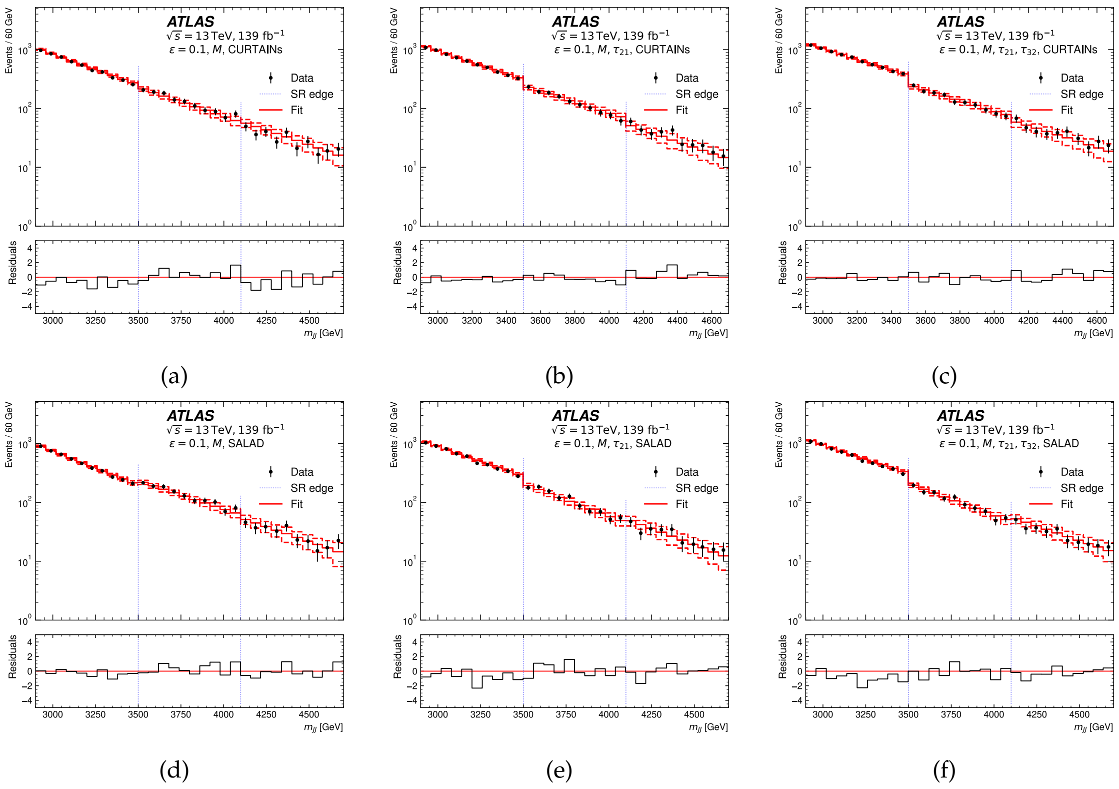
(d)

Figure B.3: Spread of significances for (a, b) CURTAINS and (c, d) SALAD at the two different selections, (a, c) $\epsilon = 0.1$ and (b, d) $\epsilon = 0.02$. Significances are shown for the M, τ_{21}, τ_{32} feature set and all m_{JJ} signal region centers. The spread is taken over ten different upsamplings of the downsampled resampled $|\Delta Y|$ sideband dataset.

