

# Implementation of Machine Learning Techniques in the Search for Emerging Jets Using the ATLAS Run II Dataset

by

**IAN RAMIREZ-BEREND**

A thesis submitted to the  
Faculty of Graduate and Postdoctoral Affairs  
in partial fulfillment of the requirements  
for the degree of

**MASTER OF SCIENCE**

in

**PHYSICS**

Carleton University  
Department of Physics  
Ottawa-Carleton Institute for Physics  
Ottawa, Ontario  
© 2022  
Ian Ramirez-Berend



# Abstract

A search for the novel experimental signature known as 'emerging jets' in the ATLAS Run II dataset is presented. The emerging jets model assumes a QCD-like dark sector in which jets of dark particles can shower and hadronize. These dark hadrons eventually decay back to Standard Model particles and produce jets at various displacements from the interaction point. These jets 'emerge' into the detectors, producing a unique signature at collider experiments. Introducing a boosted decision tree into this analysis enables sensitivity to several emerging jets models which cover a wide parameter space, allowing for the possibility of discovery, or alternatively excluding the current theoretical cross-sections. This is shown via a complete Monte-Carlo study, including a thorough evaluation of systematic uncertainties, which corresponds to proton-proton collisions at  $\sqrt{s} = 13$  TeV, and the full ATLAS Run II integrated luminosity of  $139 \text{ fb}^{-1}$ . Although the analysis is currently still blinded to data in the search region, it is validated using a selection of background-only Data which can be compared to the corresponding background Monte-Carlo.



# Acknowledgements

I would like to take the opportunity to thank the people who supported me and made this degree possible. Firstly, thank you to my supervisor, Kevin Graham, who pushed and motivated me to achieve more with this analysis than I would have initially thought feasible. I am so grateful for your continued support through the many ups and downs of this degree. Thank you to Alain Bellerive, who first gave me my start in physics, and believed in me enough to encourage me to continue working with ATLAS after my undergraduate degree.

Thank you to Carleton Emerging Jets team: Brandon, Jeremie, Jesse, and Mohsen. Our weekly meetings have not only been incredibly helpful, but have also fostered a sense of camaraderie I would not have thought possible given the circumstances; it has been a pleasure to work with you all. To the ATLAS Carleton group as a whole, thank you to the community of graduate students and faculty which have made my degree both enjoyable and a fantastic learning experience. Thank you to my committee members, Simon Viel and Bruce Wallace, for taking the time to review my thesis, and give me thoughtful and helpful feedback. Thank you to the ATLAS Emerging Jets analysis team, for remaining positive and supportive while we upended the analysis and integrated my work into the Run II paper.

Thank you to my friends: to Mari-Lisa and Michèle, who have stayed with me through so much and continue to support me; to Dom, Jenny, and Leena, it is so nice to feel at home with you, no matter how long it has been; to the Cedar Ridge Camp community, which has truly become a family over the years, and without which I would not be the person I am today; and to Richard and Keegan without whom I would not have survived my undergrad.

To my loving girlfriend and editor-in-chief, Cassandra, thank you for being with me through the challenging end of this degree, and for allowing me to start the next chapter of my life with a

whole new perspective. You truly are a wise, willful, wonderful woman.

Finally, and most importantly, to my family, thank you for all of your love and support: to my Father, who has always made the importance of an education clear, and who has paved the way for me to achieve as much as I have; to my Mother, who has fostered my love of teaching, even if she did not always understand what I was teaching her; to Megan, my twin and best friend, who has always been there for me when I needed, and who continues to inspire me to achieve my goals; and to Dani, my number one cheerleader, who is possibly more excited about my research than I am. Thank you all, I love you so much.

# Table of Contents

|                                                        |             |
|--------------------------------------------------------|-------------|
| <b>Abstract</b>                                        | <b>i</b>    |
| <b>Acknowledgements</b>                                | <b>ii</b>   |
| <b>Table of Contents</b>                               | <b>iv</b>   |
| <b>List of Figures</b>                                 | <b>viii</b> |
| <b>List of Tables</b>                                  | <b>xiii</b> |
| <b>1 Introduction</b>                                  | <b>1</b>    |
| <b>2 Theory and Background</b>                         | <b>5</b>    |
| 2.1 The Standard Model . . . . .                       | 5           |
| 2.1.1 Fermions . . . . .                               | 6           |
| 2.1.2 Bosons . . . . .                                 | 7           |
| 2.1.3 Antimatter . . . . .                             | 10          |
| 2.2 Component Theories of the Standard Model . . . . . | 10          |
| 2.2.1 Quantum Electrodynamics . . . . .                | 11          |
| 2.2.2 The Weak Interaction . . . . .                   | 12          |
| 2.2.3 Electroweak Theory . . . . .                     | 15          |
| 2.3 Quantum Chromodynamics . . . . .                   | 18          |
| 2.3.1 Color Confinement . . . . .                      | 19          |

|          |                                            |           |
|----------|--------------------------------------------|-----------|
| 2.3.2    | Asymptotic Freedom . . . . .               | 21        |
| 2.3.3    | Valence Quarks and the Quark Sea . . . . . | 21        |
| <b>3</b> | <b>Physics Beyond The Standard Model</b>   | <b>22</b> |
| 3.1      | Dark Matter . . . . .                      | 22        |
| 3.1.1    | Evidence for Dark Matter . . . . .         | 22        |
| 3.1.2    | The WIMP Model of Dark Matter . . . . .    | 24        |
| 3.2      | Hidden Valley Physics . . . . .            | 25        |
| 3.3      | Long-Lived Particles . . . . .             | 25        |
| 3.4      | Emerging Jets . . . . .                    | 26        |
| 3.4.1    | Dark QCD Model Parameters . . . . .        | 27        |
| 3.4.2    | Emerging Jets Phenomenology . . . . .      | 30        |
| <b>4</b> | <b>The ATLAS Experiment</b>                | <b>33</b> |
| 4.1      | The Large Hadron Collider . . . . .        | 33        |
| 4.2      | The ATLAS Experiment . . . . .             | 35        |
| 4.2.1    | The ATLAS Coordinate System . . . . .      | 36        |
| 4.2.2    | Inner Detector . . . . .                   | 38        |
| 4.2.3    | Calorimeters . . . . .                     | 41        |
| 4.2.4    | The Muon Spectrometer . . . . .            | 44        |
| 4.3      | Particle Identification . . . . .          | 46        |
| <b>5</b> | <b>Dataset and Physics Objects</b>         | <b>48</b> |
| 5.1      | Data and Monte Carlo Samples . . . . .     | 48        |
| 5.1.1    | The ATLAS Run II Dataset . . . . .         | 49        |
| 5.1.2    | Monte Carlo Samples . . . . .              | 50        |
| 5.1.3    | Emerging Jets Models . . . . .             | 52        |
| 5.1.4    | Event Weighting . . . . .                  | 55        |
| 5.2      | Physics Object Reconstruction . . . . .    | 57        |
| 5.2.1    | Track Reconstruction . . . . .             | 57        |
| 5.2.2    | Jet Reconstruction . . . . .               | 61        |

|          |                                                                      |            |
|----------|----------------------------------------------------------------------|------------|
| 5.2.3    | Displaced Vertex Reconstruction . . . . .                            | 65         |
| 5.3      | Event Selection . . . . .                                            | 67         |
| 5.3.1    | The ATLAS Trigger System . . . . .                                   | 67         |
| 5.3.2    | Emerging Jets Selections . . . . .                                   | 69         |
| <b>6</b> | <b>Emerging Jets Analysis</b>                                        | <b>71</b>  |
| 6.1      | The ABCD Method . . . . .                                            | 72         |
| 6.1.1    | Assumptions of the ABCD Method . . . . .                             | 73         |
| 6.1.2    | Correlation in the ABCD Plane . . . . .                              | 74         |
| 6.1.3    | The Likelihood Method . . . . .                                      | 76         |
| 6.1.4    | Model Independent Search Region . . . . .                            | 77         |
| 6.2      | Boosted Decision Trees . . . . .                                     | 81         |
| 6.2.1    | Binary Trees, Regression Trees, and Boosting . . . . .               | 83         |
| 6.2.2    | Emerging Jets BDT Structure . . . . .                                | 85         |
| 6.2.3    | Jet Level BDT . . . . .                                              | 87         |
| 6.2.4    | Event Level BDT . . . . .                                            | 91         |
| 6.2.5    | Model Dependent Search Region . . . . .                              | 100        |
| 6.3      | Additional BDT Trainings . . . . .                                   | 103        |
| 6.3.1    | Canonical Set (Canon15) Mixed Training . . . . .                     | 103        |
| 6.3.2    | Mediator Mass Mixed Training . . . . .                               | 104        |
| 6.4      | Validation Regions . . . . .                                         | 106        |
| 6.4.1    | 2,3-Jet Validation Region . . . . .                                  | 106        |
| 6.4.2    | 4-Jet Validation Region . . . . .                                    | 106        |
| 6.4.3    | Data - Monte Carlo Comparisons . . . . .                             | 108        |
| <b>7</b> | <b>Systematic Uncertainties and Results</b>                          | <b>115</b> |
| 7.1      | Systematic Uncertainties . . . . .                                   | 115        |
| 7.1.1    | Sources of Systematic Uncertainty on the Signal Efficiency . . . . . | 116        |
| 7.1.2    | Evaluating Systematic Uncertainties . . . . .                        | 127        |
| 7.1.3    | Systematic Uncertainty on the Background Estimate . . . . .          | 134        |
| 7.2      | Sensitivity Estimation . . . . .                                     | 141        |

|          |                                                      |            |
|----------|------------------------------------------------------|------------|
| 7.2.1    | Likelihood Fit . . . . .                             | 141        |
| 7.2.2    | Confidence Limits . . . . .                          | 142        |
| 7.3      | Results . . . . .                                    | 143        |
| 7.3.1    | Comparison of Methods . . . . .                      | 144        |
| 7.3.2    | Comparison of Training Scenarios . . . . .           | 145        |
| 7.3.3    | Including Systematics . . . . .                      | 146        |
| <b>8</b> | <b>Conclusions</b>                                   | <b>150</b> |
|          | <b>Appendices</b>                                    | <b>154</b> |
| <b>A</b> | <b>Physical Constants</b>                            | <b>155</b> |
| <b>B</b> | <b>PYTHIA Dijet Background Slices</b>                | <b>156</b> |
| <b>C</b> | <b>Jet-Level BDT Input Variables</b>                 | <b>158</b> |
| <b>D</b> | <b>Event-Level BDT Input Variables</b>               | <b>165</b> |
| <b>E</b> | <b>Model Dependent ABCD Variables</b>                | <b>175</b> |
| <b>F</b> | <b>Search Region Model Dependent ABCD Planes</b>     | <b>178</b> |
| <b>G</b> | <b>Validation Region Model Dependent ABCD Planes</b> | <b>182</b> |
| <b>H</b> | <b>Evaluation of Systematics</b>                     | <b>186</b> |

# List of Figures

|     |                                                                                               |    |
|-----|-----------------------------------------------------------------------------------------------|----|
| 2.1 | Table of the Standard Model particle content, organized by spin-statistics...                 | 9  |
| 2.2 | Feynman diagrams of two elementary QED processes...                                           | 12 |
| 2.3 | Feynman diagrams of the Drell-Yan process...                                                  | 13 |
| 2.4 | Schematic diagram of the flux tube formed by the gluon field between two separating...        | 19 |
| 2.5 | Schematic multi-step diagram of jet production...                                             | 20 |
| 3.1 | Rotation curve for the M33 galaxy...                                                          | 23 |
| 3.2 | Plot of several Standard Model particle lifetimes ( $c\tau$ ), in meters, as a function...    | 26 |
| 3.3 | Schematic diagram of the shared asymmetry between baryonic and dark matter...                 | 28 |
| 3.4 | Feynman diagrams of the pair-production of dark scalar mediators, $X_d$ ...                   | 30 |
| 3.5 | Feynman diagram of the full Emerging Jets process...                                          | 31 |
| 3.6 | Diagrammatic representation of two Emerging Jets, shown in a cross-section...                 | 32 |
| 4.1 | Schematic diagram of the ATLAS detector, with a partial cross-section of the...               | 36 |
| 4.2 | Depiction of the ATLAS Coordinate system...                                                   | 37 |
| 4.3 | Cross-section of the ATLAS Inner Detector, with the three main sub-detectors...               | 39 |
| 4.4 | Detailed Diagram of the Barrel section of the Inner Detector...                               | 40 |
| 4.5 | Cross-section of the ATLAS Calorimeters...                                                    | 42 |
| 4.6 | Cross-section of the ATLAS Detector highlighting the components of the Muon...                | 45 |
| 4.7 | Partial cross-section of the ATLAS Detector through the transverse plane...                   | 47 |
| 5.1 | Pileup ( $\langle\mu\rangle$ ) conditions for the full Run II dataset, with the components... | 49 |

|      |                                                                                        |     |
|------|----------------------------------------------------------------------------------------|-----|
| 5.2  | Theoretical pair-production cross-section ( $\sigma$ ) for varying values of dark...   | 52  |
| 5.3  | Schematic diagram showing two candidate vertices in black, the corresponding...        | 66  |
| 6.1  | Illustrative example of an ABCD plane with the (blinded) data distribution...          | 72  |
| 6.2  | Two possible arrangements of the ABCD plane...                                         | 75  |
| 6.3  | Detailed Map of the reconstructed displaced vertex density in the regions vetoed...    | 78  |
| 6.4  | Minimum PromptTrackFrac. ( $minPTF$ ) distributions (a) and power law fit (b)...       | 79  |
| 6.5  | EJs signal MC distributions in the search region defined by the model...               | 80  |
| 6.6  | QCD multijet background distribution in the search region defined by the...            | 81  |
| 6.7  | Schematic diagram of a general decision tree, showing the node structure...            | 82  |
| 6.8  | Jet-level Mass distributions for models...                                             | 88  |
| 6.9  | Jet-level Energy distributions for models...                                           | 88  |
| 6.10 | Jet-level $\eta$ distributions for models...                                           | 89  |
| 6.11 | Jet-level Width distributions for models...                                            | 89  |
| 6.12 | Jet-level Centroid-R distributions for models...                                       | 90  |
| 6.13 | Jet-level EM Fraction distributions for models...                                      | 90  |
| 6.14 | Jet-level BDT response distributions for the 1 <sup>st</sup> leading jet for models... | 92  |
| 6.15 | Jet-level BDT response distributions for the 2 <sup>nd</sup> leading jet for models... | 92  |
| 6.16 | Jet-level BDT response distributions for the 3 <sup>rd</sup> leading jet for models... | 93  |
| 6.17 | Jet-level BDT response distributions for the 4 <sup>th</sup> leading jet for models... | 93  |
| 6.18 | Event-level $H_T$ distributions for models...                                          | 95  |
| 6.19 | Event-level nDV distributions for models...                                            | 96  |
| 6.20 | Event-level nJet distributions for models...                                           | 96  |
| 6.21 | Event-level $S_T$ distributions for models...                                          | 97  |
| 6.22 | Event-level $\mathcal{A}$ distributions for models...                                  | 97  |
| 6.23 | Signal and Background correlation matrices for the event-level input variables...      | 99  |
| 6.24 | Signal and Background correlation matrices for the event-level input variables...      | 99  |
| 6.25 | Event-level BDT response distributions for models...                                   | 100 |
| 6.26 | Event-level $minPTF$ distributions for models...                                       | 101 |
| 6.27 | Signal and background MC distributions in the search region defined by the...          | 102 |



|      |                                                                                                       |     |
|------|-------------------------------------------------------------------------------------------------------|-----|
| 6.28 | Signal and background MC distributions in the search region defined by the...                         | 102 |
| 6.29 | QCD Background distribution for the Canon15 mixed BDT training...                                     | 103 |
| 6.30 | Signal MC distributions for models 312004 (left) and 312090 (right) tested...                         | 104 |
| 6.31 | QCD Background distributions for the mixed mediator mass BDT training...                              | 105 |
| 6.32 | QCD Background distribution in the full $H_T$ spectrum defined by the...                              | 107 |
| 6.33 | QCD Background distribution in the Low- $H_T$ spectrum defined by inverting...                        | 108 |
| 6.34 | Data-MC comparisons for the jet-level BDT input variables, in the 4-Jet...                            | 109 |
| 6.35 | Data-MC comparisons for the jet-level BDT Response, in the 4-Jet...                                   | 110 |
| 6.36 | Data-MC comparisons for the Event-level BDT input variables, in the 4-Jet...                          | 111 |
| 6.37 | Data-MC comparison for the $minPTF$ (on the left) and $minPTF$ as a 2-bin...                          | 112 |
| 6.38 | Validation ABCD planes in both background MC and Data using the BDT...                                | 113 |
| 7.1  | Up and down shifts given by the various nuisance parameters of the JES...                             | 118 |
| 7.2  | Jet mass response distributions for the canonical set, grouped by mediator...                         | 120 |
| 7.3  | Distribution of event-level pileup weights for the nominal reweighting...                             | 121 |
| 7.4  | Data-MC comparison for the number of primary vertices ( $\mu$ ) in the 4-Jet...                       | 122 |
| 7.5  | Trigger <i>Turn On</i> plot for the $L1.3j50$ , for model 312004...                                   | 123 |
| 7.6  | Data-MC comparison for the transverse DV position in the 4-Jet validation region...                   | 125 |
| 7.7  | Data-MC comparison for the transverse DV position in the 4-Jet validation region...                   | 126 |
| 7.8  | Comparison of the jet multiplicity distributions for the QCD multijet...                              | 127 |
| 7.9  | Evaluation of the effect of the Jet uncertainties (JES, JER, and JMS) on the...                       | 128 |
| 7.10 | Evaluation of the effect of the Jet uncertainties (JES, JER, and JMS) on the...                       | 129 |
| 7.11 | Evaluation of the effect of the Jet uncertainties (JES, JER, and JMS) on the...                       | 130 |
| 7.12 | Evaluation of the effect of the pileup reweighting on the event-level $\mathcal{R}'$ , in the...      | 131 |
| 7.13 | Evaluation of the effect of (a) the TrkVtx systematic on the DV multiplicity...                       | 131 |
| 7.14 | Evaluation of the effect of the full set of systematics on the event-level $\mathcal{R}'$ , in the... | 133 |
| 7.15 | Data distribution for $\mathcal{R}$ in the 4-Jet validation region...                                 | 134 |
| 7.16 | Background ABCD plane showing the control region transfer window method...                            | 139 |
| 7.17 | Transfer window estimate evolution...                                                                 | 140 |
| 7.18 | Transfer window estimate evolution for the validation region (a) data and (b) MC...                   | 140 |

|      |                                                                                                  |     |
|------|--------------------------------------------------------------------------------------------------|-----|
| 7.19 | Asymptotic confidence limit scan of the signal efficiency and the corresponding...               | 143 |
| 7.20 | Sensitivity plot showing the expected cross-sections (dashed lines) for each model...            | 144 |
| 7.21 | Sensitivity plot showing the expected cross-sections (dashed lines) for each model...            | 146 |
| 7.22 | Sensitivity plot showing the expected cross-sections (dashed lines) for each model...            | 147 |
| 7.23 | Sensitivity plot showing the expected cross-sections (dashed lines) for each model...            | 148 |
|      |                                                                                                  |     |
| C.1  | Jet Mass Distributions (Signal and Background) for the Canonical 15 Signal Points...             | 159 |
| C.2  | Jet Energy Distributions (Signal and Background) for the Canonical 15 Signal Points...           | 160 |
| C.3  | Jet Eta Distributions (Signal and Background) for the Canonical 15 Signal Points...              | 161 |
| C.4  | Jet Width Distributions (Signal and Background) for the Canonical 15 Signal Points...            | 162 |
| C.5  | Jet Centroid-R Distributions (Signal and Background) for the Canonical 15 Signal...              | 163 |
| C.6  | Jet EM Fraction Distributions (Signal and Background) for the Canonical 15 Signal...             | 164 |
|      |                                                                                                  |     |
| D.1  | BDT Response of the 1 <sup>st</sup> Leading Jet Distributions (Signal and Background) for the... | 166 |
| D.2  | BDT Response of the 2 <sup>nd</sup> Leading Jet Distributions (Signal and Background) for the... | 167 |
| D.3  | BDT Response of the 3 <sup>rd</sup> Leading Jet Distributions (Signal and Background) for the... | 168 |
| D.4  | BDT Response of the 4 <sup>th</sup> Leading Jet Distributions (Signal and Background) for the... | 169 |
| D.5  | $H_T$ Distributions (Signal and Background) for the Canonical 15 Signal Points...                | 170 |
| D.6  | $nDV$ Distributions (Signal and Background) for the Canonical 15 Signal Points...                | 171 |
| D.7  | $nJet$ Distributions (Signal and Background) for the Canonical 15 Signal Points...               | 172 |
| D.8  | $S_T$ Distributions (Signal and Background) for the Canonical 15 Signal Points...                | 173 |
| D.9  | $\mathcal{A}$ Distributions (Signal and Background) for the Canonical 15 Signal Points...        | 174 |
|      |                                                                                                  |     |
| E.1  | Event $\mathcal{R}$ Distributions (Signal and Background) for the...                             | 176 |
| E.2  | Minimum Prompt Track Fraction Distributions (Signal and Background) for the...                   | 177 |
|      |                                                                                                  |     |
| F.1  | Model Dependent ABCD Planes (Signal and Background) for the $m_{X_d} = 1400$ GeV...              | 179 |
| F.2  | Model Dependent ABCD Planes (Signal and Background) for the $m_{X_d} = 1000$ GeV...              | 180 |
| F.3  | Model Dependent ABCD Planes (Signal and Background) for the $m_{X_d} = 600$ GeV...               | 181 |
|      |                                                                                                  |     |
| G.1  | Model Dependent ABCD Planes (Data and MC) for the $m_{X_d} = 1400$ GeV...                        | 183 |
| G.2  | Model Dependent ABCD Planes (Data and MC) for the $m_{X_d} = 1000$ GeV...                        | 184 |

|     |                                                                                                       |     |
|-----|-------------------------------------------------------------------------------------------------------|-----|
| G.3 | Model Dependent ABCD Planes (Data and MC) for the $m_{X_d} = 600$ GeV...                              | 185 |
| H.1 | Evaluation of the effect of the full set of systematics on the event-level $\mathcal{R}'$ , in the... | 187 |

# List of Tables

|     |                                                                                                |     |
|-----|------------------------------------------------------------------------------------------------|-----|
| 3.1 | Model parameters for the particle content of the dark-QCD Emerging Jets model...               | 29  |
| 4.1 | LHC operating parameters for the Run II data taking period between 2015 and 2018...            | 35  |
| 5.1 | Leading jet transverse momentum ranges for the PYTHIA dijet slices...                          | 51  |
| 5.2 | Model parameters for the canonical set of 15 signal points used by the Emerging...             | 53  |
| 5.3 | Full set of possible combinations for scalar mediator mass $m_{X_d}$ and dark pion lifetime... | 54  |
| 5.4 | Table of Monte-Carlo parameter values for the relevant dijet slices 2-12 (above)...            | 56  |
| 5.5 | Loose and Tight Selection criteria...                                                          | 59  |
| 5.6 | Differences in the Standard and LRT track selections...                                        | 61  |
| 6.1 | Parameters used in the EJs implementation of the BDT. The first subsection...                  | 86  |
| 6.2 | Rank of the jet-level input variables, averaged across the canonical set of signal points...   | 91  |
| 6.3 | Rank of the event-level input variables, averaged across the canonical set of signal...        | 98  |
| 7.1 | Integrated trigger efficiencies for the canonical set. The table gives the total...            | 124 |
| 7.2 | Systematic uncertainty summary table...                                                        | 132 |
| 7.3 | Comparison of the different ABCD estimate methods for the yield in the...                      | 135 |
| 7.4 | Comparison of the T-test results for the uncorrected estimate ( $T_{est}$ ), the...            | 136 |
| 7.5 | Summary table of the ABCD estimate correction factor, the resulting...                         | 138 |
| 7.6 | Summary table of the region A yields for both signal and background...                         | 149 |
| 8.1 | Summary table of the current expected cross-section upper limits for the...                    | 151 |

A.1 Table of Physical Constants . . . . . 155

B.1 Table of Pythia Dijet Slices . . . . . 156

# Chapter 1

## Introduction

The Standard Model of particle physics [1 - 4] is widely regarded as one of the most successful physical theories, and has correctly predicted and described many of the observed phenomena in the field. Built from the component field theories of Quantum Chromodynamics and Electroweak theory, the Standard Model successfully describes the current set of known elementary particles, and their interactions via the fundamental forces, except for gravity. The Standard Model is experimentally tested using some of the world's most sophisticated experiments, including ATLAS [5] and CMS [6], situated at the world's largest and most powerful particle accelerator: the Large Hadron Collider [7].

Despite the success of the Standard Model, there are several observed phenomena which it fails to explain; one such example is that of dark matter [8], an exotic form of matter which is measured to make up roughly 83% of the matter density of the universe. Dark matter does not interact electromagnetically, but does interact gravitationally, and these gravitational interactions allow for astrophysical observations of its effects. Evidence comes from examining the rotation of galaxies and galactic clusters, as well as gravitational lensing, both of which suggest that there is more mass in these objects than can be accounted for by the visible matter.

Many models of dark matter have been proposed, however, no evidence of them has been experimentally observed yet. A novel experimental signature which may give an explanation for dark matter is given by Emerging Jets [9]; this model proposes a dark sector which behaves similarly to Quantum Chromodynamics, and interacts with the fields of the Standard Model via a heavy scalar mediator with mass on the TeV scale. This mediator, charged under both Standard Model

and dark sector symmetries, can be pair-produced in proton-proton collisions at the Large Hadron Collider, and then quickly decay into a Standard Model quark and a dark quark. The Standard Model quarks hadronize and form jets as usual, whereas the dark quark will hadronize in the dark sector into dark-jets composed of the lightest dark mesons, the dark pion, which are constrained to be GeV scale in order to conform with the current estimate of the dark matter density. The GeV-to-TeV hierarchy means that the dark pions decay exponentially back to Standard Model particles on the order of millimeters to meters, causing the resulting Standard Model particles to be variably displaced from the initial collision point. This results in the visible components of the dark-jets *emerging* into the detectors, over measurable distances, with many displaced vertices expected to be reconstructed with the jets.

This thesis describes how the implementation of a machine learning algorithm, in the form of a Boosted Decision Tree, into the search for Emerging Jets in the ATLAS Run II dataset (the data collected between 2015 and 2018), can improve the sensitivity to the Emerging Jets models. This work is also part of a larger ATLAS analysis which is looking for Emerging Jets, and is due for publication in early 2022. As such, the research is part of a collaborative effort, and credit must be given where it is due. The basis of the analysis, the first in ATLAS to look for the Emerging Jets signature, was developed by Dr. Colleen Treado (formerly NYU) and Dr. James Beacham (Duke). They were also responsible for the implementation of the *Emerging* selection into the RPVLL filter, described in section 5.3.2. The jet-level discriminator *PromptTrackFrac*. and its event-level counterpart, described in section 6.1.4, was developed by the team at the University of Oregon, namely Aaron Kilgallon, Prof. David Strom, and Dr. Giulia Gonella. The original two-step BDT code was developed by Jeremie Lepage-Bourbonnais (Carleton), however I am responsible for its further development and implementation into the Emerging Jets analysis. J. Lepage-Bourbonnais was also the first to examine the effect of the event-shape variables (section 6.2.4) on the Emerging Jets signal models, although I was responsible for including them into the BDT. Dr. Gonella and Dr. Mohsen Naseri (Carleton) were responsible for the majority of the development and validation in the 2,3-Jet validation region described in section 6.4.1. The Trigger study given in section 7.1.1 was conducted by Brandon Death (Carleton). The basis for the sensitivity machinery was first developed by Duncan Rocha (Duke), this was then further developed by Brandon and Mohsen into the version described in 7.2.

Apart from the components described above, and where otherwise cited, the work presented in this thesis is my own, and describes my main contributions to the analysis. These include: the development and implementation of the model dependent component of the analysis using a BDT to suppress background and increase the sensitivity to many signal models (section 6.2); the testing of different BDT training scenarios to try to reduce model dependence (section 6.3); the development, implementation, and testing of the 4-Jet validation region (sections 6.4.2 and 6.4.3) to allow for validation of the model dependent analysis; conducting extensive studies into the effect of correlations and biases on the ABCD method and the development of a correction factor; and the complete evaluation of systematic uncertainties (section 7.1) for both the model dependent and independent strategies. In addition to those listed above, I was also responsible for several crucial contributions which were not discussed in this thesis, including: the production of the original analysis ntuples for the JZ\*withSW ( $*$  = 0 – 12) slices; production of the pileup reweighting (PRW) files for the analysis ntuples produced with the Monte-Carlo campaigns mc16d and mc16e, which allowed for the application of the pileup reweighting systematic uncertainty; production of the signal and background ntuples with systematic shifts applied; and general clerical work including maintaining an ntuple spreadsheet with file location and metadata information, and file management on EOS. I am also the Rucio production coordinator for the Emerging Jets analysis, responsible for making the analysis ntuple permanent on the ATLAS Exotics GRID space.

In conjunction with the work described above, part of my responsibilities as a Masters student involved working on the small-strip thin gap chamber (sTGC) construction and characterization for the ATLAS New Small Wheel (NSW) upgrade. My contribution involved developing the sTGC Pulser Test with Dr. Stephen Weber (formerly Carleton) and Prof. Alain Bellerive (Carleton), implementing the test as part of the construction process with Prof. Jesse Heilman (Carlton), Dr. Chav Chhiv Chau (Carleton) and Dr. Mohsen Naseri, and the production of the sTGC Pulser database files with Prof. Kevin Graham (Carleton).

The thesis is organized as follows: Chapter 2 gives the main theoretical background, including an introduction to particles physics and the Standard Model. Chapter 3 gives the motivations for this search, including the major theoretical background of the analysis. Chapter 4 describes the ATLAS experiment, its component detectors, and the Large Hadron Collider. Chapter 5 describes the ATLAS Run II dataset, the Emerging Jets models and Monte Carlo signal grid, and the different



reconstructed objects used by the analysis. The main Structure of the Emerging Jets analysis is given in Chapter 6, as well as the developments and inclusion of the Boosted Decision Tree component. Chapter 7 describes the main sources of systematic uncertainty and their treatment, and gives the results of the work presented in this thesis. Finally, Chapter 8 gives the conclusions and outlook for the future of the analysis.

## Chapter 2

# Theory and Background

Unless otherwise stated, the information in this chapter is from [1 - 4].

### 2.1 The Standard Model

The Standard Model (SM) of particle physics is a theory which attempts to describe all of the elementary particles and the fundamental forces governing their interactions. It is constructed using Quantum Field Theory (QFT) and is based on the combination of several sub-theories which describe individual aspects of the model.

As with all QFTs, a Lagrangian,  $\mathcal{L}$ , is what sets the behaviour (kinematics and dynamics) of the SM. It is also a *gauge theory*, one where the Lagrangian does not change under a set of transformations, and as such is built on the local symmetries of its components:

$$SU(3) \times SU(2) \times U(1). \tag{2.1}$$

The  $U(n)$  is the *Unitary Group*, a symmetry group which can be represented as a set of  $n \times n$  matrices with the property that

$$U^\dagger U = U U^\dagger = I \tag{2.2}$$

where the dagger ( $^\dagger$ ) is the Hermitian conjugate (complex conjugate and transpose) and  $I$  is the Identity Matrix. The  $SU(n)$  is the *Special Unitary Group*, which can be represented as the set of

$n \times n$  unitary matrices with determinant 1. These three symmetry groups represent the component field theories of the SM; the *Strong Force* is given by  $SU(3)$ , and the  $SU(2) \times U(1)$  represents the *Weak Force* and *Electromagnetism* mixed together.

The SM describes 17 fundamental types of particles (see figure 2.1) which get sorted into two major groups (and several sub-groups) according to their characteristics, with the major division based on spin (intrinsic angular momentum). These groups are fermions, particles with half integer spin, and bosons, particles with integer spin.

### 2.1.1 Fermions

The fundamental fermions represent the particles which make up what is traditionally thought of as 'everyday matter' (i.e. atoms). Due to their spin-statistics, fermions obey the *Pauli Exclusion Principle*, which states that two identical fermions cannot occupy the same quantum state. The consequence of the exclusion principle is that gives rise to macroscopic matter built from composite fermion states. Fermions are subdivided further into two groups: quarks and leptons.

#### Quarks

Quarks are defined by the key characteristic that they exhibit color charge (interact with the strong force, see section 2.3), and are the only known elementary particles to interact with all four fundamental forces (strong, weak, electromagnetism and gravity). Quarks come in six varieties, known as *flavors*, which are organized into three *generations* based on mass. The first generation is composed of the two lightest quarks: the up (u) quark which has an electric charge of  $+\frac{2}{3}e$ , and the down (d) quark which has an electric charge of  $-\frac{1}{3}e$ .

All three generations of quarks contain both an up-type quark (an electric charge of  $+\frac{2}{3}e$ ) and a down-type quark (an electric charge of  $-\frac{1}{3}e$ ), where the key difference between the generations are the masses. The second generation is composed of the charm (c) and strange (s) quarks, and the third generation contains the top (t) and bottom (b) quarks (up- and down-type respectively in both cases). Since the quarks in the upper two generations have higher masses, they will readily decay into the first generation quarks, which are nominally stable (see section 2.3.1) and therefore make up most of the quark content of the universe.

## Leptons

Leptons include any elementary fermion which does not exhibit color charge and therefore does not interact with the strong force. Similar to quarks, leptons come in six flavors, and are also grouped into three generations of increasing mass. Each generation contains a charged lepton (which all have an electric charge of  $-1e$ ), and a corresponding neutral lepton known as a neutrino ( $\nu$ ). Charged leptons engage in three out of four of the fundamental forces (electromagnetism, weak, and gravity) while the neutrinos only engage in two (weak and gravity).

The first generation includes the electron ( $e$ ) which, along with the generation 1 quarks, make up the majority of the visible matter in the universe. The corresponding generation 1 neutral lepton is the electron neutrino ( $\nu_e$ ). Much like the quarks, the major difference between the three generations is the mass, which increases with increasing generation. The second generation contains the muon ( $\mu$ ) and the muon neutrino ( $\nu_\mu$ ), while the third generation is composed of the tau ( $\tau$ ) and tau neutrino ( $\nu_\tau$ ).

### 2.1.2 Bosons

The fundamental bosons mediate all of the interactions between the particles of the SM. Unlike fermions, they do not obey the Pauli exclusion principle, and therefore can occupy the same quantum state. The five elementary bosons of the SM fall into one of two categories, based on their spin. If a boson's spin is equal to one ( $s = 1$ ), then it is known as a *vector* boson. Whereas if the boson has spin equal to zero ( $s = 0$ ), then it is a *scalar* boson.

### Gauge Bosons

Four of the five bosons described by the SM are vector bosons; these are the force-carrying particles of the fundamental interactions and are therefore also *gauge* bosons. These bosons allow for interactions of particles within the fundamental force via exchange of the corresponding (*virtual*) gauge bosons. A virtual particle is one that exists for a sufficiently short time that it upholds the uncertainty principle:

$$\Delta E \Delta t \geq \frac{\hbar}{2} \quad (2.3)$$

where the maximum allowed time is given by the energy of the virtual particle.

The most familiar gauge boson is the force carrier of electromagnetism: the photon ( $\gamma$ ). The electromagnetic force, and by extension the photon, is responsible for most of the macroscopic phenomena that we experience. The photon is chargeless, and only interacts with those particles which have an electric charge (since charge is the intrinsic property of a particle which indicates how strongly it interacts with the electromagnetic force, see section 2.2.1). It is also massless, which has two consequences: the photon will always travel at the speed of light in a vacuum, and that the electromagnetic interaction has infinite range[10].

The force carrier of the strong interaction is the gluon ( $g$ ), which is responsible for keeping quarks together within composite matter particles known as *hadrons* (see section 2.3.1), which includes all protons and neutrons that make up atoms. Much like the photon, the gluon is also massless and electrically neutral. However, unlike photons, gluons both mediate and interact with the strong force, meaning they must exhibit color charge. This also means that gluons undergo self-interaction.

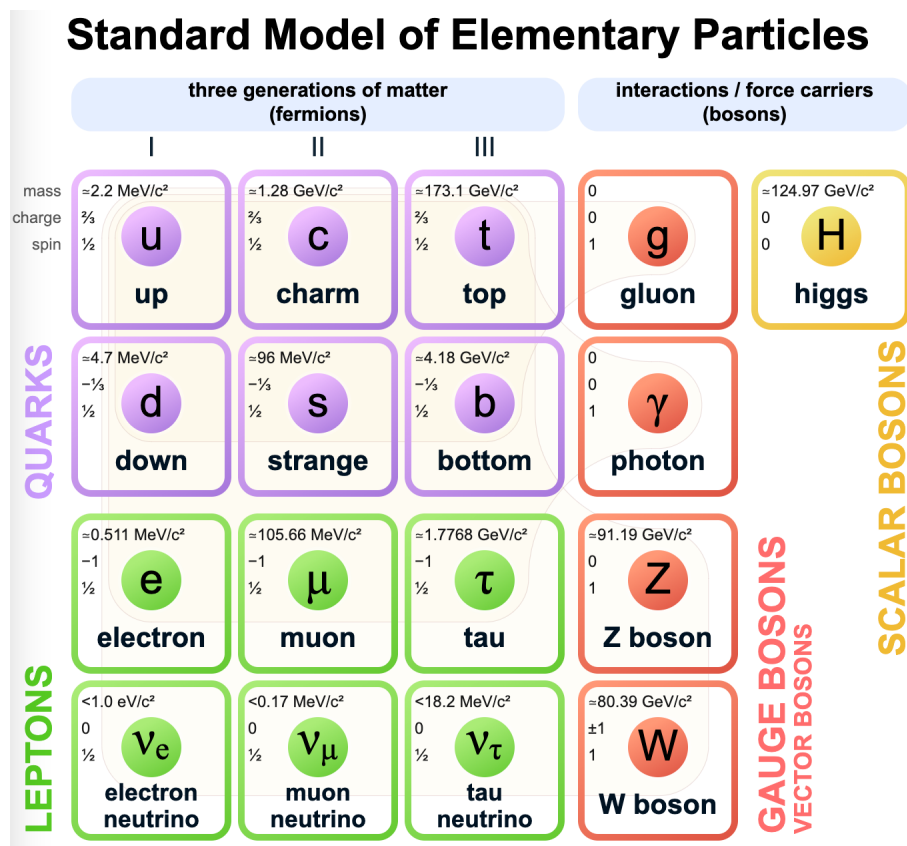
The weak force is likely the least familiar of the four forces, but it is responsible for radioactive decay, and thus plays a crucial role in the thermonuclear processes which occur in the center of the sun [11]. Unlike the other forces, it has two associated gauge bosons: the  $W^\pm$  and  $Z$  bosons. The  $W^\pm$  bosons have a charge of  $\pm 1e$ . This means that they are the only gauge bosons to interact with the electromagnetic force, and that any interaction which involves a  $W^\pm$  will change the electric charge by one unit. The  $Z$ , on the other hand, has no electric charge and can therefore only be involved in neutral interactions. Both the  $W^\pm$  and the  $Z$  have large masses, and therefore decay very rapidly. This limits the effective range of the weak interaction to very short length scales, which is what gives the force its name. The gauge bosons of the weak force can also interact with each other, so long as charge is conserved in the interactions.

## The Higgs Boson

The most recent SM particle to be discovered is the Higgs boson ( $H$ ), the only elementary scalar boson that is currently known. The Higgs boson was an important discovery as it served as evidence of the underlying Higgs field, a quantum field which is responsible for electroweak symmetry breaking (EWSB) and for giving particles their mass.

The inclusion of the Higgs field was necessary in order to keep the Standard Model consistent

with experimental evidence. Without it, the theory predicts that all the fundamental particles would be massless as a result of the symmetries of their respective field theories. However, it was known that several of these particles have mass; therefore, a mechanism was required to break the corresponding symmetries in order to explain the masses (see section 2.2.3).



**Figure 2.1:** Table of the Standard Model particle content, organized by spin-statistics (fermion-s/bosons), flavor (quark/lepton), and by generation. Each entry indicates the mass, charge, and spin of the particle in the upper left-hand corner.

In this way, analogous to how charge is the intrinsic property which indicates how strongly a particle interacts with the electromagnetic field, at the fundamental level, mass can be seen as the intrinsic property which describes how strongly a particle interacts with the Higgs field (see section 2.2.3) [12]. Unlike electromagnetism however, this only applies to particles at the elementary level; the mass of composite particles is given by their rest energy, which can be attributed to binding energy and the masses of their component particles.

### 2.1.3 Antimatter

The SM predicts that every particle has an associated antiparticle, a particle which is identical in every way except that their quantum numbers are reversed. For example, the positron (antiparticle to the electron) has the same mass as the electron but opposite electric charge. These particles are often denoted with an overline (bar), or if their charge is explicitly represented, then the sign is changed (for example:  $\bar{b}$  and  $e^+$ )

There are two classes of fermions, based on their associated antiparticle: *Majorana* fermions are those which act as their own antiparticle i.e., the particle has no quantum numbers which change sign. Alternatively, *Dirac* fermions are massive fermions which are not their own antiparticle. All the charged leptons are Dirac fermions, with their antiparticles having opposite electric charge. By convention, the charged leptons have a negative electric charge, with their antiparticles having positive electric charge. All quarks are also Dirac fermions, where their associated antiquarks have opposite electric charge and anticolor. The convention for quarks is that all up-type quarks have positive electric charge, and all down-type quarks have negative electric charge (see section 2.3 for color charge conventions). It is not currently known if neutrinos are Dirac or Majorana fermions; this represents one of the many unanswered questions in particle physics.

Unlike the fermions, the majority of the elementary bosons in the SM are their own antiparticles. This includes the photon, the  $Z$  boson, and the Higgs. Similar to the charged leptons, the two charged  $W$  bosons' antiparticles differ only by the sign of their electric charge, meaning that they are each the other's antiparticle. The gluons are more complicated; a general  $SU(n)$  gauge theory will have  $n^2 - 1$  gauge bosons which mediate the interactions (for example the 3 gauge bosons of the  $SU(2)$  weak force). So for the  $SU(3)$  strong force there are 8 gauge bosons which engage in the interaction (see section 2.3), all of which are considered gluons (as opposed to 'antigluons'). Each of these gluons in the set has an associated antiparticle, distinct to itself, that is also in the set.

## 2.2 Component Theories of the Standard Model

This section discusses the QFTs which make up the SM, with the exception of the strong interaction, which is discussed in more detail in section 2.3.

### 2.2.1 Quantum Electrodynamics

Quantum Electrodynamics (QED) is the theory which describes the fundamental interactions of electromagnetism, developed by combining Maxwell's Equations with a relativistically-invariant gauge theory (via the Dirac Equation). Simply put, QED describes how the photon interacts with charged particles.

In its simplest form, QED describes two basic processes: scattering and annihilation. Scattering, in the context of QED, describes the action in which any two charged particles interact by exchanging a photon in between them. This can include when any two like-charged particles repel each other, or when two different oppositely-charged particles attract each other. Annihilation, on the other hand, is the specific case where a particle and its associated antiparticle interact; when this happens they annihilate, and convert the entirety of their mass into energy (often in the form of photons). If this is done with enough energy (such as at a particle accelerator), the photons can pair produce a fermion and its antiparticle, even if the created particles have greater mass than the original pair.

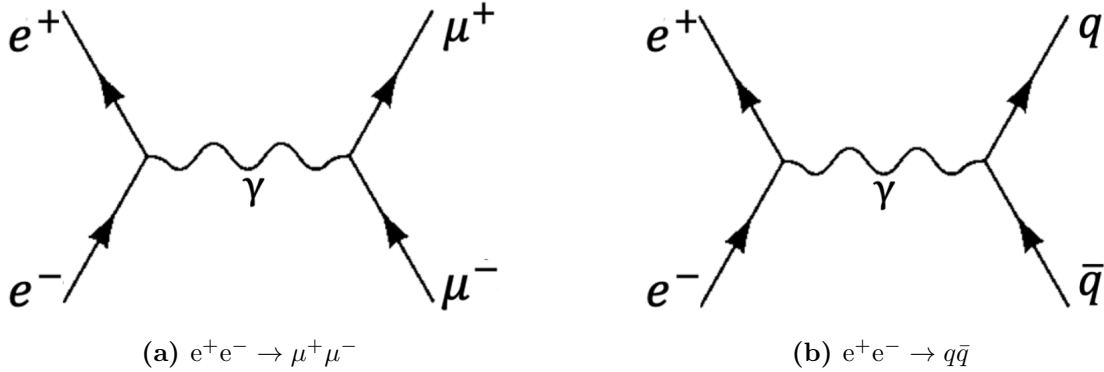
#### Coupling Constants

As the first experimentally successful quantum field theory, QED introduces several concepts which are crucial to all the component theories of the SM. One such concept is that of a *coupling constant*, a parameter which determines how strongly a particle (usually a fermion) will interact with the force-mediator (boson). The coupling constant of QED is the particle's charge,  $Q$ , so an electron will couple to the photon with a magnitude of  $e$ . These coupling constants are important because they play a role in determining *cross-sections* ( $\sigma$ ), which can be thought of as a probability for a given interaction to occur.

As an example, the cross-section for an electron and a positron annihilating to create a muon-antimuon pair (see figure 2.2a) is proportional to  $e^4$  since all four particles have charge equal to  $|e|$ :  $\sigma(e^+e^- \rightarrow \mu^+\mu^-) \propto e^4$ . Alternatively, if the electron-positron pair annihilate to form a quark-antiquark pair (see figure 2.2b), the cross-section would be:  $\sigma(e^+e^- \rightarrow q\bar{q}) \propto e^2Q^2$ , where the charge  $Q$  is that of the quarks (either  $\pm\frac{2}{3}e$  or  $\pm\frac{1}{3}e$ ). There is, of course, a lot of other information which influences a cross-section (such as the mass and energies of the particles involved), but the



coupling constant serves as a simple way to estimate the relative magnitude of the probability of a process.



**Figure 2.2:** Feynman diagrams of two elementary QED processes: (a) is the annihilation of an electron-positron pair to produce a muon-antimuon pair via a virtual photon. (b) is a similar process which instead produces a quark-antiquark pair. In both diagrams, time runs along the horizontal axis, and position runs along the vertical axis.

### 2.2.2 The Weak Interaction

The weak interaction comes in two major components: the charged and neutral current interactions. Neutral current interactions are those mediated by the  $Z$  boson. Since it has no electric charge, the  $Z$  boson can only engage in scattering processes in which the total charge of the incoming and outgoing particles is the same. Therefore, in terms of types of interactions, the  $Z$  boson behaves very similarly to the photon, or in fact as any neutral boson [13], but its large mass makes it far less likely to replace a photon in any given interaction. One such interaction where the neutral current can function in a similar manner to the photon is in high-energy processes such as the *Drell-Yan Process* [13] (see figure 2.3), where a quark-antiquark annihilate to produce a virtual photon,  $Z$  boson, or Higgs, which then pair produces leptons [13] (the process can also proceed via a charged  $W^\pm$ , but this produces a charged lepton and a neutrino [13]). However, unlike photons, the  $Z$  boson can interact with neutrinos (although very weakly) and serves as the only SM mediator for neutrino-neutrino scattering processes.

Alternatively, the charged current interactions are those mediated by the  $W^\pm$  bosons. These processes exchange particles within the fermion generations, and involve the exchange of electric

charge. Two example processes are given by:

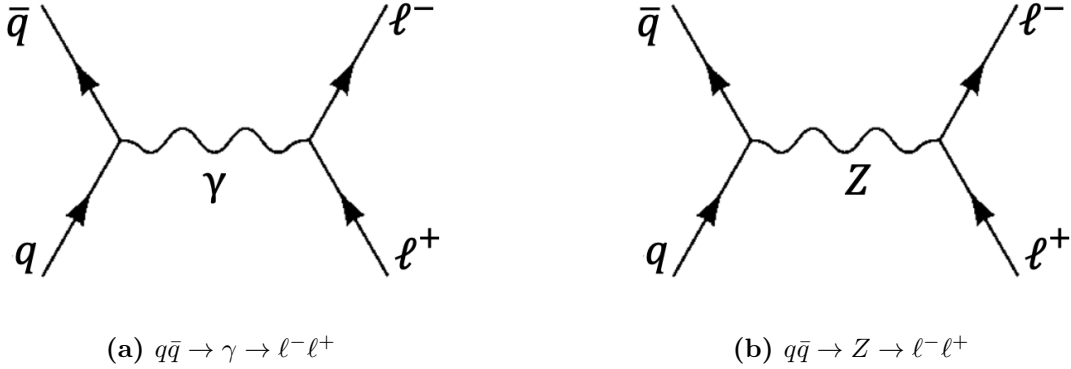
$$\mu^- \rightarrow \nu_\mu + W^- \rightarrow e^- + \bar{\nu}_e$$

$$c \rightarrow s + W^+ \rightarrow u + \bar{d}$$

where the intermediary (virtual)  $W^\pm$  boson decays to a fermion and the associated antiparticle of its generational pair such that charge is conserved.

The charged current also has the ability to exchange quarks with those outside of their generations, for example  $c \rightarrow d + W^+$ , but these interactions are suppressed (less likely) in comparison to the intragenerational mixing.

Mixing between lepton generations is forbidden due to conservation of *Lepton Flavor*, wherein every lepton generation is assigned a number and this must (in most cases, with some notable exceptions) be conserved through any process.



**Figure 2.3:** Feynman diagrams of the Drell-Yan process mediated by a photon (a) and a neutral  $Z$  boson (b). In both diagrams, time runs along the horizontal axis, and position runs along the vertical axis.

## The CKM Matrix

The probability for any quark to transition to any other quark of opposite (up/down) type is given by squaring the corresponding element in the Cabibbo-Kobayashi-Maskawa (CKM) matrix:

$$V_{CKM} = \begin{bmatrix} V_{ud} & V_{us} & V_{ub} \\ V_{cd} & V_{cs} & V_{cb} \\ V_{td} & V_{ts} & V_{tb} \end{bmatrix} \quad (2.4)$$

which, in this representation, gives the amplitude for a down-type quark to transition to an up-type quark. The diagonal elements give the transition amplitude for the intragenerational mixing and the off diagonal elements give the intergenerational mixing amplitudes. In this case, the diagonal elements all have a magnitude close to (but not equal to) 1, and the off diagonal elements all have magnitudes much less than 1.

### Chirality and Weak Isospin

An important concept, which has major implications for the weak force, is the concept of handedness; which is based on a fermion's chirality. The chirality of a fermion can be determined through action of the projection operators:

$$L = \frac{1}{2}(1 - \gamma^5)$$

$$R = \frac{1}{2}(1 + \gamma^5)$$

where  $L$  gives the left-handed component and  $R$  gives the right-handed component of the fermion field. Here  $\gamma^5 = i\gamma^0\gamma^1\gamma^2\gamma^3$  is the 'fifth' gamma matrix operator, with  $\gamma^\mu$ ,  $\mu = 0, 1, 2, 3$  being the Dirac gamma matrices. The gamma matrices are defined (in the Dirac representation) using  $2 \times 2$  block matrices as:

$$\gamma^0 = \begin{pmatrix} \mathbb{I}_2 & \mathbf{0}_2 \\ \mathbf{0}_2 & -\mathbb{I}_2 \end{pmatrix} \quad (2.5)$$

for the time-like matrix, where  $\mathbb{I}_2$  is the  $2 \times 2$  Identity matrix, and  $\mathbf{0}_2$  is the  $2 \times 2$  zero-matrix. The three space-like matrices are defined as:

$$\gamma^i = \begin{pmatrix} \mathbf{0}_2 & \sigma^i \\ -\sigma^i & \mathbf{0}_2 \end{pmatrix} \quad (2.6)$$

where the  $\sigma^i$ ,  $i = 1, 2, 3$  are the corresponding Pauli spin matrices. Therefore,  $\gamma^5$  is given by:

$$\gamma^5 = \begin{pmatrix} \mathbf{0}_2 & \mathbb{I}_2 \\ \mathbb{I}_2 & \mathbf{0}_2 \end{pmatrix} \quad (2.7)$$

which is not a true gamma matrix, but is a useful construction and crucial for determining a fermion's chirality (and therefore it's handedness).

Using this handedness, the third component of a particle's *weak isospin* ( $T^3$ ) can be determined.  $T^3$  is an additional quantum number and acts as a conserved charge for all weak interactions. As a rule, all right-handed fermions (and left-handed antifermions) have  $T^3 = 0$ . Meanwhile, left-handed fermions (and right-handed antifermions), all have  $T^3 = \pm\frac{1}{2}$ ; negatively charged fermions have negative  $T^3$ , while their generational counterparts have positive  $T^3$ , which includes both the positive quarks, and the neutral leptons, with all corresponding antifermions having opposite  $T^3$ . The charged  $W^\pm$  bosons have  $T^3 = \pm 1$  (respectively), while the  $Z$  has  $T^3 = 0$ . This means that the charged current mediators can only engage with left-handed fermions.

### 2.2.3 Electroweak Theory

Although electromagnetism and the weak force seem very different at our macroscopic level, at very high energies (such as those moments after the Big Bang [11]) the two forces are described by a unified  $SU(2) \times U(1)$  theory known as electroweak theory (EWT). This unification is characterized by an energy scale known as the electroweak scale. This determines a critical temperature,  $T_c = 159$  GeV [14], below which the theory undergoes electroweak symmetry breaking, via the Higgs mechanism, and results in the two separate forces.

EWT introduces a new quantum number, *Weak Hypercharge* ( $Y_W$ ), which is the symmetry of the  $U(1)$  component of the theory. Since the symmetries are broken by interacting with the Higgs field, neither  $Y_W$  nor  $T^3$  is totally conserved (note that weak isospin,  $T$ , is the symmetry of the  $SU(2)$  component), but the combination of the two is conserved:

$$Q = T^3 + Y_W \quad (2.8)$$

here,  $Q$  is the familiar electric charge, which is conserved even through interactions with the Higgs field.

## Electroweak Symmetry Breaking

When unified, the theory mixes together the SM gauge bosons ( $\gamma, W^\pm$ , and  $Z$ ) into massless, non-physical fields:  $W_i$  (for  $i = 1, 2, 3$ ), the three component bosons of weak isospin, and  $B$  for weak hypercharge. These gauge bosons have their own associated coupling constants,  $g$  and  $g'$ , for  $T$  and  $Y_W$  respectively.

After EWSB, the first two components of  $W_i$  become the familiar  $W^\pm$  bosons:

$$W^\pm = \frac{1}{\sqrt{2}}(W_1 \mp iW_2) \quad (2.9)$$

and acquire mass through interaction with the Higgs field. The remaining  $W_3$  boson and the  $B$  get mixed together to form two different neutral gauge bosons: one with mass and one that remains massless; these are the  $Z_\mu$  and  $A_\mu$ , the fields of the  $Z$  boson and the photon respectively. This mixing can be expressed by:

$$\begin{pmatrix} Z_\mu \\ A_\mu \end{pmatrix} = \begin{bmatrix} \cos(\theta_W) & -\sin(\theta_W) \\ \sin(\theta_W) & \cos(\theta_W) \end{bmatrix} \begin{pmatrix} W_3 \\ B \end{pmatrix} \quad (2.10)$$

where the matrix is a 2D rotation parameterized by  $\theta_W$ , the *weak mixing angle*.

It is this mixing which causes the mass difference between the three gauge bosons of the weak force. If EWT had only been described by an  $SU(2)$  symmetry, then the spontaneous symmetry breaking would have resulted in three gauge bosons: two charged and one neutral, all with the same mass. However, this would not explain the massless gauge boson of electromagnetism. The inclusion of  $U(1)$  weak hypercharge allows for the proper resultant gauge bosons, but causes an asymmetry in the weak boson masses, with their ratio given by:

$$\frac{m_W}{m_Z} = \cos(\theta_W). \quad (2.11)$$

The weak mixing angle is a useful parameterization, as  $\theta_W$  can be expressed both in terms of the EWT coupling constants and the elementary charge:

$$\cos(\theta_W) = \frac{g}{\sqrt{g^2 + g'^2}}, \quad (2.12)$$

$$\sin(\theta_W) = \frac{g'}{\sqrt{g^2 + g'^2}} = \frac{e}{g}. \quad (2.13)$$

## The Higgs Field and Standard Model Particle Masses

The Higgs field is affected by ESWB as well. Originally the scalar field,  $\phi$ , has the form of an  $SU(2)$  doublet (two component field) with a potential:

$$V(\phi) = -\mu^2 \phi^\dagger \phi + \frac{\lambda}{2} (\phi^\dagger \phi)^2 \quad (2.14)$$

where  $\lambda$  is a dimensionless coupling constant which determines the Higgs self-interaction, while  $\mu$  determines where the potential is minimized and has units of mass. If  $\mu^2 > 0$ , then the minimum of the potential is displaced from zero, which breaks the  $SU(2)$  symmetry. This causes the field to have a non-zero expectation value:

$$\langle \phi \rangle = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v \end{pmatrix} \quad (2.15)$$

where  $v$  is known as the *vacuum expectation value* for the Higgs field, given by:

$$v = \sqrt{\frac{\mu^2}{\lambda}} = \frac{1}{(\sqrt{2}G_F)^{1/2}} \approx 246 \frac{\text{GeV}}{c^2} \quad (2.16)$$

here  $G_F$  is the Fermi coupling constant.

The effect of this broken symmetry is that it leaves only one neutral, scalar boson,  $H$ , which acquires mass ( $m_H$ ), given by the self-interaction term:

$$m_H = \sqrt{2\lambda}v. \quad (2.17)$$

Since the Higgs field is responsible for the weak gauge boson masses, they too can be expressed in terms of the vacuum expectation value and a coupling constant:

$$m_W = g \frac{v}{2}, \quad (2.18)$$

$$m_Z = \sqrt{g^2 + g'^2} \frac{v}{2}. \quad (2.19)$$

Of course, the Higgs field does not couple to either electromagnetism or the strong force (the field is both electrically and color neutral), thus both the photon and the gluon remain massless.

Similarly, the mass of fermions can be expressed in terms of  $v$  and a coupling constant  $g_{f\bar{f}}$  [12], which is the general interaction term for  $H \rightarrow f\bar{f}$ , the mass of a fermion is then given by [12]:

$$m_f = g_{f\bar{f}} v. \quad (2.20)$$

If the fermion mass is well known, this can serve as a useful tool for probing the Higgs coupling to the fermion.

## 2.3 Quantum Chromodynamics

Quantum Chromodynamics (QCD) is the color-charge analog of QED and, as such, describes interactions between quarks and gluons. The theory states that all quarks are charged under one of three (physically indistinguishable) colors, usually denoted as the three primary colors: red, blue, and green ( $r, b, g$ ). With anticolor,  $(\bar{r}, \bar{b}, \bar{g})$ , being the analogous charge conjugate for QCD, meaning that antiquarks must have both the opposite electric charge and color charge. However, unlike electric charge, color charge is not an observable property, and due to a property known as *color confinement*, no colored particles can exist in nature (see section 2.3.1). The consequence of this is that quarks cannot exist in isolation. Furthermore, QCD is a *non-abelian* gauge theory; one where the elements of the symmetry group ( $SU(3)$ ) do not commute. Conversely, the  $U(1)$  symmetry group of QED is commutative, and is therefore an abelian gauge group.

Another major difference from QED is that the mediating gauge boson (the gluon) also engages in the strong interaction. In order to conserve color charge the gluon must be doubly charged with one color and one anticolor. Thus, for gluons to remain colored objects and to have both a color and anticolor charge, they must have a linear combination of different color-anticolor pairs:

$$\frac{r\bar{b} - \bar{r}b}{\sqrt{2}}. \quad (2.21)$$

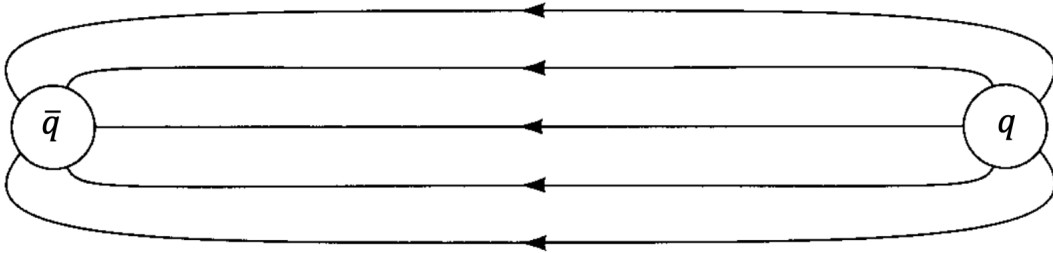
This represents one of the eight possible linearly independent combinations, corresponding to the eight physically indistinguishable gluons. These combinations are generated by:

$$\frac{\lambda_\alpha}{\sqrt{2}}, \quad \alpha = 1, \dots, 8 \quad (2.22)$$

where the  $\lambda_\alpha$  are the Gell-Mann matrices, the generalization of the Pauli spin matrices to  $SU(3)$ .

### 2.3.1 Color Confinement

Color confinement is the property of QCD which states that color-charged objects must exist in color-singlet states known as hadrons. It is the process which explains why all the particles found in nature are color neutral. In the analogy with real color (red, blue, green), confinement requires the hadrons to have the right combinations of colors which give white as the final state. There are two main classes of hadrons: *mesons* and *baryons*. Mesons are composite states of two quarks, one quark and one antiquark, which have one color charge and the corresponding anticolor charge. Baryons are three quark states (or three antiquarks) which have one of each of the three colors.



**Figure 2.4:** Schematic diagram of the flux tube formed by the gluon field between two separating color charges (quarks). Figure adapted from [1].

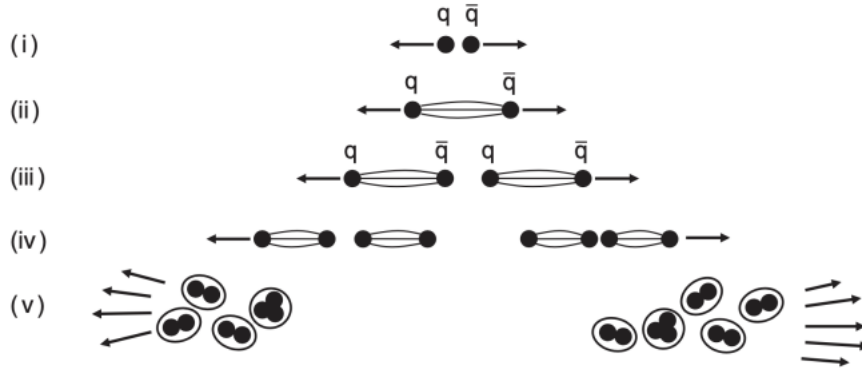
The process can be explained by the fact that gluons also carry color charge. Quarks interact by exchanging gluons, so as two quarks ( $q\bar{q}$ ) are separated, the gluon field lines (the QCD analog of the electromagnetic field) form a flux tube (see figure 2.4) which has a fixed energy density and radius. As a result, the force between the two color charges grows proportionally to the distance of the separation. Contrast this with the force between two electrically charged particles; which



is proportional to  $\frac{1}{r^2}$  (where  $r$  is the separation between the charges) [10], and therefore decreases exponentially as the distance increases. Thus, it requires increasing amounts of energy to keep the charges separated.

## Hadronization and Jets

The energy required to maintain the separation eventually exceeds the energy required to pair produce quarks. In a process known as *hadronization*, the force tube 'breaks' and this release of energy creates a quark-antiquark pair which then interact with the separating charges and leave two color neutral hadrons.



**Figure 2.5:** Schematic multi-step diagram of jet production. The quark-antiquark pair created in a high-energy particle collision undergoes repeated hadronization processes until two jets of particles recoil in the directions of the original quarks' momenta. Figure from [15].

If, instead, the two quarks ( $q\bar{q}$ ) are created (i.e. in a high energy particle collision), then their momenta will be pointing away from each other (in the center of mass (CoM) frame of reference, see figure 2.5 I). As the two quarks move away from each other, the hadronization process occurs and results in another quark-antiquark pair in between the original. This results in two new flux tubes between the new pair and the original pair (see figure 2.5 III), however, the new pair are much closer to each other than they are to the original quarks and are therefore attracted to each other as well. As the original quarks continue to separate, the hadronization process repeats, now with the two new flux tubes. This will continue until a sufficient amount of the original quarks' energy has been converted into quark-antiquark pairs, and all the quarks form hadrons. The showering and

hadronization process forms collimated sprays of particles, known as *jets*, moving in the directions of the original quark pairs (see figure 2.5 V).

### 2.3.2 Asymptotic Freedom

Along with color confinement, QCD is characterized by a phenomenon known as *asymptotic freedom*, which states that the strength of the interaction is inversely proportional to the energy transferred. This causes QCD to be *strongly coupled* at low energy, meaning that the coupling constant,  $g_s$ , is of order 1 or larger.

However, this coupling decreases as the energy scale increases (or proportionally with the distance). Thus, for very short distances (less than the width of a proton), which corresponds to large energies,  $g_s$  has a value less than 1. Whereas, at large distances (in the context of QCD, these are greater than the width of a proton) and small energies the value of  $g_s$  is much greater than 1. This is different from QED, where the coupling is weak (less than 1) at low energies, and only becomes strongly coupled at very high energies.

This change in the coupling strength of  $g_s$  is characterized by the QCD confinement scale ( $\Lambda_{QCD}$ ), with  $\mathcal{O}(\Lambda_{QCD}) \approx \mathcal{O}(m_p)$ ; at energies below  $\Lambda_{QCD}$ , QCD cannot be studied using perturbation theory (the approximation method which forms the bases of Feynman diagrams), since it requires the strength of the interactions to be small. Fortunately, perturbative QCD can still be used to study high-energy interactions, such as those from particle colliders.

### 2.3.3 Valence Quarks and the Quark Sea

Analogous to how the outermost (valence) electrons of an atom determine its properties, the quarks that determine a hadron's properties are known as *valence quarks*. However, due to gluon interactions, quark-antiquark pairs are constantly being created and annihilating within the hadron. These  $q\bar{q}$  pairs do not affect the hadron's quantum numbers, but can engage in the deep inelastic scattering processes of high-energy particle collisions.

## Chapter 3

# Physics Beyond The Standard Model

Although the Standard Model has been experimentally successful, there are still several phenomena it fails to explain. Some crucial examples include gravity, *dark matter*, and *dark energy*. Gravity is very accurately modeled by the theory of General Relativity, but all attempts to reconcile this with the Standard Model have failed. Dark matter is an exotic form of matter which is thought to make up roughly 83% of the matter density of the universe. Similarly, dark energy is an unknown source of energy which causes the expansion of the universe to accelerate and accounts for roughly 70% of the universal energy density [16]. To account for these, physicists require theories beyond the Standard Model (BSM). This thesis will not investigate gravity or dark energy, however it will investigate a dark matter model known as *Emerging Jets* (see section 3.4).

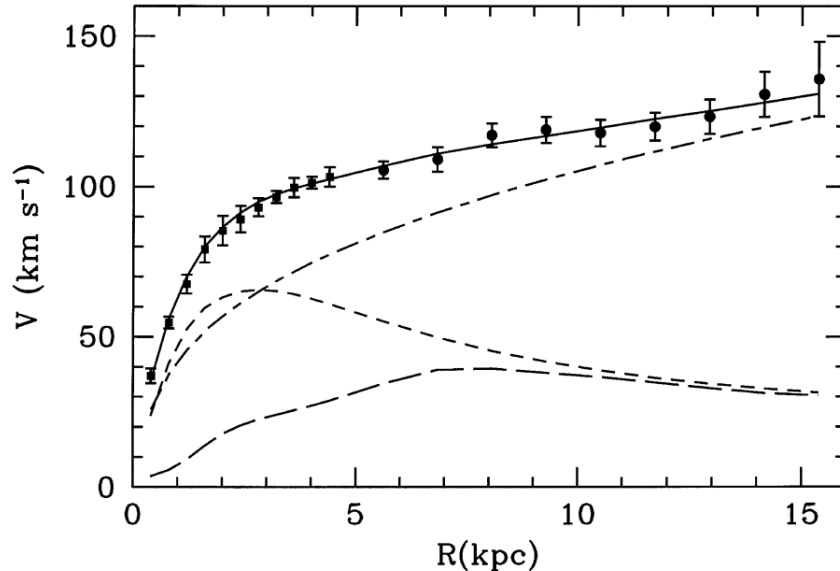
### 3.1 Dark Matter

Dark matter (DM), for the purpose of this document, is a class of matter not described by the SM and which does not interact electromagnetically. However, DM does interact gravitationally, and may, depending on the model, interact via one of the other fundamental forces or through the Higgs interaction. If it does interact with 'regular' (baryonic) matter, it must interact very weakly.

#### 3.1.1 Evidence for Dark Matter

The first evidence for DM came from astrophysical sources in the early 1930s, when astronomer Fritz Zwicky measured the rotational velocity of the Coma galaxy cluster [17]. Zwicky used the virial

theorem to determine that the luminous matter exhibited larger velocities than it should, given the corresponding gravitational potential (as measured from the luminous matter). Later evidence, pioneered by Vera Rubin [18], came from the rotation of individual galaxies; measurements of the mass density based on luminous objects suggested that most of the galactic mass resides close to the center, and decreases steadily as the radius increases [19]. Based on this distribution, the expectation was that the velocity distribution would also peak close to the center, and then decrease steadily for larger radii [19]. However, when observations of the velocity distribution were made, they found that the velocity increased steadily as the radial distance increased (see figure 3.1). This discrepancy can be explained by the addition of a dark matter halo [19]: a distribution of non-luminous matter that remains roughly uniform (radially) beyond the central bulb.



**Figure 3.1:** Rotation curve for the M33 galaxy. The data (points) and curve of best fit (solid line) are plotted in addition to the theoretical contributions from a dark matter halo (dot-dashed lines), the stellar disk (short dashed lines), and gas cloud (long dashed lines). Figure from [19].

Additional evidence came from observations of gravitational lensing [20], the phenomenon where massive objects bend the path of light around themselves, similar to an optical lens bending light to its focal point. It is important to note that lensing is a gravitational effect, not an electromagnetic one; the lens is not interacting with the light directly rather its mass curves the spacetime path which the light follows [20].

The effect of lensing is difficult to observe unless the lens (the massive object) is very large, and the effect is proportional to the mass; a more massive lens will cause the path of light from the source to bend more [20]. The effect of lensing causes distorted or repeated images of the source. Surveys of the night sky show several distant galaxies being lensed without enough visible matter to account for the distortions [21].

Between these and several other sources of observational/experimental evidence, a compelling argument was made for some non-luminous, gravitationally-interacting matter, which could not be explained using the SM.

### 3.1.2 The WIMP Model of Dark Matter

Determining the existence of dark matter and actually identifying it are two different challenges. Many candidates have been proposed to try and explain the various observational evidence; however, no model has been experimentally confirmed.

One of the earliest and most prolific models of DM are the *Weakly Interacting Massive Particles* or WIMPs. This model proposes a new elementary particle (which must interact with gravity), with the defining feature that their interactions must be roughly of the same order as the weak force, or weaker [22]. This would exclude any electromagnetic interactions which is consistent with observation, but also excludes any strong interactions. The WIMP model does not predict any one candidate for DM, but rather proposes a class of possible candidates which would help explain the observational evidence [22]. WIMP candidates would be so-called *cold dark matter*, where cold here indicates that the particles were non-relativistic when they *froze out* of thermal equilibrium in the early universe. WIMPs would also fit the estimated *relic density*: the relative abundance of dark matter particles which were created in the Big Bang [22].

Due to the WIMP model's generality, it has been able to integrate with several other theories to give more specific DM candidates [8]. One such theory is *Supersymmetry*, which predicts that each SM particle has a partner with similar properties but opposite spin-statistics (i.e. fermions become bosons and vice-versa) and larger masses [23]. It has been proposed that the lightest, stable superpartner could be a candidate for dark matter [8,23].

Despite its flexibility, there has yet to be any experimental confirmation of the WIMP model, even after many years of searching. These null results do not exclude WIMPs [24]; in fact there

remain several *Direct-Detection* experiments, detectors which look for evidence of nuclear recoil from collisions with DM particles [22], which are searching for WIMPs. However, the lack of evidence for WIMPs does allow for increasingly complex models of DM to be explored with more confidence.

## 3.2 Hidden Valley Physics

The *hidden valley* or *hidden sector* refers to an extension of the SM using an additional gauge group with its own particles and interactions [25]. These new particles would have no interactions with the forces of the SM, and therefore be ‘invisible’ to all the SM particles. Likewise, all SM particles would be neutral under these new charges (the gauge symmetries of the hidden sector)[25].

The hidden valley is connected to the SM via a *mediator* ( $X$ ), a new particle which is charged under both SM and hidden symmetries, allowing it to interact with particles from both sectors [25]. These mediator particles would likely be heavy bosons with masses  $m_X \geq m_H$ , in order to be consistent with experiment [26].

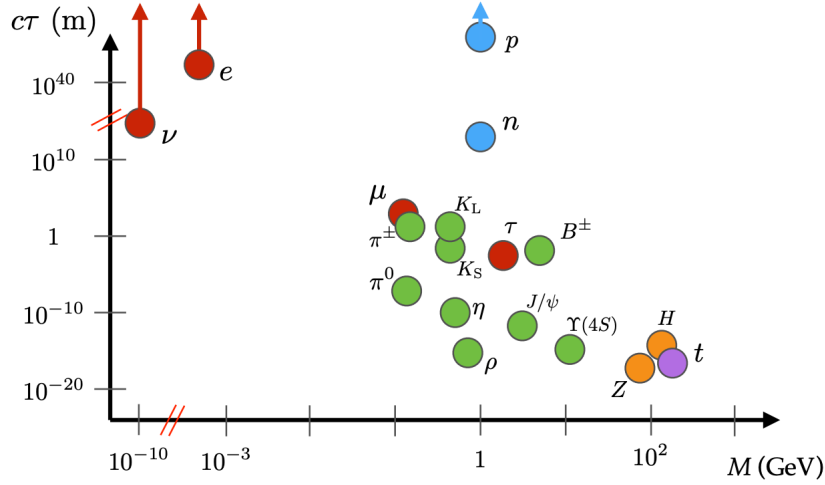
This framework provides a natural extension to the SM which can be used to try and explain various BSM phenomena. For example, if the hidden sector was the dark sector (i.e. the new particles were DM particles), then the hidden valley gives a novel avenue to search for DM [27], beyond the traditional Direct-Detection experiments.

## 3.3 Long-Lived Particles

A long-lived particle (LLP) is any fundamental massive particle whose lifetime is sufficiently long that it decays within a macroscopic distance, generally  $c\tau \geq \mathcal{O}(\mu m)$  [28]. Here,  $\tau$  is the proper lifetime for the particle, and the multiplication by  $c$  converts the value to a distance. For relativistic particles produced in collider experiments, whose velocities are very close to the speed of light, this gives a good approximation for their decay length.

This definition includes many SM particles such as muons and several hadrons with  $c\tau$  of  $\mathcal{O}(m - km)$ , and those particles which are stable for all practical purposes, the *ultra*-long lived particles (i.e. electrons, protons and neutrinos) [28]. This is in contrast to those particles such as

the  $Z$  or the Higgs whose lifetimes are so short that they decay in practically the same position that they were created [28].



**Figure 3.2:** Plot of several Standard Model particle lifetimes ( $c\tau$ ), in meters, as a function of their mass, in GeV. Vertical arrows indicate value is a lower limit on stable particle lifetimes. Figure from [29].

LLPs are also a common feature of many BSM theories, usually as a natural consequence of a component of the theory, for example: small interaction strengths, variable mass scales, or weak couplings to lighter particles [28-31]. If these new LLPs decay to SM particles within a measurable scale (generally  $\mathcal{O}(10^{-4} - 10^7 m)$ ), then they provide a novel tactic with which to search for BSM physics [31]. Depending on their mass and lifetime, charged (either electrically or colored) LLPs may be difficult to distinguish from SM particles in experimental searches since the types of signatures they would produce would be very similar; although, sufficiently massive charged LLPs may be a distinguishable signature in high-energy particle colliders [32]. Alternatively, neutral LLPs would give relatively unique decays; the LLP itself would not interact with the detectors of the experiment, but the decay products would. This leaves a signal in the detectors that is displaced from the origin of the LLP [28-31].

### 3.4 Emerging Jets

The *Emerging Jets* (EJs) theory presents a novel model for dark matter which combines both long-lived particles and the hidden valley. The model proposed by Schwaller, Stolarski, and Weiler

in [9] represents the theoretical basis for this section, unless expressly stated otherwise.

The theory assumes a QCD-like dark sector, which interacts with the SM via a heavy scalar mediator ( $X_d$ ). These dark mediators interact with both SM quarks and dark quarks ( $Q_d$ ), and like QCD, the dark quarks shower and hadronize into composite states of dark mesons and baryons.

As the lightest stable state, the dark analogue of the proton would serve as an excellent candidate for dark matter; that is, all the composite states of the dark sector would technically be dark matter, but as a stable particle, the dark proton ( $p_d$ ) would help explain the astrophysical and cosmological evidence. The dark proton's stability comes from conservation of *dark baryon number* ( $B_d$ ), the dark analogue of the conserved SM *baryon number* ( $B = \frac{1}{3}(n_q - n_{\bar{q}})$ ), where all baryons have  $B = 1$ , antibaryons have  $B = -1$  and everything else has  $B = 0$  [2]. The conservation of  $B$  (or  $B_d$ ) ensures that the lightest baryon does not decay; for the case of dark-QCD this means that the dark proton will not decay into SM particles.

Alternatively, many of the dark hadrons will be unstable, and therefore decay via the dark mediator into SM quarks. Some of these dark hadrons would decay very quickly, but others would have non-trivial lifetimes causing them to act as LLPs.

### 3.4.1 Dark QCD Model Parameters

In order for the dark-QCD model to be consistent with observational evidence, certain aspects of the theory must be constrained: namely, the dark matter number density ( $n_{DM}$ ) and energy density ( $\Omega_{DM}$ ).

For the hidden valley implementation of Emerging Jets, a new  $SU(N_d)$  symmetry group is added to the SM, where  $N_d$  is the number of dark colors, with  $N_d \geq 2$ . The new dark sector is confined by an energy scale ( $\Lambda_d$ ), analogous to the QCD confinement scale. If this scale shares a similar relationship with the dark proton, then we can say  $\Lambda_d \approx m_{p_d}$ . It is assumed that dark matter exhibits the same asymmetry as baryonic matter (i.e. there is more matter than antimatter), and if the mechanism which causes the asymmetry is common to both, then it can be assumed that the dark matter number density is similar to the baryonic number density:  $n_{DM} \approx n_B$ . The current estimate for  $\Omega_{DM}$  from [33] gives the ratio:

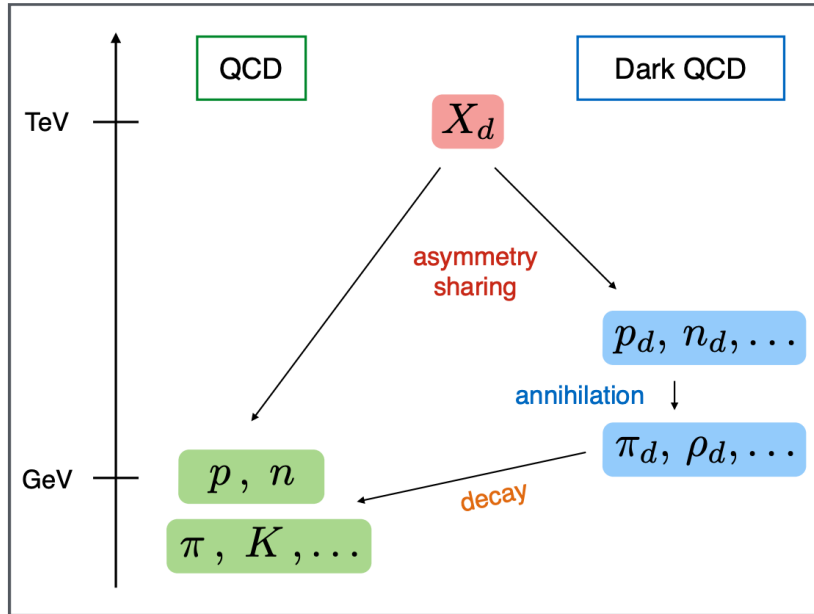


$$\frac{\Omega_{DM}}{\Omega_B} = 5.36 \pm 0.05. \quad (3.1)$$

This gives an estimate for the dark proton mass since:

$$\frac{\Omega_{DM}}{\Omega_B} = \frac{m_{p_d} \cdot n_{DM}}{m_p \cdot n_B} \approx \frac{m_{p_d} \cdot n_B}{m_p \cdot n_B} = \frac{m_{p_d}}{m_p} \approx 5. \quad (3.2)$$

Thus,  $m_{p_d} \approx 5 \cdot m_p$ . This, gives an approximation for the dark confinement scale:  $\Lambda_d \approx \mathcal{O}(1 - 10 \text{ GeV})$ . This is the range in which the majority of the dark hadron masses reside.



**Figure 3.3:** Schematic diagram of the shared asymmetry between baryonic and dark matter for a TeV scale dark mediator. The dark baryons decay down to dark mesons (pions and rhos) which eventually decay down to SM hadrons. Figure from [9].

The lightest dark hadron is the corresponding pseudo-Goldstone boson of dark-QCD, the dark pion ( $\pi_d$ ). A Goldstone boson is a massless scalar boson whose corresponding symmetry is spontaneously broken; for example the three  $W_i$  bosons of electroweak theory. If instead the symmetry is both spontaneously and explicitly broken, then the boson acquires mass and is a pseudo-Goldstone boson. In QCD the pion (the lightest meson) is the pseudo-Goldstone boson of chiral symmetry, and is responsible for mediating the strong force between the nucleons of an atom [1].

Being the lightest hadrons these dark pions would have masses with  $m_{\pi_d} < \Lambda_d$ , and being mesons there is no conserved quantum number to ensure their stability [2], so they will decay back to SM particles with a lifetime given by:

$$c\tau = \frac{c\hbar}{\Gamma(\pi_d \rightarrow d\bar{d})} \approx \frac{80mm}{\kappa^4} \left( \frac{2 \text{ GeV}}{f_{\pi_d}} \right)^2 \left( \frac{100 \text{ MeV}}{m_d} \right)^2 \left( \frac{2 \text{ GeV}}{m_{\pi_d}} \right) \left( \frac{m_{X_d}}{1 \text{ TeV}} \right)^4 \quad (3.3)$$

where,  $\Gamma(\pi_d \rightarrow d\bar{d})$  is the decay width for a dark pion to decay to a  $d\bar{d}$  pair,  $\kappa$  is the Yukawa coupling,  $f_{\pi_d}$  is the dark pion decay constant,  $m_d$  is the mass of the down quark, and  $m_{X_d}$  is the mass of the dark mediator.

Current cosmological models constrain the dark pion lifetime to be less than  $\mathcal{O}(1s)$ , which gives a rough upper limit on the dark mediator mass:  $m_{X_d} \leq 100 \text{ TeV}$ . However, in order to have comparable confinement scales between dark and SM QCD ( $\Lambda_d \approx \Lambda_{QCD}$ ), a stricter limit on the dark mediator mass is applied:  $m_{X_d} \approx \mathcal{O}(\text{TeV})$ . Therefore, dark pions with masses  $m_{\pi_d} \approx \mathcal{O}(\text{GeV})$  will have lifetimes which give decay lengths of  $\mathcal{O}(10^{-2} - 1m)$ .

| Field | $SU(3) \times SU(2) \times U(1)$ | $SU(N_d)$ | Mass                      | Spin          |
|-------|----------------------------------|-----------|---------------------------|---------------|
| $Q_d$ | $(1, 1, 0)$                      | $N_d$     | $\mathcal{O}(\text{GeV})$ | $\frac{1}{2}$ |
| $X_d$ | $(3, 1, \frac{1}{3})$            | $N_d$     | $\mathcal{O}(\text{TeV})$ | 0             |

**Table 3.1:** Model parameters for the particle content of the dark-QCD Emerging Jets model.

For the purpose of this document we will take  $N_d = 3$  and use the parameters listed in Table 3.1 as the basis for the Emerging Jets model. For a dark mediator with hypercharge  $\frac{1}{3}$ , the Lagrangian for the Yukawa coupling is given by:

$$\mathcal{L}_\kappa = \kappa_{ij} \bar{Q}_{di} q_j X_d + h.c. \quad (3.4)$$

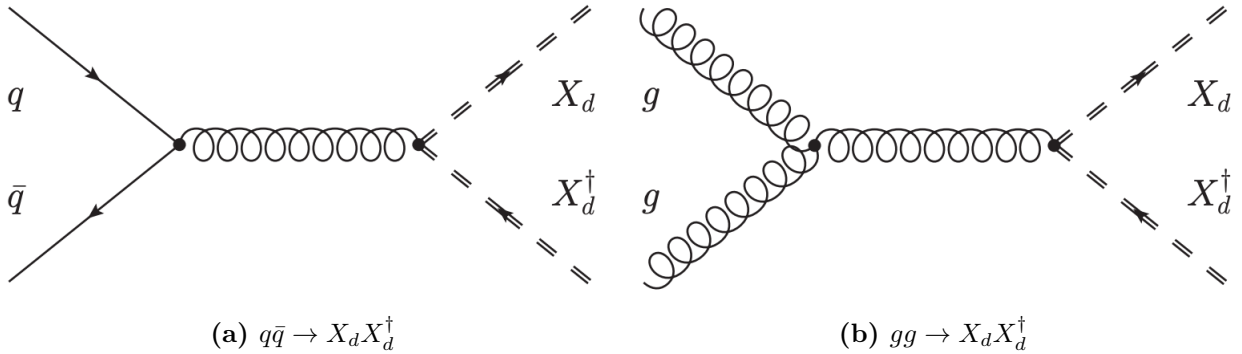
where  $\kappa_{ij}$  is the  $n_f \times 3$  matrix of Yukawa couplings (where  $n_f$  is the number of dark flavors), the  $q_j$  are the left-handed down-type SM quarks, and  $h.c.$  represents the Hermitian conjugate term. Thus, the full Lagrangian for dark-QCD is given by:

$$\mathcal{L} \supset \bar{Q}_{di}(\not{D} - m_{Q_{di}})Q_{di} + (D_\mu X_d)(D^\mu X_d)^\dagger - m_{X_d}^2 X_d X_d^\dagger - \frac{1}{4} G_d^{\mu\nu} G_{\mu\nu,d} + \mathcal{L}_\kappa + \mathcal{L}_{SM} \quad (3.5)$$

where  $G_d^{\mu\nu}$  is the field strength tensor for the dark gluon,  $D_\mu$  is the covariant derivative which contains the couplings to gauge fields, with:  $\not{D} := \gamma^\mu D_\mu$ , and  $\mathcal{L}_{SM}$  is the full SM Lagrangian.

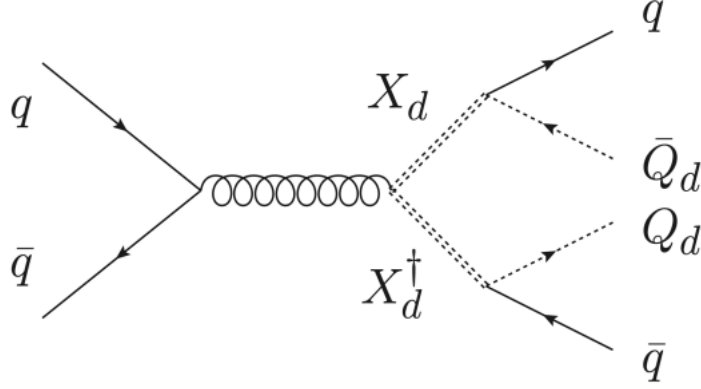
### 3.4.2 Emerging Jets Phenomenology

The EJs model would give a readily discernible signature in a high-energy particle collider. A TeV scale dark mediator could be pair-produced via a intermediary gluon in hadronic collisions,  $pp \rightarrow X_d^\dagger X_d$  (see figure 3.4). These mediators would instantly decay into a light SM quark and a dark quark,  $X_d^\dagger X_d \rightarrow \bar{q} Q_d \bar{Q}_d q$ , which can all be thought of as a single process (see figure 3.5).



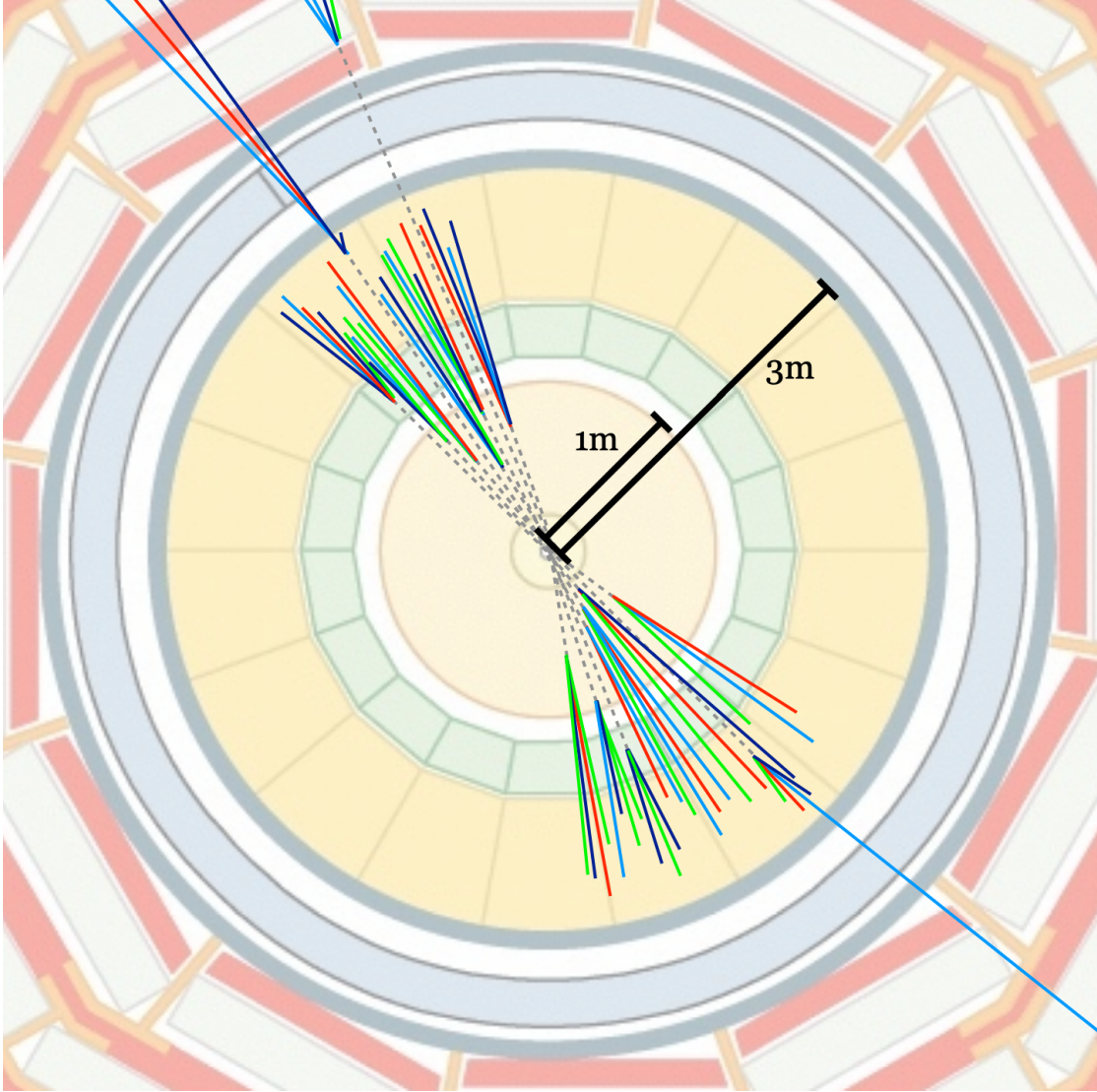
**Figure 3.4:** Feynman diagrams of the pair-production of dark scalar mediators,  $X_d$ , from initial quark (a) and gluon (b) states. Figures from [9]

The SM quarks will hadronize into QCD jets as usual, whereas the dark quarks will hadronize in the dark sector on a time scale of  $\Lambda_d^{-1} \ll \tau_{\pi_d}$ . This process of dark-hadronization is theoretically very similar to QCD hadronization, and produces dark jet-like objects which are invisible to the detectors. Much like QCD, the production of dark baryons is suppressed in relation to the production of dark mesons, so the dark jets will be built largely from unstable states. Heavier dark mesons with masses of  $\mathcal{O}(\Lambda_d)$  will decay down to the lightest dark mesons,  $\pi_d$ , also on the time scale of  $\Lambda_d^{-1}$ . This means that the dark jets will be composed mainly of dark pions.



**Figure 3.5:** Feynman diagram of the full Emerging Jets process. A dark mediator pair is produced from a virtual gluon, which each decay to a SM and dark quark pair. Figure from [34].

These dark pions eventually decay to SM quarks, exponentially, over length scales characterized by  $\beta\gamma c\tau_{\pi_d}$ , where  $\beta\gamma$  is the momentum-dependent boost factor for each pion. This causes each decay to be at varying distances from the initial interaction. As the dark pions decay, they produce smaller sub-jets, which contain a few SM hadrons all originating at the same displaced vertex (DV). As more decays occur, the sub-jets begin to merge into an object that looks very similar to a standard QCD jet, but which does not originate at the interaction point (the position of the  $X_d$  decay). The final result, as shown in figure 3.6, is a jet which is comprised of many DVs. As one examines the radial profile, the jet contains very little visible energy close to the interaction point, and gradually increases as the radius increases, thus is said to emerge into the detectors.



**Figure 3.6:** Diagrammatic representation of two Emerging Jets, shown in a cross-section perpendicular to the beam-line. The initial interaction is located at the center with a general detector layout around it. The dark hadrons are represented by dashed lines, and the solid colored lines represent the SM hadrons. Note that this schematic does not include the SM jets which would nominally be produced by the process given in 3.5 Figure from [9].

## Chapter 4

# The ATLAS Experiment

### 4.1 The Large Hadron Collider

The Large Hadron Collider (LHC), located at the European Organization for Nuclear Research (CERN), is the world's largest particle accelerator, capable of reaching the highest energies of any particle collider on record. Straddling the Franco-Swiss border, and roughly 100 m underground, the LHC consists of two 27 km super-conducting beamlines. Housed within a single ring, each beamline accelerates proton beams in opposite directions, and collides them at one of four points along the ring, at a center-of-mass energy of up to 14 TeV [35][36]. Instead of continuous beams of single protons, the beams are made from bunches of roughly 100 billion protons [37]. Several thousand bunches circulate around the rings, spaced out such that the collision rate is high, but not so high that the detectors at the collision points are overwhelmed.