Appendix C

Histograms

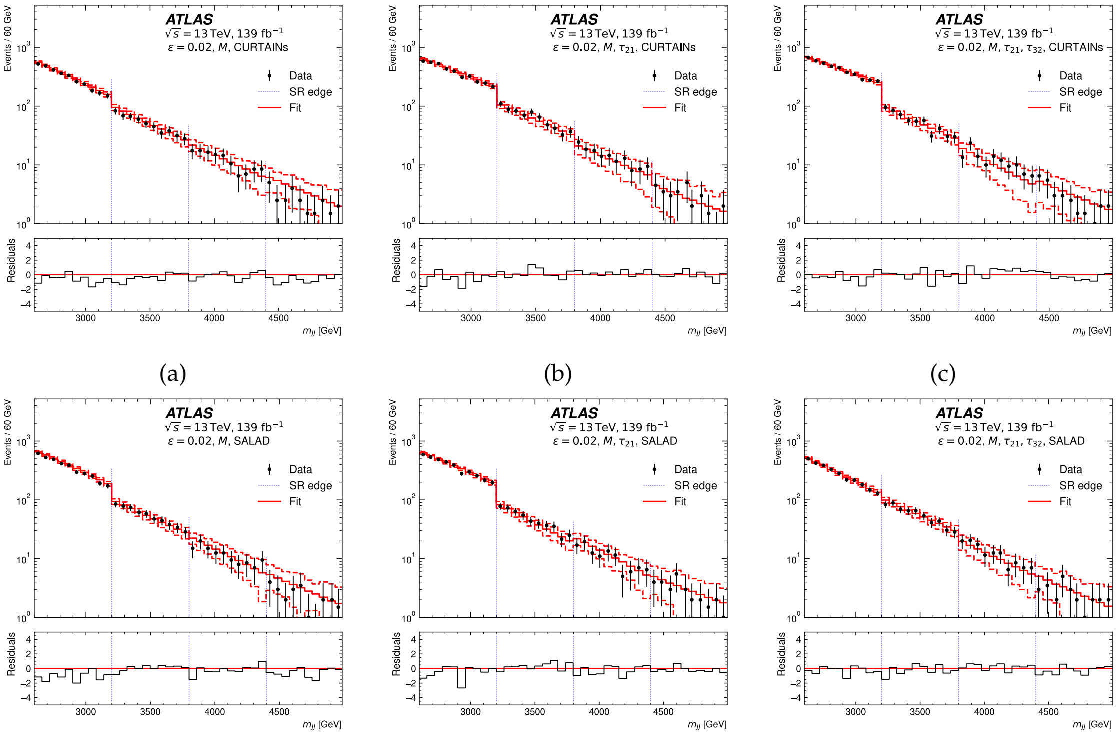
The histograms of m_{JJ} in the second set of non-overlapping m_{JJ} signal regions on all feature sets at the $\epsilon = 0.1$ classifier selection on one (down)upsampled validation set are shown in Figure 10.10. The histograms of m_{JJ} in the first (second) set of non-overlapping m_{JJ} signal regions on all feature sets at the $\epsilon = 0.02$ classifier selection on one (down)upsampled validation set are shown in Figure C.2 (Figure C.3).



(a) (b) (c)

(d) (e) (f)

Figure C.1: Histograms of m_{JJ} in the second set of non-overlapping m_{JJ} signal regions on all feature sets at the $\epsilon = 0.1$ classifier selection on one (down)upsampled validation set. Dashed histograms represent the fit uncertainty. The rows show different methods: (a, b, c) Curtains and (d, e, f) Salad. The columns show different feature sets: (a, d) is the result of $\mathcal{T} = M$, (b, e) is the result of $\mathcal{T} = M, \tau_{21}$ and (c, f) is the result of $\mathcal{T} = M, \tau_{21}, \tau_{32}$. The fit is derived from the background-only fit interpolated from the sidebands and the uncertainty on the fit comes from the fit uncertainty and the Poisson statistical uncertainty. The uncertainty on the observed counts is the Poisson uncertainty plus the uncertainty from the mass bin ensembling procedure. The vertical dashed lines mark the edges of each signal region in m_{JJ}. The lower panel in each plot shows the Gaussian-equivalent significance of the deviation between the fit and data.



(a)

(b)