The LHC itself is supplied by a successive series of accelerators which start with ionized hydrogen gas, stripped of the electrons, that bring the proton bunches to higher and higher energies [36]. By the time the bunches are injected into the LHC ring, each bunch has been accelerated to 450 GeV, the LHC then brings them up to their final collision energy [36]; for Run II (the data taking period from 2015 to 2018), each proton bunch circulated at 6.5 TeV, such that when two bunches collided, they had a center-of-mass energy of 13 TeV [37].

Around the four collision points are the four major experiments of the LHC: ATLAS, CMS, ALICE, and LHCb [35]. The two largest, ATLAS (A Toroidal LHC ApparatuS, see section 4.2) and CMS (Compact Muon Solenoid), are general purpose experiments, designed to discover the

Higgs Boson, which they did in 2012 [38], to test the Standard Model, and to look for evidence of new physics (BSM). The other two detectors are more specialized; ALICE (A Large Ion Collider Experiment) which takes advantage of special LHC runs where lead nuclei are accelerated instead of protons, and LHCb (LHC Beauty) which specializes in studying B mesons (mesons which contain b quarks).

Like all particle accelerators, the LHC is characterized by certain beam parameters. One such parameter is luminosity ( $\mathcal{L}$ ), which is given by the ratio of the event rate ( $\frac{dN}{dt}$ ) to the cross-section [39].

$$\mathcal{L} = \frac{1}{\sigma} \frac{dN}{dt}. \quad (4.1)$$

This takes the units of  $m^{-2}s^{-1}$ , although is more commonly quoted in units of barns ( $b = 10^{-28}m^2$ ), therefore,  $[\mathcal{L}] = [b^{-1}s^{-1}]$ . The luminosity can be integrated with respect to time to get the *integrated luminosity*:

$$\mathcal{L}_{int} = \int \mathcal{L} dt = \frac{N}{\sigma} \quad (4.2)$$

which gives the expected number of events for a certain cross-section [39].

Increasing integrated luminosity is crucial for optimizing physics programs of accelerators like the LHC, as it provides the experiments with more data for analysis. This, along with higher beam energies give the best chances for discovering new physics. The LHC parameters for Run II are shown in table 4.1.

The two general purpose detectors, ATLAS and CMS, are designed to take advantage of the full LHC luminosity; however, this presents a problem known as *pileup* ( $\mu$ ). When proton bunches collide, there is a chance that multiple protons in the bunch will interact, producing several low energy (soft) interactions in conjunction with the main hard scatter event. The average number of interactions per bunch crossing,  $\langle\mu\rangle$ , gives a rough measure of the number of primary interactions (vertices) in an event ( $N_{PV}$ ), which is known as *in time* pileup. Since the response time of the detectors is longer than the bunch spacing, there may also be interactions from subsequent bunch crossings which get reconstructed with the initial event; this is known as *out of time* pileup.

| Parameter                                  | Run II Value          |
|--------------------------------------------|-----------------------|
| Bunch Energy [TeV]                         | 6.5                   |
| Center of Mass Energy ( $\sqrt{s}$ ) [TeV] | 13                    |
| Protons per Bunch                          | $1.15 \times 10^{11}$ |
| Average Number of Bunches                  | 2470                  |
| Bunch Spacing [ns]                         | 25                    |
| Collision Rate [MHz]                       | 40                    |
| Design Luminosity [ $cm^{-1}s^{-1}$ ]      | $1.0 \times 10^{34}$  |
| Peak Luminosity [ $cm^{-1}s^{-1}$ ]        | $2.1 \times 10^{34}$  |
| Integrated Luminosity [ $fb^{-1}$ ]        | 156                   |

**Table 4.1:** LHC operating parameters for the Run II data taking period between 2015 and 2018 [37][40][41].

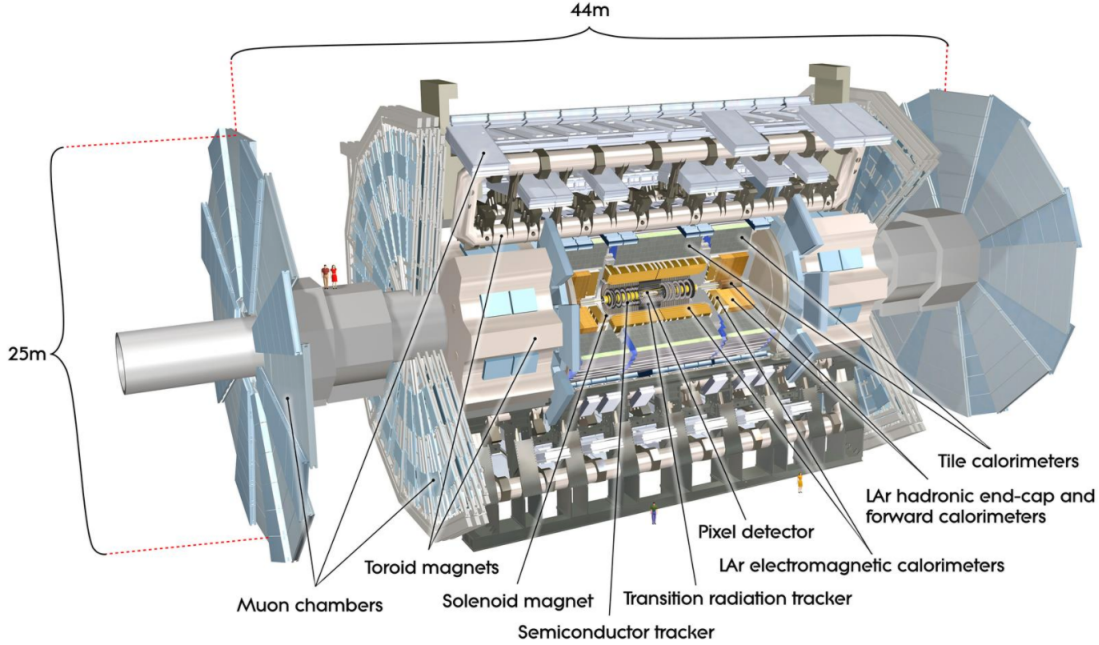
On average, only one proton collision in a given bunch crossing, if any, will correspond to an interesting physics event. Both in time, and out of time pileup represent noise in the detectors, which make it more difficult for experiments to reconstruct the underlying event. It should be noted that multiple parton (quark/gluon) interactions from a single proton are not considered pileup.

## 4.2 The ATLAS Experiment

The ATLAS experiment is the largest of the four experiments on the LHC, with a cylindrical design measuring 44 m across and 25 m in diameter. The experiment is built from layers of different specialized detectors, centered on the interaction point, which measure certain aspects of the collision products. Unless otherwise stated, the information for this section is from [42 - 51].

At the center of ATLAS is the Inner Detector (see section 4.2.2); a collection of tracking detectors which are designed to track and measure the properties of outgoing particles. After the tracking detectors are the Calorimeters (see section 4.2.3), which are designed to stop most of the collision products and measure their energy. The outermost detectors are the Muon Spectrometers (see section 4.2.4), designed to track and measure the energy of muons, which can pass through the calorimeters without being stopped.



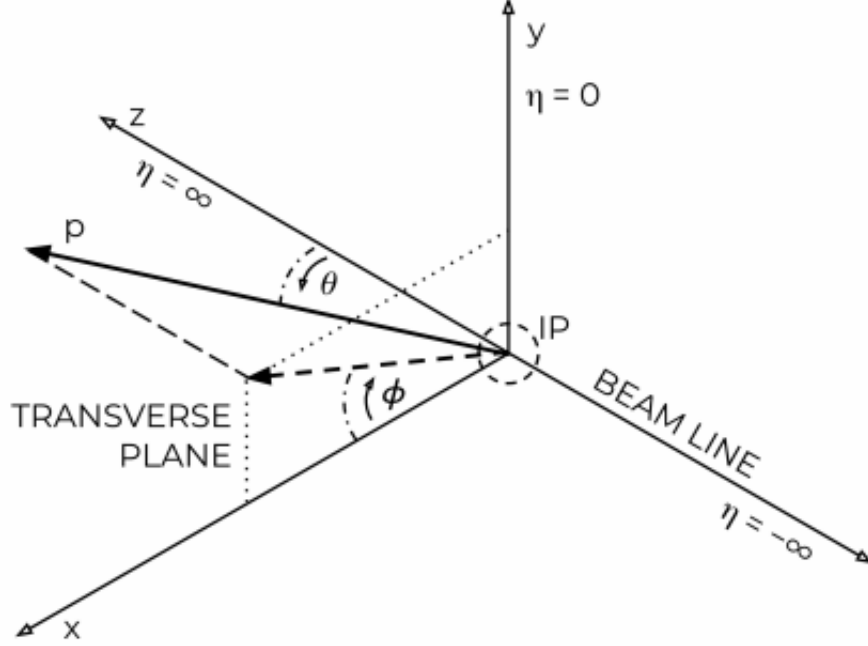


**Figure 4.1:** Schematic diagram of the ATLAS detector, with a partial cross-section of the central detectors. Several key components are highlighted, with humans shown for scale. Figure from [42].

Incorporated between the layers of detectors are the ATLAS Magnet Systems. ATLAS employs both an inner solenoid and outer toroidal magnet to bend the path of charged particles, which allows the detectors to measure their momenta. Located between the Inner Detector and the calorimeters is the 2T solenoid which provides the uniform magnetic field required by the tracking detectors. The toroidal magnet system, which gives ATLAS its name, sits between the calorimeters and the Muon Spectrometers, and is made up of three main components: eight toroidal coils which extend the full length of the central detectors, and two toroids at either end of the detectors. Due to its large size and geometry, the toroid provides a non-uniform magnetic field, and thus its field strength varies across its enclosed volume, with an average value of 4T.

#### 4.2.1 The ATLAS Coordinate System

ATLAS uses a right-handed cartesian coordinate system, centered on the interaction point, as shown in figure 4.2. The z-axis runs along the beamline, with the x and y axes defining the transverse plane (the plane perpendicular to the beamline). The azimuthal angle ( $\phi$ ) wraps clockwise around



**Figure 4.2:** Depiction of the ATLAS Coordinate system. The three principal axes are shown, with the x-axis pointing into the LHC ring, and the y-axis pointing up out of the top of ATLAS. A sample particle's three-momentum vector is shown with the bold arrow. Figure from Dr. S. Weber.

the z-axis, while the polar angle ( $\theta$ ) is defined from the +z-axis to the -z-axis. Due to its shape, ATLAS often uses the cylindrical coordinate system, which employs  $r = \sqrt{x^2 + y^2}$  as the radial distance from the interaction point.

Instead of using the standard polar angle, it is more common to use the *pseudorapidity* ( $\eta$ ):

$$\eta = -\ln \left[ \tan \left( \frac{\theta}{2} \right) \right] = \frac{1}{2} \ln \left( \frac{|\mathbf{p}| + p_L}{|\mathbf{p}| - p_L} \right) \quad (4.3)$$

where  $\eta = 0$  along the transverse plane, with  $|\eta| \rightarrow \infty$  along the beamline. Here,  $p_L$  is the longitudinal (along the beamline) component of the particle's momentum (the dashed line in figure 4.2). Pseudorapidity is an approximation for the true rapidity:

$$y = \frac{1}{2} \ln \left( \frac{E + p_L}{E - p_L} \right) \quad (4.4)$$

where in the limit that the particle is massless, or that  $m \ll |\mathbf{p}|$  (which is the case for most collision products),  $\eta \approx y$ .

Rapidity and pseudorapidity are important measures in particle colliders since  $\Delta y$  and  $\Delta \eta$  are Lorentz invariant for boosts along the beamline. In general, the pseudorapidity is preferred since it depends on the particle's momentum, not its energy. Furthermore, pseudorapidity is useful as it relates to a particle's transverse momentum,  $p_T$ , or the component of its momentum perpendicular to the beamline:

$$p_T = \frac{|p|}{\cosh \eta} \quad (4.5)$$

which is shown by the dashed arrow in figure 4.2.

Since differences in  $\eta - \phi$  are invariant under longitudinal boosts, ATLAS generally uses a standard definition for angular separation between two objects given by:

$$\Delta R = \sqrt{(\Delta \eta)^2 + (\Delta \phi)^2}. \quad (4.6)$$

ATLAS can generally be split into two main sections: the Barrel section, generally defined within  $|\eta| < 2.5$ , which covers the central region of the detector, and the End-Cap, for  $|\eta| > 2.5$ , which covers the 'forward' region of the detector (the area close to the beamline).

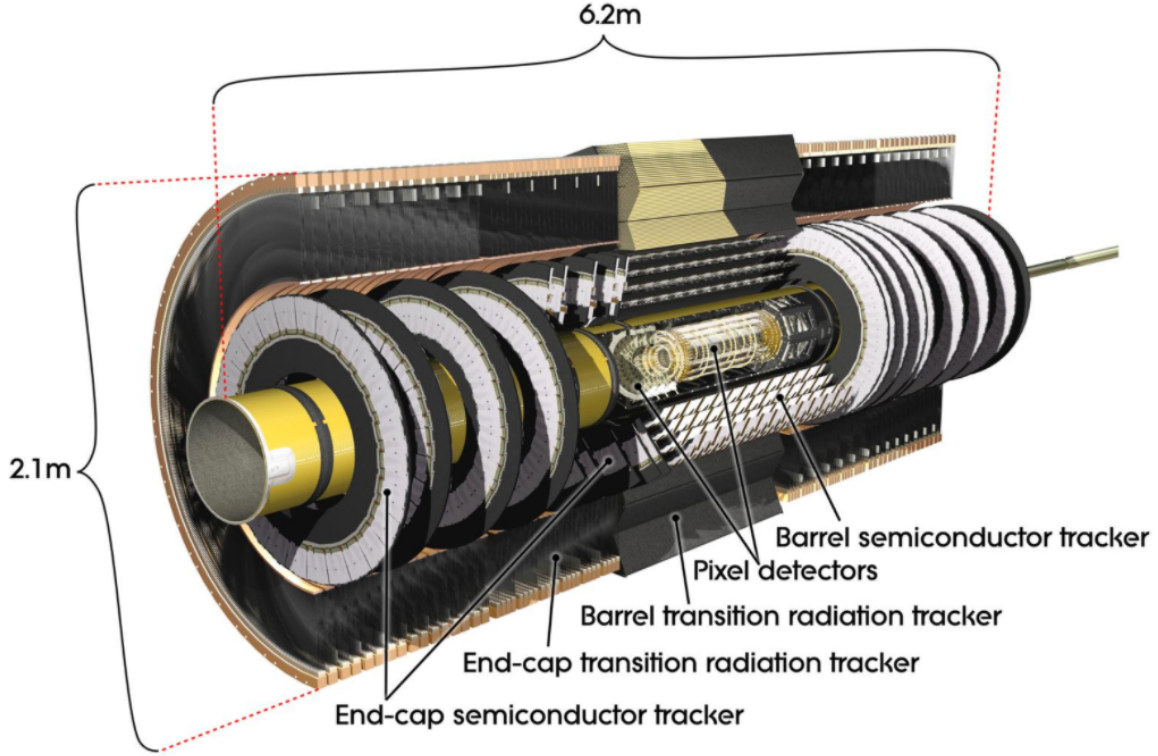
### 4.2.2 Inner Detector

The Inner Detector (ID) is the innermost section of ATLAS, built from three sub-detector components, the Pixel Detector, the Semiconductor Tracker (SCT), and the Transition Radiation Tracker (TRT), each with a barrel and end-cap section.

#### The Pixel Detector

The Pixel detector is a high granularity silicon detector, with a channel size of  $50 \times 400 \mu m^2$ , or  $50 \times 250 \mu m^2$  in the innermost Insertable B Layer (IBL). The detector is made up of four pixel layers in the barrel (with the first being the IBL) that run parallel to the beamline, and three end-cap disks placed perpendicular to the beamline. The Pixel detector is designed to give a high-precision location of a particle as it passes through the detector (based on hit patterns in the pixels), as well as measure of the energy of the particle based on a Time-Over-Threshold (ToT) measurement; where the amount of time that the deposited charge is above some threshold can be converted into

an energy measurement. The first two layers can also provide b-tagging information, by identifying the decays of hadrons which contain b quarks.



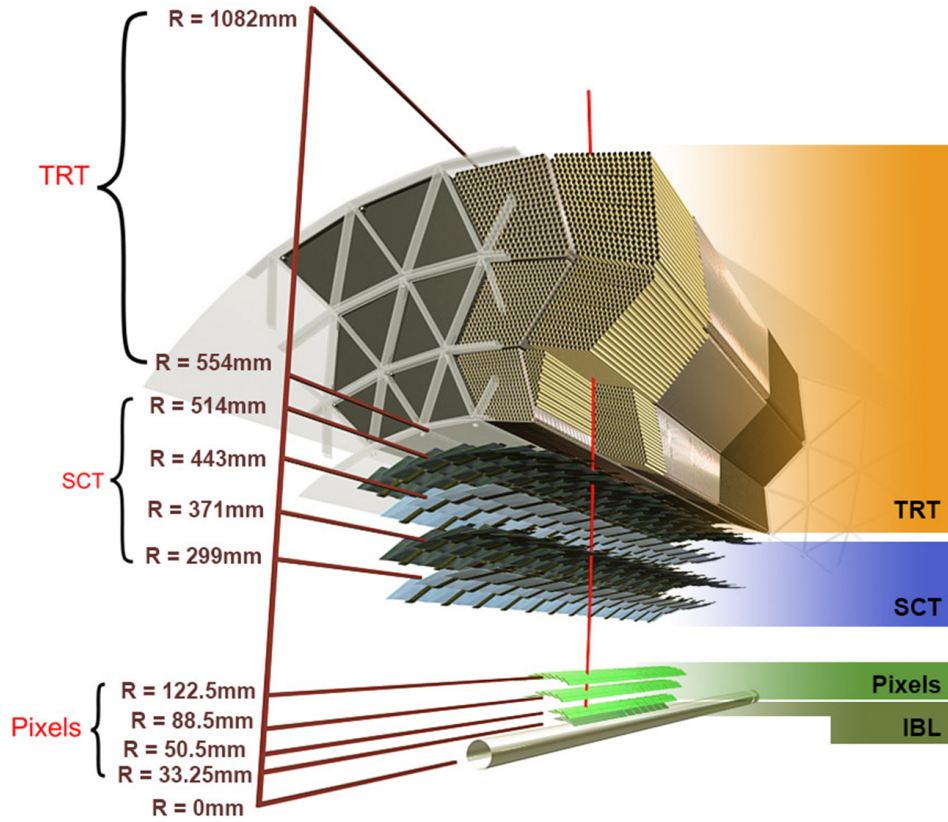
**Figure 4.3:** Cross-section of the ATLAS Inner Detector, with the three main sub-detectors highlighted. Figure from [42].

### The Semiconductor Tracker

The SCT is a silicon micro-strip tracker, which completely surrounds the Pixel detectors. Built from 4088 rectangular modules, with each module measuring  $12\text{cm}$  in length, with an  $80\mu\text{m}$  pitch, and  $124\text{mm}$  strips across their surface. In the barrel, the modules are aligned parallel to the beamline, and in the end-cap they are perpendicular to it. This gives the SCT high resolution in only one direction, so to combat this, each module is double-sided, giving two measurements for each strip pair. The hits from each strip pair are combined to give a single spatial measurement, giving the SCT a resolution of  $17\mu\text{m}$  per layer (in the transverse plane). Unlike the Pixel detector, the SCT only gives a binary hit, and not a ToT measurement.

## The Transition Radiation Tracker

The TRT is a collection of 300,000 drift tubes known as straws. Each straw is made up of a  $4\text{mm}$  diameter cylindrical chamber with a thin gold-plated tungsten wire at its center. The inner walls of the tube are held at a high potential difference with respect to the wire, so that when charged particles enter the chamber, they ionize the enclosed gas. The resulting positive and negative ions drift towards the walls and central wire, respectively, and deposit a signal.



**Figure 4.4:** Detailed Diagram of the Barrel section of the Inner Detector. The radial distances of the different sub-detector elements are shown. Figure from [47].

Like the SCT, the straws of the barrel are parallel to the beamline, and the end-cap straws are perpendicular. Since the entire straw acts as the readout element, each section loses resolution of one spatial dimension: the barrel has no ability to reconstruct the  $z$  position, and the end-caps cannot resolve a radial distance. To make up for this, the TRT uses many more layers than the previous two sub-detectors, with 73 layers in the barrel, and 160 in each end-cap.

In addition to an  $r - \phi$  measurement, the TRT can also be used for particle identification. Energy is deposited in the straws as relativistic charged particles pass through the inhomogeneous media of the detector, this is known as transition radiation. The energy deposits can then be integrated and compared to a predefined threshold; comparing what fraction of the straws had hits with energy above this threshold. If the fraction is high, this corresponds to an electron track, while a low fraction corresponds to a pion track.

The ID as a whole is designed to have a track momentum resolution of:

$$\frac{\sigma(p_T)}{p_T} = \left(0.05 \frac{p_T}{\text{GeV}} \oplus 1\right)\% \quad (4.7)$$

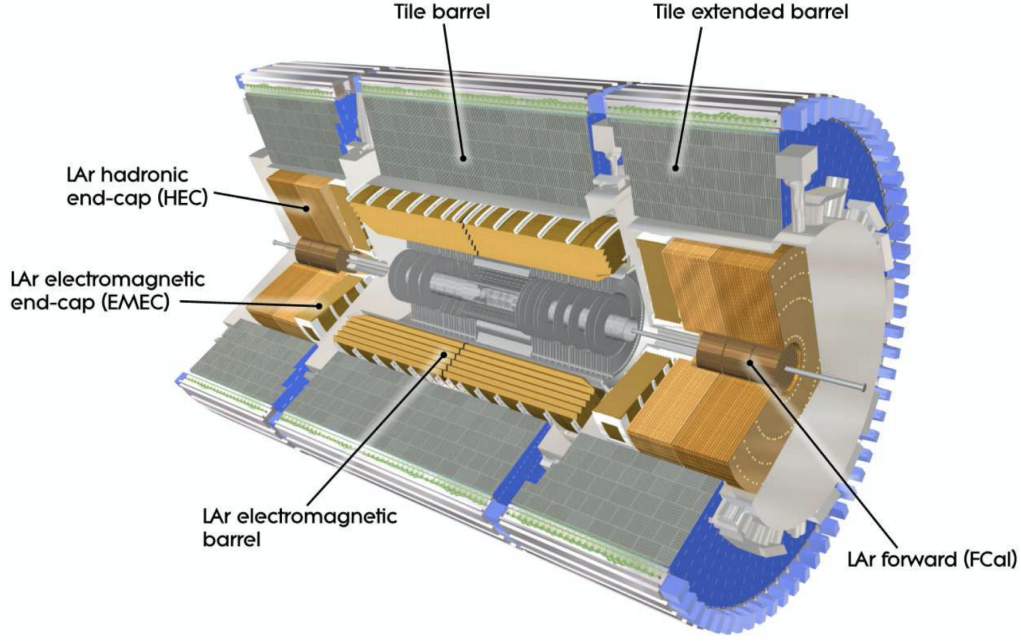
which decreases with increasing  $p_T$  since higher momenta particles bend less in the magnetic field.

### 4.2.3 Calorimeters

The ATLAS calorimeters surround the inner solenoid, providing coverage for  $|\eta| < 4.9$  and the full  $\phi$  around the beampipe. They are all sampling calorimeters, and are designed to completely stop and measure the energy of the majority of particles which leave the Inner Detector.

The general design of the calorimeters is the same: sections of active and passive material interlaced together. The passive sections are generally a high-density material designed to slow and eventually stop high energy particles through nuclear interactions which causes the particles to lose energy in the form of a particle shower. The active material is designed to measure the energy deposited by the shower. Because some energy is deposited in the passive material, the active material can only sample the energy from the shower, so the calorimeter must be calibrated in order to properly estimate the total energy of the shower. The calorimeters must also be large enough to fully contain the shower, otherwise energy will be lost out the back of the calorimeter (*punch-through*) and the energy estimate will be miscalculated.

There are two main types of calorimeters, electromagnetic calorimeters (ECals) and hadronic calorimeters (HCals). The former is designed to contain the showers of particles which interact electromagnetically, mainly electrons and photons. HCals are designed to stop particles which interact more through the strong force, such as proton, neutrons, and pions.



**Figure 4.5:** Cross-section of the ATLAS Calorimeters. The diagram highlights the different calorimeter technologies: the LAr ECal (both barrel and end-cap), the LAr HCal (end-cap), and the Tile HCal (barrel). It also shows the LAr FCal which has both ECal and HCal components. Figure from [42].

ATLAS employs two main calorimeter technologies, the Liquid Argon (LAr) calorimeters for both ECals and HCals, and the Tile hadronic calorimeter (TileCal). The LAr uses liquid argon as its active material, and various passive materials depending on its use. The TileCal uses scintillator as its active material and steel as its passive material.

## Electromagnetic Calorimeters

The ATLAS ECals are located closest to the Inner Detector, and use the LAr technology, with lead as the passive material, for both the barrel section ( $|\eta| < 1.475$ ) and the end-caps ( $1.375 < |\eta| < 3.2$ ). The layers of the ECal are folded into an accordion shape which naturally allows for full  $\phi$  coverage.

ECals are characterized in terms of the radiation length ( $X_0$ ); a property of the material which gives the average distance an energetic electron will travel before losing all but  $\frac{1}{e}$  of its energy.  $X_0$  is also a useful measure since the mean free path for a photon to pair produce is approximately one radiation length.



The barrel section of the ECals are divided into three layers, with an additional presampling layer at its center. The presampling layer is a thin layer of active material, which is used to determine if the shower started before it enters the ECal. The first layer is thin,  $4X_0$ , but has the highest granularity in order to distinguish between photons and neutral pions (which very quickly decay into two photons). The second layer is  $16X_0$  thick, designed to contain the majority of the EM shower. The third layer is the thinnest, only  $2X_0$ , and is used to help distinguish between EM showers and hadronic showers (which start in the ECal).

The end-cap ECals are divided into two sections, an inner and outer wheel, which transition at  $|\eta| = 2.5$ . They are thicker and coarser, but roughly follow the same design as the barrel. The inner wheel only has the first two layers, while the outer wheel has all three layers (plus the presampling layer), with a slightly thicker third layer for a total thickness of  $24X_0$ . The ECal has a granularity of  $\Delta\eta \times \Delta\phi = 0.025 \times 0.025$ .

## Hadronic Calorimeters

The HCals completely surround the ECals, and use two different calorimeter technologies, the steel-scintillator TileCal for the barrel (and extended barrel), and copper-LAr for the end-caps. Unlike the ECals, both the LAr end-caps and the TileCal use flat sections of active and passive material.

Analogous to the radiation length, hadronic calorimeters are measured using the interaction length ( $\lambda$ ); which gives the mean free path for a hadron between inelastic collisions, and thus can be used to characterize the size of a hadronic shower. In general, an interaction length will be longer than a radiation length, so we expect a hadronic shower to be larger than an EM shower. Because particles have to pass through the ECals, the hadronic showers start before they reach the HCals, so the ECals contribute roughly  $2 - 3\lambda$  to the total depth.

The barrel TileCal is composed of two main parts, the central barrel which covers  $|\eta| < 1$ , and a barrel extension on either side which encircles the end-cap calorimeters. Each section is divided into 64 wedges in  $\phi$ , and then separated into three layers of varying thickness, for a total depth of  $9.7\lambda$ . Each end-cap consists of two separate wheels, divided into 32 sections in  $\phi$  and two layers per wheel. Together, both wheels have a thickness of  $11\lambda$ . The HCal has a granularity of  $\Delta\eta \times \Delta\phi = 0.1 \times 0.1$ .



## Forward Calorimeter

The forward calorimeter (FCal) is a special calorimeter which is located between the beampipe and the inner radius of the HCal end-caps, covering the region of  $3.1 < |\eta| < 4.9$ . Due to its density, each FCal end-cap is able to provide roughly ten interaction lengths in an area smaller than the corresponding HCal.

Each FCal is composed of three sections, all of which use the LAr technology. The first section (innermost in  $z$ ), uses copper as its passive material and acts as an electromagnetic calorimeter. The second two layers use tungsten as their passive material and are designed to act as HCals. The FCal has a granularity of  $\Delta\eta \times \Delta\phi = 0.15 \times 0.15$ .

The energy resolutions of the ATLAS calorimeters depends on the technology (LAr or TileCal). The LAr resolution is:

$$\frac{\sigma(E)}{E} = \left( \frac{10\%}{\sqrt{E}} \oplus \frac{170\text{MeV}}{E} \oplus 0.7\% \right) \quad (4.8)$$

and is the same for the ECal, HCal, and FCal. The TileCal has a resolution of:

$$\frac{\sigma(E)}{E} = \left( \frac{150\%}{\sqrt{E}} \oplus 6\% \right). \quad (4.9)$$

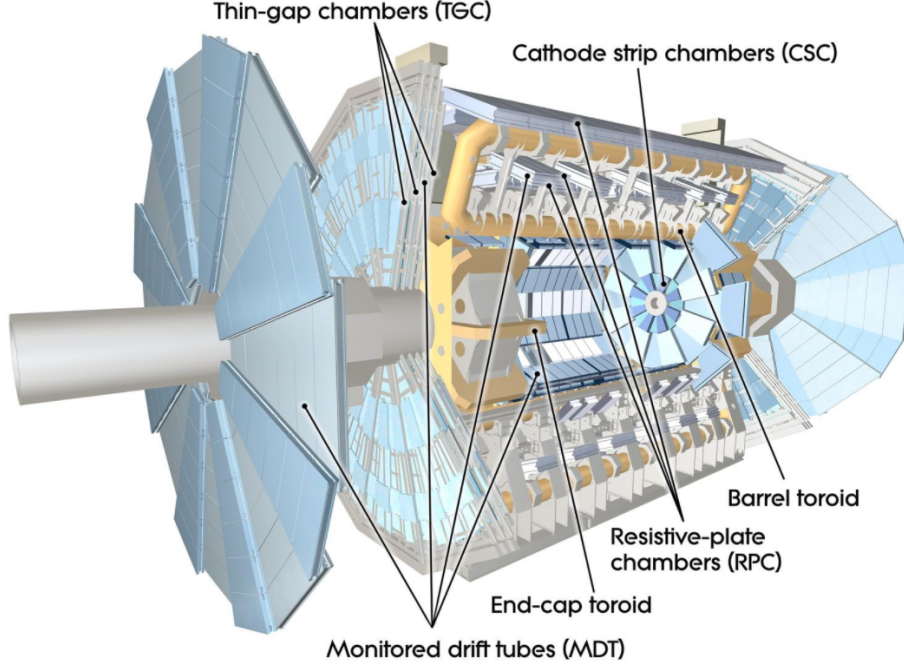
### 4.2.4 The Muon Spectrometer

The Muon Spectrometer (MS) is the outermost detector system in ATLAS, designed to track and measure the energy of muons, since they can pass through the calorimeters without depositing much energy (so-called minimum ionizing particles). The detectors cover the region of  $|\eta| < 2.7$ , with trigger (event selection) capabilities in  $|\eta| < 2.4$ .

The MS is divided, as usual, into the barrel and end-cap sections, with the barrel detectors placed both within and surrounding the toroidal coils, and radial detector disks or *wheels* placed both before and after the toroids in the end-caps. There are four main detector technologies employed by the MS: monitored drift tubes (MDT), cathode strip chambers (CSC), thin gap chambers (TGC), and resistive plate chambers (RPC).

The main tracking detectors are the MDTs, with both barrel and end-cap components, generally interspersed with the other detector technologies. The design of the chambers is relatively simple;

several layers of  $3\text{cm}$  cathode tubes aligned perpendicular to the beamline. Each tube has a tungsten anode wire at its center, and is filled with a gas-gain mixture to encourage ionization when an energetic muon passes through. The MDTs provide very good  $\eta$  resolution, while only giving coarse  $\phi$  resolution due to the tubes running the full length of the chambers.



**Figure 4.6:** Cross-section of the ATLAS Detector highlighting the components of the Muon Spectrometer. The diagram shows the locations of the four main detector technologies: MDTs, CSCs, RPCs and TGCs, as well as the location of the toroid magnets. The ATLAS calorimeters and Inner Detector have not been included in the diagram. Figure from [42].

There are three layers of MDTs in the barrel: the first layer is closest to the calorimeters, with the second and third layers being surrounded by RPCs. There are four layers in each end-cap: two partial wheels, one encircling the CSCs to form the *small wheel* (the innermost end-cap of the MS) and one encircling the end-cap toroid, and one layer in each *big wheel* (the outermost end-caps of the MS which cover the full 25m diameter of ATLAS).

The small wheel is formed by a ring of 32 CSCs (innermost in  $r$ ) and then three layers of MDTs. The CSCs are a type of multi-wire proportional chambers, which use several thin tungsten wires as anodes, sandwiched between two cathode planes segmented into strips. The strips and wires run perpendicular to each other, giving independent  $\eta$  and  $\phi$  measurements with good resolution.

The trigger is handled by the RPCs in the barrel, and the TGCs in the end-caps. These detectors can provide timing measurements for muons with a resolution of  $15 - 25ns$ , which is sufficient to determine in which bunch crossing the muon originated. They can also provide rough position measurements, which gives coarse momentum resolution, and can provide higher  $\phi$  granularity than the MDTs.

The RPCs are arranged in three layers, including the outermost barrel layers. They are made of a  $2mm$  gas-filled gap contained within two parallel plate electrodes, which give coarse  $z - \phi$  measurements. Each end-cap has four layers of TGCs, three of which are in the first (innermost in  $z$ ) big wheel, and one directly in front of the small wheel. They are multi-wire proportional chambers, similar to the CSCs, which give  $\eta - \phi$  measurements. Unlike the CSCs, the TGCs are designed for fast readout as opposed to precision tracking.

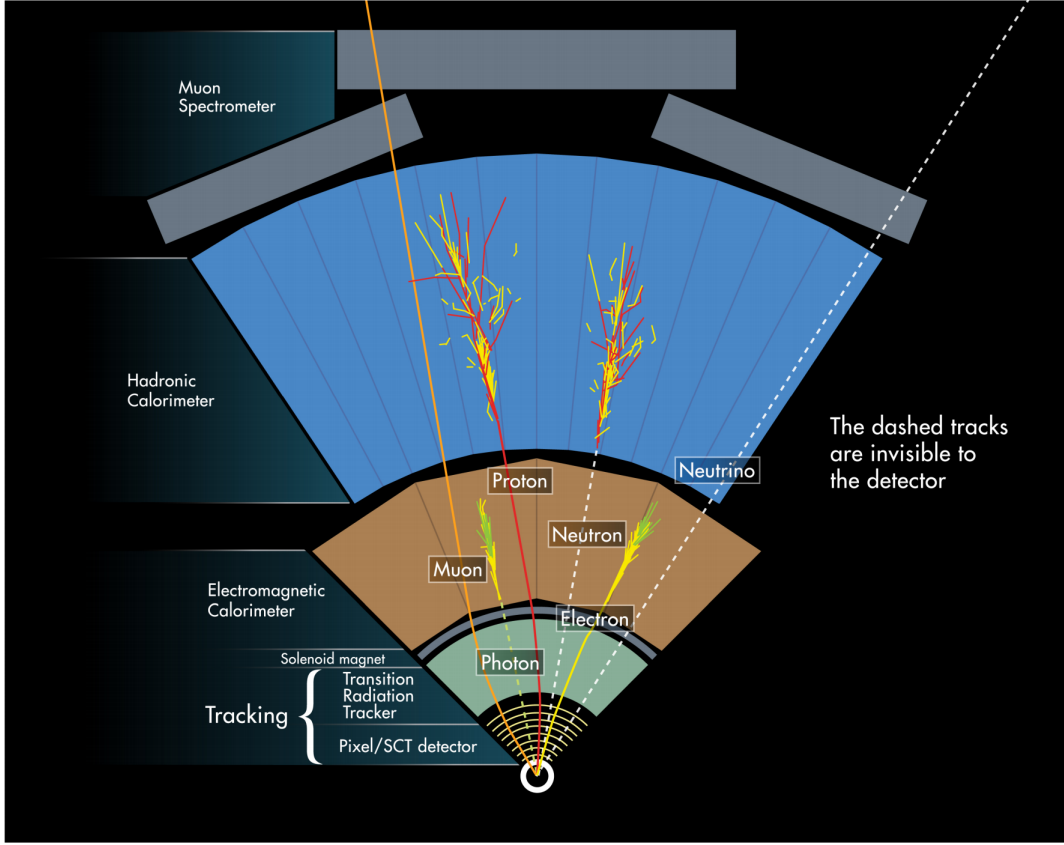
### 4.3 Particle Identification

Particle identification in ATLAS is achieved by combining characteristic signatures from the various detector components. It starts with the Inner Detector, where charged particles will deposit hits which form tracks. Tracks get matched to other detector signatures in order to distinguish between charged and neutral particles. They can also be used to determine the charge and momentum of the particles due to the magnitude and direction of the bending. See section 5.2.1 for more information on track reconstruction.

Electrons can be reconstructed from a shower in the ECal which is matched to a track in the Inner Detector. Photons have similar calorimeter responses, but have no available matched tracks. Neutral pions which decay into photons also do not leave tracks, but can be distinguished from standalone photons by the first layer of the ECal.

HCal showers can be used to reconstruct hadrons such as pions, protons and neutrons. Protons and neutrons can be distinguished because protons have associated tracks, whereas neutrons pass through the tracking detectors without leaving hits. Charged pions can be reconstructed from both calorimeter information and TRT threshold measurements. Significant energy deposits in the HCal are reconstructed as jets, see section 5.2.2.

In general, the only particles which are not expected to deposit the majority of their energy



**Figure 4.7:** Partial cross-section of the ATLAS Detector through the transverse plane. The diagram shows different detector signatures left by particles commonly produced in collisions, in each of the different detector components. Figure from [51].

in the calorimeters are muons and neutrinos. Muons are reconstructed from Inner Detector tracks which are matched to Muon Spectrometer tracks, whereas neutrinos can pass through the entirety of ATLAS without interacting with the detectors.

Massive particles whose lifetimes are sufficiently short that they do not interact with the detectors can also sometimes be reconstructed by their decay products. For example, in the case where the  $Z$  Boson decays to two muons, the  $Z$  can be reconstructed from the invariant mass of the two reconstructed muons, where the invariant mass is given by:

$$m^2 = (E_i + E_j)^2 - (\mathbf{p}_i + \mathbf{p}_j)^2 \approx 2p_{T,i}p_{T,j}(\cosh(\eta_i - \eta_j) - \cos(\phi_i - \phi_j)). \quad (4.10)$$

Here the approximation is valid for high energy particle colliders where the reconstructed particles satisfy  $E \gg m$ .

## Chapter 5

# Dataset and Physics Objects

The analysis presented in this thesis is part of a larger effort being conducted within the ATLAS collaboration. As such, the information in this chapter does not represent work which I have done, but rather forms the basis for the analysis. All of the work here was either conducted by the collaboration as a whole (such as collecting the data, implementation of triggers, etc.) or by the analysis team (such as designing the signal samples, applying specific event selection criteria, etc.), and has been cited where appropriate.

### 5.1 Data and Monte Carlo Samples

The analysis is, to first order, based on the Run II dataset collected by ATLAS during the period of 2015-2018. A special filtering called RPVLL (R-Parity Violating and LLP) has been applied to the dataset. This filter includes large radius tracking (section 5.2.1) and displaced vertex (section 5.2.3)reconstructions, which are computationally expensive and therefore not applied to the dataset as a whole.

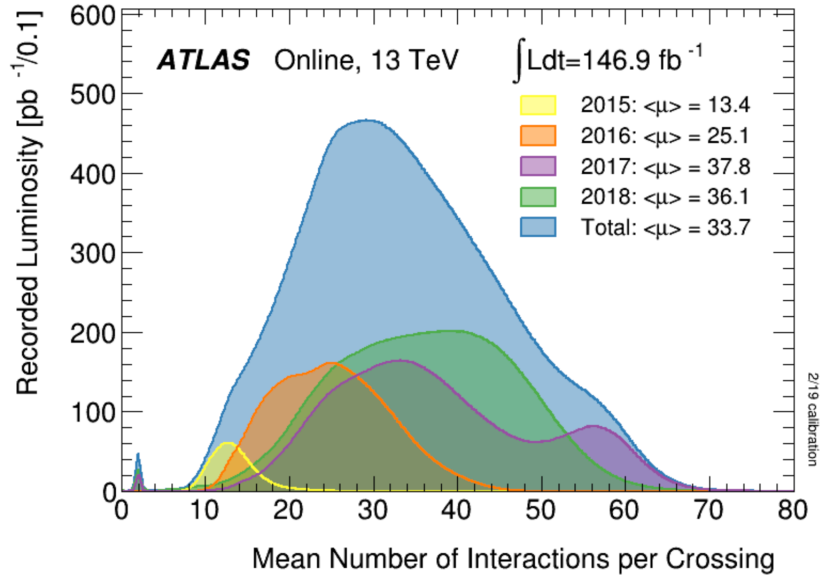
In addition to the data, Monte Carlo (MC) simulations have been produced for several Emerging Jets model scenarios, as well as a QCD background sample. The MC is used to study the ATLAS detector's potential sensitivity to EJs-like events, as well as develop and optimize the specific analysis strategies.

### 5.1.1 The ATLAS Run II Dataset

Of the  $156 \text{ fb}^{-1}$  of data collected by the LHC during Run II, ATLAS selected  $139 \text{ fb}^{-1}$  which it determined to be physics-quality. This data was collected using strictly proton-proton collisions at a center of mass energy of  $\sqrt{s} = 13 \text{ TeV}$  [40].

The dataset is generally divided into three components: the set from 2015 and 2016 which corresponds to roughly 26% of the data, the set from 2017 which makes up roughly 32%, and the set from 2018 which makes up the final 42% of the Run II dataset [40].

The ongoing Emerging Jets analysis is still blinded to the data, meaning that the events which may correspond to the search region (sections 6.1.4 and 6.2.5) cannot be looked at until the analysis strategy is finalized. Therefore, the only data used for this thesis corresponds to the *validation* regions (section 6.4); which use an orthogonal set of events, expected to contain only background.



**Figure 5.1:** Pileup ( $\langle\mu\rangle$ ) conditions for the full Run II dataset, with the components from the different years overlayed. The peak at  $\langle\mu\rangle = 2$  corresponds to a special low-pileup run conducted in 2017 and 2018. The integrated luminosity of  $146.9 \text{ fb}^{-1}$  corresponds to the total recorded by ATLAS during Run II, of which  $139 \text{ fb}^{-1}$  were deemed physics-quality. Figure from [40].

Validation is done using the 2017 set (Data17), which contains  $44.3 \text{ fb}^{-1}$  of data. This corresponds to an above average  $\langle\mu\rangle$  of 37.8 interactions per bunch-crossing, and a peak  $\langle\mu\rangle$  of 65 [40]. The pileup conditions for 2017, as well as the Run II dataset as a whole are shown in figure 5.1.

### 5.1.2 Monte Carlo Samples

Two sets of Monte Carlo simulation samples have been produced for the Emerging Jets analysis: a set of signal samples produced with the PYTHIA Hidden Valley implementation, and a set of background samples using the PYTHIA QCD Dijet slices.

MC samples are produced according to three campaigns: mc16a, mc16d, and mc16e, which correspond to the data runs for 2015+2016, 2017, and 2018, respectively. These campaigns help account for differences in detector layouts and pileup conditions which can change from year to year. Because the Data17 set is being used for validation, the corresponding MC campaign (mc16d) is used for the majority of the MC studies.

#### PYTHIA Hidden Valley Implementation

The hidden valley implementation in PYTHIA 8.230 [52] allows for the simulation of dark sector particles, their showering, and subsequent decay into SM particles. In general, the hidden sector has many free parameters which change the dynamics of the simulation; for the case of the EJs signal models, many of them have been set to reflect a QCD-like dark sector:

- The number of dark colors,  $N_d$ , of the  $SU(N_d)$  gauge group, has been set to 3
- The number of dark flavors,  $n_{f_d}$ , corresponding to the number of dark quarks, is set to 7
- The gauge coupling is allowed to run so that it more accurately represents QCD-like behavior

There are 12 dark partners to the SM fermions,  $F_d$ , which are charged under both the dark sector and their corresponding SM charges. These are pair-produced by either gluon-gluon fusion or quark-antiquark annihilation. The  $F_d$  eventually decay to their SM partners and a hidden quark  $Q_d$ :  $F_d \rightarrow f Q_d$ . The dark quarks are only charged under the dark sector and are all mass-degenerate, meaning that they are all identical except for their flavor. The Yukawa couplings are negligible except for those which connect to the SM down quark, so the dark partner to the down quark,  $D_d$ , is taken as the scalar mediator  $X_d$ .

QCD-like hadronization and showering in the dark sector is slightly simplified in the implementation; the dark quarks will not form dark baryons, only vector ( $\rho_d$ ) and pseudoscalar ( $\pi_d$ ) mesons. Both of these can decay to SM quarks, however the  $\rho_d$  is far more likely to decay to two dark pions.

The mass of the dark quark is equal to the dark confinement scale,  $\Lambda_d$ , which also sets the masses for the dark mesons. The specific choices of the various free parameters of the models ( $m_{X_d}$ ,  $\Lambda_d$ ,  $c\tau_{\pi_d}$ , etc.) are discussed in section 5.1.3.

### PYTHIA Dijet Background Samples

The background MC samples are produced using the PYTHIA dijet slices, which breaks up the SM QCD jet  $p_T$  spectrum into different slices. These slices are labeled JZ\*withSW where  $*$  = 0 – 12, each of which represents a specific  $p_T$  range, given by table 5.1, into which the leading jet must fall.

| DSID   | Dijet Slice | $p_T$ Range [GeV] |
|--------|-------------|-------------------|
| 364700 | JZ0withSW   | 0-20              |
| 364701 | JZ1withSW   | 20-60             |
| 364702 | JZ2withSW   | 60-160            |
| 364703 | JZ3withSW   | 160-400           |
| 364704 | JZ4withSW   | 400-800           |
| 364705 | JZ5withSW   | 800-1300          |
| 364706 | JZ6withSW   | 1300-1800         |
| 364707 | JZ7withSW   | 1800-2500         |
| 364708 | JZ8withSW   | 2500-3200         |
| 364709 | JZ9withSW   | 3200-3900         |
| 364710 | JZ10withSW  | 3900-4600         |
| 364711 | JZ11withSW  | 4600-5300         |
| 364712 | JZ12withSW  | 5300-7000         |

**Table 5.1:** Leading jet transverse momentum ranges for the PYTHIA dijet slices. The DataSet ID (DSID) is used to refer to the specific slice.

The events in each slice are given a specific weight such that, when the slices are combined, the usual falling  $p_T$  spectrum is resolved. The analysis uses slices 2-12 to estimate the full QCD background. See appendix B for more information on the PYTHIA dijet slices.



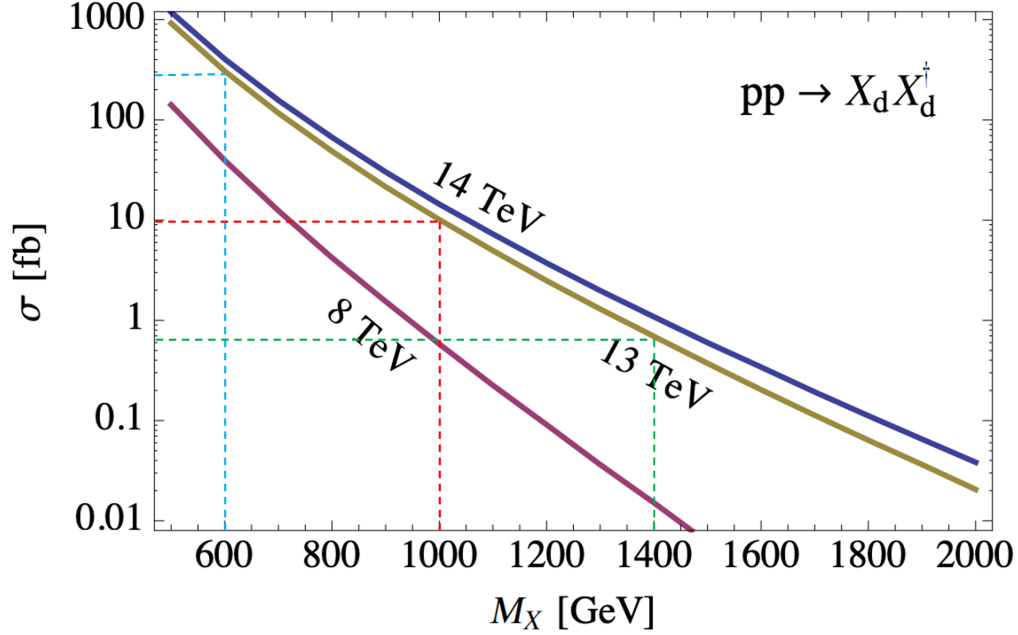
### 5.1.3 Emerging Jets Models

The EJs signal samples are a set of benchmark models, based on the value of the dark confinement scale,  $\Lambda_d$ , since this sets the mass scales for the dark mesons:

$$m_{\pi_d} = \frac{\Lambda_d}{2} \quad (5.1)$$

$$m_{\rho_d} = 2\Lambda_d \quad (5.2)$$

Five values of  $\Lambda_d$  spanning the range from 1.6 – 40 GeV set the baseline model space. Several signal points are then defined based on the values of  $m_{X_d}$  and  $c\tau_{\pi_d}$ , which are still free parameters in the models.



**Figure 5.2:** Theoretical pair-production cross-section ( $\sigma$ ) for varying values of dark mediator mass ( $M_X$ ). Curves are presented for three different center of mass energies of proton-proton collisions which correspond to the Run I, Run II, and designed peak values of 8 TeV, 13 TeV, and 14 TeV, respectively. The dashed lines represent the selected EJs mediator mass cases and the corresponding cross-sections for the Run II analysis. Figure modified from [9]

Three TeV-scale mediator masses are selected to cover a wide range of scenarios, these are: 600, 1000, and 1400 GeV. As the mass of the scalar mediator increases, the pair-production cross-section decreases, while the generator filter efficiency increases, both of which are independent of

the benchmark models or  $c\tau_{\pi_d}$ . The production cross-sections for these cases are given in figure 5.2. The MC generator (PYTHIA) filter selects generated events which have at least 4 jets with  $p_T > 100$  GeV and  $|\eta| < 2.7$  before detector-level reconstruction is applied. This selection is used to reflect the *Emerging* component of the RPVLL filter (see section 5.3.2 for the filter criteria), which is applied to data and used to define to search regions for the analysis (see sections 6.1.4 and 6.2.5). The filter efficiency, therefore, corresponds to how well the generator produces events which pass the selection filter; a higher filter efficiency means that fewer initial events need to be produced, in comparison to a lower efficiency, in order to see a proportionate number of filtered events:  $N_{final} = \epsilon_{filter} \times N_{generated}$ .

| DSID   | Model    | $\Lambda_d$ [GeV] | $m_{\pi_d}$ [GeV] | $m_{\rho_d}$ [GeV] | $m_{X_d}$ [GeV] | $c\tau_{\pi_d}$ [mm] |
|--------|----------|-------------------|-------------------|--------------------|-----------------|----------------------|
| 312004 | <b>A</b> | 10                | 5                 | 20                 | 1400            | 20                   |
| 312008 |          |                   |                   |                    | 1000            | 150                  |
| 312017 |          |                   |                   |                    | 600             | 1                    |
| 312022 | <b>B</b> | 4                 | 2                 | 8                  | 1400            | 20                   |
| 312028 |          |                   |                   |                    | 1000            | 5                    |
| 312031 |          |                   |                   |                    | 600             | 300                  |
| 312039 | <b>C</b> | 20                | 10                | 40                 | 1400            | 75                   |
| 312046 |          |                   |                   |                    | 1000            | 5                    |
| 312052 |          |                   |                   |                    | 600             | 2                    |
| 312060 | <b>D</b> | 40                | 20                | 80                 | 1400            | 2                    |
| 312066 |          |                   |                   |                    | 1000            | 1                    |
| 312067 |          |                   |                   |                    | 600             | 300                  |
| 312075 | <b>E</b> | 1.6               | 0.8               | 3.2                | 1400            | 75                   |
| 312080 |          |                   |                   |                    | 1000            | 150                  |
| 312090 |          |                   |                   |                    | 600             | 0.5                  |

**Table 5.2:** Model parameters for the canonical set of 15 signal points used by the Emerging Jets analysis. The DataSet ID (DSID) is used to refer to the corresponding set of parameters for a given signal point.

For each mediator mass, there are 6 possible dark pion lifetimes, which span the range from  $0.5 - 300\text{mm}$  in order to test the detector sensitivity. This combination of parameters gives a grid of 90 possible signal points. Of these, there are 36 signal points available, and a subset of canonical 15 signal points are used for this thesis, which sufficiently span the grid. The details of these models are given in table 5.2.

Models A through E represent the benchmark models, with corresponding dark pion masses which determine the hidden sector dynamics. Heavier dark pions result in a larger phase space for producing SM particles, and therefore result in higher track multiplicity DVs. However, higher mass  $\pi_d$  also mean there is a smaller phase space for DV production, resulting in lower DV multiplicity events. Lighter dark pions also tend to be more boosted, therefore having a larger transverse decay distance than a heavier  $\pi_d$  for a given lifetime. Thus, for a relatively long-lifetime, there is a higher probability that a light dark pion will decay after the calorimeters and not be reconstructed into the emerging jet.

| Mediator Mass $m_{X_d}$ [GeV] | Dark Pion Lifetime $c\tau_{\pi_d}$ [mm] |
|-------------------------------|-----------------------------------------|
| 1400                          | 300                                     |
|                               | 150                                     |
|                               | 75                                      |
|                               | 20                                      |
|                               | 5                                       |
|                               | 2                                       |
| 1000                          | 300                                     |
|                               | 150                                     |
|                               | 75                                      |
|                               | 5                                       |
|                               | 2                                       |
|                               | 1                                       |
| 600                           | 300                                     |
|                               | 150                                     |
|                               | 20                                      |
|                               | 2                                       |
|                               | 1                                       |
|                               | 0.5                                     |

**Table 5.3:** Full set of possible combinations for scalar mediator mass  $m_{X_d}$  and dark pion lifetime  $c\tau_{\pi_d}$  which define the full Emerging Jets signal grid.

Models A, B, and C represent the intermediate mass regimes; these are used to determine the

general kinematics of the dark sector. Models D and E represent the so-called *edge cases*; they represent the limits of the detectors capabilities to reconstruct emerging jets. Model D gives the high mass regime, where the majority of the hidden sector energy is contained within one or very few dark pions, which results in more of a displaced jet signature than a characteristic emerging jet. Model E gives the lower mass limit, where the lighter dark pions have smaller track multiplicity DVs, making them more difficult to reconstruct, and resulting in a more QCD-like jet.

The remaining 21 available signal points maintain the same five benchmark models, and simply have a different mediator mass - dark pion lifetime combination. The full set of available  $(m_{X_d}, c\tau_{\pi_d})$  combinations are given in table 5.3. These 18 combinations are applied to each of the 5 benchmarks in order to recover the full 90 point signal grid.

#### 5.1.4 Event Weighting

In general, the number of generated MC events depends on the MC filter efficiency, but it may represent many more events than the corresponding data. Therefore, the MC events, for both signal and background, must be weighted appropriately. The event weight comes from many sources, each of which can be placed into one of two categories: overall scale factors, and event-level weights.

Event-level weights can be different for each event in a set, and are used to change the shape of the distributions. Alternatively, the scale factors have no impact on the shape of the distributions, but affect the overall scale, these are used to normalize the events after applying the event-level weights, and change the scale, i.e. to match a specific integrated luminosity. The total event weight is given by:

$$W_i = \frac{\sigma \times \epsilon_{filter} \times w_i^{mc} \times w_i^{pileup} \times \mathcal{L}_{int}}{\sum_i w_i^{mc}} \quad (5.3)$$

where the event-level weights are  $w_i^{mc}$  and  $w_i^{pileup}$ . The first is the MC event weight; in the background MC this is used to weight each event from each slice properly such that the higher slices contribute less to the overall  $p_T$  spectrum and the lower slices contribute more. In the signal MC,  $w_i^{mc}$  is set to 1 for all events so that each event is simply weighted to the theoretical cross-section. The second event-level weight,  $w_i^{pileup}$ , is the pileup weight, this is used to try and correct  $N_{PV}$  (number of primary vertices) distribution in the MC to match  $\langle\mu\rangle$  (average number of interactions per bunch-crossing) distribution from the data.

| DSID   | $N_{final}$ | $\sigma$ [fb]           | $\epsilon_{filter}$ | $\sum_i w_i^{mc}$        |
|--------|-------------|-------------------------|---------------------|--------------------------|
| 364702 | 683385      | $2.4330 \times 10^{12}$ | 0.0098              | 7384.87                  |
| 364703 | 11183277    | $2.6450 \times 10^{10}$ | 0.0117              | 145.457                  |
| 364704 | 14826619    | $2.5461 \times 10^8$    | 0.0134              | 4.45373                  |
| 364705 | 7378375     | $4.5529 \times 10^6$    | 0.0145              | 0.166351                 |
| 364706 | 7369200     | $2.5754 \times 10^5$    | 0.0095              | 0.0278025                |
| 364707 | 3753137     | $1.6215 \times 10^4$    | 0.0111              | 0.00382369               |
| 364708 | 3959144     | $6.2506 \times 10^2$    | 0.0102              | 0.00110877               |
| 364709 | 1955776     | 19.639                  | 0.0121              | 0.000214898              |
| 364710 | 513869      | 1.1962                  | 0.0059              | $2.65035 \times 10^{-5}$ |
| 364711 | 585442      | $4.2263 \times 10^{-2}$ | 0.0027              | $1.40227 \times 10^{-5}$ |
| 364712 | 559707      | $1.0367 \times 10^{-3}$ | 0.0004              | $8.15004 \times 10^{-6}$ |
| 312004 | 79656       | 1.0786                  | 0.9353              | 80000                    |
| 312008 | 98112       | 15.195                  | 0.8681              | 100000                   |
| 312017 | 276422      | $4.3009 \times 10^2$    | 0.6056              | 280000                   |
| 312022 | 79629       | 1.0783                  | 0.9425              | 80000                    |
| 312028 | 89526       | 15.216                  | 0.8905              | 90000                    |
| 312031 | 178433      | $4.3017 \times 10^2$    | 0.6787              | 190000                   |
| 312039 | 98838       | 1.0779                  | 0.9436              | 100000                   |
| 312046 | 129206      | 15.183                  | 0.8662              | 130000                   |
| 312052 | 393338      | $4.2997 \times 10^2$    | 0.5731              | 400000                   |
| 312060 | 109554      | 1.0775                  | 0.9690              | 110000                   |
| 312066 | 158959      | 15.188                  | 0.9087              | 160000                   |
| 312067 | 434946      | $4.3001 \times 10^2$    | 0.6318              | 500000                   |
| 312075 | 59579       | 1.0773                  | 0.9479              | 60000                    |
| 312080 | 78668       | 15.184                  | 0.9057              | 80000                    |
| 312090 | 138646      | $4.3026 \times 10^2$    | 0.7283              | 140000                   |

**Table 5.4:** Table of Monte-Carlo parameter values for the relevant dijet slices 2-12 (above) and the canonical 15 signal points (below). The values for the total unweighted MC events ( $N_{final}$ ), the cross-section ( $\sigma$ ), the generator filter efficiency( $\epsilon_{filter}$ ), and the initial sum of weights ( $\sum_i w_i^{mc}$ ) are given.

The rest of the components are overall scale factors. The cross-section,  $\sigma$  (generally given in fb), is used to scale the event to the corresponding physics process; in the background it represents the QCD process which gives multi-jet events of a certain  $p_T$  range, and is therefore different for each slice. In the signal MC,  $\sigma$  represents the theoretical cross-section, given by the mediator mass of the corresponding signal point. The MC filter efficiency,  $\epsilon_{filter}$ , is used to scale the events based on how they perform in the MC generation. The denominator is the initial sum of weights, which is given by the sum of the MC event weights for each event in the set. This is used to normalize the weighting so that the distribution is not scaled by the MC event weight. By contrast, the pileup weight is not included in the sum because the weights are designed in such a way that  $\sum_i w_i^{pileup} \approx N_{final}$ , where  $N_{final}$  is the total number of unweighted MC events in the set.

The final scale factor is the integrated luminosity,  $\mathcal{L}_{int}$ , which can be changed to scale the MC to match a given dataset. For the purpose of this thesis, the common values are the full Run II luminosity of  $139 \text{ fb}^{-1}$ , or the Data17 luminosity of  $44.3 \text{ fb}^{-1}$ . Table 5.4 gives the values for the overall scale factors for each background slice and signal point.

## 5.2 Physics Object Reconstruction

In order to make sense of a collision event, the raw detector responses must be converted into digital signals which can be used in reconstruction. Objects like jets, tracks and vertices are used as proxies to represent physical objects, but should not be confused with the actual particles. These *Physics Objects* are extremely useful for understanding the underlying event; as such it is very important to have a good reconstruction algorithm.

The details of the reconstruction algorithms can be found in [47], and [51 - 56].

### 5.2.1 Track Reconstruction

Track reconstruction is preformed using hits in the Inner Detector, which sits within the 2T inner solenoid. This provides a uniform magnetic field parallel to the beamline, which bends the path of charged particles as they pass through the sub-detectors, with the force given by the Lorentz force law:

$$\vec{F} = q\vec{v} \times \vec{B} \tag{5.4}$$

where  $q$  is the particle's charge,  $\vec{v}$  is the particle's velocity, and  $\vec{B}$  is the magnetic field strength vector. Because the magnetic field is aligned along the z-axis, the bending will occur in the transverse plane.

The degree to which the different particles bend is due to their momentum: very high momentum particles bend very little, while low momentum particles will bend much more. Therefore, if the amount of bending can be quantified, the particle's momentum can be determined. This is achieved using the relationship with the radius of curvature  $R$ :

$$R = \frac{p_T}{qB}. \quad (5.5)$$

The solenoid field is sufficiently strong that even the most energetic particles will be bent. However, this also means that particles with very low momenta will be bent so strongly that they do not exit the tracking detectors and therefore cannot be accurately reconstructed. This puts a baseline  $p_T$  cut of 0.5 GeV for a track to be properly reconstructed. Very energetic particles will pass through the tracking detectors without bending into a complete circle, therefore  $R$  must be inferred from the partial arc of the path:

$$R = \frac{s}{2} + \frac{L^2}{8s} \approx \frac{L^2}{8s} \quad (5.6)$$

where  $s$  is the sagitta of the arc, or the distance from the top of the arc to its base, and  $L$  is the base length of the arc. In general, the approximation in 5.6 is valid for a particle with very high momentum, since it will have a longer base length and a smaller sagitta.

## Track Selection

Track reconstruction starts by identifying hits in the silicon detectors (either the SCT or the pixels), and requires three hits as seeds (enough to properly reconstruct  $R$ ), which start closest to the Interaction Point (IP) and move out. Seeds are selected by finding the physical position of the detector hits with the largest energy deposits. These seed hits are then extended into the TRT to form a primary track candidate. This is known as *inside-out* selection. Secondary tracks are seeded from TRT hits, and then extended back into the silicon detectors. This is known as

*outside-in* selection. If a secondary track cannot be extended into the silicon detectors, then it is known as a TRT Standalone track.

There are two main sets of selection criteria defined for reconstructed tracks: *Loose* and *Tight* track selections. The Loose selection starts with the base kinematic selection of  $p_T > 500$  MeV and fiducial selection of  $|\eta| < 2.5$ . In addition to these cuts, the selection imposes some requirements on the hits within the detectors, as well as holes and shared hits. Shared hits are locations where two or more track candidates have hits, and holes are detector modules which are within the path of the track, but where hits have not been recorded. The Tight selection is applied on top of the Loose selection, with more stringent requirement on the detector hits. The selection criteria for both the Loose and Tight track selections are detailed in table 5.5.

| Loose Track Selection:    | Tight Track Selection:                          |
|---------------------------|-------------------------------------------------|
| $N_{Si}^{hits} \geq 7$    | $N_{Si}^{hits} \geq 9$ if $ \eta  < 1.65$       |
| $N_{Si}^{shared} \leq 1$  | $N_{Si}^{hits} \geq 11$ if $ \eta  > 1.65$      |
| $N_{Si}^{hole} \leq 2$    | $N_{Pixel}^{hits} \geq 1$ in the inner 2 layers |
| $N_{Pixel}^{hole} \leq 1$ | $N_{Pixel}^{hole} = 0$                          |

**Table 5.5:** Loose and Tight Selection criteria.  $N_{Si}$  represents the number of hits (or holes) in both the Pixel layers and SCT, whereas  $N_{Pixel}$  refers to the Pixel detector exclusively.

The Loose selection is the standard track reconstruction designed to give fast and efficient track reconstruction, at the cost of introducing more fake tracks. The Tight selection helps reduce the fake rate, giving higher quality tracks at the cost of reconstruction efficiency.

## Primary Vertices

In addition to track reconstruction, the Inner Detectors are also designed to reconstruct primary vertices (PVs), or the reconstructed location of the individual proton collisions in an event. PVs are seeded using the z position of Tight reconstructed tracks, then nearby tracks are added to the seeds in order to form a vertex. Tracks which are sufficiently displaced in z from a seed track are used to seed additional vertices.

Each vertex is required to have at least two tracks, with each associated track contributing to



the reconstructed location of the PV. An iterative  $\chi^2$  fit is used to match tracks to the seed vertex, where the tracks with the lowest  $\chi^2$  score have a larger influence on the position of the vertex, and tracks with higher  $\chi^2$  scores have less influence.

Vertexing is crucial for discriminating the key vertex in an event from the pileup vertices. The vertex with the largest  $p_T$  sum is selected as *the Primary Vertex*, with all other reconstructed vertices taken as pileup. The total number of PVs gives a useful measure of the amount of pileup ( $\mu$ ) in the event.

## Track Parameters

Each track candidate is characterized by five parameters:

$$\mathbf{track} = \{q/p, \theta, \phi, d_0, z_0\} \quad (5.7)$$

which are defined with respect to the track's *pedigree*, the closest approach to the associated primary vertex.  $\frac{q}{p}$  is the ratio of the particle's charge to the magnitude of its momentum; this is measured from the radius of curvature for the track. The angular location is given by  $(\theta, \phi)$ , where  $\theta$  is the polar angle between the particle's momentum vector and the z-axis,  $\phi$  is the azimuthal angle of the particle's momentum vector.

The spatial position is given by  $d_0$  and  $z_0$ , which are the transverse and longitudinal impact parameters, respectively.  $d_0$  gives the distance in the transverse plane between the track's pedigree and its associated PV, whereas  $z_0$  gives the corresponding z-axis distance.

Because the standard tracking is designed to find particles which originate from the primary vertex, it applies cuts on the impact parameters before applying any Tight or Loose selections. The relatively strict cuts of  $|d_0| < 10mm$  and  $|z_0| < 250mm$  is applied to all track candidates, in order to reconstruct primary particles. These cuts are useful for reconstructing the majority of tracks without much strain on computing systems, however they eliminate most track candidates which may correspond to LLP decays. In order to account for this, LLP searches can apply an additional track selection known as Large Radius Tracking (see section 5.2.1), which can help recover tracks from potential LLP decays.

## Large Radius Tracks

Large radius tracking (LRT) is an additional track selection, applied after the standard tracking, as a third pass for unused hits in the Inner Detector. Because many LLP searches look for particles which decay on the order of  $mm$  to  $cm$ , the LRT selection reduces the severity of the track candidate criteria, particularly on the impact parameters  $d_0$  and  $z_0$ . The key differences between the standard and LRT track selections are detailed in table 5.6. In addition to the relaxed spatial and hit requirements, the LRT selection also increases the minimum track  $p_T$  to be considered a candidate from 500 MeV to 900 MeV.

| Criteria                 | Standard Selection: | LRT Selection: |
|--------------------------|---------------------|----------------|
| Max. $ d_0 $             | 10mm                | 30cm           |
| Max. $ z_0 $             | 25cm                | 1.5m           |
| Max. $ \eta $            | 2.7                 | 5              |
| Max. $N_{Si}^{shared}$   | 1                   | 2              |
| Min. $N_{Si}^{unshared}$ | 6                   | 5              |
| Min. $N_{Si}^{hits}$     | 7                   | 7              |

**Table 5.6:** Differences in the Standard and LRT track selections [53].

The LRT selection follows a similar procedure to the inside-out selection, starting in the silicon detectors with a simplified algorithm which avoids creating multiple track candidates for a given set of hits. These tracks are extended into the TRT in the same way, although with the loosened criteria. Due to the increased cost in computational resources, the LRT selection is not applied to the entire ATLAS dataset, but only to a subsection intended for LLP searches. After the LRT selection is applied, it is combined with the set of tracks from the standard tracking.

### 5.2.2 Jet Reconstruction

Jets are reconstructed objects used to represent the final state of a quark or gluon produced in the collision, or in the decay of a massive particle. The collimated stream of particles produced by hadronization interact with the calorimeters and produce particle showers. These showers leave

energy deposits in the calorimeters which are roughly localized around the original direction of the parton.

## Topological Clustering

Jet reconstruction in ATLAS starts with topological (topo) clustering, which combines energy deposits in the calorimeters into 3D clusters. The clustering uses seed cells, which are determined according to the estimated level of noise ( $\sigma$ ) in any given cell. The topo-clustering algorithm proceeds as follows:

1. The collection of seed cells is determined as the set of all calorimeter cells with a deposited energy greater than  $4\sigma$  above the noise.
2. All cells adjacent to the seed (in 3D space) with energies greater than  $2\sigma$  are then added to the cluster
3. This process is repeated until no other adjacent cells have at least  $2\sigma$  of deposited energy
4. Finally, all adjacent cells are added to the cluster, regardless of the energy

The clustering algorithm uses noise suppression in this way to ensure it does not reconstruct arbitrarily large topo clusters.

## Small-R and Large-R Jets

In general, jets are reconstructed with a given angular distance parameter  $R = \sqrt{(\eta)^2 + (\phi)^2}$ , known as the jet radius, which roughly dictates the size of the object. This parameter is generally determined by the origin of the jet: If it originated from a quark or gluon, the resulting jet tends to be more highly collimated. Therefore, it can be fully reconstructed with a smaller  $R$ . However, if the jet is produced from the hadronic decay of a massive particle, then the jet tends to be more diffuse, requiring a larger  $R$  to reconstruct. ATLAS generally uses  $R = 0.4$  for small-R jets, and  $R = 1$  for large-R jets.

## Sequential Recombination Algorithms

After clustering, the next step in jet reconstruction is the jet-building algorithm. There are two main classes of algorithms used to create jets: fixed cone, and sequential recombination algorithms. ATLAS uses the latter for jet reconstruction, so that will be the focus of this section.

The general algorithm is based upon two distance parameters,  $d_{ij}$  and  $d_{iB}$ , which are the distance between a reference object  $i$  and another object  $j$ , and the distance between the reference object and the beamline, respectively. These distances are defined as:

$$d_{ij} = \min(p_{T,i}^{2k}, p_{T,j}^{2k}) \frac{\Delta_{ij}^2}{R^2} \quad (5.8)$$

$$d_{iB} = p_{T,i}^{2k} \quad (5.9)$$

where  $\Delta_{ij}^2 = (y_i - y_j)^2 + (\phi_i - \phi_j)^2$ . Expressed this way, the distances only depend on two defined parameters,  $R$  which is the jet radius, and  $k \in [-1, 1]$  which corresponds to which algorithm is being used. The reference object  $i$  could be a topo cluster, or a pseudo-jet (a group of clusters already combined by the algorithm), depending on the stage of the sequence. The first reference object however, is always the hardest (highest  $p_T$ ) topo cluster in the set.

Independent of  $k$ , the basic operating principals for sequential recombination are as follows:

1. Calculate the distance  $d_{ij}$
2. Compare  $d_{ij}$  with distance  $d_{iB}$
3. if  $d_{ij} < d_{iB}$  then combine the two objects  $i$  and  $j$ , otherwise, remove  $i$  from the list of entities and label it a jet
4. Continue until there are no remaining objects in the set, each time starting the sequence with the hardest remaining cluster

Although the steps are the same, changing the parameter  $k$  will change how the distances are calculated, so it can drastically change the behaviour of the algorithm. Taking  $k = 0$  gives the Cambridge-Aachen algorithm, which is independent of the cluster  $p_T$ . Here  $d_{iB} = 1$ , so  $d_{ij} = \frac{\Delta_{ij}^2}{R^2}$  must be less than 1 in order to be included in the jet.

Values of  $k$  in the range 0 to 1 have similar behaviour to  $k = 1$ , which gives the  $k_t$  algorithm. Here  $d_{ij}$  will always be scaled to the  $p_T$  of object  $j$  (since  $i$  is either the hardest cluster or a pseudo-jet which contains it), this makes the jet susceptible to soft radiation, which gives the resulting jets irregular boundaries.

Negative values of  $k$  behave similarly to  $k = -1$ , which gives the anti- $k_t$  algorithm. This algorithm is known as *soft resilient*, since  $d_{ij}$  will always be scaled to the reference object's  $p_T$ . This results in regularly shaped conical jets, centered around the hardest cluster. Jets in ATLAS are generally reconstructed with the anti- $k_t$  algorithm.

## Jet Calibration

Energy scale determination is the first step in calibration, and it happens before the jet-building. There are generally two options: the raw electromagnetic scale, and the hadronic scale. For jets calibrated to the EM scale, the energy of the clusters is simply taken as the sum of the energy in each cell. These clusters are known as EM-topo clusters, and are usually applied to small-R jets. However, if the jet is calibrated to the hadronic scale, then a local cell weighting (LCW) is applied to the cells within a cluster, these are known as LCW-topo clusters and are generally used for large-R jets.

The second step in jet calibration comes after the building stage, and depends on the type of jet. For small-R jets, there are four main steps: 1. Origin and Pileup Correction 2. MC-Based Energy Calibration 3. Global Sequential Calibrations 4. Residual *in-situ* Calibrations. Origin and pileup correction (1) first ensures that the jet is pointing to the correct primary vertex and not a pileup vertex, then it scales the energy based on  $\langle\mu\rangle$  and  $N_{PV}$ . (2) uses Monte-Carlo based jet information to try and account for information in hadronic showers that is not available to the detectors. The global sequential calibrations (3) attempt to account for shower development, punch through and flavor differences (changes in the jet behavior based on the original quark flavor). The final step (4) uses the well-defined central jets with  $\eta < 1$  to successively calibrate more forward jets.

The calibrations for large-R jets are similar, with some important differences. Instead of the origin and pileup corrections (1), large-R jets use jet *grooming*; each jet is reclustered using a similar algorithm but a smaller  $R$  parameter so that the original jet is now made of smaller sub-jets, then any sub-jet with a  $p_T$  less than some kinematic cut is eliminated and the remaining sub-jets are

regrouped into a new trimmed jet. Jet grooming helps large-R jets with pileup suppression, since they are more susceptible to including clusters from pileup vertices and not the main collision vertex. Other key differences are that the MC (2) and in-situ (4) calibrations are applied both to the jet energy and the jet mass. There is no global sequential calibration applied to large-R jets.

This analysis uses EM-topo jets clustered using the anti- $k_t$  algorithm with  $R = 0.4$ .

### 5.2.3 Displaced Vertex Reconstruction

Displaced or secondary vertices (DVs) are common to many LLP searches, where the particles decay over macroscopic distances which are displaced from the initial collision but still contained within the Inner Detector. Because the lifetimes of the potential LLPs span such a large range, DV reconstruction requires the use of both standard and LRT tracks.

DV reconstruction starts by selecting seed track pairs, those which may be consistent with the decay of an LLP. In order to be included in a potential pair, each track must first pass a set of quality cuts. The set of all tracks which pass these cuts are labelled as *selected* tracks, and are passed to the pair finding algorithm. The track selections are designed to find a balance between efficiently reconstructing LLP decays, and not including fake vertices, as such, they generally sit between the Loose selection of the standard tracking, and the LRT selections.

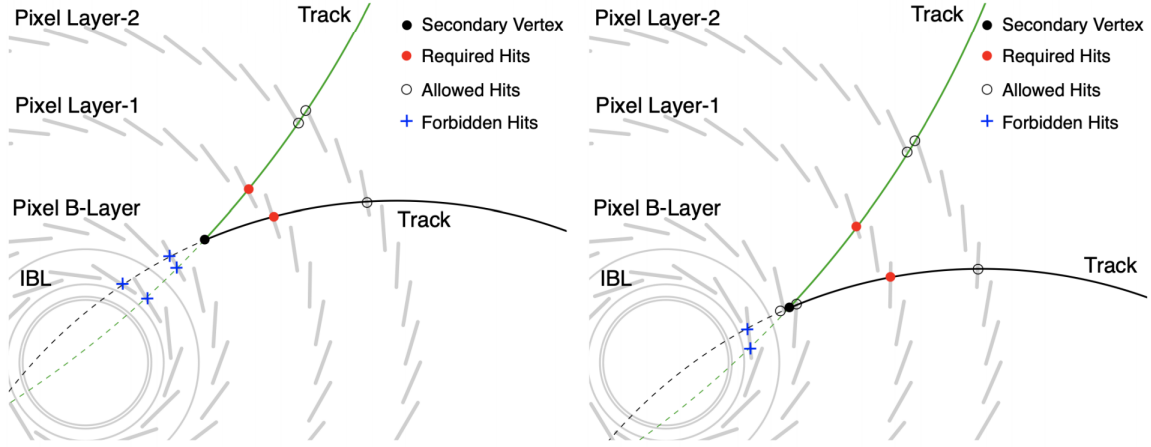
Seed pair formation proceeds in two passes. The first groups together all possible pairs of tracks based on the initial track parameters (5.7), this allows for an estimation of the seed position. Any seeds are removed if either of the tracks in the pair are sufficiently displaced, or if the radial position of the seed is outside of the silicon detectors. The second pass takes the remaining seeds and applies precision track fitting to more accurately locate the seed position. This fit minimizes a  $\chi^2$ , to estimate the quality of the seed. If it fails a goodness of fit, or if the updated vertex position is outside the SCT, then it is removed.

Once the initial set of seeds is finalized, the corresponding tracks must pass a hit-pattern consistency test, which has three basic components:

- Tracks are required to have hits on the ID layers which are at larger radii than the seed.
- Tracks are required to have at least one hit in the ID layer which immediately follows the seed.

- Tracks are forbidden from having hits on any ID layer located at smaller radii than the seed.

This test is designed to ensure the vertex is consistent with the decay of an LLP; the first point ensures the basic logic of the vertex formation is followed (particles typically travel radially outward, so tracks should leave hits after the vertex, not before it). The second point ensures proper reconstruction of the vertex location, and the third point ensures it is from an LLP. Figure 5.3 shows the hit-pattern for two candidate vertices. The vertex on the left is close enough to a detector layer that there is an allowed hit in the adjacent module, even if it is at a smaller radius than the vertex. However, hits on the layer interior to the vertex are still forbidden.



**Figure 5.3:** Schematic diagram showing two candidate vertices in black, the corresponding required hits and further allowed hits in white, as well as the forbidden hits in blue. Figure from [56]

The two-track vertices which pass the hit-pattern consistency test then go through multi-track vertex forming; where all possible associated tracks are matched to the vertex, and their compatibility is assessed to determine which tracks should be matched to which vertices. If any tracks are matched to multiple vertices in this process, the goodness of associated is determined, and the track is associated to the vertex with the best score.

There is a possibility that the reconstruction algorithm up to this point may produce multiple vertices which correspond to the same LLP decay, therefore *splitting* the actual vertex. These split vertices cause a discrepancy in the track multiplicity and invariant mass of the actual vertex, which makes it more difficult to properly reconstruct the LLP, so an attempt is made to merge them. Candidates for merging are ordered in ascending track multiplicity, as the merging process

will attempt to match a low multiplicity vertex ( $V_{LM}$ ) to a high multiplicity vertex ( $V_{HM}$ ), since a  $V_{HM}$  will have a better position resolution. There are three ways in which vertices can be merged:

- The  $V_{LM}$  is seeded with the location of the  $V_{HM}$ , and then the vertex is refit.
- The selected tracks from the  $V_{HM}$  are iteratively associated to the  $V_{LM}$ , for each track the vertex is again seeded with the location of the  $V_{HM}$  and refit.
- The tracks from both the vertices are used to form a new vertex, again seeded with the location of the  $V_{HM}$  and refit.

In each case the two vertices are merged if the new location of the  $V_{LM}$  is compatible with the location of the  $V_{HM}$ . In order to be compatible, the two vertices must be separated by a distance less than  $10\sigma$ , where  $\sigma$  is the uncertainty on the differences of their locations. If there are any vertices which do not pass the merging criteria, but are within  $1mm$  of each other, they are forced to be merged, and all merged vertices are refit using the tracks from both vertices.

Until this point, the only tracks that have been used are those which have passed the set of quality requirements. So, the final step in DV reconstruction is to associate any additional tracks that did not pass the initial selection but which may be compatible. These associated tracks must pass a less stringent set of quality criteria, and a modified hit-pattern consistency test. Additional tracks can only be associated to a single vertex, so once they are matched they are removed from the list of available tracks.

## 5.3 Event Selection

Event selection for the EJs analysis begins with the ATLAS trigger system, which allows ATLAS to determine which events to record, and which to discard. Next, the *Emerging* component of the RPVLL filter is applied to select events which contain emerging-jet-like signatures.

### 5.3.1 The ATLAS Trigger System

The collision rate delivered by the LHC during Run II was  $40MHz$ , which corresponds to an event every  $25ns$ . Since ATLAS has over 100 million detector readout channels, a single event



takes up roughly  $1.5MB$  of computational storage. Therefore, if ATLAS recorded and stored every single event it would consume  $60TBs^{-1}$ , enough to very quickly consume the global storage supply. Furthermore, only a small fraction of the events actually represent interesting physics processes, so the vast majority need not be kept. As such, it is crucial for ATLAS to be able to quickly and effectively select out any events which may actually contain interesting physics, and discard the rest. It achieves this through the use of a two-tiered *trigger* system.

### **The Level-1 Trigger (L1T)**

The first tier of the trigger system is the Level-1 Trigger (L1T), a hardware based trigger which uses information from the muon spectrometers and the calorimeters to quickly decide if an event is worth keeping. Because the L1T runs at the same time that events are occurring, it is referred to as an *online* selection.

The L1T is designed to reduce the decision rate from  $40MHz$  down to roughly  $100kHz$ . To do this, it needs to be able to make a decision in less than  $2.5ms$ . Because of this quick decision time, the L1T uses only some components of the detector information, and simplified reconstruction algorithms to efficiently define a Region of Interest (RoI) in the detector components.

RoIs are used to determine locations (in  $\eta$  and  $\phi$ ) where the transverse energy ( $E_T$ ) exceeds some pre-determined threshold. The muon spectrometer component of the L1T uses the trigger detectors (RPCs and TGCs) to find RoIs which may correspond to energetic muons. The calorimeter component uses both the ECal and HCal to create trigger towers; high granularity regions of the calorimeters where energy is deposited. Depending on the trigger menu, the predetermined list of trigger selections which correspond to different physics processes, the L1T may select on single, high energy towers, or multiple lower energy towers. After the initial selections, the L1T passes the RoIs on the next tier of the trigger, in order to make the final decision.

### **High Level Trigger (HLT)**

Events which pass the L1T are sent to the High Level Trigger (HLT), which uses the full information from all of the detector components. The software-based HLT uses more complete reconstruction, similar to the full *offline* reconstruction, in order to reduce the final decision rate to roughly  $1kHz$ , with an event decision every  $200ms$ .

Using the full granularity and precision, the HLT reconstructs objects within the RoIs, for the jet objects in particular, this is done using the same topological clustering algorithm as in section 5.2.2. Jet energies are computed in the same way as the offline reconstruction, but calibrated without track information. This gives very good agreement between HLT jet energies and fully reconstructed jet energies.

### 5.3.2 Emerging Jets Selections

Not all events which pass the HLT are used for the analysis; events must pass the *Emerging* component of the RPVLL filter. The filter applies two selections, one on signal-like events which populate the search region (see sections 6.1.4 and 6.2.5) and the other on dijet background-like events which populate the validation region (see section 6.4).

The signal-like selection takes events which pass the lowest unrescaled four-jet trigger out of the following list:

- *HLT\_4j90*
- *HLT\_4j100*
- *HLT\_4j110*
- *HLT\_4j120*
- *HLT\_4j130*
- *HLT\_4j140*
- *HLT\_4j150*

The first part of the trigger name indicates that it is from the HLT, not the L1T, the second indicates how many jets (4), and the number at the end indicated the  $p_T$  threshold in GeV.

The signal selection looks for events which contain at least four jets that all have  $p_T \geq 120$  GeV and  $|\eta| < 2.5$ . This means that events are expected to at least pass the *HLT\_4j120* (in reality, events may have at least four jets with  $p_T \geq 120$  GeV, but may not trigger the HLT), as well as all of the lower triggers. There may also be events which pass the lower triggers but do not meet the kinematic requirements.

The background-like selection uses a heavily prescaled single-jet trigger, the *HLL\_j110*. This picks out events which have at least one jet with  $p_T \geq 110$  GeV. The background-like events are those which have at least 2 jets, with same kinematic requirements on  $p_T$  and  $\eta$  as the signal selection. This means that there may be several events which pass the trigger but not the additional selection.

## Chapter 6

# Emerging Jets Analysis

To date, the Emerging Jets analysis has implemented a two-pronged approach to its search: one method which uses a cut-based selection and the *ABCD Method* (see section 6.1) to provide a *Model Independent* result, as well as a machine learning implementation known as a *Boosted Decision Tree* (see section 6.2) to improve on background suppression and provide a *Model Dependent* result.

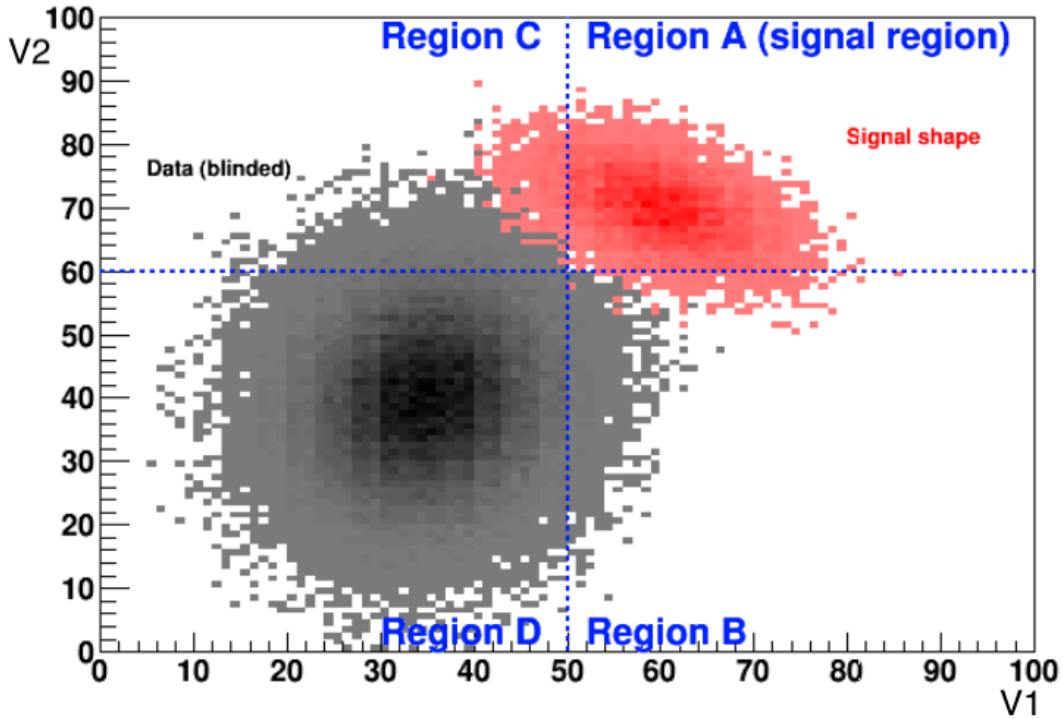
For a search-based analysis, the end result is generally either a *Confidence Limit* or a *Discovery Sensitivity*. A Confidence Limit (CL) is an upper limit on the theoretical cross-section for the process in question (in this case it would be the pair-production of the dark mediator  $X_d$ ), which is given in the event that no statistically significant evidence of a signal is found in the current search. A Discovery Sensitivity is the corresponding result if the opposite is true, and there is evidence of the physics process; in this case the statistical significance is given in terms of the uncertainty on the background prediction. Since this thesis presents only a Monte-Carlo based study of the signal efficiency, the results will only be given in terms of a CL. However, after unblinding, the corresponding ATLAS analysis may find a statistically significant excess of events and would therefore have to provide a discovery sensitivity.

Model independence is useful for theoretical reinterpretation of the final result, in the case of a CL. Since the predicted cross-section is independent of the benchmark model (i.e. the probability of pair-producing dark mediators is the same regardless of what the mediators decay to), if the CL set by the analysis is independent of the model as well, then the result can be applied to a broader scope of models which involve some heavy scalar mediator, but may not share many of the model specifics. Model dependence makes this reinterpretation more difficult, since it becomes harder to

separate the result from the specifics of the model. However, it tends to use more of the available information in a more efficient manner, which aids in background suppression. This allows for more stringent tests of the theoretical cross-sections (lower upper limits), and improves the possibility of seeing a signal.

## 6.1 The ABCD Method

The ABCD method [58] is a data-driven background estimation method. Two (ideally uncorrelated) variables ( $V_1, V_2$ ) are plotted on a 2D plane, which is then divided into four regions by a cut on each variable ( $C_1, C_2$ ). The goal of the method is that the signal event will fall mainly in one of the four regions, which is labeled the A region, and the background events will fall mainly in the other three *control* regions. The two neighboring control regions to A are labeled B and C, and the region diagonally adjacent to A is labeled D.



**Figure 6.1:** Illustrative example of an ABCD plane with the (blinded) data distribution in black, and the potential signal distribution in red. The plane is defined using the two variables  $V_1$  and  $V_2$ , with the cuts  $C_1$  and  $C_2$  given by the two dashed blue lines. Figure from [58].

If the two variables are sufficiently uncorrelated for the background distribution, then the ratio of background events in regions C and D should be equal to the ratio of events in regions A and B (within statistical uncertainties):

$$\frac{N_A^{Bkg}}{N_B^{Bkg}} = \frac{N_C^{Bkg}}{N_D^{Bkg}}. \quad (6.1)$$

Thus, the number of background events in region A (the signal region) can be estimated using only the background events in the control regions:

$$N_{A,est}^{Bkg} = N_B^{Bkg} \frac{N_C^{Bkg}}{N_D^{Bkg}}. \quad (6.2)$$

If we assume that the majority of the signal events are in region A (which is the goal of the method), and that the number of background events outweighs the number of signal events, as is the case for a background-dominated analysis, then the number of background event in each control region should be roughly equal to the total number of events in that region, that is  $N_i \approx N_i^{Bkg}, i = B, C, D$ . Therefore, equation 6.2 becomes:

$$N_{A,est}^{Bkg} = N_B \frac{N_C}{N_D}. \quad (6.3)$$

This is crucial if the analysis is blinded in the signal region (A) but has the data for the control regions. Once the analysis unblinds, it can compare the actual counts in region A to the estimate to determine if an excess of signal is observed.

### 6.1.1 Assumptions of the ABCD Method

The ABCD method is based upon two crucial assumptions: the first is that the two variables are uncorrelated (section 6.1.2), and the second is that the signal events are contained within region A. These assumptions may not always be valid, and should always be checked. Furthermore, it is important to understand their effects and how to deal with them.

## Signal Leakage

If the signal distribution has a significant portion of events in the control regions (B, C, and D), then the assumption that  $N_i \approx N_i^{Bkg}$  becomes less reliable. If a prediction exists for the signal events in the control regions, then it can simply be subtracted to obtain the number of background events:  $N_i^{Bkg} = N_i - N_i^{Sig}$ . However, if there is no reliable signal prediction (as is the case for most searches for new physics), then it must be assumed that the signal contribution is negligible.

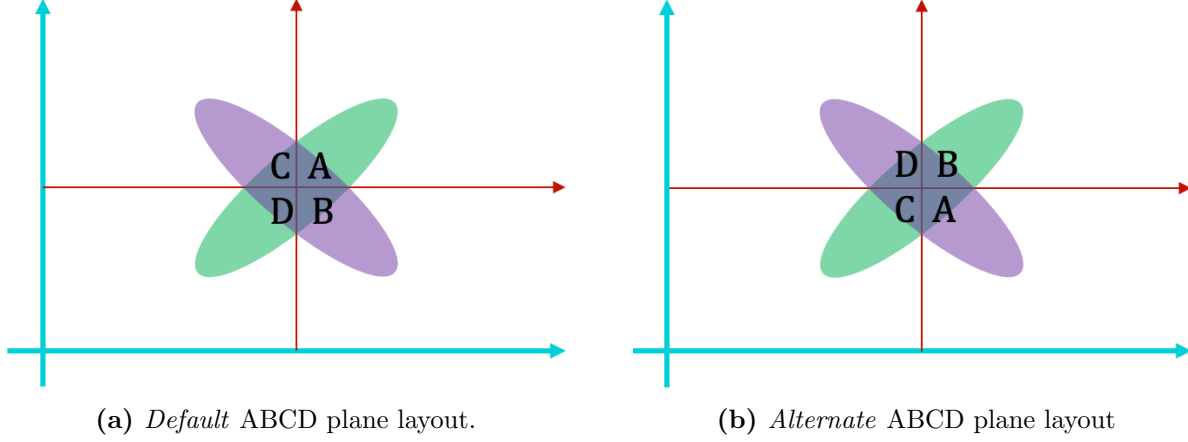
Even if the proportion of signal events is large in the control regions, the assumption remains valid if the analysis is background-dominated; the scale of the signal distribution in the control regions will be significantly smaller than the background in these regions. That is,  $N_i = N_i^{Bkg} + \epsilon_i^{Sig} \approx N_i^{Bkg}$ , where  $\epsilon$  represents a small number of events.

The profile likelihood method (see section 6.1.3) can account for signal leakage by using the shape of the signal distribution in the control regions to estimate the signal contribution. Dealing with signal leakage introduces some model dependence in the background estimation, even if the background distribution itself is model independent. The degree of signal leakage will, in general, be different for various signal models, therefore the signal subtraction will give a different background estimate depending on which signal model is being tested.

### 6.1.2 Correlation in the ABCD Plane

The main assumption of the ABCD method is that the two variable are uncorrelated, such that the proportion given by 6.1 holds. However, any correlation ( $\rho$ ) will cause this proportionality to be wrong, and therefore cause a discrepancy between the actual counts in region A ( $N_A$ ) and the estimate ( $N_{A,est.}$ ) given by 6.2.

The effect of the correlation depends not only on its magnitude, but both on the arrangement of the ABCD plane, and the sign of the correlation as well. Figure 6.2 shows two possible layouts of the ABCD plane. In the *default* layout (a), region A is defined as being larger than the two cuts ( $V_1 > C_1$  and  $V_2 > C_2$ ). The *alternate* layout (b) is defined by inverting the selection on the y-axis ( $V_2 < C_2$ ). Note that the two other possible arrangements can be achieved by inverting the selection on the x-axis in both cases, but the effects of the correlations will be the same. Note that the alternate arrangement can be turned into the default arrangement by taking  $V_2 \rightarrow -V_2$ .