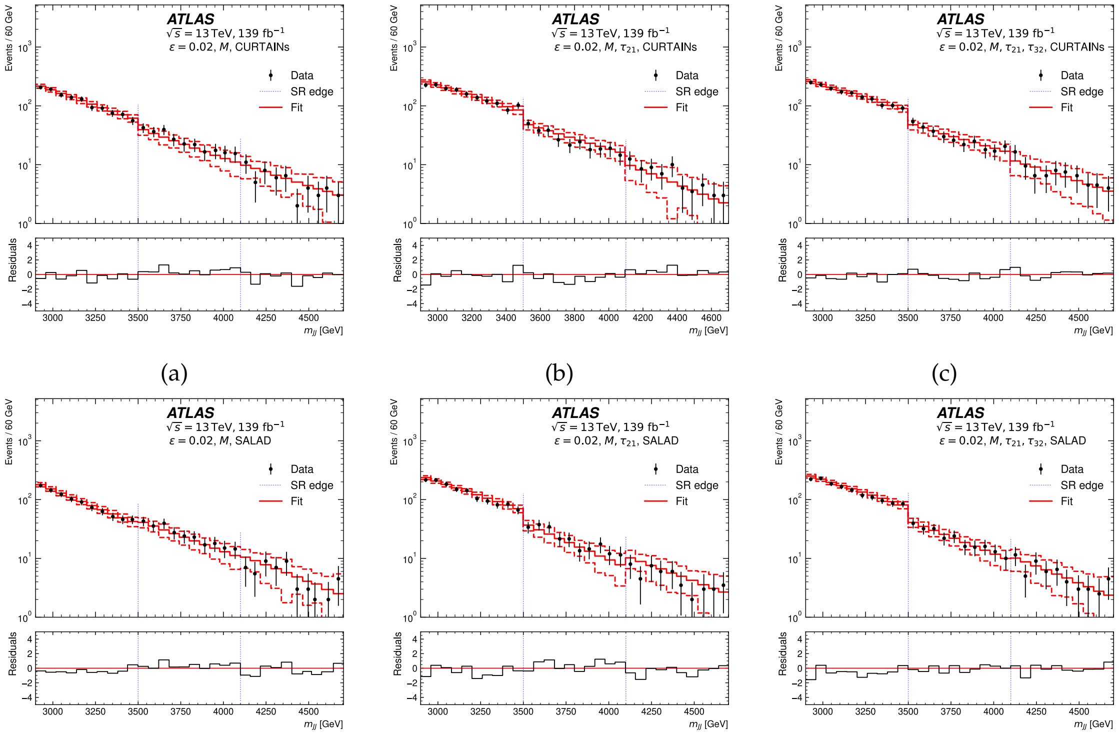
(c)

(d)

(e)

(f)

Figure C.2: Histograms of m_{JJ} in the second set of non-overlapping m_{JJ} signal regions on all feature sets at the $\epsilon = 0.02$ classifier selection on one (down)upsampled validation set. Dashed histograms represent the fit uncertainty. The rows show different methods: (a, b, c) CURTAINS and (d, e, f) SALAD. The columns show different feature sets: (a, d) is the result of $\mathcal{T} = M$, (b, e) is the result of $\mathcal{T} = M, \tau_{21}$ and (c, f) is the result of $\mathcal{T} = M, \tau_{21}, \tau_{32}$. The fit is derived from the background-only fit interpolated from the sidebands and the uncertainty on the fit comes from the fit uncertainty and the Poisson statistical uncertainty. The uncertainty on the observed counts is the Poisson uncertainty plus the uncertainty from the mass bin ensembling procedure. The vertical dashed lines mark the edges of each signal region in m_{JJ}. The lower panel in each plot shows the Gaussian-equivalent significance of the deviation between the fit and data.



(a) (b) (c)