**Figure 6.2:** Two possible arrangements of the ABCD plane. In each case, the same two cuts (red arrows) define the four regions, but the regions are labelled differently to optimize the signal region. Two correlated distributions are shown, one with a large positive correlation (green), and one with a large negative correlation (purple).

In the case of the default arrangement, a positive correlation skews events from the B and C regions, and into the A and D regions. This causes 6.2 to underestimate the actual counts in region A. The effects of a given correlation in the default arrangement is given by:

$$\rho > 0 \rightarrow N_A > N_{A,est.} \quad (6.4)$$

$$\rho < 0 \rightarrow N_A < N_{A,est.} \quad (6.5)$$

While for the alternate arrangement it is a negative correlation which pulls events into the A and D regions, so the opposite is true:

$$\rho > 0 \rightarrow N_A < N_{A,est.} \quad (6.6)$$

$$\rho < 0 \rightarrow N_A > N_{A,est.} \quad (6.7)$$

The degree of the misestimate is mostly determined by the magnitude of the correlation, where a larger  $|\rho|$  will, in general, give a larger disagreement between  $N_A$  and  $N_{A,est.}$ .

This simple correlation effect gives one example of a bias in the ABCD method, which may cause the background estimate to differ from the actual counts in region A. Others may include bin-to-bin correlations between the four bins which correspond to the ABCD regions. An evaluation of these effects, and an attempt to correct for a non-uniform response is given in section 7.1.3.



### 6.1.3 The Likelihood Method

The above description of the ABCD method gives the standard arithmetic approach. However, in the cases where the signal leakage is non-negligible, a profile likelihood fit using the basic model of the ABCD method can be implemented in order to give a more robust background estimation.

The basic ABCD model, expressed in terms of free parameters of the fit (indicated by the  $\sim$ ), is given by:

$$N_A = \tilde{m}\tilde{N}_B \quad (6.8)$$

$$N_C = \tilde{m}\tilde{N}_D \quad (6.9)$$

where  $\tilde{m}$  is known as the *Transfer Factor*, and represents the ratios given by  $\frac{N_A}{N_B}$  and  $\frac{N_C}{N_D}$ . This allows the full ABCD relationship to be described by three parameters:  $\tilde{m}$ ,  $\tilde{N}_B$ , and  $\tilde{N}_D$ . Thus, the likelihood for observing the data given these parameters is:

$$L(\mathbf{data}|\tilde{m}, \tilde{N}_B, \tilde{N}_D) = \text{Poisson}(N_{Total}|\tilde{N}_{Total}) \times \prod_{N_A} \frac{\tilde{m}\tilde{N}_B}{\tilde{N}_{Total}} \prod_{N_B} \frac{\tilde{N}_B}{\tilde{N}_{Total}} \prod_{N_C} \frac{\tilde{m}\tilde{N}_D}{\tilde{N}_{Total}} \prod_{N_D} \frac{\tilde{N}_D}{\tilde{N}_{Total}} \quad (6.10)$$

where **data** is the set of actual counts in the four regions:  $\{N_A, N_B, N_C, N_D\}$ ,  $N_{Total}$  is the sum of all four regions:  $N_A + N_B + N_C + N_D$ , and  $\tilde{N}_{Total}$  is the corresponding value from the fit:  $\tilde{m}\tilde{N}_B + \tilde{N}_B + \tilde{m}\tilde{N}_D + \tilde{N}_D$ .

### Signal Leakage in the Likelihood Method

In the case that the signal distribution leaks into the control regions, the fit takes on an additional parameter,  $\tilde{\mu}$ , which represents the signal efficiency. The estimated counts in the four regions are then given by four fit parameters  $(\tilde{m}, \tilde{N}_B, \tilde{N}_D, \tilde{\mu})$ , and four corresponding values from the signal Monte Carlo ( $N_{i,MC}^{Sig}$ ):

$$N_A = \tilde{m}\tilde{N}_B^{Bkg} + \tilde{\mu}N_{A,MC}^{Sig} \quad (6.11)$$

$$N_B = \tilde{N}_B^{Bkg} + \tilde{\mu}N_{B,MC}^{Sig} \quad (6.12)$$

$$(6.13)$$

$$N_C = \tilde{m}\tilde{N}_D^{Bkg} + \tilde{\mu}N_{C,MC}^{Sig} \quad (6.14)$$

$$N_D = \tilde{N}_D^{Bkg} + \tilde{\mu}N_{D,MC}^{Sig}. \quad (6.15)$$

Here the signal efficiency ( $\tilde{\mu}$ ) acts as a normalization for the signal counts ( $N_{i,MC}^{Sig}$ ). Allowing  $\tilde{\mu}$  to be a free parameter of the fit only works when all four regions are used. If a blinded version of the fit is used (i.e the counts in region A are not estimated), then  $\tilde{\mu}$  cannot be a free parameter, but can still be used if it is based on the predicted cross-section (and scaled to a corresponding luminosity).

#### 6.1.4 Model Independent Search Region

The EJs search has implemented a cut-based selection and the ABCD method for its model independent, data-driven background estimation method. The base event selection is made using the 4 leading (highest  $p_T$ ) jets, and it matches the signal-like selection of the *Emerging* component of the RPVLL filter:

- Jet Multiplicity ( $n_{Jet}$ )  $\geq 4$
- Passes Lowest Unprescaled 4-Jet Trigger from:  $HLT_{4j90-150}$
- $p_T^{4 \text{ Lead Jets}} \geq 120 \text{ GeV}$
- $|\eta|^{4 \text{ Lead Jets}} > 2.5$

The search region is then defined by a selection on the  $H_T$ , the scalar sum of jet  $p_T$ :

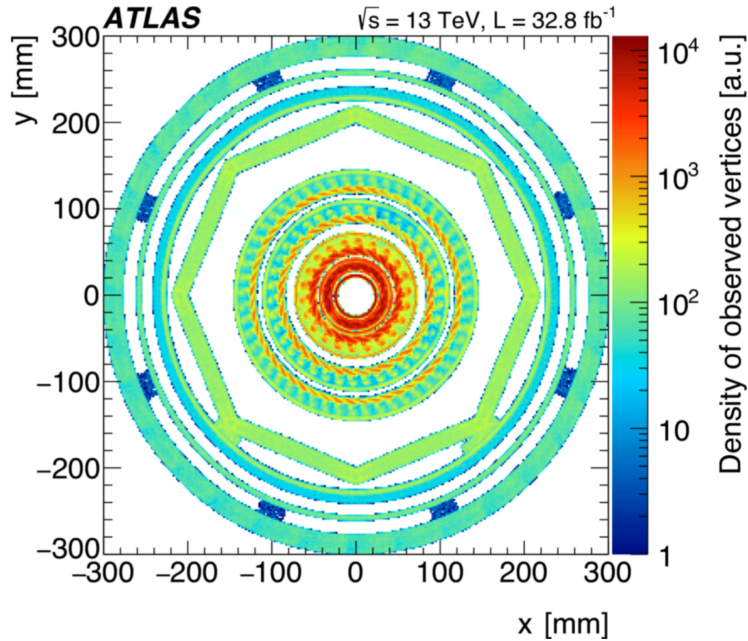
$$H_T^{4 \text{ Lead Jets}} = \sum_{i=1}^4 p_{T,i} \geq 1000 \text{ GeV}. \quad (6.16)$$

These criteria are collectively known as the *pre-selection* cuts, which are used to select all search-region events for both the model dependent and independent components of the analysis.

## Displaced Vertex Selection

In addition to the pre-selection, the model independent ABCD plane applies an additional selection on the displaced vertex multiplicity ( $nDV$ ).

Event selection is designed to optimize for the EJs signature described in section 3.4, where many displaced vertices are expected to be contained within the jet cone. As such, DV selection begins with jet matching, where a DV is matched to a reconstructed jet if it is within  $\Delta R = 0.4$  of the jet's 4-momentum vector.



**Figure 6.3:** Detailed Map of the reconstructed displaced vertex density in the regions vetoed by the Inner Detector material map. Figure from [59].

In addition to jet-matching, the selected DVs must pass a Material Map Veto. Collision products can interact with the dense material of the detectors and produce particles which are separate from the actual event. These secondary particles can get reconstructed as a displaced vertex, which looks similar to a DV produced in an LLP decay. Therefore, a detailed map of the Inner Detector material (figure 6.3) is used to determine regions where these DVs are more common, so that reconstructed DVs in these areas can be effectively vetoed.

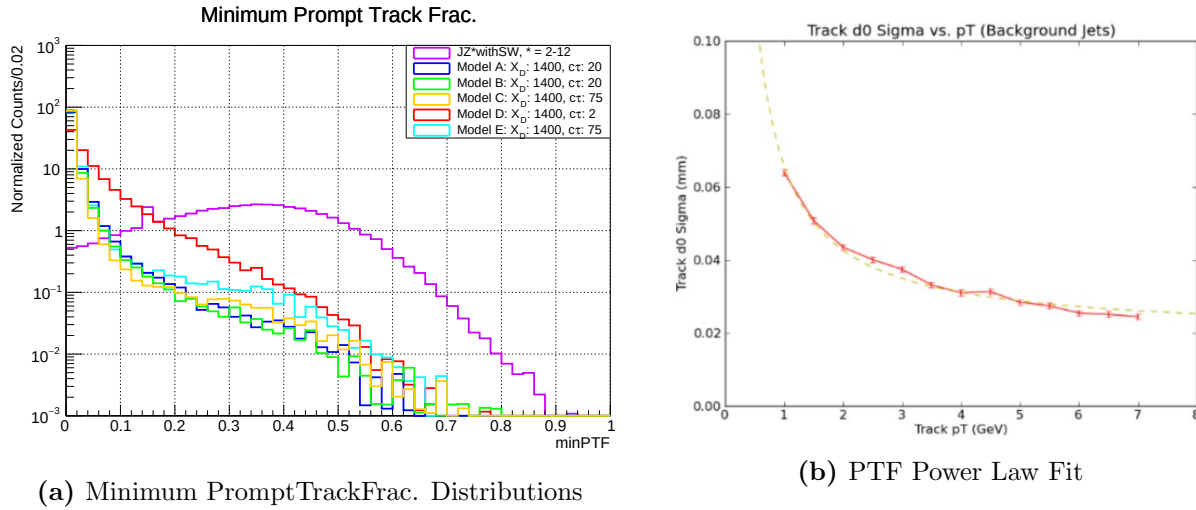
The model independent ABCD method requires events with at least 1 reconstructed, jet-matched displaced vertex which passes the material veto.

## Prompt Track Fraction

The y-axis of the model independent ABCD plane is an event-level version of the *PromptTrackFrac.* (*PTF*); a measure of what fraction of a jet's  $p_T$  comes from prompt tracks. The *PTF* is defined at jet-level as:

$$PTF = \sum_i \frac{p_{T,i}^{track}(d_0 < 2.5f(\sigma_{d_0}; p_T^{track}))}{p_T^{jet}} \quad (6.17)$$

where the tracks must be jet-matched, and have  $p_T \geq 2$  GeV in order to be considered in the sum.



**Figure 6.4:** Minimum PromptTrackFrac. ( $minPTF$ ) distributions (a) and power law fit (b). The distributions in (a) are given for the QCD multijet background (JZ\*withSW, \* = 2-12) and for the 5 signal points in the canonical set of 15 with the 1400 GeV mediator mass. The fitted function in (b) gives the power law fit which determines the prompt criteria when calculating the PTF. Note that, for the sake of comparison, the distributions in (a) have been normalized to unit area. Figure (b) from A. Kilgallon.

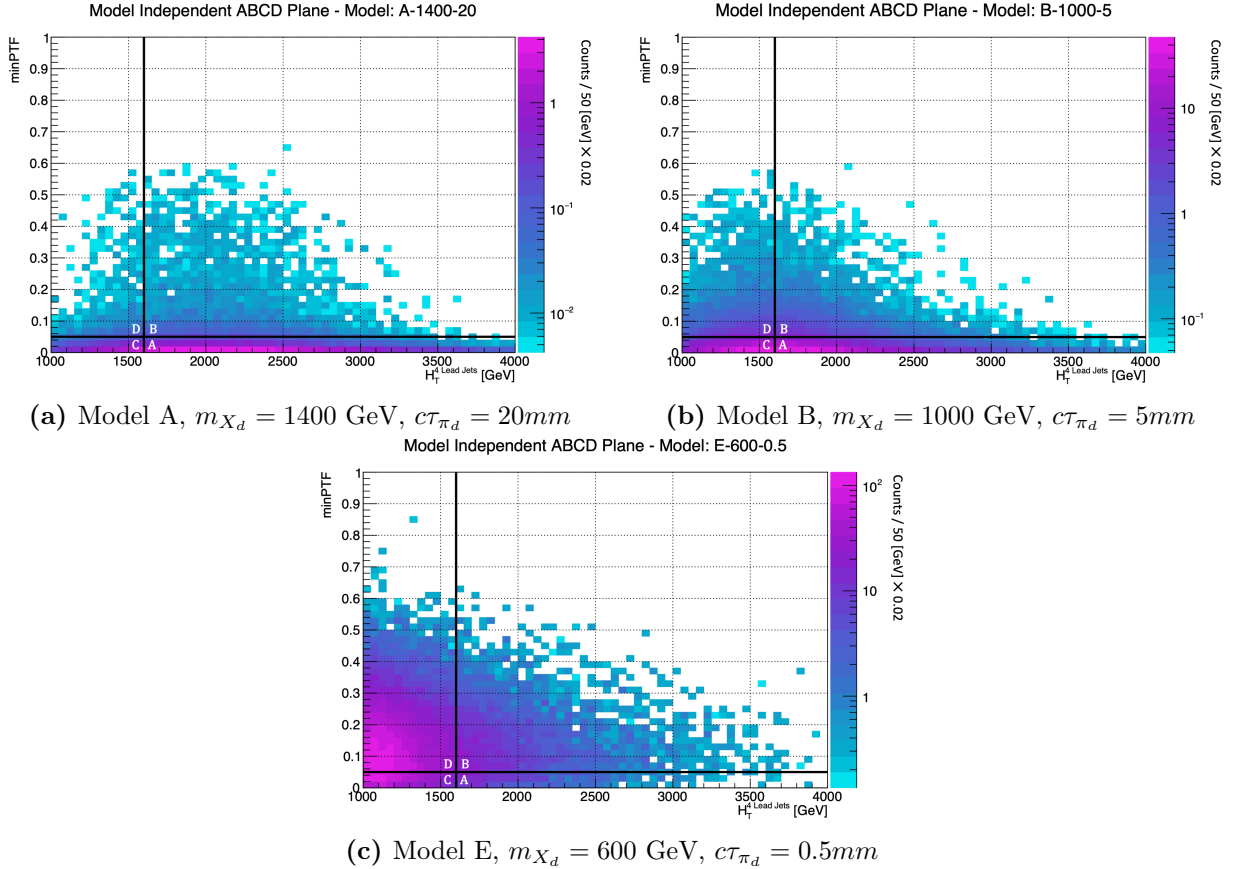
The prompt criteria is determined using  $f(\sigma_{d_0}; p_T^{track}) = \frac{0.0463}{p_T^{track}} + 0.0195$ ; a power law fit (shown in figure 6.4b) which determines the uncertainty of a track's  $d_0$  for a given  $p_T$ .  $f(\sigma_{d_0}; p_T^{track})$  was determined using a sample of  $Z$ +jets events and is designed to model background QCD jets in multijet events. If the measured  $d_0$  of the track is less than 2.5 times this fit uncertainty, then the track is considered prompt. In order to make this an event-level variable, the minimum PTF ( $minPTF$ ) of the 4 leading jets is taken.

The expectation is that QCD jets will largely be composed of prompt tracks. Emerging jets, which first shower in the dark-sector, are expected to have a larger proportion of displaced tracks.

Thus, even the minimum  $PTF$  of a QCD event should be higher than that of an event with an emerging-like jet, as can be seen in figure 6.4a.

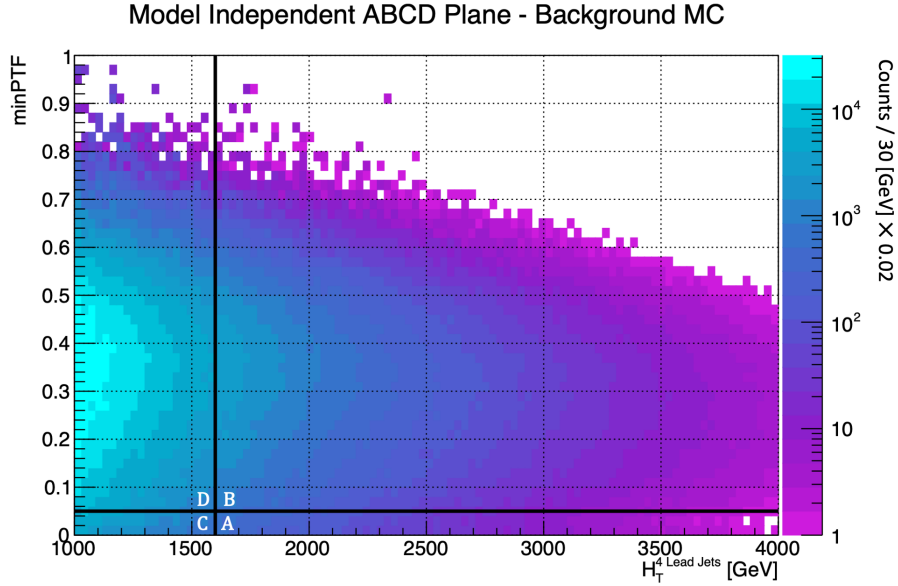
### Model Independent ABCD Plane

The model independent ABCD plane then uses the event  $H_T$  and the  $minPTF$  as its x and y axes respectively. The plane follows the alternate arrangement shown in figure 6.2b, with regions defined by the cuts  $H_T^{ALeJets} = 1600$  GeV and  $minPTF = 0.05$ . Example ABCD planes for the QCD background and three representative signal points are shown in figures 6.6 and 6.5. Because the ABCD plane is defined the same way for both signal and background, the ABCD method will give one single background estimate, regardless of the signal point, and is therefore model independent.



**Figure 6.5:** EJs signal MC distributions in the search region defined by the model independent ABCD plane. Three signal points (312004 (a), 312028 (b), 312090 (c)) are shown for each of the three dark mediator masses. The cuts of  $H_T^{ALeJets} = 1600$  GeV and  $minPTF = 0.05$  (shown in black) define the four regions, which follow the alternate ABCD arrangement. The distributions have been scaled to the full Run II luminosity of  $139 \text{ fb}^{-1}$ .

The z-axis scales of the distributions reflect the background dominance of the analysis, which demonstrates the need for significant background suppression in region A. There are two main competing factors which present challenges for the model independent method: the higher mediator mass ( $m_{X_d} = 1400$  GeV) models have a much smaller cross-section, so despite being more easily separated from the background, any potential signal will be small. The lower mass mediator ( $m_{X_d} = 600$  GeV) cases have cross-sections which are roughly 2 orders of magnitude larger; however, these models tend to behave more SM-like, and are therefore more difficult to separate from the QCD background (see figure 6.5c). These are the main motivations behind the inclusion of the model dependent approach in the analysis.

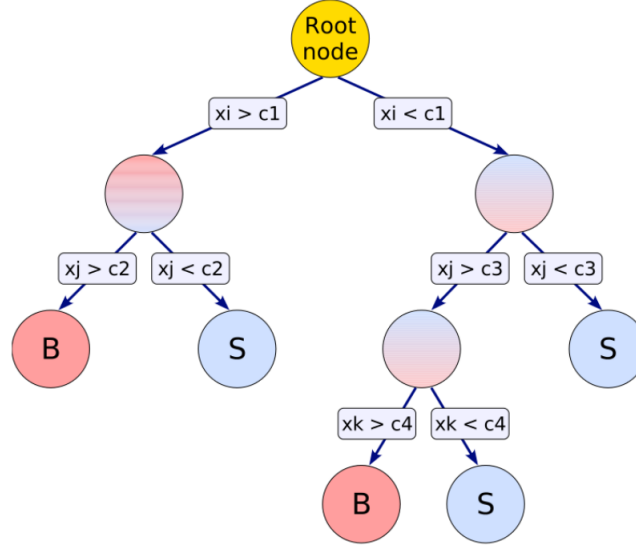


**Figure 6.6:** QCD multijet background distribution in the search region defined by the model independent ABCD plane. The cuts of  $H_T^{4LeadJets} = 1600$  GeV and  $minPTF = 0.05$  (shown in black) define the four regions, which follow the alternate ABCD arrangement. The distribution has been scaled to the full Run II luminosity of  $139 \text{ fb}^{-1}$ .

## 6.2 Boosted Decision Trees

A Boosted Decision Tree (BDT) is a machine learning algorithm which uses a successive series of decision-making nodes (leaves) in order to classify a dataset based on several inputs [60]. The decisions are made using variable cuts on the inputs, with each layer of nodes making decisions about just one input variable. Each node is given a weight, which helps determine how important

that node's decision is to the end result. These weights can then be adjusted in order to change the impact of that node. For example, if one of the input variables is more powerful at separating the data set into the desired outcomes, the node which corresponds to that variable will be weighted more strongly.



**Figure 6.7:** Schematic diagram of a general decision tree, showing the node structure. The variables  $x_{i,j,k}$  correspond to a given node, with the cuts  $c_{1-4}$  used for decision-making. The final node is labelled S (signal) or B (background) depending on the decision it represents. Image from [60].

The final decision is made after the data has passed through each node, which classifies the data as either signal-like or background-like. Here signal and background do not necessarily correspond to physics events, but rather refer to whether the data is sufficiently close to a target value or not. The benefit of the BDT is that it can take in several weakly-separating inputs, and return one more strongly separating classifier.

As with all machine learning algorithms, a BDT is developed by separating a dataset into a training set and a testing set. In both cases, the target value for each data in the set should be known. The training set is used to optimize the classifier (decide on cuts, set the weights for each node, etc.). Here, the algorithm uses the target value to determine how accurately the classifier sorts the data. The testing set is then used to see how well the classifier performs on an independent dataset. Here, the target value is used to determine the success rate, but no additional changes to

the classifier can be made. The testing set can also be used to evaluate if the classifier is *overtrained*. This can occur when the training set does not have sufficient statistics; in these cases, the classifier learns from particular features of the events in the set, instead of the set of events as a whole. Overtraining can produce a classifier which may exhibit better separating power than it actually has, and when tested against the test set, may have a poor performance.

### 6.2.1 Binary Trees, Regression Trees, and Boosting

The BDT model is based on two major improvements to a *Binary Tree*: these are the *Regression Tree*, and *Boosting*.

A Binary tree is a simple classification model where each node makes a binary decision based on the given cut. The final decision is made based on the collection of each node's decisions, and sorts the data into either signal (reached the target value), or background (did not reach the target value).

Binary trees are limited in their ability to handle ambiguous cases; since each decision is binary, there is no ability to determine the degree of certainty in any given decision.

#### Regression Trees

Regression trees help address this shortcoming by allowing each decision to be continuous in a certain range (say  $[0, 1]$ ). This allows the final decision to be made as an aggregate of the node decisions. This allows for better ambiguity handling since the data can fall anywhere in the range. A simple cut can be used then to determine whether the data is signal-like or not, or several cuts can be implemented on the range to give several cases.

#### Boosting

Boosting is an optimization method which re-trains the decision tree several times. For each training a cost function is calculated; this function determines how successfully the tree classified the dataset. The weights are then adjusted in a manner which attempts to decrease the cost, and then the classifier is retrained. Note that, while an individual cost is calculated for each piece of data in the set, the weight adjustment is done based on the set as a whole, so that no one data point can have too much influence over the outcome of the classifier. Weight adjustment can proceed via



one of several methods including *gradient descent* or *adaptive boosting* (AdaBoost) [61], with the latter being used for the current implementation.

The general algorithm for a given training ( $\mu$ ) using AdaBoost is as follows:

- Determine the set of weights at the beginning of the training  $W^\mu = \{w_\nu^\mu\}$ , where  $\nu$  runs over every data point in the set  $X = \{x_\nu\}$ .

- Define the weight factor:

$$Q^\mu = \frac{W^\mu}{\sum_\nu w_\nu^\mu} \quad (6.18)$$

- Calculate the classifier's response  $R_\mu(x_\nu)$  as the output on the spectrum of allowed values.
- Define the cost function,  $C_\mu$ , for the training. This compares the response to the corresponding target value,  $t_\nu$ , for each data point:

$$C_\mu(X) = \sum_\nu Q_\nu^\mu |R_\mu(x_\nu) - t_\nu| \quad (6.19)$$

- Define the *boost weight*,  $\beta_\mu$ , this is used to adjust the set of weights in order to improve the response (reduce the cost):

$$\beta_\mu = \frac{C_\mu}{1 - C_\mu} \quad (6.20)$$

- Weights are then adjusted according to:

$$w_\nu^{\mu+1} = w_\nu^\mu \beta_\mu^{1-|R_\mu(x_\nu)-t_\nu|} \quad (6.21)$$

The algorithm will run until  $\mu$  goes to some defined  $\mu_{max}$ , at which point the final response,  $R_{\mu_{max}}(X)$  is determined according to:

$$R_{\mu_{max}}(X) = \begin{cases} 1, & \text{if } \sum_\mu R_\mu(X) \log\left(\frac{1}{\beta_\mu}\right) \geq \frac{1}{2} \sum_\mu \log\left(\frac{1}{\beta_\mu}\right) \\ 0, & \text{else} \end{cases} \quad (6.22)$$

This final binary decision is then renormalized according to a weighted average of  $R_\mu(X)$ , to give the *output response*,  $\mathcal{R}(X)$ , on the allowed range of continuous values. Overtraining is determined

by comparing the  $\mathcal{R}(Test)$  and  $\mathcal{R}(Train)$  distributions using a Kolmogorov-Smirnov (KS) test. The KS test quantifies the distance between the two distributions to determine if they are consistent.

### 6.2.2 Emerging Jets BDT Structure

The EJs search has implemented a BDT component which is designed to be complimentary to the model independent method. The BDT uses the ROOT TMVA [60] implementation of machine learning tools, the specifics of which can be found in table 6.1. It was introduced into the analysis as a means of suppressing the QCD multijet background, and to increase the sensitivity of the analysis to emerging-jet-like signatures.

The major benefit of the BDT is its separating power, especially for the signal points which tend to behave more SM-like, and are therefore difficult to distinguish from the QCD background. However, A new BDT is trained for each of the 15 canonical signal points, each one producing a unique set of weights. Therefore, each time the background is tested against a different signal point, it produces a different response distribution for the background. Thus, there will be a set of 15 different background distributions (in BDT response) to be used for background estimation.

The BDT is also more susceptible to the modelling of the input variables, than a cut-based selection. Therefore, a careful evaluation of the systematic uncertainties must be conducted (see section 7.1).

The general structure of the EJs BDT is divided into two parts: the first uses jet-level information to attempt to distinguish emerging-like jets from standard QCD jets, the second uses event-level information to try and classify emerging-jet-like event signatures. The event-level response distribution for each signal point is then used as an axis of an ABCD plane, in order to perform a model dependent, data-driven background estimate.

The training and testing datasets are divided up in a way that aims to reduce the effects of overtraining, while maximizing the available statistics after passing through both classifiers. Because the jet-level BDT response is used as an input to the event-level classifier, the training sets of the jet- and event-level BDTs must be statistically independent. Therefore, the datasets are divided in the following way: first, the set of available events (after preselection) is broken into the jet-level training and testing set; Next, from jet-level test events, the training and testing sets for the event-level BDT are selected. For the background, roughly one quarter of the available events

from each background slice are used for training, leaving  $\frac{3}{4}$  of the events for testing. For the signal cases, the number jet-level training events were selected such that there would be at least 20,000 events left for testing, 10,000 of which are used in the event-level BDT training; this means that the jet-level training set corresponds to roughly 20% of the available signal events.

| Parameter                                                          | EJs BDT Implementation |
|--------------------------------------------------------------------|------------------------|
| Forest Size ( $\mu_{max}$ )                                        | 800                    |
| Tree Size (Max. Depth)                                             | 3                      |
| Boosting Algorithm                                                 | AdaBoost               |
| Response Range: [b,s]                                              | $[-1, 1]$              |
| Jet-Level Input Variables ( $N_{input}^{jet}$ )                    | 6                      |
| Jet-Level Signal Training Events ( $N_{train}^{jet,Sig}$ )         | 5000 per Signal Model  |
| Jet-Level Background Training Events ( $N_{train}^{jet,Bkg}$ )     | 1215000                |
| Event-Level Input Variables ( $N_{input}^{event}$ )                | 9                      |
| Event-Level Signal Training Events ( $N_{train}^{event,Sig}$ )     | 10000 per Signal Model |
| Event-Level Background Training Events ( $N_{train}^{event,Bkg}$ ) | 50000                  |

**Table 6.1:** Parameters used in the EJs implementation of the BDT. The first subsection relates to the ROOT::TMVA default parameters. The second and third subsections refer to the specific parameters used by the jet- and event-level BDTs respectively.

It should be noted that the choice of the BDT as the machine learning algorithm was somewhat arbitrary; since the ultimate goal of this analysis is to test the emerging jets models, and not simply an exploration of machine learning applications in physics, at some point a decision was made to optimize a method which had been shown to be effective, rather than determine which was the most effective method. An investigation into a Neural Network was conducted, however, this was eventually rejected since it could not provide the same level of separation between signal and background as the BDT, and there was concern that the effects of the systematic uncertainties would be exaggerated.

### 6.2.3 Jet Level BDT

The jet-level BDT uses the 4 leading jets in each event which pass the pre-selection cuts. The following variables are used:

- Jet Mass; calculated using the energy ( $E_i$ ) and momentum ( $\vec{p}_i$ ) deposited in the calorimeter cells ( $i$ ):

$$M_{Jet} = \sqrt{\left(\sum_i E_i\right)^2 - \left(\sum_i \vec{p}_i\right)^2} \quad (6.23)$$

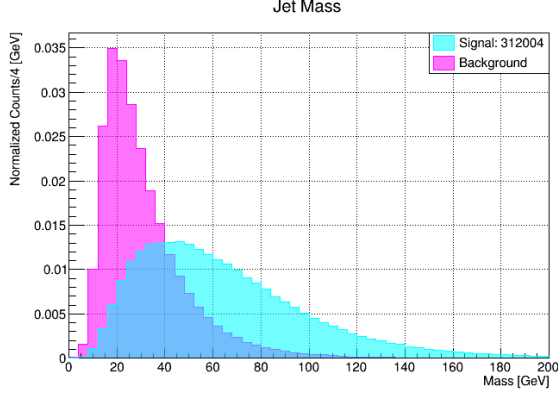
- Jet Energy; measured directly from the calorimeter deposits
- Jet Pseudorapidity; measured at the central axis of the jet
- Jet Width; energy-weighted angular spread of the constituents ( $i$ ) which characterizes the size of the jet:

$$W_{Jet} = \frac{1}{p_{T,jet}} \sum_i p_{T,i} \Delta R_i \quad (6.24)$$

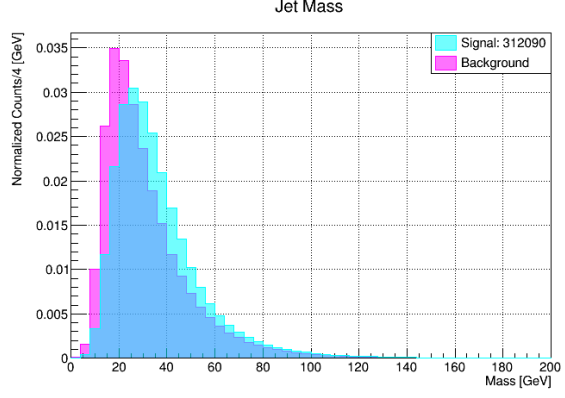
- Jet Centroid-R; transverse co-ordinate of the energy-weighted barycentre of the jet
- Jet EM Frac.; fraction of a jet's energy deposited in the electromagnetic calorimeters

The distributions for these variables are given in figures 6.8-6.13. Two characteristic signal models are shown, 312004 (Model A,  $m_{X_d} = 1400$  GeV,  $c\tau_{\pi_d} = 20mm$ ) which tends to be the most well-separated case in the canonical set, and 312090 (Model E,  $m_{X_d} = 600$  GeV,  $c\tau_{\pi_d} = 0.5mm$ ) which tends to behave the most SM-like in the set, and is therefore the most difficult to separate from the QCD Background. In each case, the corresponding background distribution is shown for comparison. The multijet background is built from the JZ slices 3-12, where JZ2withSW is omitted due to low statistics in the search region.

What the distributions in figures 6.8-6.13 show is that none of the jet-level discriminators are particularly powerful, especially when considered across the full canonical set. This is where the benefit of the BDT becomes apparent; it is able to produce a much better classifier from several poorly separating distributions.

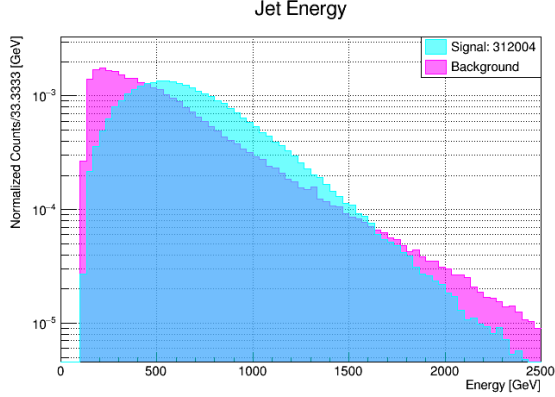


(a) Jet Mass - Model 312004

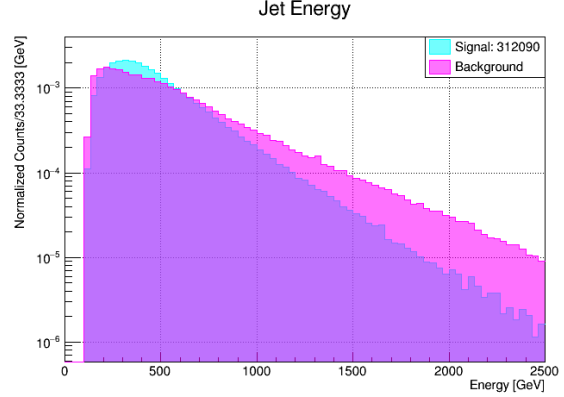


(b) Jet Mass - Model 312090

**Figure 6.8:** Jet-level Mass distributions for models 312004 (Model A,  $m_{X_d} = 1400$  GeV,  $c\tau_{\pi_d} = 20mm$ ) on the left and 312090 (Model E,  $m_{X_d} = 600$  GeV,  $c\tau_{\pi_d} = 0.5mm$ ) on the right. The QCD multijet background has also been plotted for comparison in each case. Both distributions have been normalized to unit area for the sake of comparison.

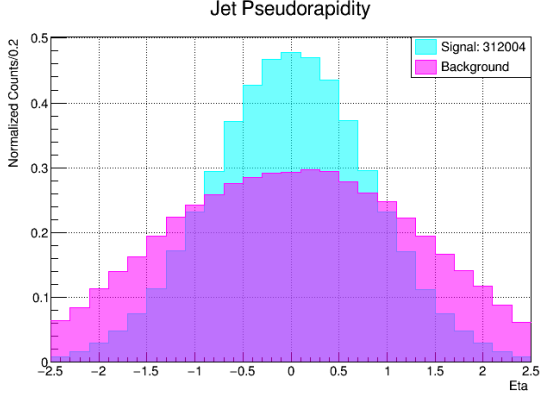


(a) Jet Energy - Model 312004

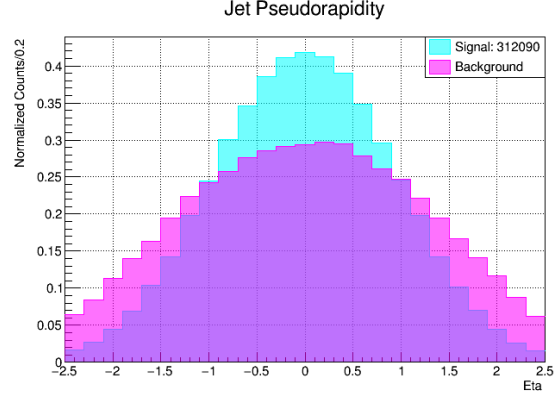


(b) Jet Energy - Model 312090

**Figure 6.9:** Jet-level Energy distributions for models 312004 (Model A,  $m_{X_d} = 1400$  GeV,  $c\tau_{\pi_d} = 20mm$ ) on the left and 312090 (Model E,  $m_{X_d} = 600$  GeV,  $c\tau_{\pi_d} = 0.5mm$ ) on the right. The QCD multijet background has also been plotted for comparison in each case. Both distributions have been normalized to unit area for the sake of comparison.

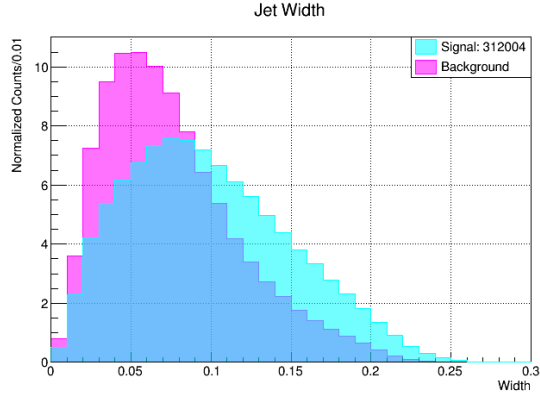


(a) Jet Pseudorapidity - Model 312004

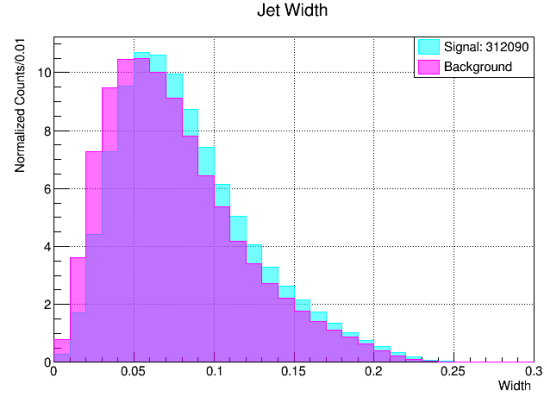


(b) Jet Pseudorapidity - Model 312090

**Figure 6.10:** Jet-level  $\eta$  distributions for models 312004 (Model A,  $m_{X_d} = 1400$  GeV,  $c\tau_{\pi_d} = 20mm$ ) on the left and 312090 (Model E,  $m_{X_d} = 600$  GeV,  $c\tau_{\pi_d} = 0.5mm$ ) on the right. The QCD multijet background has also been plotted for comparison in each case. Both distributions have been normalized to unit area for the sake of comparison.

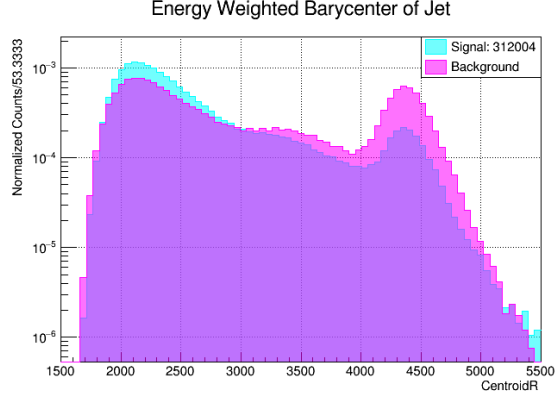


(a) Jet Width - Model 312004

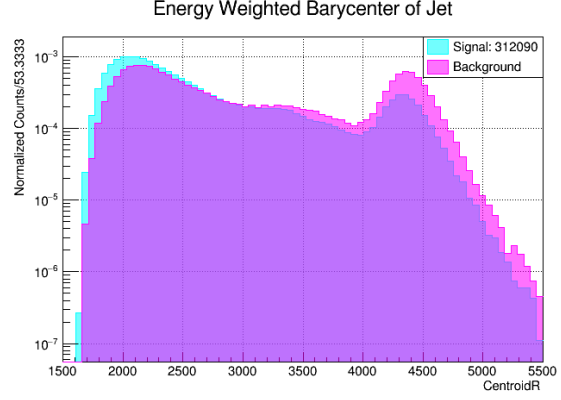


(b) Jet Width - Model 312090

**Figure 6.11:** Jet-level Width distributions for models 312004 (Model A,  $m_{X_d} = 1400$  GeV,  $c\tau_{\pi_d} = 20mm$ ) on the left and 312090 (Model E,  $m_{X_d} = 600$  GeV,  $c\tau_{\pi_d} = 0.5mm$ ) on the right. The QCD multijet background has also been plotted for comparison in each case. Both distributions have been normalized to unit area for the sake of comparison.

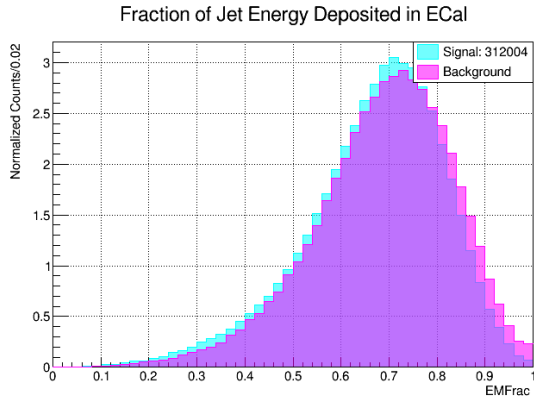


(a) Jet Centroid-R - Model 312004

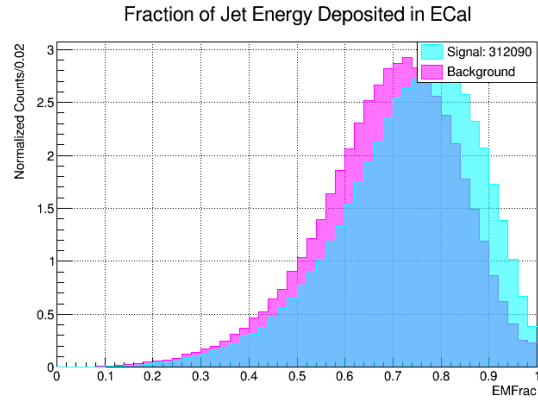


(b) Jet Centroid-R - Model 312090

**Figure 6.12:** Jet-level Centroid-R distributions for models 312004 (Model A,  $m_{X_d} = 1400$  GeV,  $c\tau_{\pi_d} = 20mm$ ) on the left and 312090 (Model E,  $m_{X_d} = 600$  GeV,  $c\tau_{\pi_d} = 0.5mm$ ) on the right. The QCD multijet background has also been plotted for comparison in each case. Both distributions have been normalized to unit area for the sake of comparison.



(a) Jet EM Frac. - Model 312004



(b) Jet EM Frac. - Model 312090

**Figure 6.13:** Jet-level EM Fraction distributions for models 312004 (Model A,  $m_{X_d} = 1400$  GeV,  $c\tau_{\pi_d} = 20mm$ ) on the left and 312090 (Model E,  $m_{X_d} = 600$  GeV,  $c\tau_{\pi_d} = 0.5mm$ ) on the right. The QCD multijet background has also been plotted for comparison in each case. Both distributions have been normalized to unit area for the sake of comparison.

The BDT will also rank each input variable based on separation power and importance. This ranking can, in general, be different for each signal point the BDT is trained against. The average rank of each jet-level variable is given in table 6.2.

| Input Variable | Average Rank |
|----------------|--------------|
| Jet Energy     | 1            |
| Jet Mass       | =2           |
| Jet Width      | =2           |
| Jet Centroid-R | 4            |
| Jet EMFrac.    | 5            |
| Jet $\eta$     | 6            |

**Table 6.2:** Rank of the jet-level input variables, averaged across the canonical set of signal points.

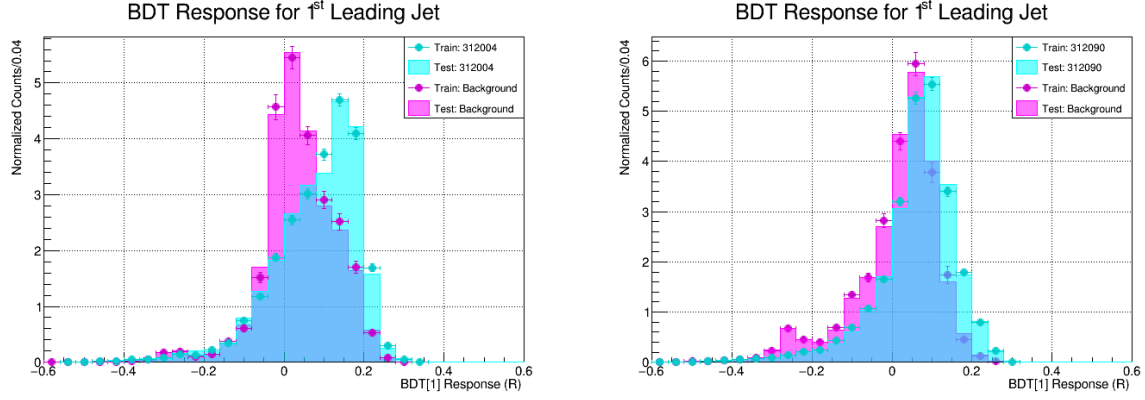
#### 6.2.4 Event Level BDT

The BDT response from the four leading jets is then used as an input variable to the event-level BDT. These distributions are shown in figures 6.14-6.17, for the two characteristic signal models, 312004 (left) and 312090 (right). Here, both the  $\mathcal{R}_{jet}(Train)$  and  $\mathcal{R}_{jet}(Test)$  distributions are plotted to check for overtraining. In general there is good agreement between the training and testing sets for the jet-level BDT. For many of the signal points, the  $\mathcal{R}_{jet}$  is not sufficient for effectively separating the signal and background. Therefore, an additional level of classification is required to try and improve the signal efficiency.

In addition to improved separating power, the next BDT has the benefit of being an event-level classifier; sorting events into emerging-like or SM-like. Since the EJs model is more of an event signature than a specific dark matter model, it is useful to be able to classify events across a large range of model parameters.

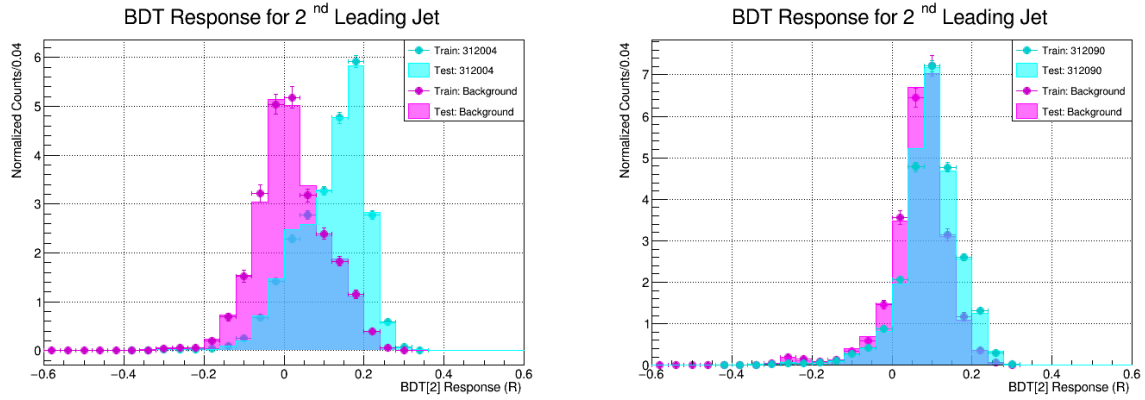
The event-level BDT also uses several event-level discriminators to try and improve the separation between signal and background; the  $H_T$  of the 4 leading jets, and their associated displaced vertex multiplicity, as well as the jet multiplicity (the number of reconstructed jets in the event), and two *event shape* variables built from the isotropy measures.





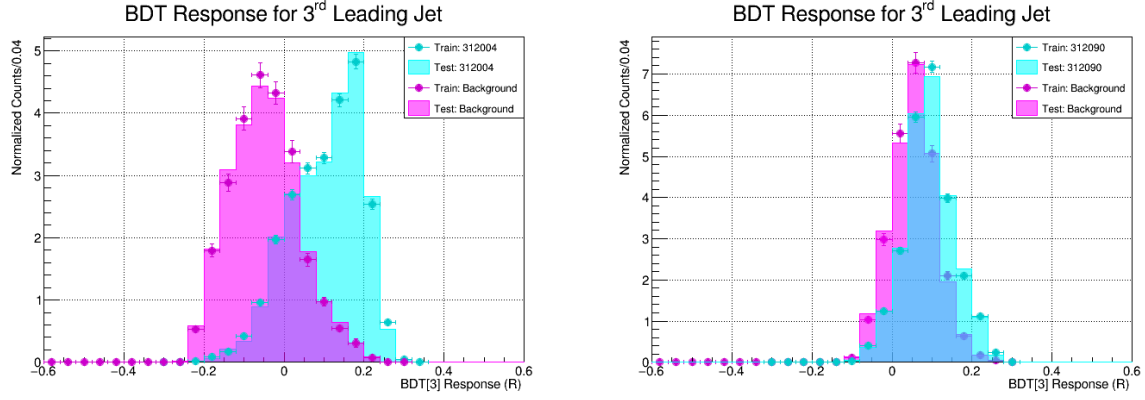
(a) BDT Response: 1<sup>st</sup> leading jet - Model 312004 (b) BDT Response: 1<sup>st</sup> leading jet - Model 312090

**Figure 6.14:** Jet-level BDT response distributions for the 1<sup>st</sup> leading jet for models 312004 (Model A,  $m_{X_d} = 1400$  GeV,  $c\tau_{\pi_d} = 20mm$ ) on the left and 312090 (Model E,  $m_{X_d} = 600$  GeV,  $c\tau_{\pi_d} = 0.5mm$ ) on the right. The QCD multijet background has also been plotted for comparison in each case. Both distributions have been normalized to unit area for the sake of comparison.



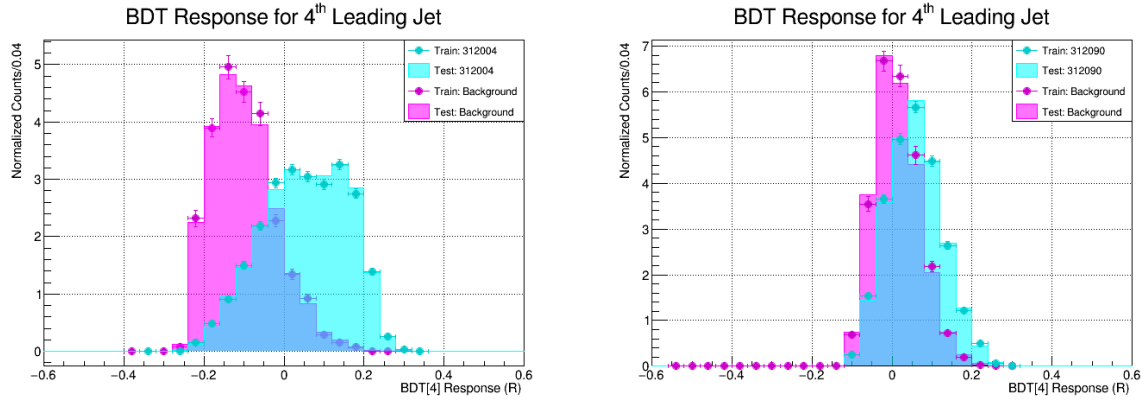
(a) BDT Response: 2<sup>nd</sup> leading jet - Model 312004 (b) BDT Response: 2<sup>nd</sup> leading jet - Model 312090

**Figure 6.15:** Jet-level BDT response distributions for the 2<sup>nd</sup> leading jet for models 312004 (Model A,  $m_{X_d} = 1400$  GeV,  $c\tau_{\pi_d} = 20mm$ ) on the left and 312090 (Model E,  $m_{X_d} = 600$  GeV,  $c\tau_{\pi_d} = 0.5mm$ ) on the right. The QCD multijet background has also been plotted for comparison in each case. Both distributions have been normalized to unit area for the sake of comparison.



(a) BDT Response: 3<sup>rd</sup> leading jet - Model 312004 (b) BDT Response: 3<sup>rd</sup> leading jet - Model 312090

**Figure 6.16:** Jet-level BDT response distributions for the 3<sup>rd</sup> leading jet for models 312004 (Model A,  $m_{X_d} = 1400$  GeV,  $c\tau_{\pi_d} = 20mm$ ) on the left and 312090 (Model E,  $m_{X_d} = 600$  GeV,  $c\tau_{\pi_d} = 0.5mm$ ) on the right. The QCD multijet background has also been plotted for comparison in each case. Both distributions have been normalized to unit area for the sake of comparison.



(a) BDT Response: 4<sup>th</sup> leading jet - Model 312004 (b) BDT Response: 4<sup>th</sup> leading jet - Model 312090

**Figure 6.17:** Jet-level BDT response distributions for the 4<sup>th</sup> leading jet for models 312004 (Model A,  $m_{X_d} = 1400$  GeV,  $c\tau_{\pi_d} = 20mm$ ) on the left and 312090 (Model E,  $m_{X_d} = 600$  GeV,  $c\tau_{\pi_d} = 0.5mm$ ) on the right. The QCD multijet background has also been plotted for comparison in each case. Both distributions have been normalized to unit area for the sake of comparison.

## Event Shape Variables

The event-level BDT employs two event shape variables [62] which try to characterize the energy distribution in the event. These variables are built from the eigenvalues of the jet momentum tensor:

$$\mathcal{M}_{ij} = \frac{\sum_k (p_i p^j)_k}{\sum_k (p_i p^i)_k} \quad (6.25)$$

where the explicit sum over  $k$  runs over all of the jets in the event, and the implicit sums over  $i, j$  run over the components of the jet three-momentum. Defined this way, the tensor has the corresponding normalized eigenvalues  $\lambda_1, \lambda_2, \lambda_3$  such that:  $\lambda_1 > \lambda_2 > \lambda_3$  and  $\lambda_1 + \lambda_2 + \lambda_3 = 1$ . These are then used to create the three event shape variables: *Sphericity* ( $S$ ), *Transverse Sphericity* ( $S_T$ ), and *Aplanarity* ( $\mathcal{A}$ ):

$$S = \frac{3}{2}(\lambda_2 + \lambda_3) \quad (6.26)$$

$$S_T = \frac{2\lambda_2}{\lambda_1 + \lambda_2} \quad (6.27)$$

$$\mathcal{A} = \frac{3}{2}\lambda_3 \quad (6.28)$$

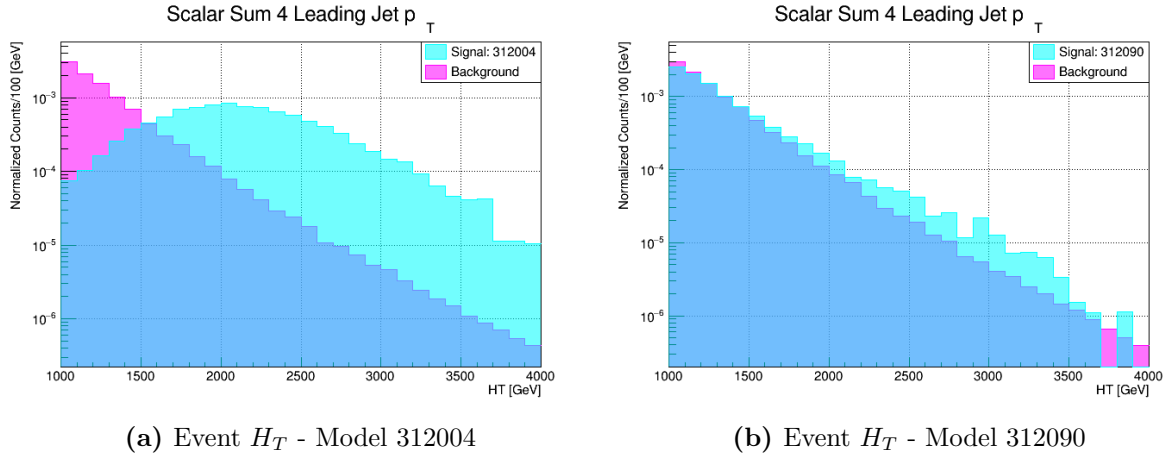
The event-level BDT uses the transverse sphericity and aplanarity as input variables.

The transverse sphericity (and sphericity) starts by defining an axis through the leading jet, known as the sphericity axis.  $S_T$  is then a measure of how much of the event's  $p_T$  is in the plane perpendicular to this axis, as contributed by jets. The transverse sphericity takes values in the range  $0 \leq S_T < 1$ , where an event with a value closer to 0 has most of its  $p_T$  along the sphericity axis, and an event with a value closer to 1 has more  $p_T$  in the plane perpendicular to it. For a background multijet event, where the majority of the  $p_T$  is contained within the 2 leading jets (which should be roughly back-to-back), it is expected to have a value of  $S_T \approx 0$ . Conversely, signal events tend to have a higher jet multiplicities, with jets arranged more isotropically, therefore are expected to have values of  $S_T$  closer to 1.

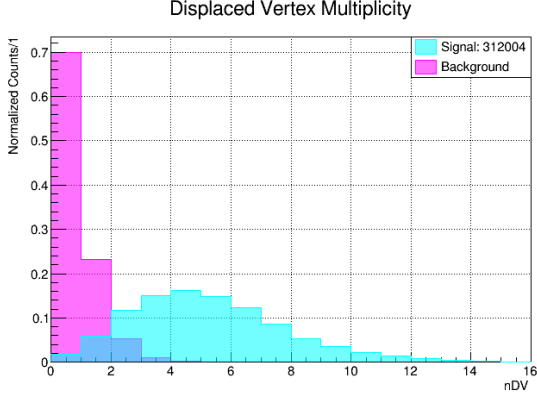
Aplanarity defines a plane using the 2 leading jets, and measures the amount of  $p_T$  in the event that is outside this plane, again as contributed by jets.  $\mathcal{A}$  has values in the range from 0 to  $\frac{1}{2}$ , although due to it being made only from the smallest eigenvalue, it tends to fall more on the range  $0 \leq \mathcal{A} < 0.3$ . Higher values of  $\mathcal{A}$  represent events which have a higher fraction of  $p_T$  outside of the

plane, whereas values closer to 0 represent more planar events. For background events, where the 2 leading jets are likely collinear, the sphericity axis should be along the plane defined between the 2 leading jets, so again the expected value is  $\mathcal{A} \approx 0$ . Similarly, since the signal events tend to have more jets, arranged isotropically, the expected value of  $\mathcal{A}$  is higher.

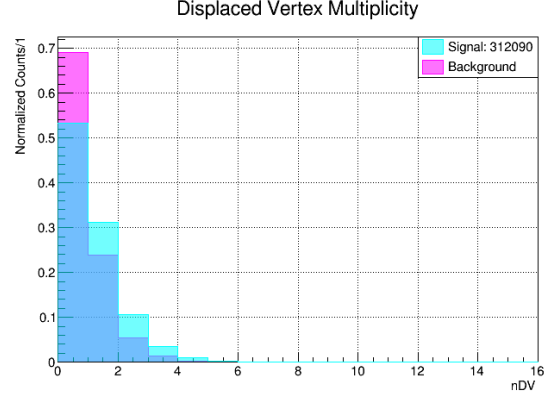
The event-level input variable distributions are shown for the characteristic signal models (312004 and 312090) in figures 6.18-6.22. In the best cases, the event-level input variables provide good separation, but for many others, the individual variables are difficult to distinguish from the QCD background. However, when combined with the jet-level BDT response into the event-level BDT, the result is a classifier with much better separating power. The rank of the event-level input variables is given in table 6.3.



**Figure 6.18:** Event-level  $H_T$  distributions for models 312004 (Model A,  $m_{X_d} = 1400$  GeV,  $c\tau_{\pi_d} = 20mm$ ) on the left and 312090 (Model E,  $m_{X_d} = 600$  GeV,  $c\tau_{\pi_d} = 0.5mm$ ) on the right. The QCD multijet background has also been plotted for comparison in each case. Both distributions have been normalized to unit area for the sake of comparison.

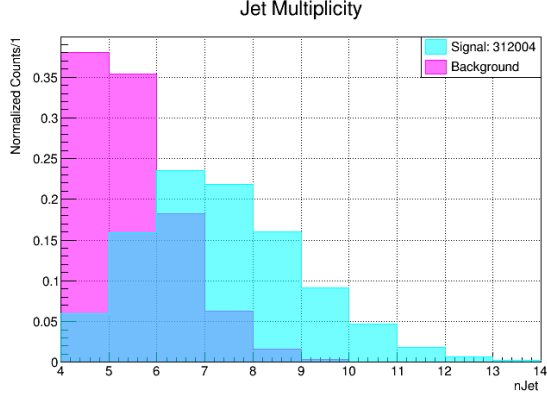


(a) Event nDV - Model 312004

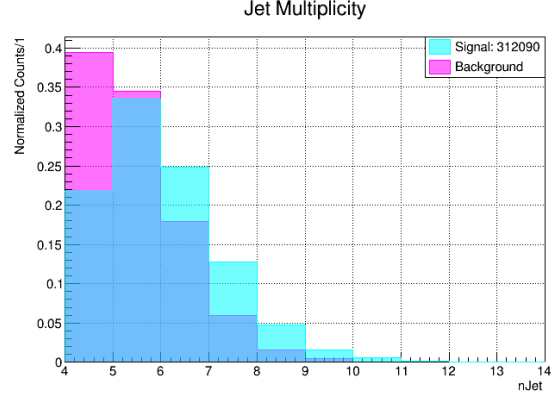


(b) Event nDV - Model 312090

**Figure 6.19:** Event-level nDV distributions for models 312004 (Model A,  $m_{X_d} = 1400$  GeV,  $c\tau_{\pi_d} = 20mm$ ) on the left and 312090 (Model E,  $m_{X_d} = 600$  GeV,  $c\tau_{\pi_d} = 0.5mm$ ) on the right. The QCD multijet background has also been plotted for comparison in each case. Both distributions have been normalized to unit area for the sake of comparison.

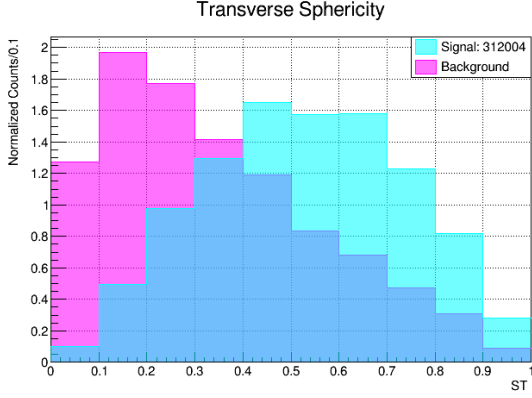


(a) Event nJet - Model 312004

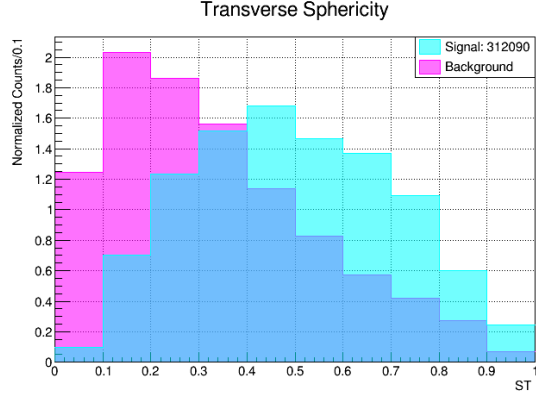


(b) Event nJet - Model 312090

**Figure 6.20:** Event-level nJet distributions for models 312004 (Model A,  $m_{X_d} = 1400$  GeV,  $c\tau_{\pi_d} = 20mm$ ) on the left and 312090 (Model E,  $m_{X_d} = 600$  GeV,  $c\tau_{\pi_d} = 0.5mm$ ) on the right. The QCD multijet background has also been plotted for comparison in each case. Both distributions have been normalized to unit area for the sake of comparison.

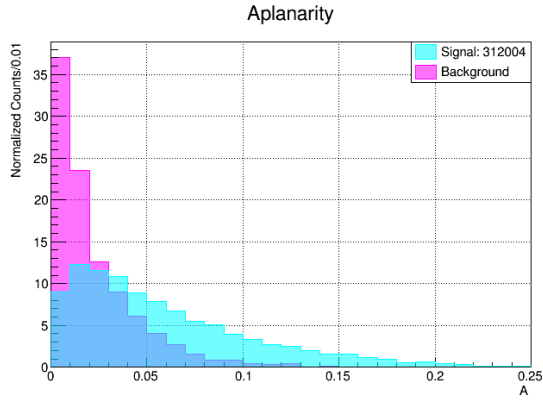


(a) Event  $S_T$  - Model 312004

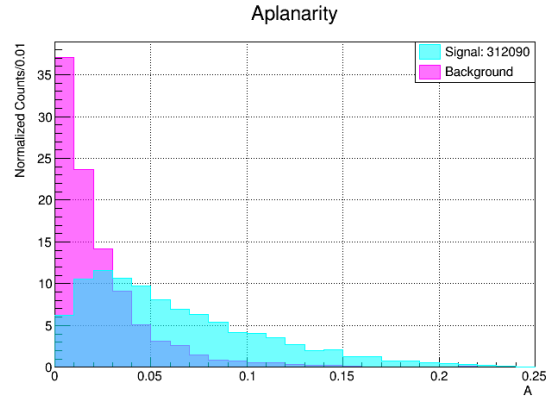


(b) Event  $S_T$  - Model 312090

**Figure 6.21:** Event-level  $S_T$  distributions for models 312004 (Model A,  $m_{X_d} = 1400$  GeV,  $c\tau_{\pi_d} = 20mm$ ) on the left and 312090 (Model E,  $m_{X_d} = 600$  GeV,  $c\tau_{\pi_d} = 0.5mm$ ) on the right. The QCD multijet background has also been plotted for comparison in each case. Both distributions have been normalized to unit area for the sake of comparison.



(a) Event  $\mathcal{A}$  - Model 312004



(b) Event  $\mathcal{A}$  - Model 312090

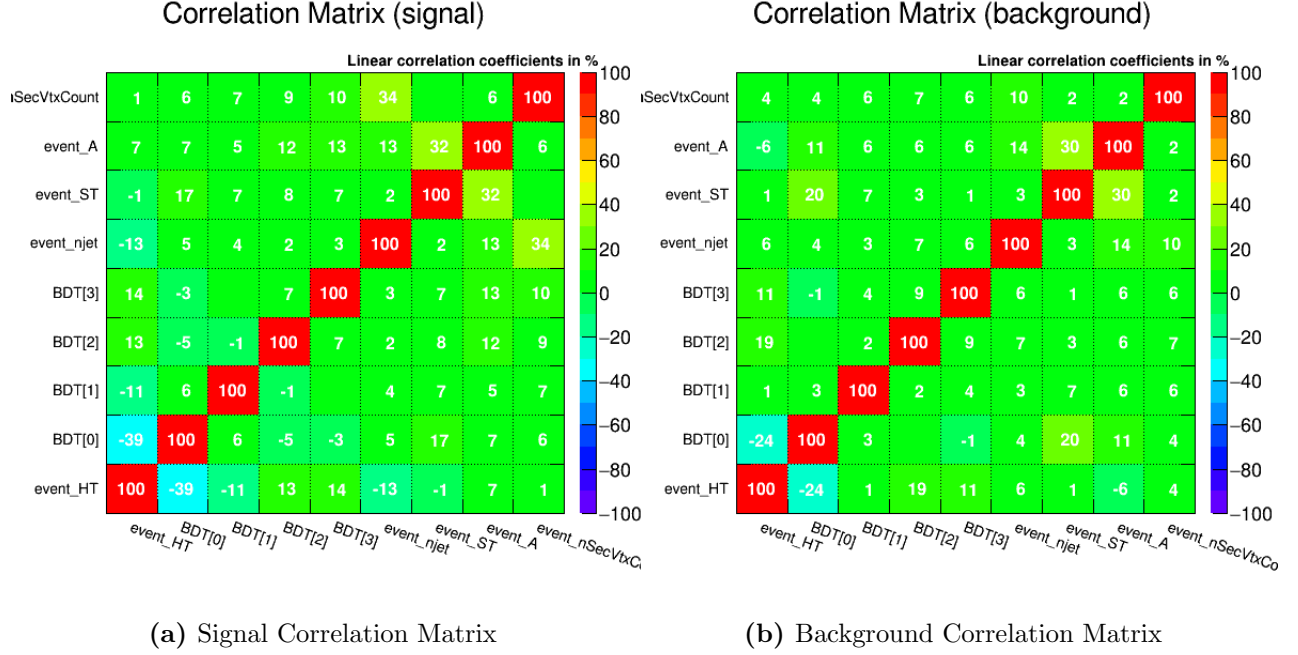
**Figure 6.22:** Event-level  $\mathcal{A}$  distributions for models 312004 (Model A,  $m_{X_d} = 1400$  GeV,  $c\tau_{\pi_d} = 20mm$ ) on the left and 312090 (Model E,  $m_{X_d} = 600$  GeV,  $c\tau_{\pi_d} = 0.5mm$ ) on the right. The QCD multijet background has also been plotted for comparison in each case. Both distributions have been normalized to unit area for the sake of comparison.

In addition to ranking the variables, the TMVA implementation also generates correlations matrices for the input variables. Correlations between variables indicates that there is some redundant information in the BDT. For large correlations this means that one of the variables could likely be removed without affecting separating power of the classifier. Removing redundant variables can benefit the BDT, since each variable may increase the systematic uncertainty. Thus, removing any variables which does not help increase the separating power also reduces the chance of having a variable with a high associated systematic uncertainty.

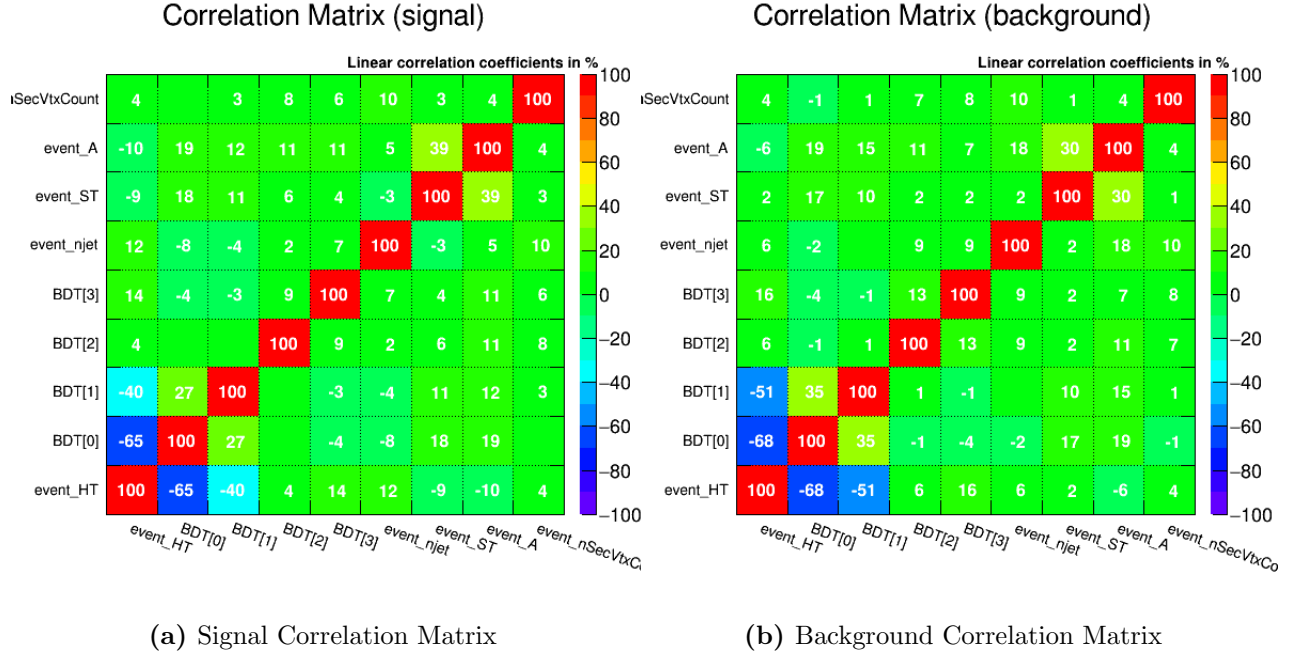
It should be noted here that a high correlation between variables does not guarantee a high systematic contribution; rather any variable can, in general, have a high associated systematic uncertainty. Therefore, removing any variable may help decrease the systematic uncertainty. The correlation matrices (both signal and background) for models 312004 and 312090 are given in figures 6.23 and 6.24.

| Input Variable        | Average Rank |
|-----------------------|--------------|
| $\mathcal{R}_{jet}^3$ | 1            |
| $\mathcal{R}_{jet}^4$ | 2            |
| $\mathcal{A}$         | 3            |
| $\mathcal{R}_{jet}^2$ | 4            |
| $S_T$                 | 5            |
| nJet                  | 6            |
| $H_T$                 | =7           |
| $\mathcal{R}_{jet}^1$ | =7           |
| nDV                   | 9            |

**Table 6.3:** Rank of the event-level input variables, averaged across the canonical set of signal points.



**Figure 6.23:** Signal and Background correlation matrices for the event-level input variables. Matrices are given for model 312004 (Model A,  $m_{X_d} = 1400$  GeV,  $c\tau_{\pi_d} = 20mm$ ) and The QCD multijet background.

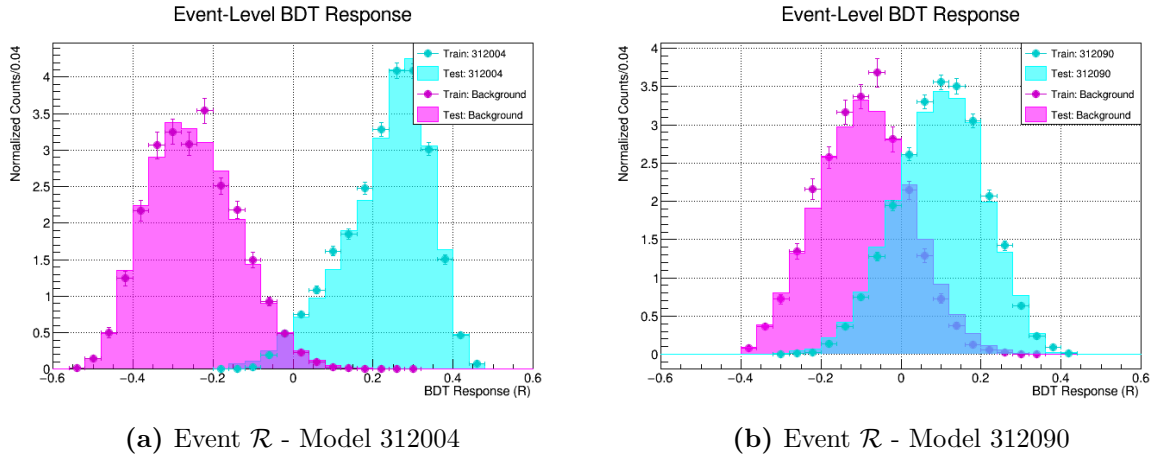


**Figure 6.24:** Signal and Background correlation matrices for the event-level input variables. Matrices are given for model 312004 (Model E,  $m_{X_d} = 600$  GeV,  $c\tau_{\pi_d} = 0.5mm$ ) and The QCD multijet background.



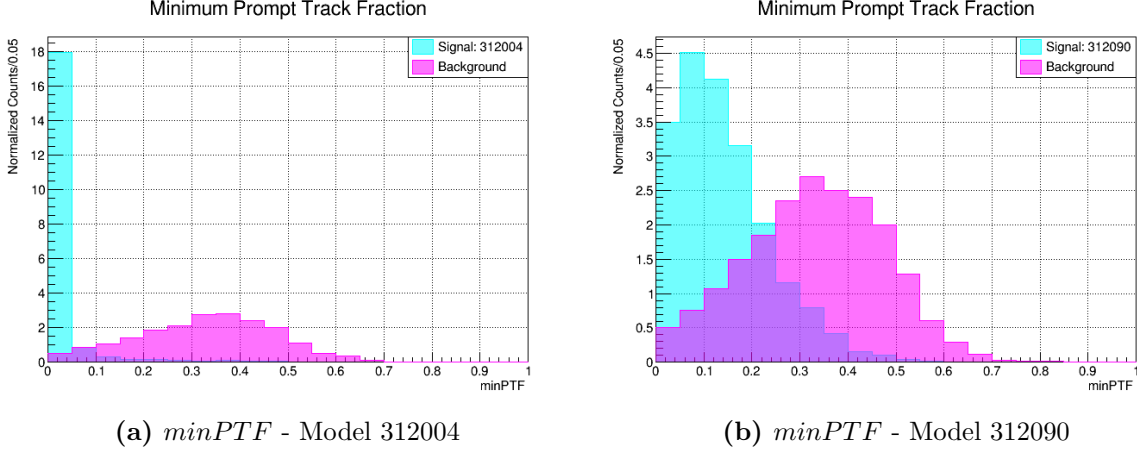
### 6.2.5 Model Dependent Search Region

The model dependent search region is then defined very similarly to the model independent version, the major difference is that it exchanges the  $H_T$  as the x-axis variable for  $\mathcal{R}$ . This distribution is shown in figure 6.25, here both the  $\mathcal{R}(Train)$  and  $\mathcal{R}(Test)$  distributions are plotted to check for overtraining. The comparison in figure 6.25b exhibits some overtraining, with the training distributions showing better separation than the corresponding test distributions. This could be caused by low MC statistics in the training set.



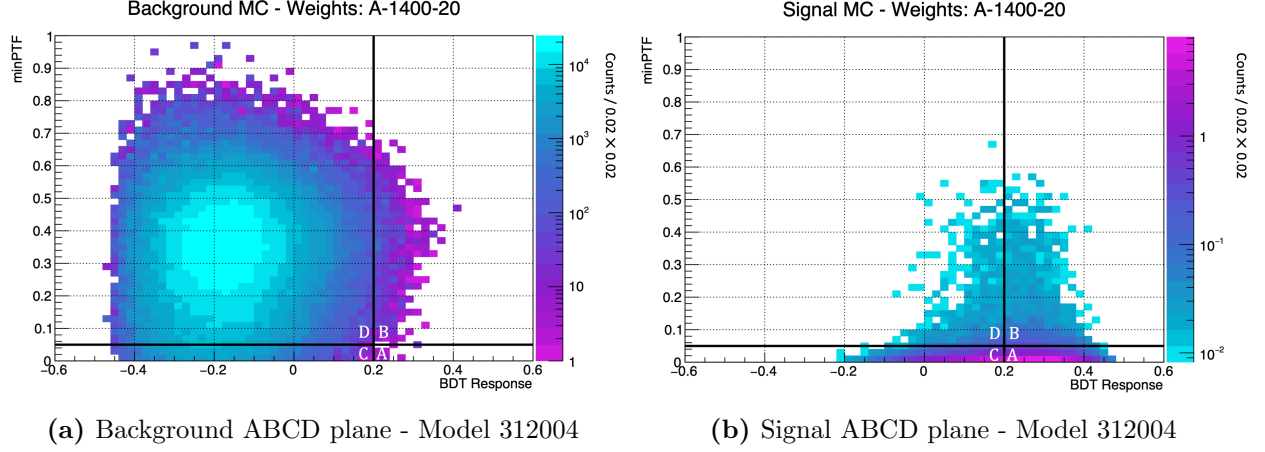
**Figure 6.25:** Event-level BDT response distributions for models 312004 (Model A,  $m_{X_d} = 1400$  GeV,  $c\tau_{\pi_d} = 20mm$ ) on the left and 312090 (Model E,  $m_{X_d} = 600$  GeV,  $c\tau_{\pi_d} = 0.5mm$ ) on the right. The QCD multijet background has also been plotted for comparison in each case. Both distributions have been normalized to unit area for the sake of comparison.

The ABCD y-axis ( $minPTF$ ) is shown in figure 6.26. The corresponding ABCD plane requires the same pre-selection cuts as the model independent plane. Similarly, the cut of  $minPTF = 0.05$  is maintained from the model independent plane. The cut on the BDT response was optimized to give the best sensitivity for the majority of the signal points, with the cut  $\mathcal{R} = 0.2$  being selected for the canonical set. This then defines the ABCD regions which also follow the alternate arrangement.

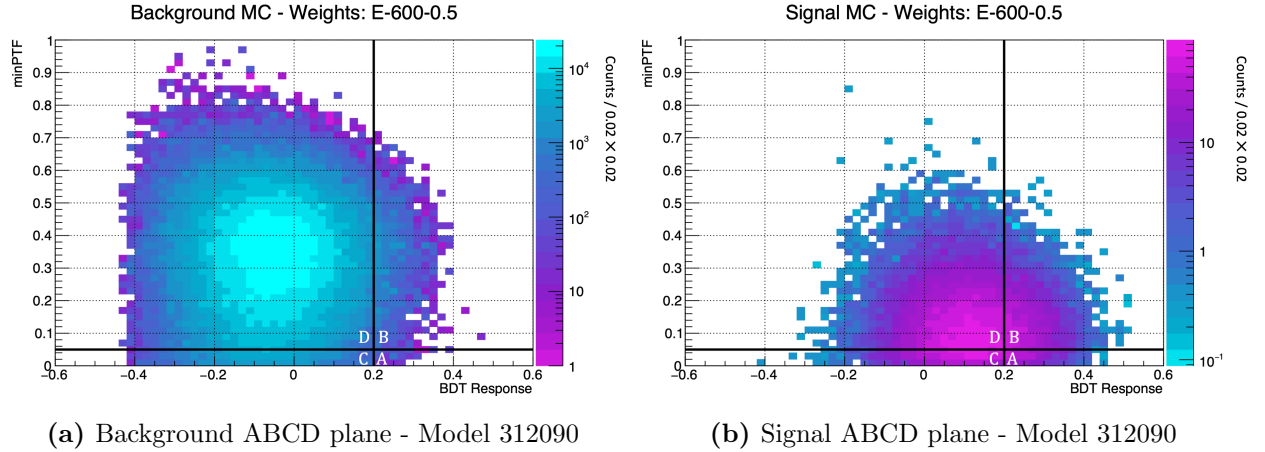


**Figure 6.26:** Event-level  $\text{minPTF}$  distributions for models 312004 (Model A,  $m_{X_d} = 1400$  GeV,  $c\tau_{\pi_d} = 20\text{mm}$ ) on the left and 312090 (Model E,  $m_{X_d} = 600$  GeV,  $c\tau_{\pi_d} = 0.5\text{mm}$ ) on the right. The QCD multijet background has also been plotted for comparison in each case. Both distributions have been normalized to unit area for the sake of comparison.

The example ABCD planes for models 312004 and 312090 (both signal and background), which are shown in figures 6.27 and 6.28, demonstrate both the major benefit and disadvantage of the BDT implementation. Since  $\mathcal{R}$  is a much better classifier than  $H_T$  alone, the signal models (especially those that behave very SM-like) are more easily separated from the QCD background. However, since the training will produce a different background  $\mathcal{R}$  distribution for each signal point, there will have to be 15 different data-driven estimates using the ABCD method.



**Figure 6.27:** Signal and background MC distributions in the search region defined by the model dependent ABCD plane for model 312004 (Model A,  $m_{X_d} = 1400$  GeV,  $c\tau_{\pi_d} = 20mm$ ). The cuts of  $\mathcal{R} = 0.2$  and  $minPTF = 0.05$  (shown in black) define the four regions, which follow the alternate ABCD arrangement. The distributions have been scaled to the full Run II luminosity of  $139 \text{ fb}^{-1}$ .



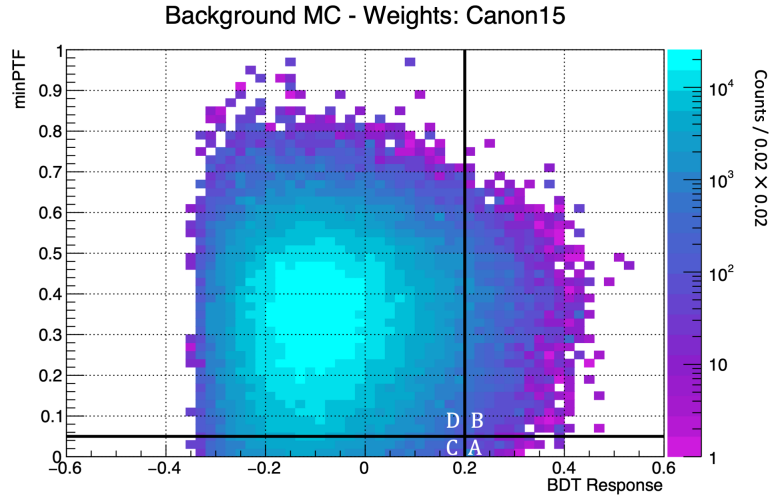
**Figure 6.28:** Signal and background MC distributions in the search region defined by the model dependent ABCD plane for model 312090 (Model E,  $m_{X_d} = 600$  GeV,  $c\tau_{\pi_d} = 0.5mm$ ). The cuts of  $\mathcal{R} = 0.2$  and  $minPTF = 0.05$  (shown in black) define the four regions, which follow the alternate ABCD arrangement. The distributions have been scaled to the full Run II luminosity of  $139 \text{ fb}^{-1}$ .

## 6.3 Additional BDT Trainings

In an attempt to mitigate the model dependence, two other training scenarios were tested. These trainings combine the signal points in different configurations in order to reduce the number of different background distributions, and therefore the number of background estimates.

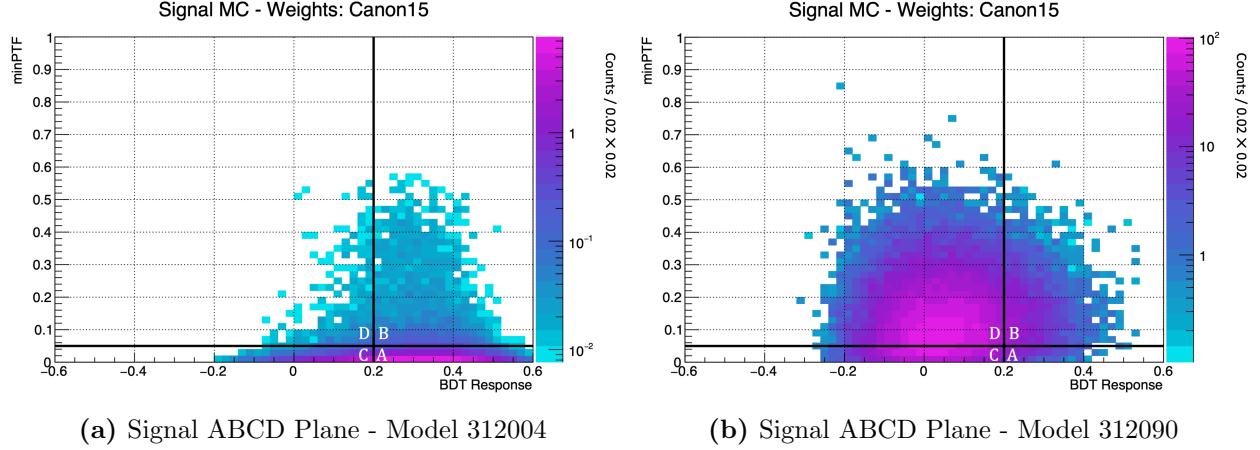
### 6.3.1 Canonical Set (Canon15) Mixed Training

The Canon15 training scenario combined all 15 of the canonical signal points in equal proportion, and then tests each signal point individually against the mixed weights. This method produces one single pseudo-model-independent background distribution, as shown in figure 6.29. Thus, requiring only one data-driven background estimate.



**Figure 6.29:** QCD Background distribution for the Canon15 mixed BDT training. The cuts of  $\mathcal{R} = 0.2$  and  $minPTF = 0.05$  (shown in black) define the four ABCD regions, which follow the alternate arrangement. The distribution has been scaled to the full Run II luminosity of  $139 \text{ fb}^{-1}$ .

The background distribution, when compared to figures 6.27a and 6.28a, is shifted to the right in  $\mathcal{R}$ , moving more events into the signal region. Similarly, the signal distributions (see figure 6.30) get smeared slightly in  $\mathcal{R}$ , in general broadening the distributions. These two effects compound and reduce the sensitivity of all of the signal points.



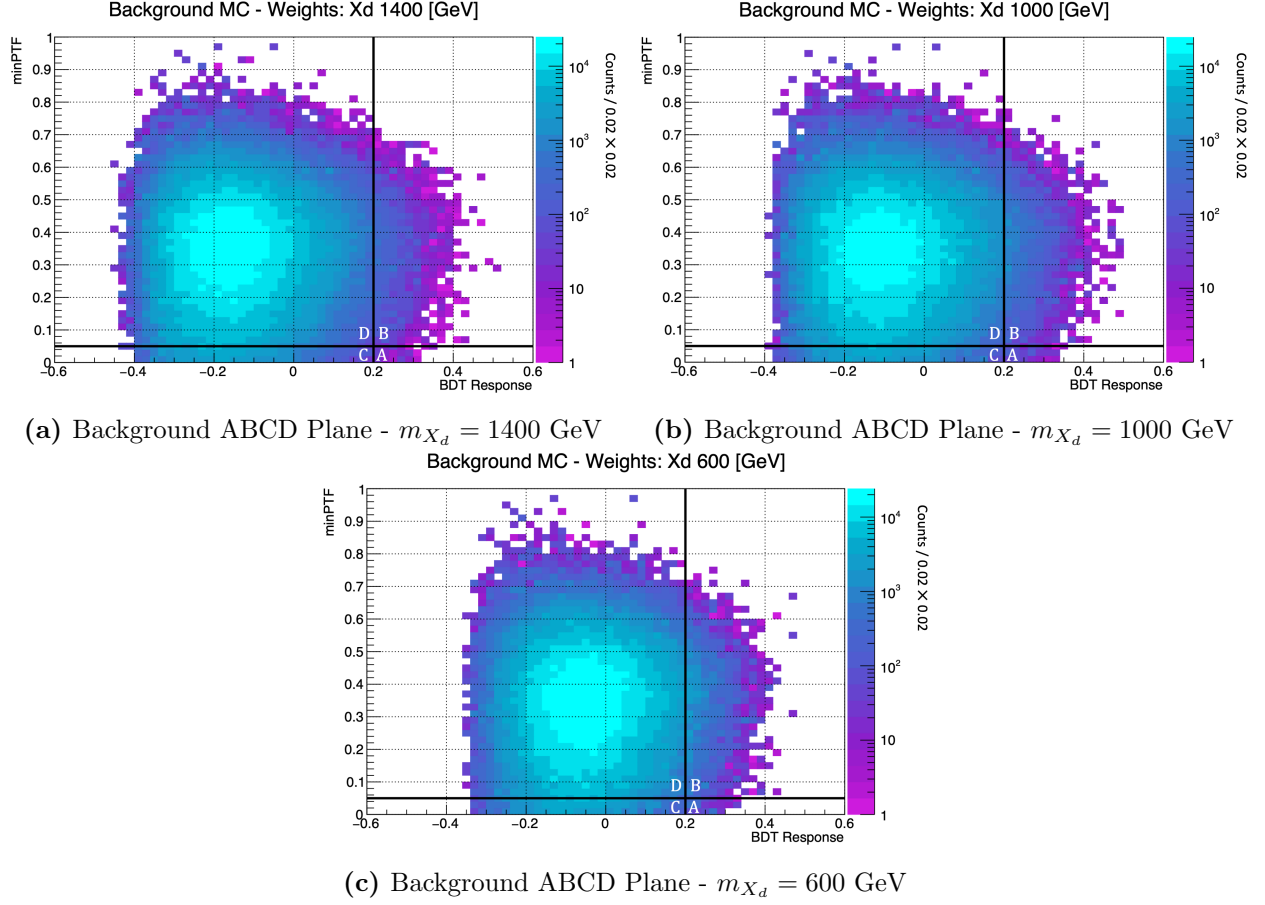
**Figure 6.30:** Signal MC distributions for models 312004 (left) and 312090 (right) tested against the Canon15 mixed BDT training. The cuts of  $\mathcal{R} = 0.2$  and  $minPTF = 0.05$  (shown in black) define the four ABCD regions, which follow the alternate arrangement. The distribution has been scaled to the full Run II luminosity of  $139 \text{ fb}^{-1}$ .

### 6.3.2 Mediator Mass Mixed Training

The second training scenario grouped signal points by their mediator mass. For this case, an additional 21 signal points were included in the training so that each mediator mass group had 12 signal points. However, only the canonical set were tested against the mixed weights so that a fair comparison could be made. This method produces three distinct background distributions (one for each mediator mass), and therefore requires three data-driven estimates.

In all cases, the background distributions have a right side edge which is higher in  $\mathcal{R}$  than any of the nominally (individually) trained background distributions. This means that there are more background events which populate the signal region. The left side edge of the distributions show which group of mediator masses has more separating power; because events from the  $m_{X_d} = 1400 \text{ GeV}$  group look the least SM-like, more background events will be pushed to the left in  $\mathcal{R}$ , giving a broader distribution. The opposite is true for the  $m_{X_d} = 600 \text{ GeV}$  group, which behaves more SM-like and, therefore, has a left side edge that is further to the right in  $\mathcal{R}$ .

Similar to the Canon15 mixed training, the signal distributions are broadened in  $\mathcal{R}$ . Again, this compounds with the effect of the background and results in a general loss in sensitivity for the signal models.



**Figure 6.31:** QCD Background distributions for the mixed mediator mass BDT training. Each distribution corresponds to a given mediator mass. The cuts of  $\mathcal{R} = 0.2$  and  $\min PTF = 0.05$  (shown in black) define the four ABCD regions, which follow the alternate arrangement. The distribution has been scaled to the full Run II luminosity of  $139 \text{ fb}^{-1}$ .

From a technical perspective, the mixed BDT training require less resources, since each training takes time. However, each signal point is still tested individually, so the difference in computational resources is not so dramatic. The other benefit of mixed training is the reduction in model dependence; the BDT will always depend on the models it is trained on, but grouping signal points together means that it does not learn from the features of any one parameter of the signal models ( $m_{\pi_d}, c\tau_{\pi_d}$ ), but rather from the collection as a whole. Thus, for a sufficiently large group (for example using the full grid of 90 signal points), the BDT approaches a more model independent classifier.

The disadvantage to mixed trainings is the loss in sensitivity. A mixed BDT will always have less separating power, for a given signal point, than a BDT trained individually on that signal

point. For a background-dominated analysis where the potential signal is small, it can be desirable to optimize the classifier to maximize signal sensitivity.

## 6.4 Validation Regions

In addition to the search regions defined by both the model independent and model dependent planes, the EJs analysis also defines two validation regions, generally identified by jet multiplicity.