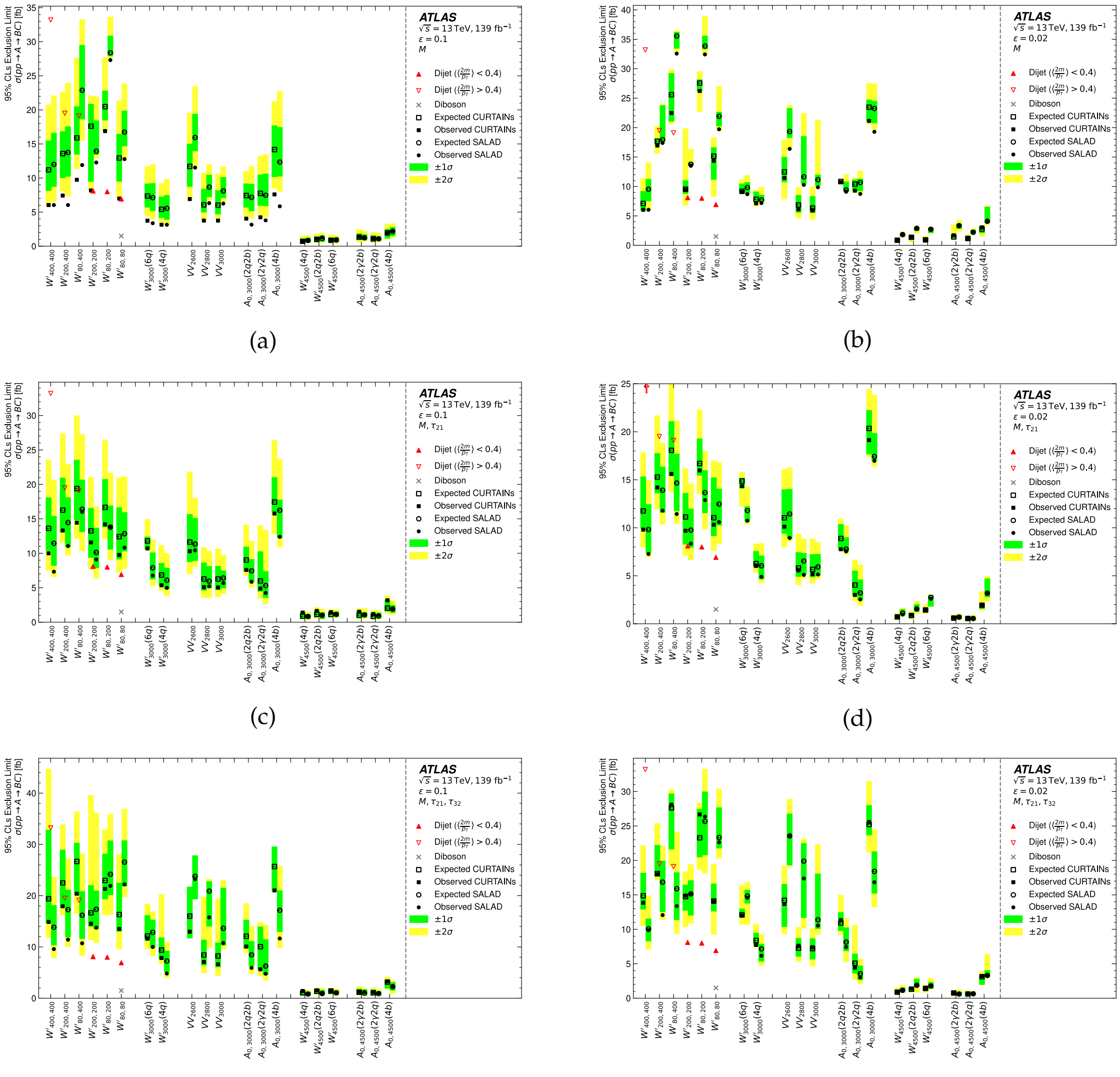
(d) (e) (f)

Figure C.3: Histograms of m_{JJ} in the second set of non-overlapping m_{JJ} signal regions on all feature sets at the $\epsilon = 0.02$ classifier selection on one (down)upsampled validation set. Dashed histograms represent the fit uncertainty. The rows show different methods: (a, b, c) CURTAINS and (d, e, f) SALAD. The columns show different feature sets: (a, d) is the result of $\mathcal{T} = M$, (b, e) is the result of $\mathcal{T} = M, \tau_{21}$ and (c, f) is the result of $\mathcal{T} = M, \tau_{21}, \tau_{32}$. The fit is derived from the background-only fit interpolated from the sidebands and the uncertainty on the fit comes from the fit uncertainty and the Poisson statistical uncertainty. The uncertainty on the observed counts is the Poisson uncertainty plus the uncertainty from the mass bin ensembling procedure. The vertical dashed lines mark the edges of each signal region in m_{JJ}. The lower panel in each plot shows the Gaussian-equivalent significance of the deviation between the fit and data.

Appendix D

Limits

The full set of limits on all signal regions and selections for Curtains and Salad are shown in Figure D.1. The comparison of the limits set by Curtains and Salad using the best possible derived limit and the limit the analysis reported is shown in Figure D.2. The best possible limits, including expected and observed limits, for Curtains and Salad are shown in Figure D.3.