The validation regions are defined in a way which selects a set of background-only events to use for Data-MC comparisons. Because the regions are background-only, the analysis is not blinded to the data in the validation regions. The comparisons are made to determine if the MC adequately models the data for the variables which define the analysis. The validation regions can be used to evaluate which distributions are well modelled in the MC, and which may require a specific systematic uncertainty associated to the modelling. The regions can also be used to determine if the systematics adequately cover the differences between the data and MC.

### 6.4.1 2,3-Jet Validation Region

The 2,3-Jet validation region is defined by an orthogonal jet multiplicity selection, requiring either 2 or 3 jets for each event. Because of this, a four-jet trigger cannot be used, so validation events are selected using the heavily prescaled single-jet trigger *HLT-j110*.

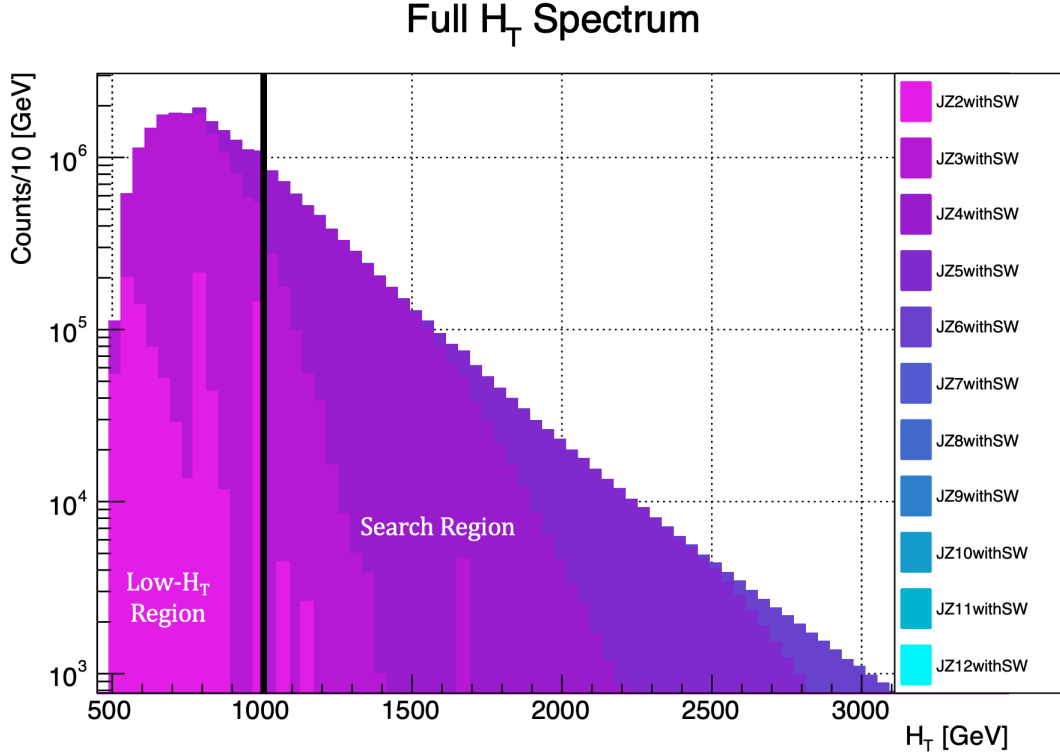
Each of the jets in the validation events are required to have the same kinematic and fiducial requirements as the pre-selection cuts:  $p_T^{Jet} \geq 120$  GeV and  $|\eta^{Jet}| < 2.5$ .

Because the trigger is prescaled, data events in the 2,3-Jet validation region must be unprescaled (weighted according to the trigger prescale) so that it corresponds to the correct luminosity.

### 6.4.2 4-Jet Validation Region

The 4-Jet validation region follows a similar selection to the pre-selection, with a reciprocal cut on  $H_T$ ; the selection is inverted, such that the region is defined with  $H_T < 1000$  GeV. In theory, the full *Low- $H_T$*  region can be used as a 4-Jet validation region. However, for  $H_T$  values below roughly 775 GeV, the agreement between Data and MC becomes quite poor. Therefore, to ensure the validation is done in a region where data and MC agree, the 4-Jet validation region is restricted

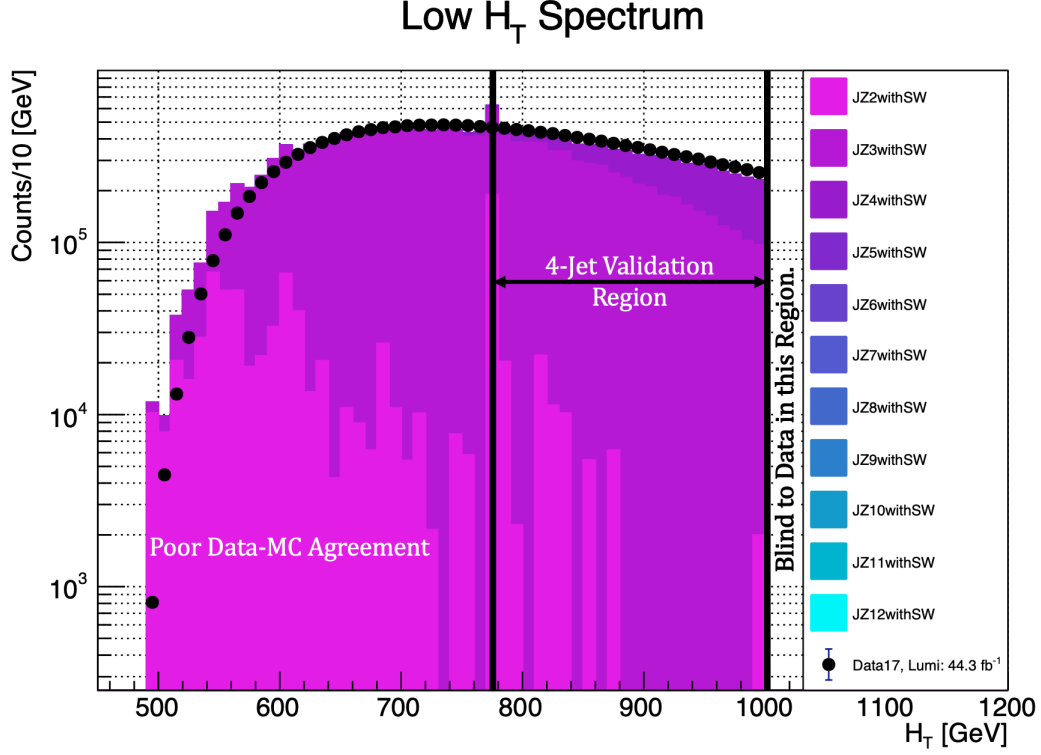
to  $775 \text{ GeV} \leq H_T < 1000 \text{ GeV}$ . Because JZ2withSW has a non-negligible contribution in the lower  $H_T$  range, it is included in the background when conducting validation studies. However, the cut of  $H_T < 1000 \text{ GeV}$  excludes any slices higher than JZ4withSW, so even though the validation background is composed of slices 2-12, the only slices which contribute are 2, 3, and 4.



**Figure 6.32:** QCD Background distribution in the full  $H_T$  spectrum defined by the *Emerging* component of the RPVLL filter. The cut at  $H_T = 1000 \text{ GeV}$  defines both the search and Low- $H_T$  regions. The distribution has been scaled to  $44.3 \text{ fb}^{-1}$ , which corresponds to the full Data17 Luminosity.

The major benefit of including the 4-Jet validation region is that the BDT response distributions can be validated there. Since the BDT is trained on the 4 leading jets, it requires at least four jets in order to produce a response. Thus, the 2,3-Jet validation does not have sufficient jet multiplicity to actually calculate a BDT response.



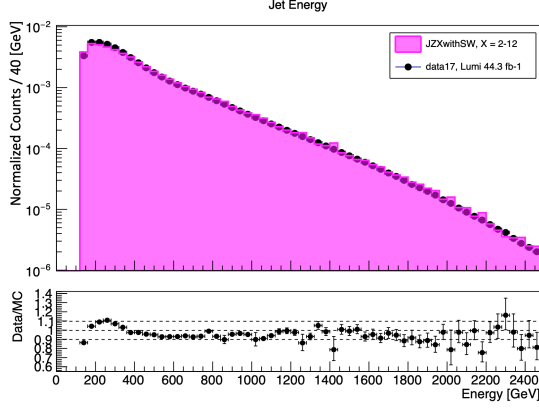


**Figure 6.33:** QCD Background distribution in the Low- $H_T$  spectrum defined by inverting the search region selection. Data17 has been plotted on top, to show the agreement between Data and MC. The different sub-regions of the Low- $H_T$  spectrum, including the 4-Jet validation region, are indicated on the plot.

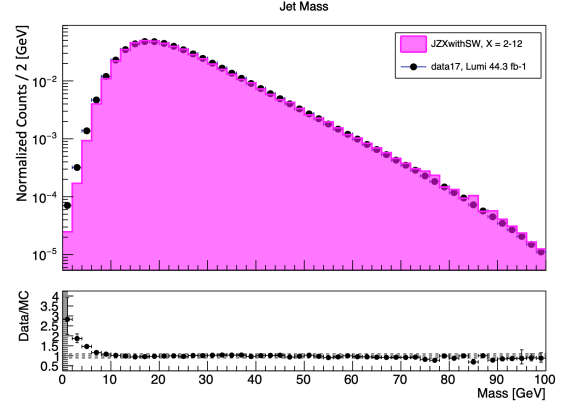
#### 6.4.3 Data - Monte Carlo Comparisons

Because the BDT will use the shape of the input variables in order to separate signal and background, it is important to ensure that the modelling of these variables in MC agrees with the data. Thus, data-MC comparisons are made for each of the input variables, as well as the BDT response distributions and the ABCD axis variables. However, it is important to remember that the final background estimate will be data-driven, so any mismodelling in the background MC will only impact the estimated signal efficiencies.

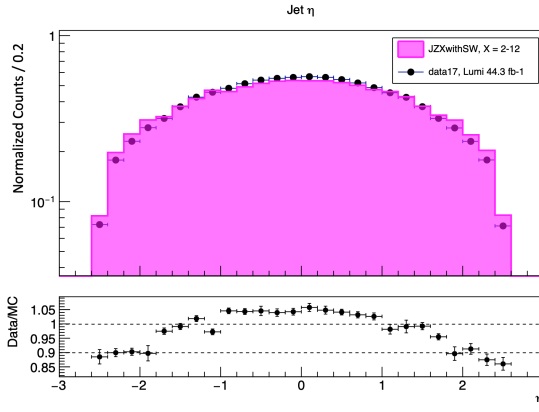
For the purpose of this thesis, validation is done using the full set of Data17, which corresponds to  $44.3 \text{ fb}^{-1}$  of integrated luminosity. However, there is an offset of  $\sim 13\%$  in the scale when comparing data with the MC from mc16d (the source of which is still under investigation), so for the sake of proper shape comparison, the data-MC comparisons will use normalized distributions, where each is scaled to unit area.



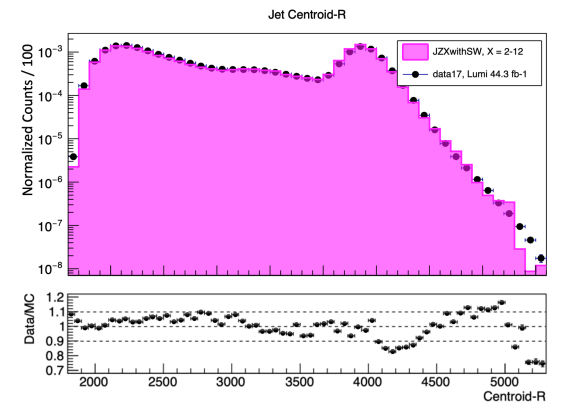
(a) Data-MC Comparison: Jet Energy



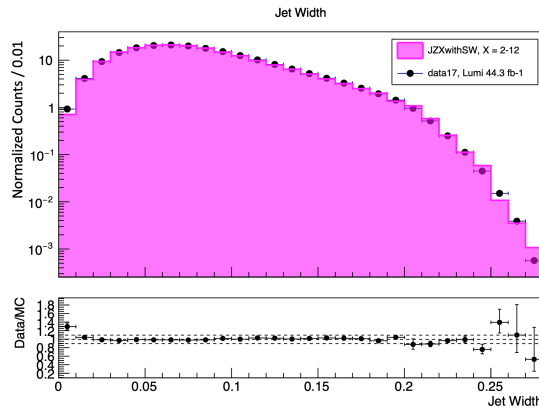
(b) Data-MC Comparison: Jet Mass



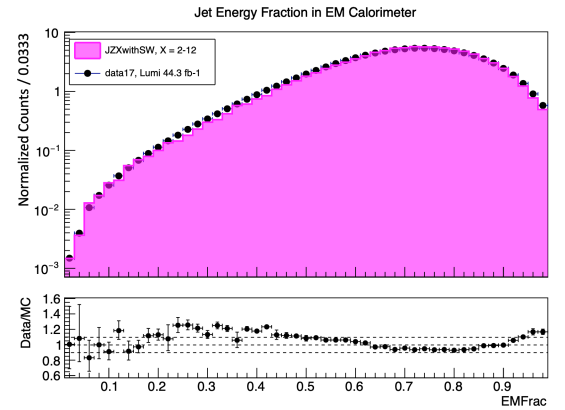
(c) Data-MC Comparison: Jet  $\eta$



(d) Data-MC Comparison: Jet Centroid-R



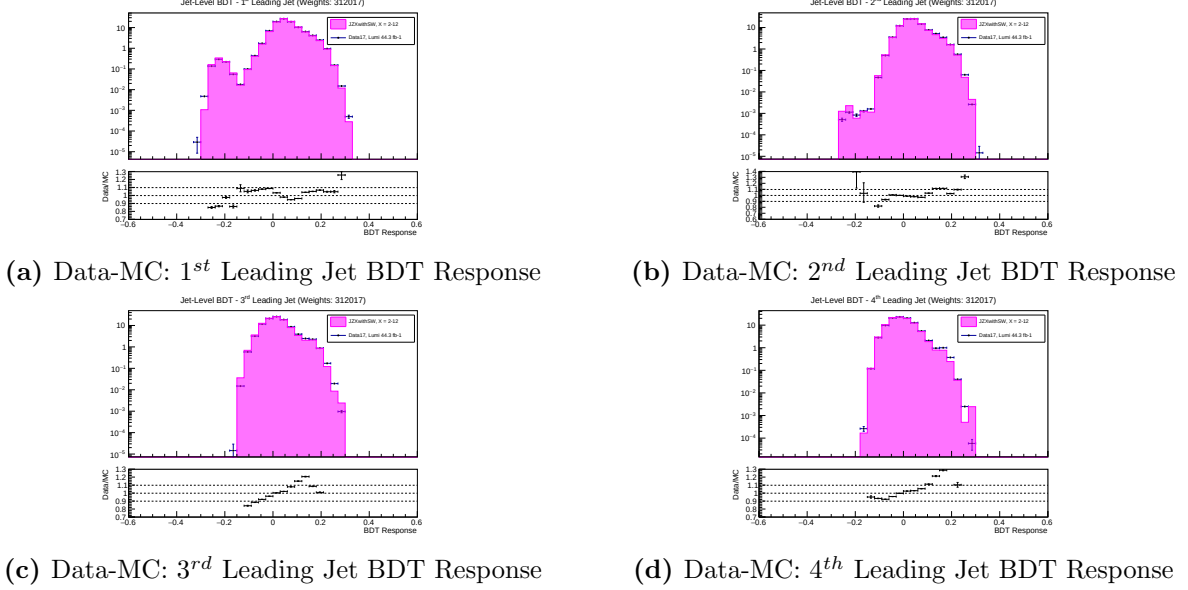
(e) Data-MC Comparison: Jet Width



(f) Data-MC Comparison: Jet EM Frac.

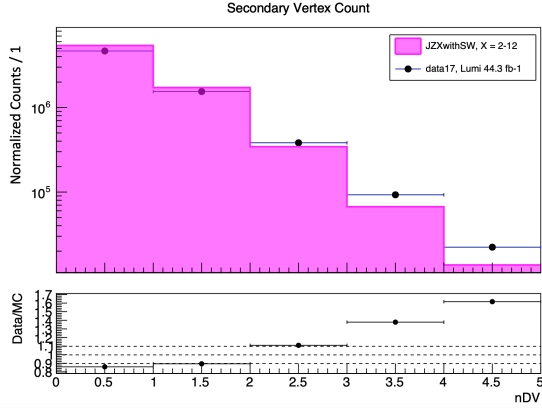
**Figure 6.34:** Data-MC comparisons for the jet-level BDT input variables, in the 4-Jet validation region. The data (black points) is the full set of Data17, and the MC (filled histogram) is weighted to the corresponding luminosity, then both distributions are normalized to unit area.

The Jet-Level BDT response distributions are shown in figure 6.35. The validation BDT response is calculated using the weights from each of the signal points. As an example, the distribution for 312017 (Model A,  $m_{X_d} = 600$  GeV,  $c\tau_{\pi_d} = 1mm$ ) is shown.

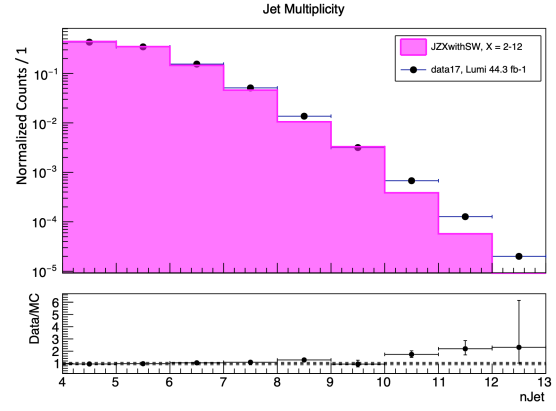


**Figure 6.35:** Data-MC comparisons for the jet-level BDT Response, in the 4-Jet validation region. The data (black points) is the full set of Data17, and the MC (filled histogram) is weighted to the corresponding luminosity, then both distributions are normalized to unit area. The validation distributions are tested against the weights for signal point 312017: (Model A,  $m_{X_d} = 600$  GeV,  $c\tau_{\pi_d} = 1mm$ ).

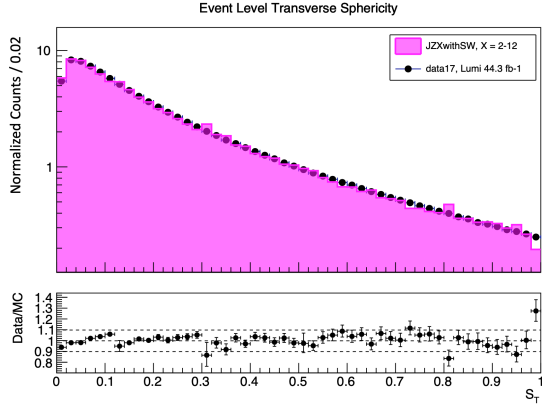
Figure 6.36 shows the data-MC comparisons for the event-level BDT input variables, as well as the validation  $\mathcal{R}$  distribution (6.36f). The event  $H_T$ ,  $S_T$ , and  $\mathcal{A}$  all have relatively good agreement between data and MC and, like the jet-level input variables, the differences may be covered by the shifts given from the jet systematics. The jet and DV multiplicities show more obvious disagreement, especially in the tails of the distributions; these variables are also not covered by the jet systematics, and thus require more careful treatment when evaluating the systematic uncertainties.



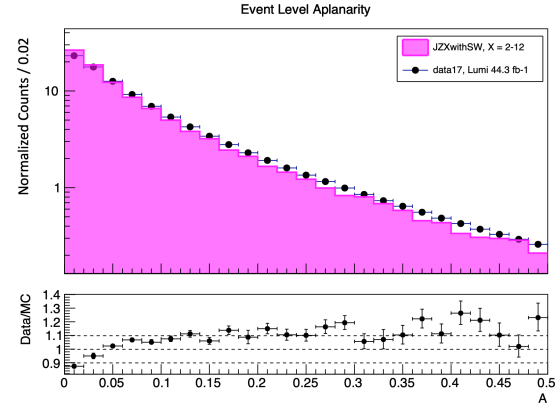
(a) Data-MC Comparison: Event  $n_{DV}$



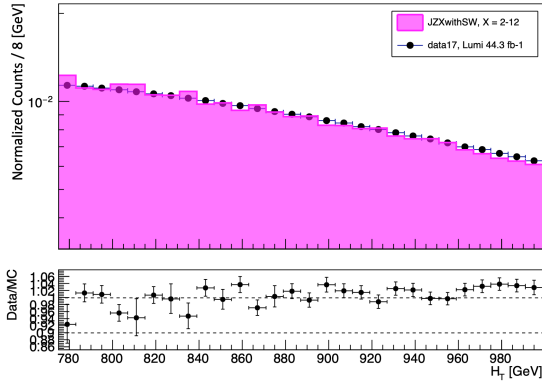
(b) Data-MC Comparison: Event  $n_{Jet}$



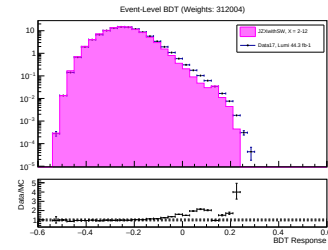
(c) Data-MC Comparison: Event  $S_T$



(d) Data-MC Comparison: Event  $\mathcal{A}$



(e) Data-MC Comparison: Event  $H_T$



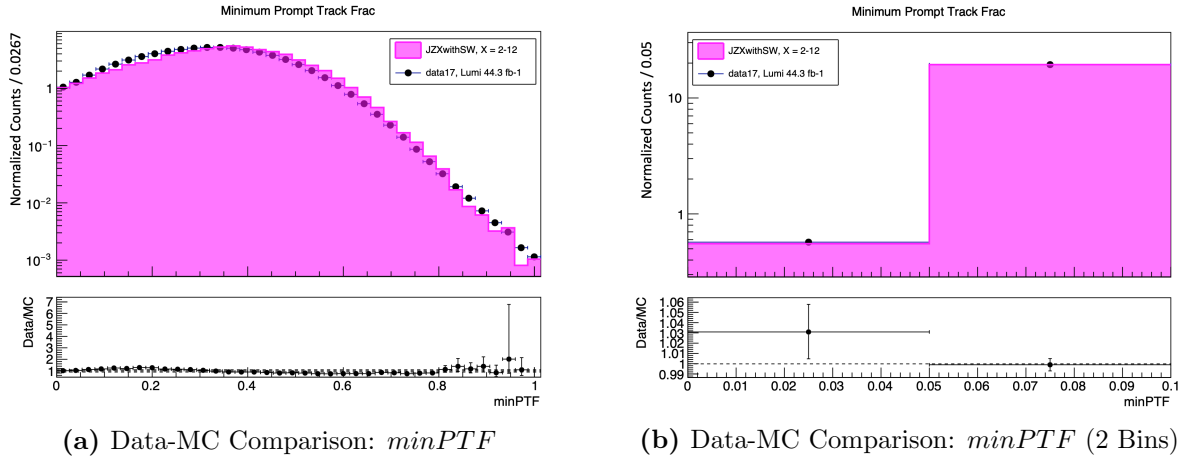
(f) Data-MC Comparison: Event  $\mathcal{R}$

**Figure 6.36:** Data-MC comparisons for the Event-level BDT input variables, in the 4-Jet validation region. The data (black points) is the full set of Data17, and the MC (filled histogram) is weighted to the corresponding luminosity, then both distributions are normalized to unit area.

The validation  $\mathcal{R}$  shows some disagreement in the high-side tail; examining the trends in figures 6.19a/6.19b and 6.20a/6.20b, higher DV and jet multiplicities indicate more signal-like events. As such, the higher tails in the data may be skewing the data  $\mathcal{R}$  distribution to the right, thus indicating that the event-level BDT is more susceptible to the mismodelling of its input variables. This highlights the importance of a careful systematic evaluation for the different BDT inputs.

### Modelling PromptTrackFrac.

The last variable which requires a thorough evaluation is the minimum promptTrackFrac., since it acts as an axis of the ABCD plane for both the model independent and model dependent approaches. Figure 6.37a shows some slight mismodelling, with the data again being skewed in a more signal-like manner (in this case to lower  $minPTF$ ).

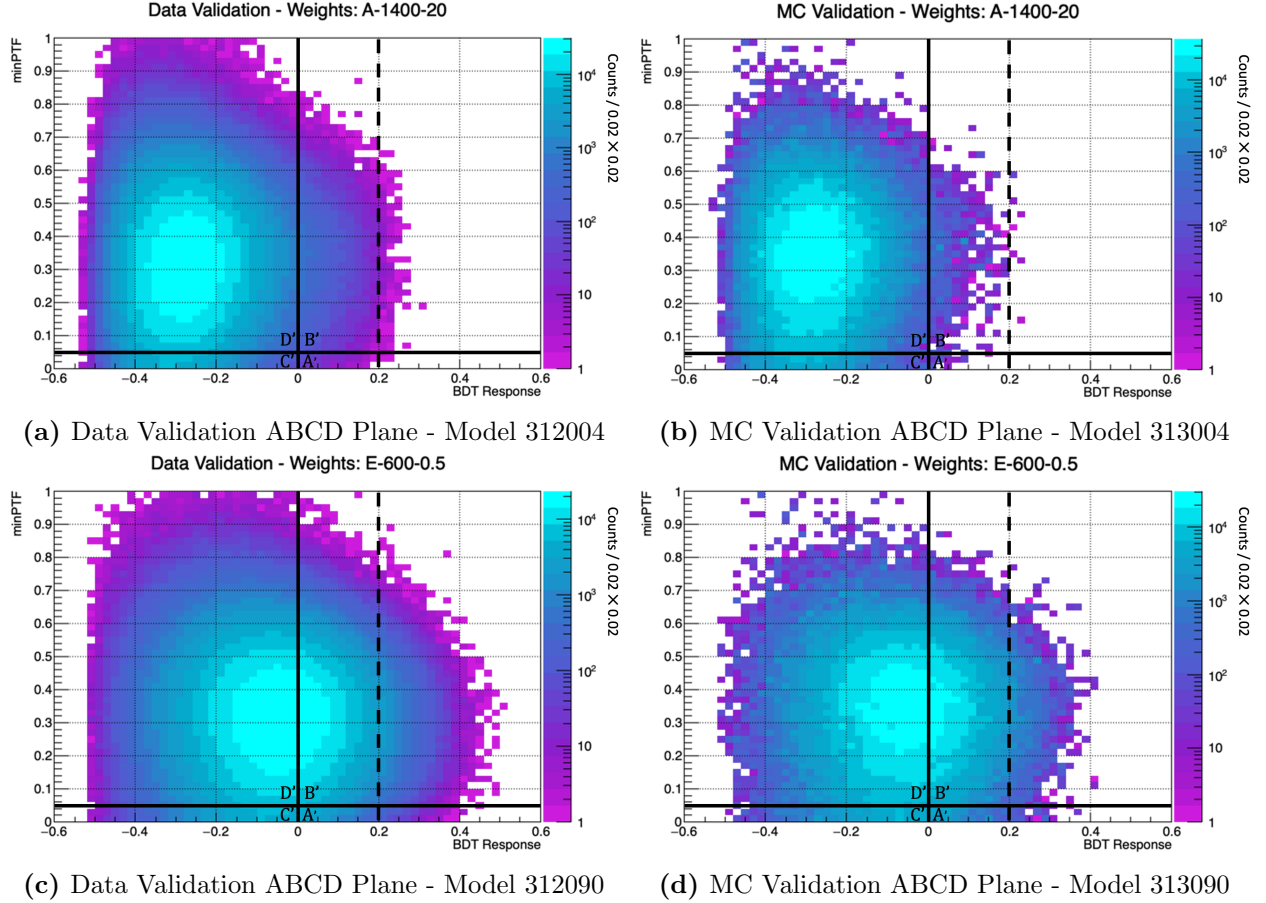


**Figure 6.37:** Data-MC comparison for the  $minPTF$  (on the left) and  $minPTF$  as a 2-bin distribution, in the 4-Jet validation region. In the 2-bin plot the first bin contains only events with  $minPTF \leq 0.05$ , and the second has all events with  $minPTF > 0.05$  (note that the axis is labeled to 0.1, but extends to 1). The data (black points) is the full set of Data17, and the MC (filled histogram) is weighted to the corresponding luminosity, then both distributions are normalized to unit area.

However, since  $minPTF$  is an ABCD axis variable, it acts more as a 2-bin distribution, separated by the cut at 0.05. When examined this way, figure 6.37b shows much better agreement between data and MC. Furthermore, since the actual background estimate will be data-driven, the modelling of the  $minPTF$  in the background MC is not so important.

## Validation ABCD Planes

In addition to data-MC comparisons, the 4-jet validation region can be used to create corresponding ABCD planes, both for the background, and the data. These planes can be used to validate the ABCD method, and develop corrections to the background estimate (see section 7.1.3).



**Figure 6.38:** Validation ABCD planes in both background MC and Data using the BDT weights from models 312004 (Model A,  $m_{X_d} = 1400$  GeV,  $c\tau_{\pi_d} = 20mm$ ) and 312090 (Model E,  $m_{X_d} = 600$  GeV,  $c\tau_{\pi_d} = 0.5mm$ ). The modified cuts of  $minsPTF = 0.05$  and  $\mathcal{R} = 0$  are shown (solid black lines) with the corresponding regions following the alternate ABCD arrangement. The nominal BDT cut of  $\mathcal{R} = 0.2$  is shown (dashed black line) for comparison. The data planes represent the full Data17, and the MC planes have been scaled to the corresponding  $44.3 \text{ fb}^{-1}$  of integrated luminosity.

Because the validation region contains events which are meant to be more background-like, the resulting  $\mathcal{R}$  distributions are skewed to the left, as compared to the search region background distributions. To account for this, the cut on the BDT response is lowered to  $\mathcal{R} = 0$ . Figure 6.38 shows the validation ABCD planes (both MC and data) using the weight from models 312004 and

312090. These distributions dramatically show the effect of the BDT; for the well separated cases (312004) the validation  $\mathcal{R}$  is much farther to the left, while the poorly separated cases (312090) have BDT responses which are more similar to the corresponding search region.

## Chapter 7

# Systematic Uncertainties and Results

### 7.1 Systematic Uncertainties

The EJs analysis will use the profile likelihood method for evaluating the data-driven background estimate; this computes the yields in the ABCD regions by performing a fit in all four regions simultaneously. Because of this, the effects of each systematic uncertainty ( $\sigma_{syst}$ ) must be evaluated as a percent change to the yield in all four ABCD regions:

$$\sigma_{syst}[\%] = \frac{N_i^{Nominal} - N_i^{Shifted}}{N_i^{Nominal}} \times 100\%, \quad i = \{A, B, C, D\} \quad (7.1)$$

where ( $N_i^{Shifted}$ ) is determined by using systematically shifted distributions when building the ABCD plane distributions. However, for the majority of the systematics, the corresponding uncertainty is on the signal efficiency, so the quoted uncertainty is given by the shift in region A.

In general, the distributions corresponding to a given systematic may be shifted both up or down, resulting in two shifts for a systematic. If the shifts are asymmetric, then they can be symmetrized to give a single  $\sigma_{syst}$ . For a conservative estimate, the prescription is:

$$\sigma_{syst} = \max(|\sigma_{syst}^{up}|, |\sigma_{syst}^{down}|) \quad (7.2)$$

which will give the larger of the two for asymmetric shifts, or simply return the value of symmetric shifts.

Similarly, the method for evaluating the systematic uncertainties using the BDT is the same



regardless of the source of the systematic. Shifted distributions are tested against the BDT using the nominal (unshifted) weights. This method produces a shifted BDT response ( $\mathcal{R}'$ ), without retraining the BDT weights on the shifted distributions. The  $\mathcal{R}'$  distribution is then used as the x-axis of the ABCD plane, and the standard evaluation of  $\sigma_{syst}$  is conducted.

### 7.1.1 Sources of Systematic Uncertainty on the Signal Efficiency

The following systematics represent uncertainties on the signal efficiency, and therefore the values of  $\sigma_{syst}$  are only given for the signal points. Because the final result will use a data-driven background estimation, the corresponding shifts to the background MC are less relevant to the analysis.

The prescription for evaluating the uncertainties on the jet energy scale and resolution, the jet mass scale, and pileup all come from the relevant combined performance groups in ATLAS. The *jet uncertainties* (jet energy scale, jet energy resolution, and jet mass scale) all come from the Jet  $E_T^{Miss}$  (JetEtMiss) group, which deals with jets and missing transverse energy. The pileup prescription comes from the ATLAS analysis software group, which develop the physics analysis tools (PAT), including the pileup reweighting tool.

The other systematics have no predefined prescription, therefore the methods for evaluating the uncertainties were developed for this thesis.

## Jet Energy Scale

The *Jet Energy Scale* (JES) [63] is the measurement which is used to calibrate the reconstructed energy of jets in the ATLAS detectors. This measurement has a corresponding uncertainty which is used to create the systematic uncertainty associated with the JES.

The JES uncertainty is  $p_T$  and  $\eta$  dependent, meaning that different jets may have different associated uncertainties. The systematic is then determined by shifting the jet's 4-momentum vector by the corresponding JES uncertainty.

The JES systematic is built from several different nuisance parameters (NPs) which correspond to different components of the uncertainty, where each NP represents a different shift to the jet 4-momentum. The full set of JES shifts contains roughly 100 NPs, and is aimed at the most

precise analyses. These NPs can be combined in different scenarios to give smaller uncertainties for analyses which are less dependent on the JES.

These combinations are done in a way which minimizes correlations between the NPs such that each one can be evaluated as a standalone effect. Each NP has a corresponding up and down shift, which are symmetric at jet level, but may be asymmetric at event level, and thus are symmetrized using the prescription in eqn. 7.2. In the end, the effect of each NP is combined together to give one total systematic uncertainty from the JES.

## Jet Energy Resolution

Similar to the JES, the *Jet Energy Resolution* (JER) [63] is used to calibrate the reconstructed jet energy. However, the JER aims to improve (minimize) the resolution ( $\mathbf{R}$ ) of the jet, that is:

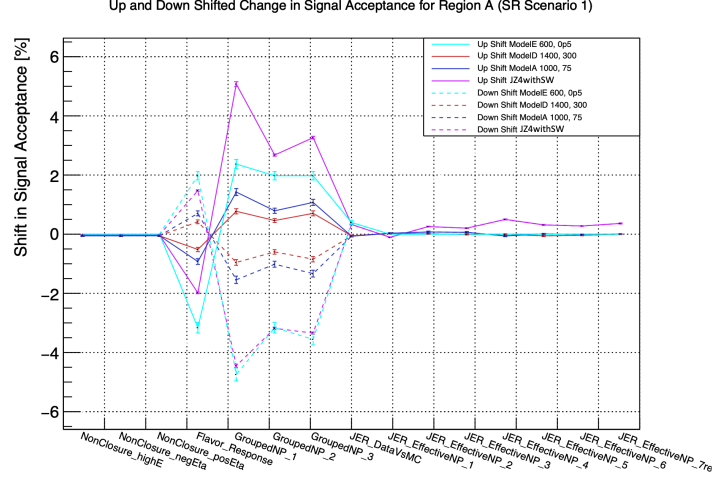
$$\mathbf{R} = \frac{\sigma(E)}{E} \approx \frac{\sigma(p_T)}{p_T} \quad (7.3)$$

where the approximation holds for constant rapidity. The corresponding systematic is given by the residual uncertainty.

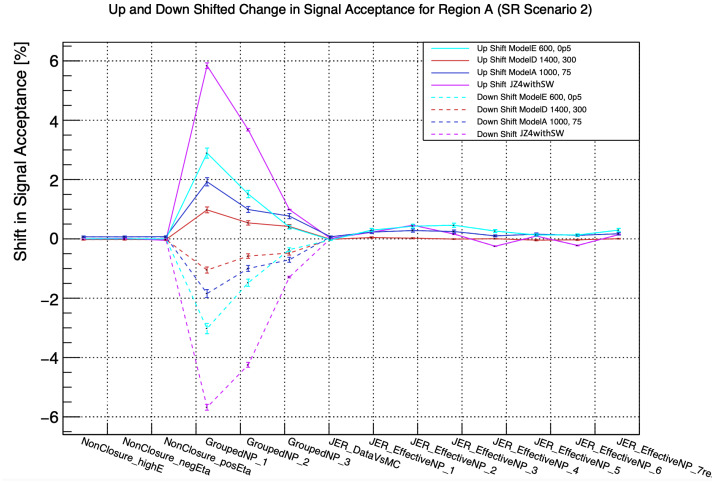
Like the JES, the JER uncertainties affect the full 4-momentum of the jet, and is built from a collection of NPs. However, the NPs give an absolute shift, so there is no up and down variation. The full JER systematic is a set of 34 NPs, which can also be combined in an uncorrelated way to give a smaller set of NPs for analyses which are less sensitive to the JER. Again, these NPs must also be combined to give an overall JER uncertainty.

## Comparing Strong Reduction Scenarios

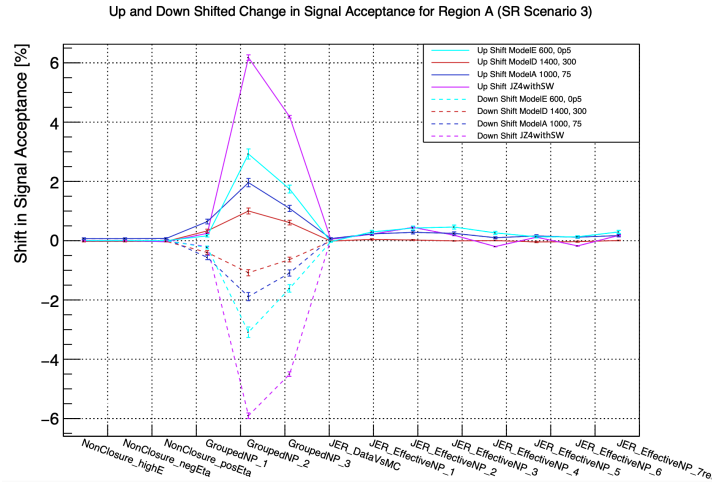
The smallest set of JES NPs is known as the *Strong Reduction* (SR), which combines the set of 100 into 8 NPs. This is intended only for analysis which can demonstrate that they are insensitive to the effects of the JES. In order to determine if the analysis is sufficiently insensitive, three varying SR scenarios are tested to see if a significant change in the systematic uncertainty is observed. It should be noted that the effects of each individual NP may change between scenarios, but the overall effect of the JES, after combinations, should remain roughly the same.



(a) Strong Reduction - Scenario 1



(b) Strong Reduction - Scenario 2



(c) Strong Reduction - Scenario 3

**Figure 7.1:** Up and down shifts given by the various nuisance parameters of the JES Strong Reduction (SR) scenarios and the Simple JER. Shifts are given individually for each NP, and as a percent change to the counts in region A. Three example signal points (312008, 312060, 312090), and the dominant background slice (JZ4withSW), are shown in each case.

For the different SR scenarios, there is only one combination scheme for the JER. The so-called *Simple JER* reduces the full set of 34 NPs into a set of eight. The different scenarios may also change the effect of the individual JER NPs, however, the total JER uncertainty should also remain roughly the same.

The effect of each NP in the Strong Reduction and Simple JER is given for all three scenarios in figure 7.1. The test was conducted for three signal points (one from each mediator mass group) and the dominant background slice (JZ4withSW). The main contributing NP changes between the three scenarios, however the overall scale remains the same. It should be noted that the *Flavor Response* NP is not in the 2<sup>nd</sup> and 3<sup>rd</sup> scenarios. In the end it was concluded that the analysis is sufficiently insensitive, so it will use the SR scenario 1, with the Simple JER to evaluate the effect of the systematic uncertainties from the JES and JER.

## Jet Mass Scale

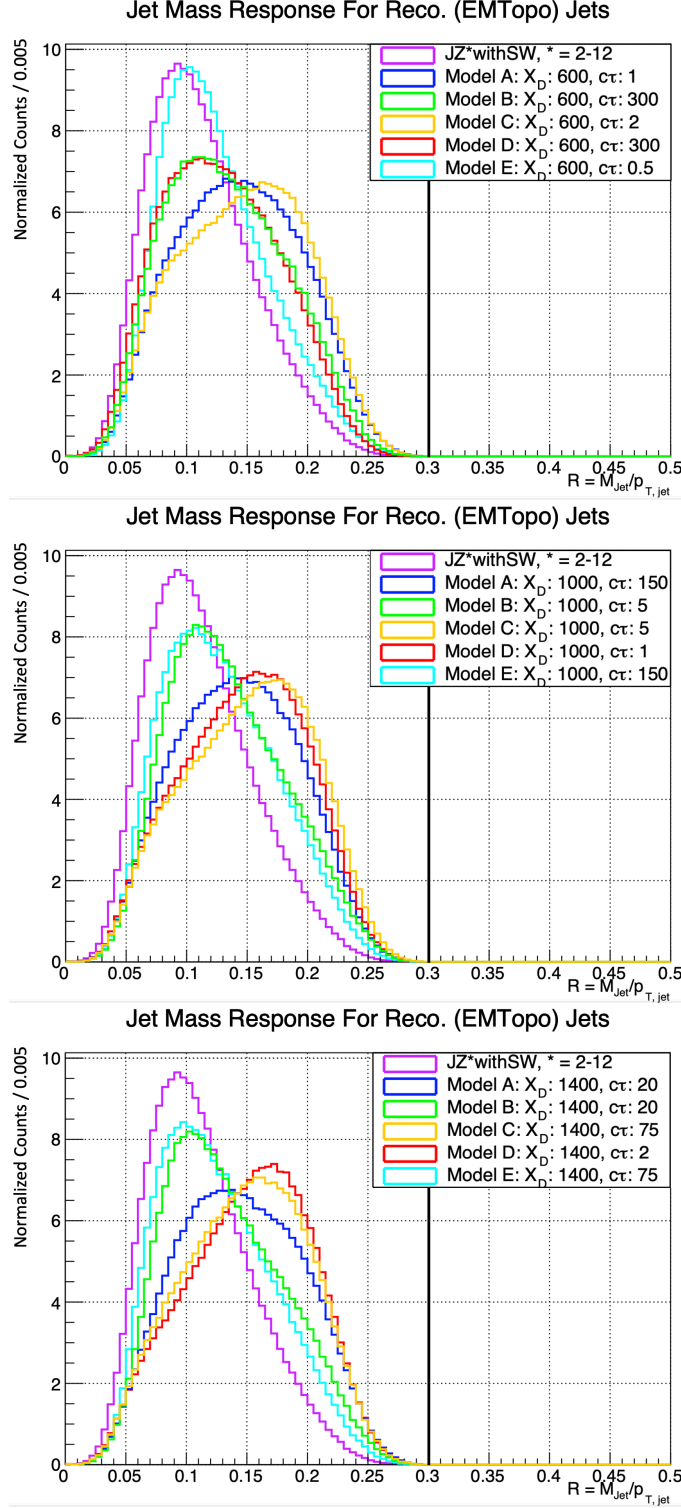
The *Jet Mass Scale* (JMS) [64] is an additional uncertainty applied after the JES, for analyses which depend on the reconstructed mass of the jet. The JMS uncertainties are built from track information, as opposed to the calorimeter information used by the JES and JER.

The JMS provides two configurations of NPs, depending on the analysis' sensitivity to the jet mass response:

$$R = \frac{M_{jet}}{p_{T,jet}}. \quad (7.4)$$

For analyses which are sensitive to values of  $R$  above 0.3, the scaled configuration must be used (which linearly scales the JMS uncertainty after  $R = 0.25$ ). Alternatively, for analyses which are insensitive to  $R = 0.3$ , the frozen configuration can be used (which takes the JMS uncertainty at  $R = 0.25$  for all values of  $R \geq 0.25$ ).

Figure 7.2 shows the response distributions for jets in the canonical set, as well as the QCD background; this shows that the analysis is insensitive to  $R \geq 0.3$  and thus the frozen configuration can be used. Much like the other jet uncertainties, each NP provides an up and down shift which must be symmetrized and combined to give the final JMS uncertainty.

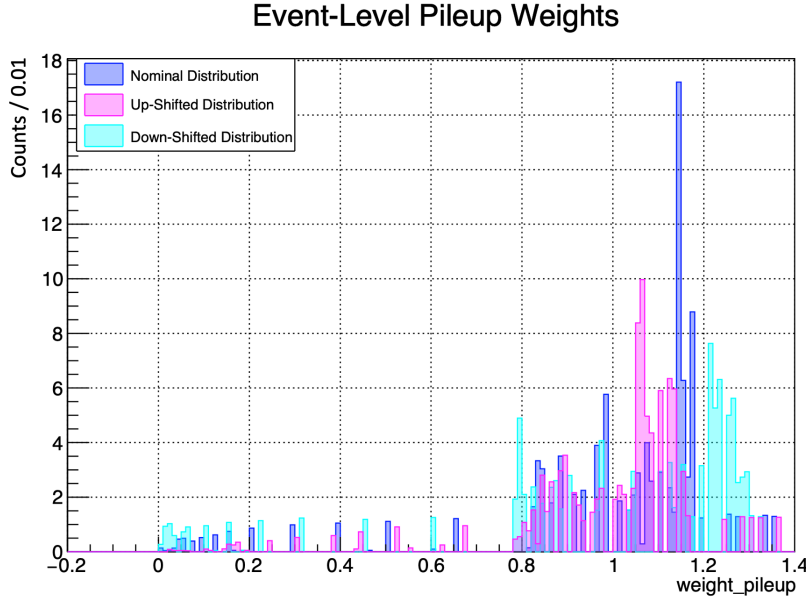


**Figure 7.2:** Jet mass response distributions for the canonical set, grouped by mediator mass:  $m_{X_d} = 600$  GeV (left),  $m_{X_d} = 1000$  GeV (center),  $m_{X_d} = 1400$  GeV (right). In each case the QCD background has been plotted for comparison. The cut of  $R = 0.3$  is shown in black. All distributions have been normalized to unit area for the sake of comparison.

## Pileup

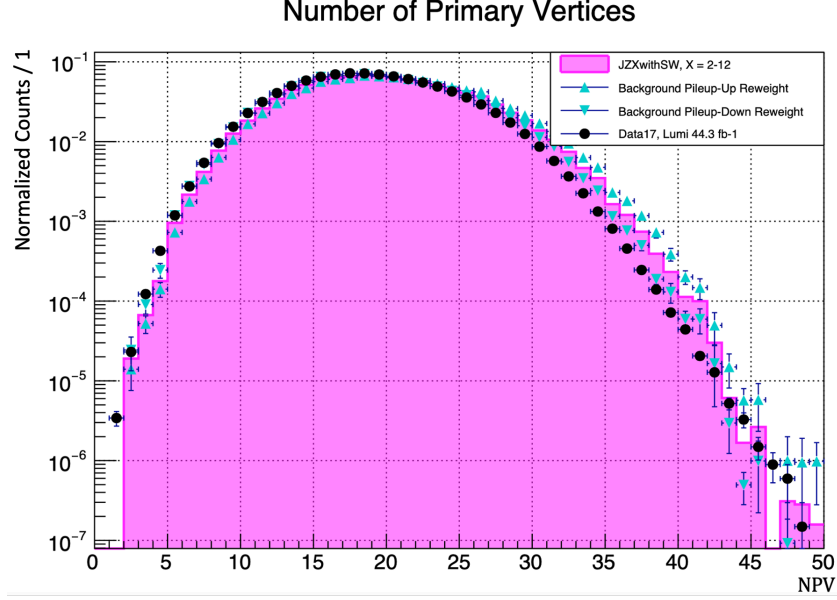
Pileup reweighting [65] is applied to MC in order to match the corresponding average interactions per bunch crossing ( $\langle\mu\rangle$ ) distribution from data. This adds the event-level weight (described in section 5.1.4), which is applied to MC distributions, and may weight a given event to be greater than or less than one.