(a)

(b)

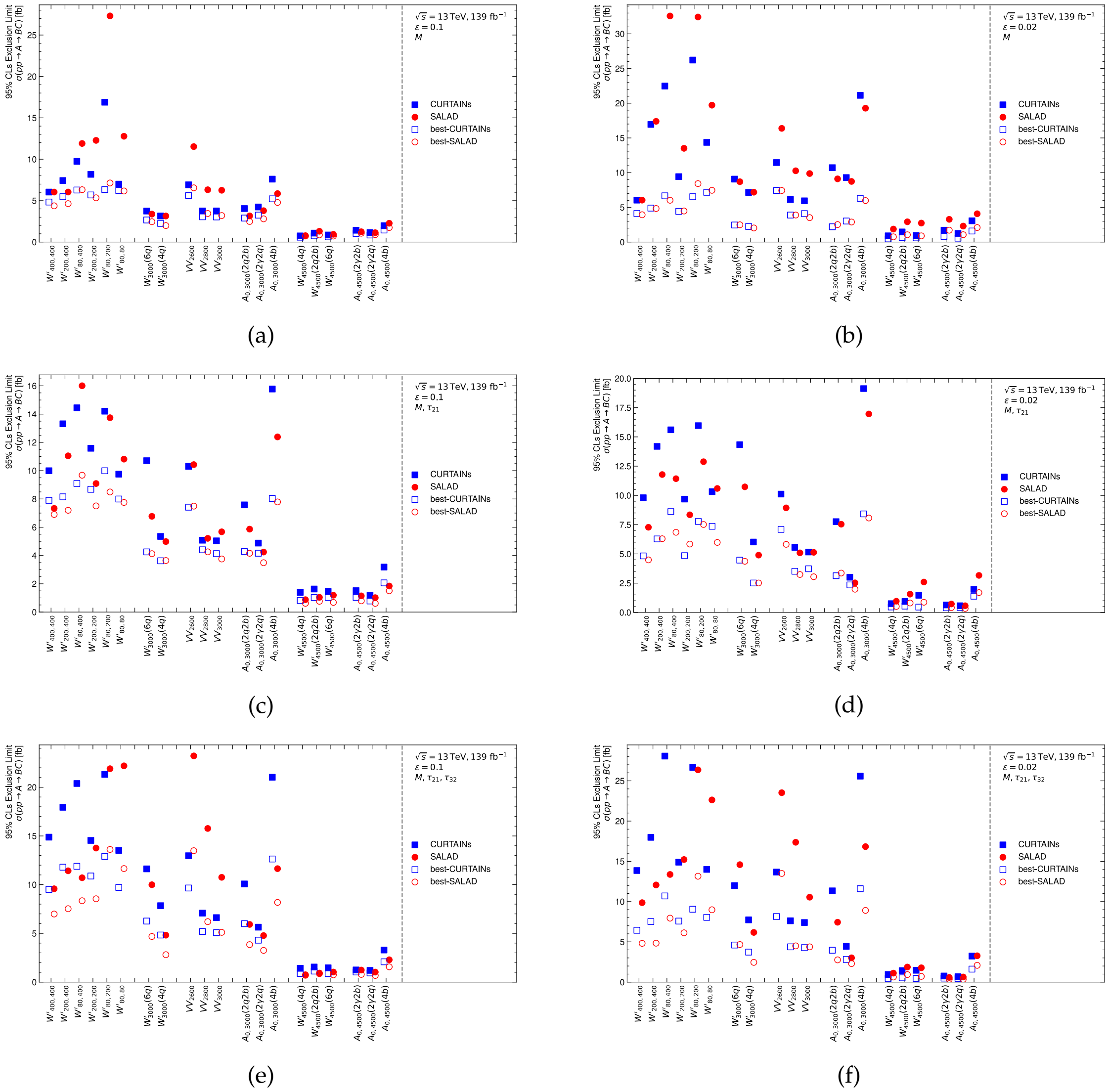
(c)

(d)

(e)

(f)

Figure D.1: Limits for SALAD and CURTAINS with (a, b) $\mathcal{T} = M$, (c, d) $\mathcal{T} = M, \tau_{21}$, and (e, f) $\mathcal{T} = M, \tau_{21}, \tau_{32}$ and (a, c, e) at $\epsilon = 0.1$ and (b, d, f) at $\epsilon = 0.02$.



(a)

(b)

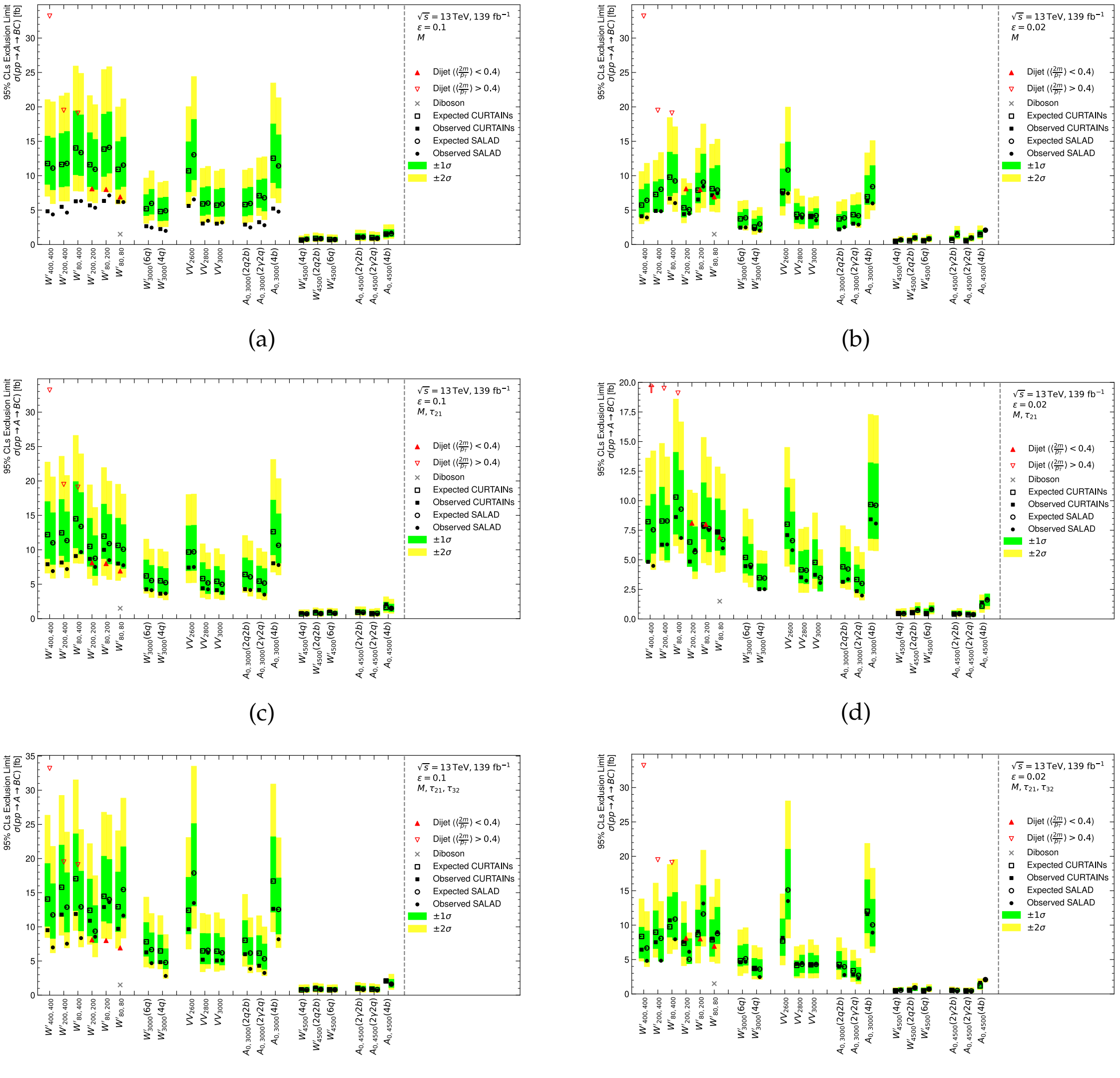
(c)

(d)

(e)

(f)

Figure D.2: Comparison of the limits set by CURTAINs and SALAD using the best possible derived limit and the limit the analysis reported.



(a)

(b)

(c)

(d)

(e)

(f)

Figure D.3: Best possible limits for SALAD and CURTAINS with (a, b) $\mathcal{T} = M$, (c, d) $\mathcal{T} = M, \tau_{21}$, and (e, f) $\mathcal{T} = M, \tau_{21}, \tau_{32}$ and (a, c, e) at $\epsilon = 0.1$ and (b, d, f) at $\epsilon = 0.02$.