Because there is no difference between  $\mu$  and  $\langle\mu\rangle$  in MC, but there is a difference in data ( $\langle\mu\rangle$  is averaged across all the collisions in a run), there is a discrepancy when trying to reweight the  $\langle\mu\rangle$  for MC using the actual  $\mu$  from data. To account for this difference, the pileup weights get shifted up and down, while still maintaining the properties discussed in section 5.1.4; namely that  $\sum_i w_i^{pileup} \approx N_{final}$  and  $\langle w_i^{pileup} \rangle \approx 1$  for both the up and down shifted weights.



**Figure 7.3:** Distribution of event-level pileup weights for the nominal reweighting, and both up and down shifted weights.

To evaluate the systematic from pileup, the up and down shifted weights are used in the event weighting, and the corresponding change in the signal efficiency is determined. This is different to the systematic shifts of the jet uncertainties, which changes the unweighted distributions; the shifted weights can only indirectly change the distributions by weighting certain events more or less strongly.



**Figure 7.4:** Data-MC comparison for the number of primary vertices ( $\mu$ ) in the 4-Jet validation region. The shifted MC distributions, using the up and down shifted pileup weights, have been plotted for comparison. The data (black points) is the full set of Data17, and the MC (filled histogram) is weighted to the corresponding luminosity, then all distributions are normalized to unit area.

In contrast to the NP sets of the jet uncertainties, the pileup systematic only has one shift up and down, so once symmetrized, it gives the final uncertainty.

Figure 7.4 gives the data-MC comparison of the actual  $\mu$  distribution, with corresponding shifted  $\mu$  distributions using the up and down shifter pileup weights. In the very low tail ( $\mu < 5$ ) there is some clear discrepancy between the data and the MC (both shifted and nominal). Then, for much of the middle of the distribution,  $5 \leq \mu \leq 25$ , the discrepancy between the nominal MC and the data is covered by the shifted pileup weights. There is another section for  $25 < \mu < 42$  where again the discrepancy between the data and the nominal MC is not covered by the shifted weights, but then in differences are covered again in the high side tail ( $\mu \geq 42$ ).

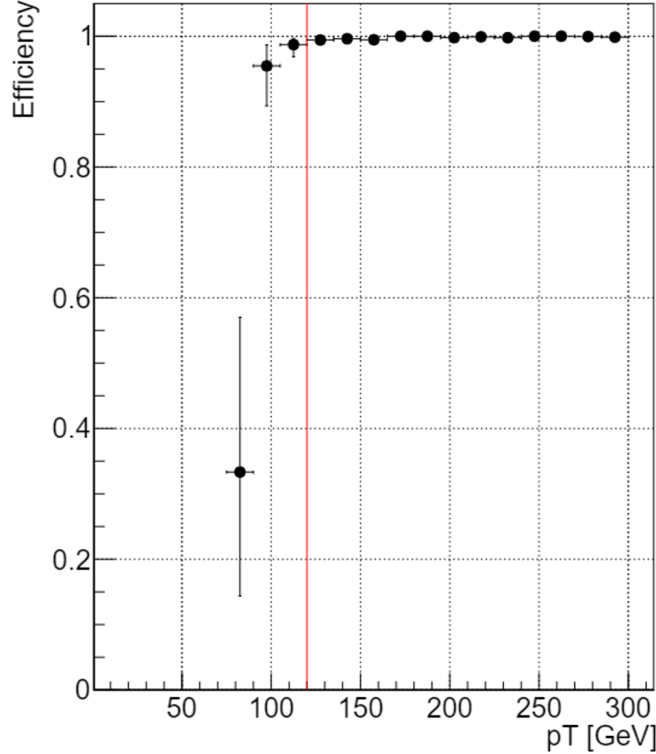
## Trigger

The EJs selection requires 4 jets with  $p_T \geq 120$  GeV, and is therefore dependent on the 4 jet trigger  $HLT\_4j120$ . In order to evaluate a systematic, the efficiency of the trigger must be determined.

To calculate the trigger efficiency, two other triggers are needed: a probe trigger ( $T_{probe}$ ) and a reference trigger ( $T_{ref}$ ). The probe trigger for the  $HLT\_4j120$  is the  $L1\_3j50$ , which seeds the 4 jet trigger in the HLT. The reference trigger, in this case the  $L1\_j30$ , is one that is assumed to be fully efficient for events which pass the probe trigger. Because the probe trigger requires at least three jets with  $p_T \geq 50$  GeV, while the reference trigger requires only one jet with  $p_T \geq 30$  GeV, it is expected that any events which pass the  $L1\_3j50$  will also pass the  $L1\_j30$ .

For any event which passes the  $L1\_3j50$ , the  $p_T$  of the third leading jets is determined, and similarly for the  $L1\_j30$ . The efficiency is then calculated as the ratio of these distributions, in bins of  $p_T$ :

$$\epsilon(p_T) = \frac{N(T_{probe})}{N(T_{ref})} \bigg|_{p_T^{3^{rd} \text{ Leading Jet}}} \quad (7.5)$$



**Figure 7.5:** Trigger *Turn On* plot for the  $L1\_3j50$ , for model 312004 (Model A,  $m_{X_d} = 1400$  GeV,  $c\tau_{\pi_d} = 20mm$ ). The plot shows the efficiency ( $\epsilon(p_T)$ ) for the given  $3^{rd}$  leading jet  $p_T$ . The selection of  $p_T = 120$  GeV is shown in red, with the trigger being highly efficient after this line. Figure from B. Death.



Since the trigger of interest is the *HLT\_4j120*, the efficiency is integrated for  $p_T \geq 120$  GeV to get the integrated efficiency ( $E$ ):

$$E = \int_{120}^{\infty} \epsilon(p_T) dp_T. \quad (7.6)$$

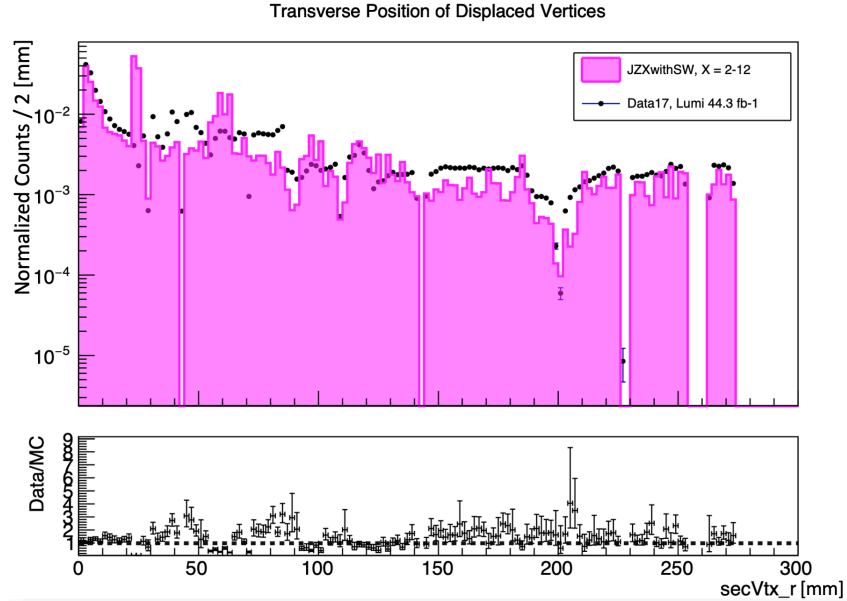
| Model  | Events  | $N(T_{probe})$ | $N(T_{ref})$ | Efficiency ( $E$ ) |
|--------|---------|----------------|--------------|--------------------|
| 312004 | 60,000  | 59,049         | 59,079       | 0.9995             |
| 312008 | 72,000  | 59,631         | 59,693       | 0.9990             |
| 312017 | 15,000  | 6,636          | 6,640        | 0.9994             |
| 312022 | 60,000  | 59,337         | 59,352       | 0.9997             |
| 312028 | 70,000  | 66,943         | 66,963       | 0.9997             |
| 312031 | 150,000 | 28,730         | 28,785       | 0.9981             |
| 312039 | 78,000  | 75,802         | 75,846       | 0.9994             |
| 312046 | 99,000  | 86,985         | 87,049       | 0.9993             |
| 312052 | 320,000 | 123,385        | 123,474      | 0.9993             |
| 312060 | 90,000  | 86,607         | 86,640       | 0.9996             |
| 312066 | 130,000 | 107,057        | 107,110      | 0.9995             |
| 312067 | 410,000 | 119,971        | 120,070      | 0.9992             |
| 312075 | 50,000  | 48,205         | 48,241       | 0.9993             |
| 312080 | 60,000  | 42,845         | 42,933       | 0.9980             |
| 312090 | 110,000 | 71,113         | 71,134       | 0.9997             |

**Table 7.1:** Integrated trigger efficiencies for the canonical set. The table gives the total number of events for a given model, as well as the number of events which pass the probe and reference triggers with the 3<sup>rd</sup> leading jet  $p_T \geq 120$  GeV. The integrated efficiencies are above 99% in all cases.

The integrated efficiencies are given in table 7.1, with the lowest efficiency at 99.8%. Since the trigger is sufficiently efficient at  $p_T \geq 120$  GeV, it was deemed unnecessary to include a systematic uncertainty from the trigger.

## Tracking-Vertexing

The tracking and vertexing (TrkVtx) systematic attempts to correct the discrepancy seen in the data-MC comparisons of nDV. To do this, the data-MC comparison of the radial (transverse) position of the DVs is used to create per-DV weights. The weights are based off of the relative difference (ratio) between the data and MC (where both distributions are normalized).



**Figure 7.6:** Data-MC comparison for the transverse DV position in the 4-Jet validation region. The data (black points) is the full set of Data17, and the MC (filled histogram) is weighted to the corresponding luminosity, then both distributions are normalized to unit area.

The assumption here is that the majority of the discrepancy is coming from the material veto; somehow it is not vetoing on DVs at certain radial distances, giving rise to the discrepancy seen in figure 7.6. Therefore, DVs in regions where the MC underestimates the data will be weighted more strongly, and DVs in regions where the MC overestimates the data will have their weights reduced.

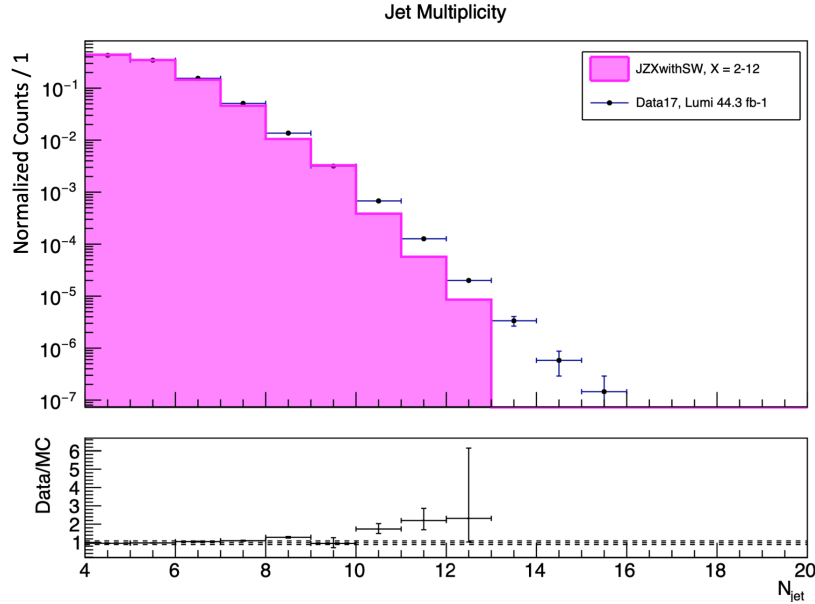
These per-DV weights are then combined for each DV in an event to make a per-event weight. These weights are then applied to the MC at event level in order to shift the nDV distribution so that it more closely resembles the data, particularly in the tail.

It should also be noted that the prescription described here is just a first attempt at an approximate estimation. This will likely serve as a starting point for the ongoing ATLAS analysis, which will likely implement an additional check on the tracking and vertexing uncertainty. To do this,

the per-DV weights from the radial position will be used to determine a relative uncertainty on the DV reconstruction. The DVs will then be reconstructed after randomly removing a selection of the tracks based on this relative uncertainty. Removing tracks may change how some DVs are reconstructed, therefore changing the nDV distribution.

## Jet Multiplicity

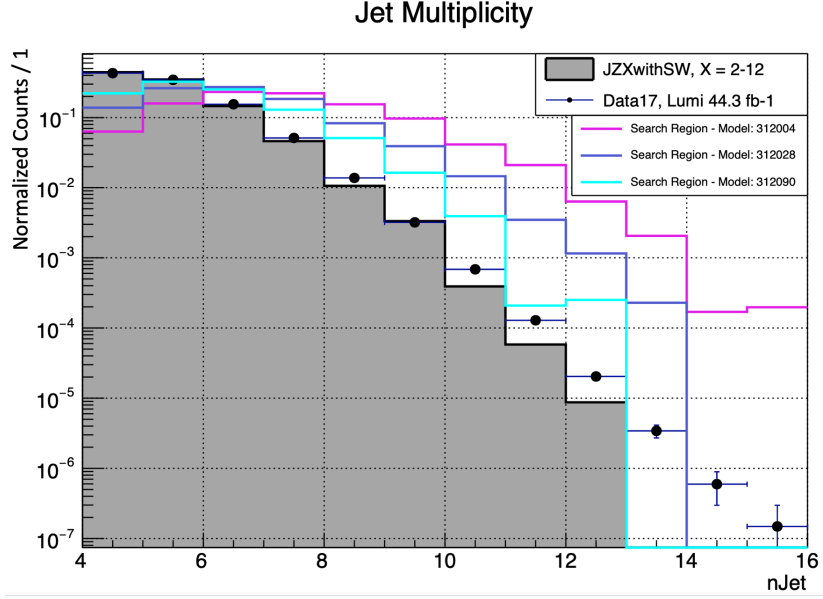
Like the TrkVtx, Jet Multiplicity (nJet) systematic attempts to correct the discrepancy in the data-MC comparison of the nJet distribution. Figure 7.7 shows the nJet distribution extended out to higher jet multiplicities, where there is a clear discrepancy in the tail since there is data at  $nJet \geq 13$  but no MC.



**Figure 7.7:** Data-MC comparison for the extended jet multiplicity in the 4-Jet validation region. The high side tail shows where there is data but no MC. The data (black points) is the full set of Data17, and the MC (filled histogram) is weighted to the corresponding luminosity, then both distributions are normalized to unit area.

This becomes a concern since the signal models tend to exhibit higher jet multiplicities (in the search region) than the QCD background. So, if the data is trending to higher nJet values than the MC in the validation region, then the separating power of the nJet distribution in the search region may be overestimated. Figure 7.8 shows the same comparison, with the nJet distribution of three signal points (312004, 312028, and 312090) in the search region overlaid on top for comparison.

In all three cases, the signal  $n\text{Jet}$  distributions exhibit higher jet multiplicities than the QCD background, suggesting that signal-like events have higher  $n\text{Jet}$  distributions.

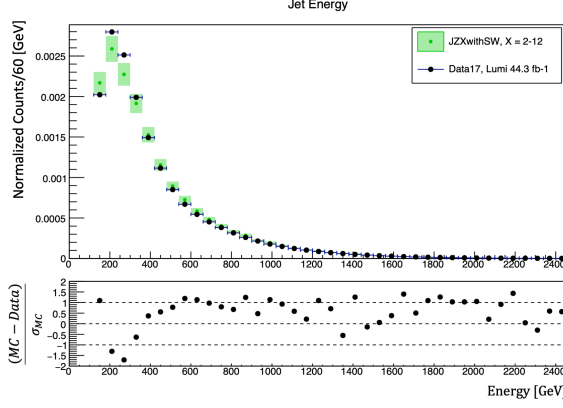


**Figure 7.8:** Comparison of the jet multiplicity distributions for the QCD multijet background and data in the validation region, and three signal points in the search region. All MC distributions were scaled to the Data17 luminosity of  $44.3 \text{ fb}^{-1}$ , before being normalized to unit area.

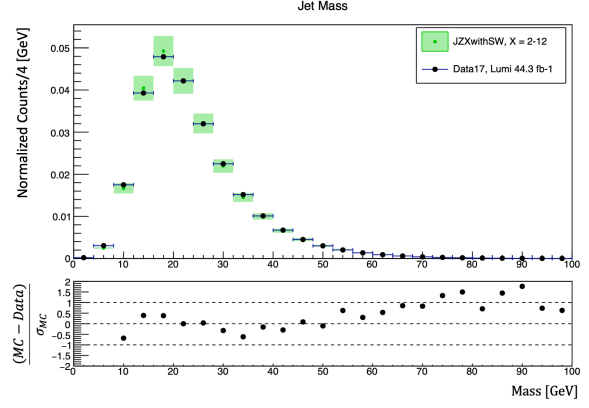
To evaluate the systematic in this case, event-level weights are built directly from the relative difference between data and MC. Since there is no MC for  $n\text{Jet} \geq 13$ , a constant uncertainty built from the average of the last three bins ( $n\text{Jet} = 10, 11, 12$ ) is applied at higher multiplicities. These are then applied to the MC order to shift the  $n\text{Jet}$  distribution so that it more closely resembles the data.

### 7.1.2 Evaluating Systematic Uncertainties

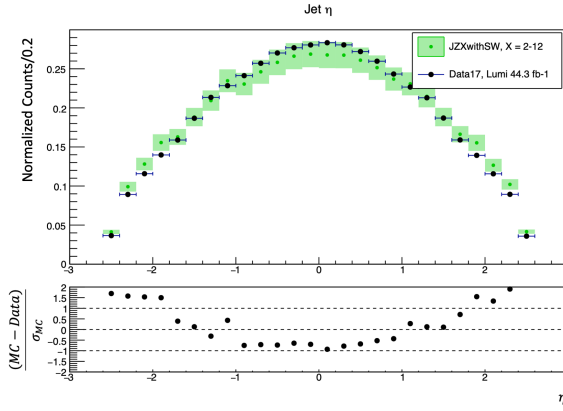
In addition to determining the systematic uncertainty, an evaluation of the systematics can be conducted by comparing the nominal and shifted distributions. This evaluation is conducted in the 4-Jet validation region to determine if the shifts cover the difference between the data and the MC.



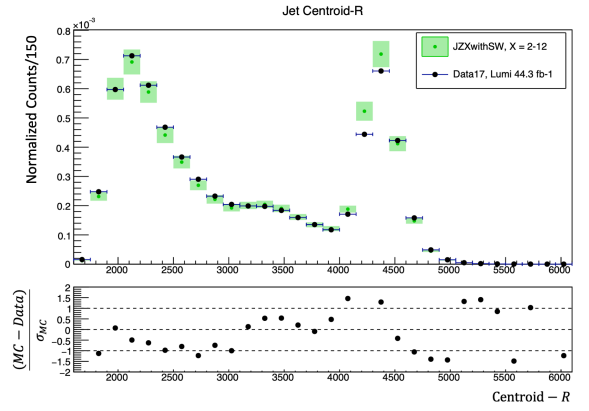
(a) Jet Systematics: Jet Energy



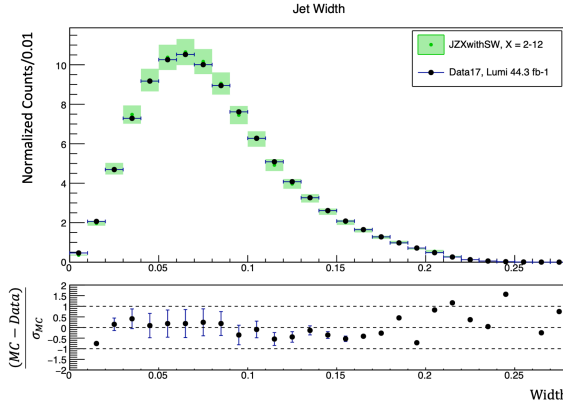
(b) Jet Systematics: Jet Mass



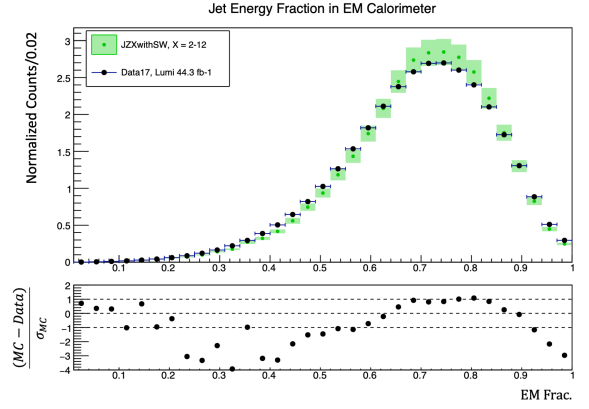
(c) Jet Systematics: Jet  $\eta$



(d) Jet Systematics: Jet Centroid-R



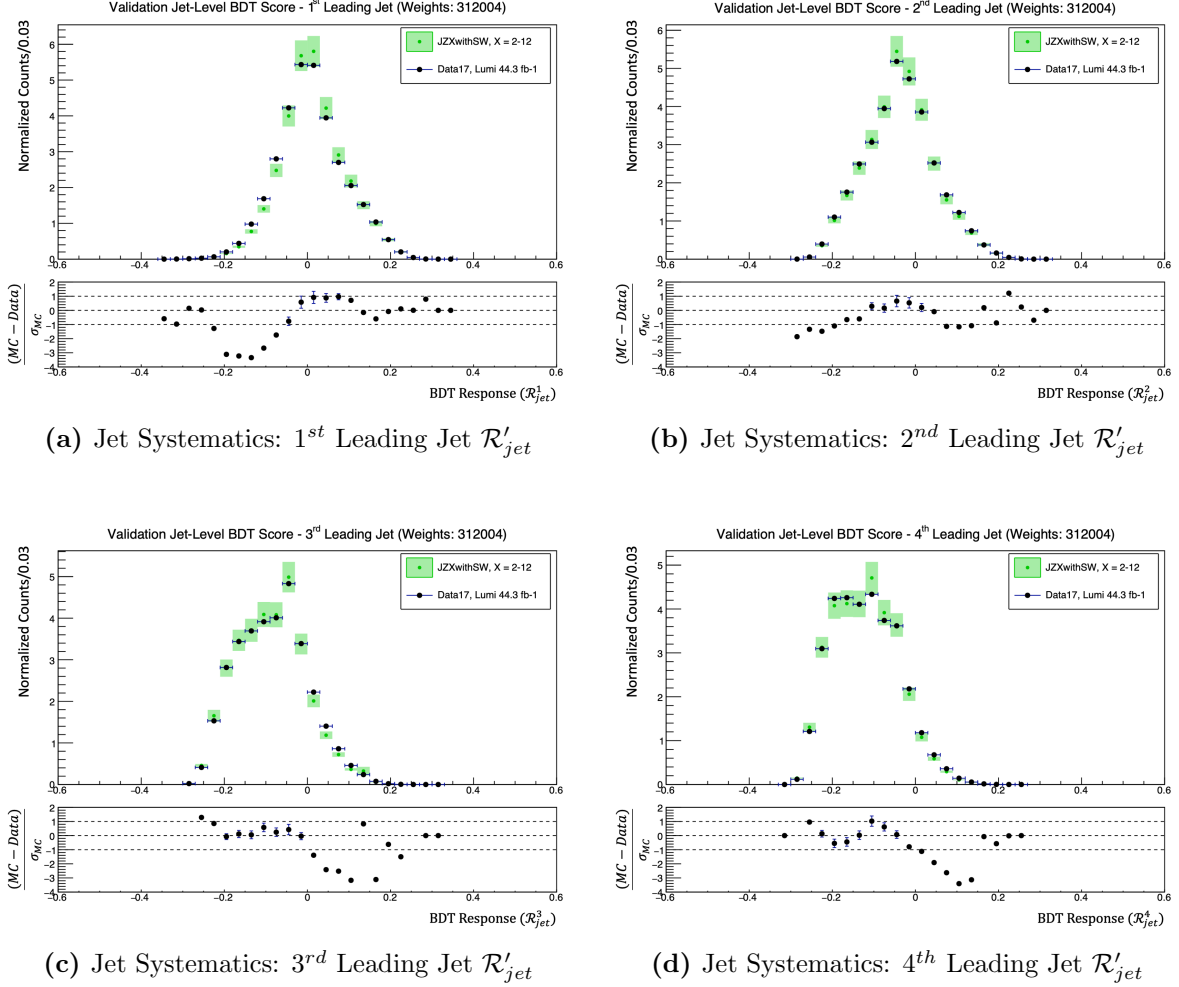
(e) Jet Systematics: Jet Width



(f) Jet Systematics: Jet EM Frac.

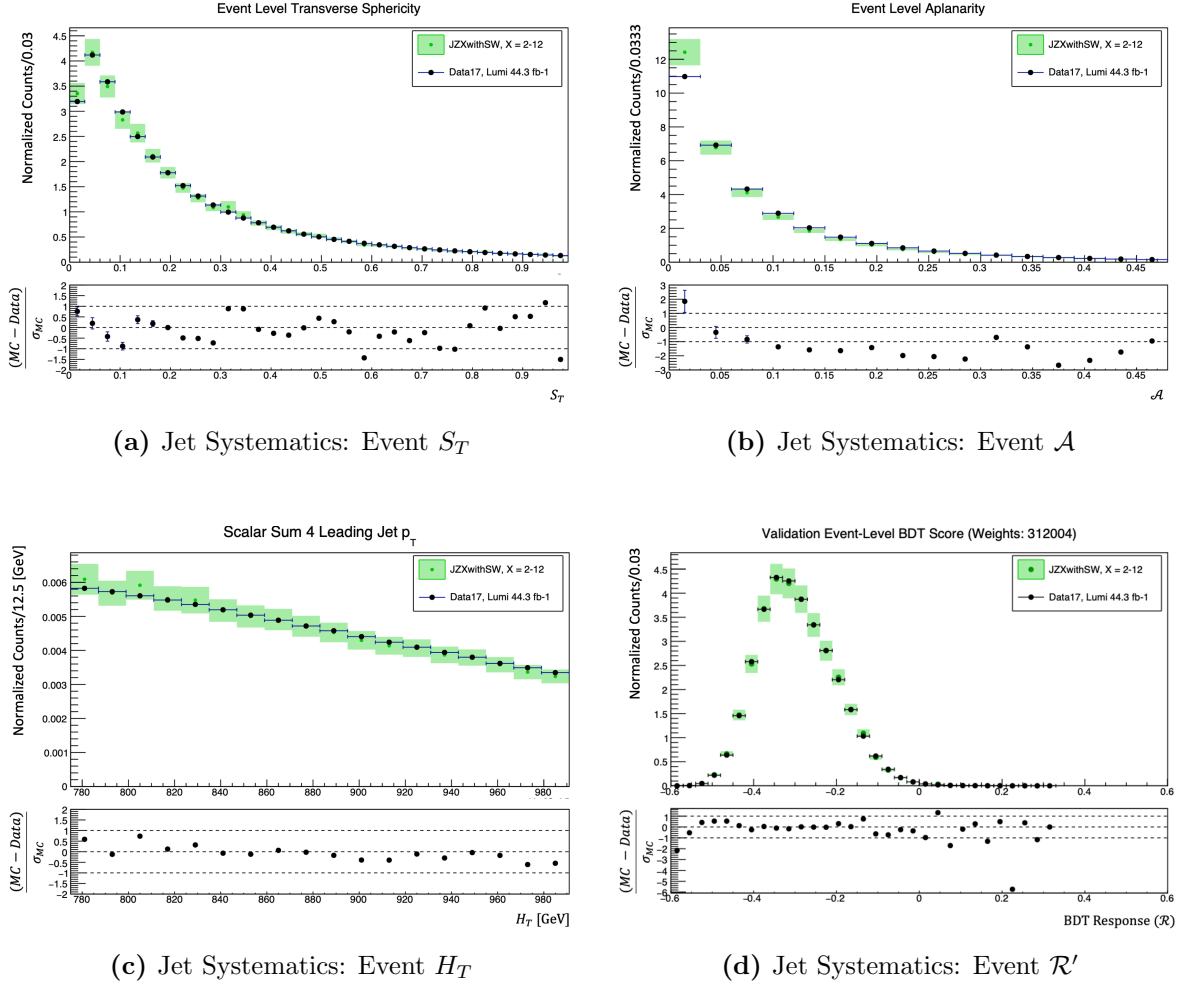
**Figure 7.9:** Evaluation of the effect of the Jet uncertainties (JES, JER, and JMS) on the jet-level BDT input variables, in the 4-Jet validation region. The green bands give the combined uncertainty,  $\sigma_{MC} = \sqrt{\sigma_{stat,MC}^2 + \sigma_{syst}^2}$ , where the dominant effect is from the systematic uncertainty. The comparison in the bottom panel gives the significance of the difference between data and MC. All distributions have been normalized to unit area for comparison.

Figure 7.9 shows the effect of the Jet uncertainties (JES, JER, and JMS) on the input variables of the jet-level BDT. In the plots the green bands give the combined uncertainty,  $\sigma_{MC} = \sqrt{\sigma_{stat,MC}^2 + \sigma_{syst}^2}$ , where in most cases  $syst = \{JES, JER\}$ , except for the jet mass distribution in which case  $syst = \{JES, JER, JMS\}$ . To account for the systematic uncertainty, the comparison in the bottom panel gives the significance of the difference between data and MC, and the error bars correspond to the data statistical uncertainty, which in some are too small to see.



**Figure 7.10:** Evaluation of the effect of the Jet uncertainties (JES, JER, and JMS) on the jet-level BDT response ( $\mathcal{R}'_{jet}$ ) for the four leading jets, in the 4-Jet validation region. The green bands give the combined uncertainty,  $\sigma_{MC} = \sqrt{\sigma_{stat,MC}^2 + \sigma_{syst}^2}$ , where the dominant effect is from the systematic uncertainty. The comparison in the bottom panel gives the significance of the difference between data and MC. All distributions have been normalized to unit area for comparison.

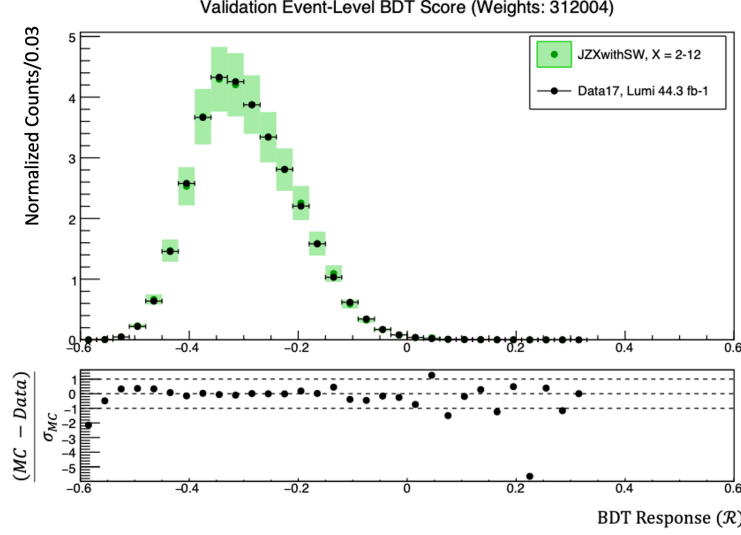
Figures 7.10 and 7.11 give similar comparisons for the jet-level BDT response and the event-level BDT inputs (which are sensitive to the Jet uncertainties), respectively. The latter also shows the shifted event-level BDT response ( $\mathcal{R}'$ ). In these cases,  $syst = \{JES, JER, JMS\}$  for both the jet and event-level BDT responses, since the jet mass is used in the training. However,  $syst = \{JES, JER\}$  for  $S_T$ ,  $\mathcal{A}$ , and  $H_T$  since jet mass is not involved.



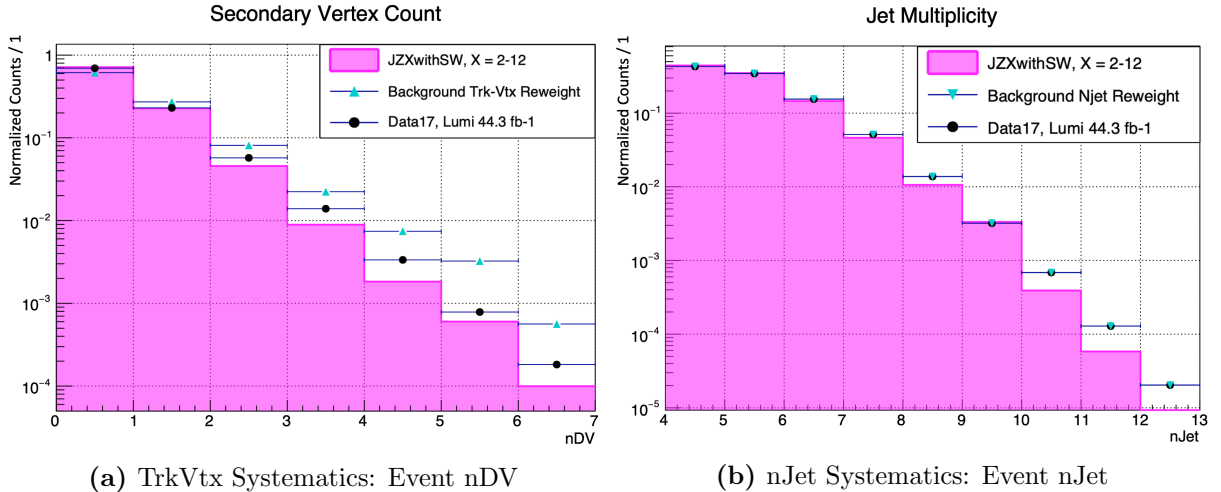
**Figure 7.11:** Evaluation of the effect of the Jet uncertainties (JES, JER, and JMS) on the event-level input variable which are affected ( $S_T$ ,  $\mathcal{A}$ ,  $H_T$ ) as well as the  $\mathcal{R}'$  distribution, in the 4-Jet validation region. The green bands give the combined uncertainty,  $\sigma_{MC} = \sqrt{\sigma_{stat,MC}^2 + \sigma_{syst}^2}$ , where the dominant effect is from the systematic uncertainty. The comparison in the bottom panel gives the significance of the difference between data and MC. All distributions have been normalized to unit area for comparison.

Pileup reweighting will not affect the input distributions directly, however the weights are shape-affecting so it may impact the final signal efficiency. The distribution where the pileup is expected to

have the biggest effect (in terms of the systematic uncertainty) is in the event  $\mathcal{R}'$ , which is shown in figure 7.12. In this case the green bands give the combined uncertainty,  $\sigma_{MC} = \sqrt{\sigma_{stat,MC}^2 + \sigma_{syst}^2}$ , where  $syst = \{Pileup\}$ .



**Figure 7.12:** Evaluation of the effect of the pileup reweighting on the event-level  $\mathcal{R}'$ , in the 4-Jet validation region. The green bands give the combined uncertainty,  $\sigma_{MC} = \sqrt{\sigma_{stat,MC}^2 + \sigma_{syst}^2}$ , where the dominant effect is from the systematic uncertainty. The comparison in the bottom panel gives the significance of the difference between data and MC. The distribution have been normalized to unit area for comparison.



**Figure 7.13:** Evaluation of the effect of (a) the TrkVtx systematic on the DV multiplicity, and (b) the nJet systematic on the jet multiplicity, in the 4-Jet validation region. All distributions have been normalized to unit area for comparison.



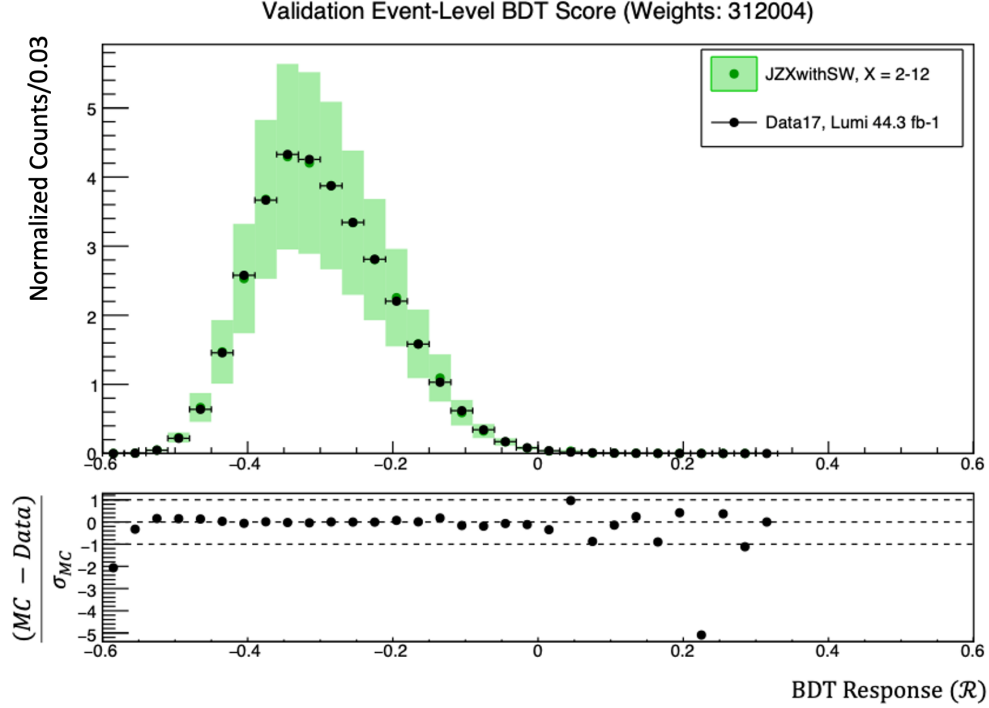
The reweighted distributions from the systematics implementation for the jet and DV multiplicities are shown in figure 7.13. Because these systematics are intentionally built to address discrepancies in the data-MC comparison, both of the shifted distributions sufficiently cover the differences. Since the TrkVtx uncertainty uses per-DV weights from the radial position to build event-level weights, the shifted distribution in 7.13a does not perfectly match the data. Alternatively, since the nJet systematic directly reweights according to the data-MC comparison of the jet multiplicity, the shifted distribution in 7.13b perfectly matches the data in the validation region.

| <b>Model</b> | $\sigma_{JES}[\%]$ | $\sigma_{JER}[\%]$ | $\sigma_{JMS}[\%]$ | $\sigma_{Pileup}[\%]$ | $\sigma_{TrkVtx}[\%]$ | $\sigma_{nJet}[\%]$ | $\sigma_{syst}^{Total}[\%]$ |
|--------------|--------------------|--------------------|--------------------|-----------------------|-----------------------|---------------------|-----------------------------|
| 312004       | 3.31               | 3.34               | 3.18               | 4.70                  | 28.32                 | 4.23                | 29.57                       |
| 312008       | 3.52               | 3.26               | 2.46               | 9.28                  | 30.64                 | 18.04               | 37.14                       |
| 312017       | 7.08               | 5.79               | 5.57               | 26.51                 | 14.44                 | 0.19                | 32.04                       |
| 312022       | 0.93               | 0.20               | 0.15               | 4.98                  | 35.82                 | 8.30                | 37.12                       |
| 312028       | 1.28               | 1.10               | 1.36               | 6.09                  | 39.24                 | 5.94                | 40.21                       |
| 312031       | 12.15              | 9.15               | 8.90               | 23.28                 | 14.67                 | 5.39                | 33.11                       |
| 312039       | 2.83               | 2.34               | 2.72               | 6.52                  | 31.02                 | 11.00               | 33.86                       |
| 312046       | 2.38               | 1.95               | 1.83               | 3.08                  | 29.12                 | 2.11                | 29.58                       |
| 312052       | 3.65               | 1.93               | 2.48               | 11.18                 | 25.94                 | 0.19                | 28.66                       |
| 312060       | 2.13               | 1.99               | 2.12               | 4.82                  | 21.87                 | 11.56               | 25.46                       |
| 312066       | 1.46               | 0.27               | 0.22               | 12.25                 | 9.58                  | 10.02               | 18.56                       |
| 312067       | 11.65              | 9.72               | 10.75              | 26.76                 | 10.89                 | 7.45                | 35.15                       |
| 312075       | 5.88               | 5.05               | 4.33               | 12.08                 | 28.26                 | 12.58               | 34.38                       |
| 312080       | 6.22               | 4.23               | 5.36               | 22.50                 | 9.38                  | 2.99                | 26.25                       |
| 312090       | 9.30               | 7.84               | 6.64               | 11.47                 | 17.15                 | 11.71               | 27.47                       |

**Table 7.2:** Systematic uncertainty summary table. The percent uncertainty is given for each source which contributes to the signal efficiency, as well as the total combined systematic uncertainty.

Table 7.2 gives a full summary of the signal systematics, listing each component individually, as well as giving the total combined systematic uncertainty for each signal point in the canonical set. It should be noted here that the total systematic uncertainty on the signal efficiency is determined by

adding together the individual systematics in quadrature, which ignores any potential correlations between the different sources of systematic uncertainties.

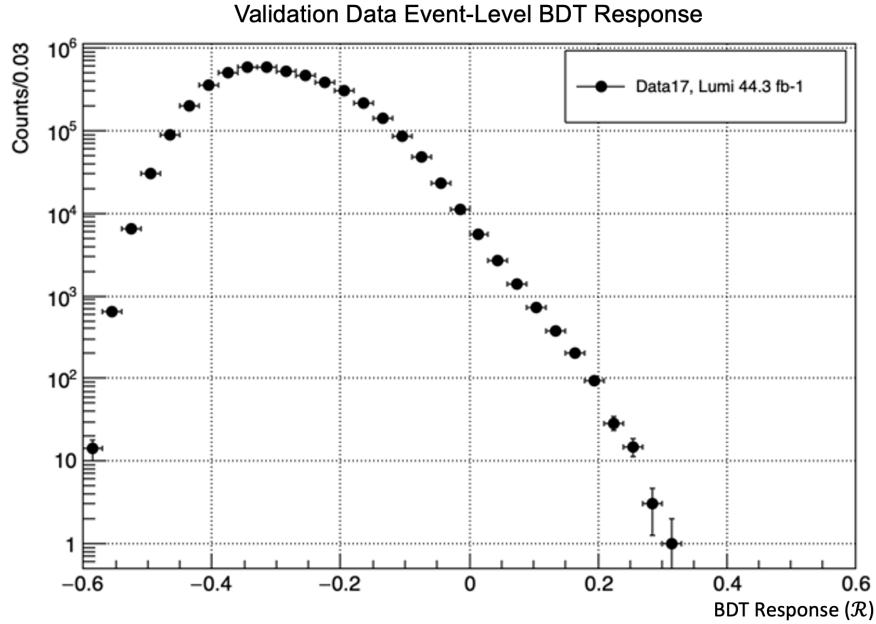


**Figure 7.14:** Evaluation of the effect of the full set of systematics on the event-level  $\mathcal{R}'$ , in the 4-Jet validation region. The green bands give the total combined uncertainty,  $\sigma_{MC} = \sqrt{\sigma_{stat,MC}^2 + \sigma_{syst}^2}$ , where the dominant effect is from the systematic uncertainty. The comparison in the bottom panel gives the significance of the difference between data and MC. The distribution have been normalized to unit area for comparison.

Figure 7.14 gives the evaluation of the full set of systematic on the event-level  $\mathcal{R}'$  distribution. In this case the green bands give the total combined uncertainty,  $\sigma_{MC} = \sqrt{\sigma_{stat,MC}^2 + \sigma_{syst}^2}$ , where  $syst = \{JES, JER, JMS, Pileup, TrkVtx, nJet\}$ . It is clear that the systematics sufficiently cover the differences between Data and MC in the peak of the distribution. However, the high side tail still shows significant differences; this is somewhat expected since the individual systematics did not cover the mismodelling in this region. Still, there is some improvement when the full set is examined together.

Figure 7.15 gives the corresponding data distribution, which is sharply falling above the nominal cut at  $\mathcal{R} = 0.2$ . Therefore, even though the residual difference is large, the numbers of events shifted is fairly small. However, this difference is still concerning as it seems to suggest that data from

what should be a background-only (validation) region will be classified as more signal-like than the background MC would suggest. It is possible that for the ATLAS analysis, a more detailed evaluation of some systematics, or the inclusion of others, may be necessary to try and cover these differences. There may also be concern regarding over-estimation of some systematics, particularly in the peak. In this case, the alternate evaluation of the TrkVtx systematic may help to alleviate this. Of course, it is important to remember that the background estimate will be data-driven, so mismodelling of the background MC is not necessarily the dominant concern, as it only effects the signal efficiencies.



**Figure 7.15:** Data distribution for  $\mathcal{R}$  in the 4-Jet validation region. The points correspond to the 44.3  $\text{fb}^{-1}$  of Data17. The figure has been set in logarithmic scale to give a sense of the statistics in the tails.

### 7.1.3 Systematic Uncertainty on the Background Estimate

The systematics above are all related to the signal efficiency, since background modelling does not affect a data-driven estimate. Therefore, the only systematic that is applied to the background is related to the estimate method, specifically to *non-closure* of the ABCD estimate due to biases. These biases may be caused by correlations in the plane, or some non-uniformity in the response.

The correction method attempts to address flaws in the underlying assumption of the ABCD

method (given by eqn. 6.1), and as such can be used to correct either the simple arithmetic estimate, or the fit estimate. The major benefit of the fit is that it is performed simultaneously on both the background and signal, and thus is able to correct for signal leakage in the control regions, however, the resulting estimate is still based on the assumption that  $\frac{N_A}{N_B} = \frac{N_C}{N_D}$  and is therefore susceptible to the same biases that affect the simple arithmetic estimate.

| Model  | $N_A \pm \sigma_A^{stat}$ | $N_{A,est}^{Sig+Bkg} \pm \sigma_{A,est}^{stat}$ | $N_{A,est}^{Bkg} \pm \sigma_{A,est}^{stat}$ | $N_{A,est}^{Fit} \pm \sigma_{A,est}^{stat}$ |
|--------|---------------------------|-------------------------------------------------|---------------------------------------------|---------------------------------------------|
| 312004 | 46.0 $\pm$ 26.5           | 102.6 $\pm$ 3.4                                 | 102.4 $\pm$ 3.4                             | 102.4 $\pm$ 3.3                             |
| 312008 | 180.0 $\pm$ 57.8          | 434.3 $\pm$ 7.1                                 | 431.5 $\pm$ 7.0                             | 431.5 $\pm$ 6.7                             |
| 312017 | 642.6 $\pm$ 159.9         | 1261.5 $\pm$ 18.3                               | 1188.1 $\pm$ 16.5                           | 1188.1 $\pm$ 18.1                           |
| 312022 | 45.2 $\pm$ 10.1           | 137.0 $\pm$ 1.9                                 | 136.8 $\pm$ 1.8                             | 136.8 $\pm$ 1.8                             |
| 312028 | 243.8 $\pm$ 71.5          | 506.9 $\pm$ 7.2                                 | 500.9 $\pm$ 7.1                             | 500.9 $\pm$ 6.7                             |
| 312031 | 1099.4 $\pm$ 291.3        | 972.0 $\pm$ 17.7                                | 957.5 $\pm$ 14.7                            | 957.5 $\pm$ 13.2                            |
| 312039 | 42.7 $\pm$ 26.3           | 73.5 $\pm$ 1.7                                  | 73.3 $\pm$ 1.7                              | 73.3 $\pm$ 1.7                              |
| 312046 | 140.8 $\pm$ 61.6          | 202.4 $\pm$ 4.8                                 | 196.7 $\pm$ 4.6                             | 196.7 $\pm$ 4.7                             |
| 312052 | 469.2 $\pm$ 113.1         | 1043.7 $\pm$ 16.4                               | 973.2 $\pm$ 14.4                            | 973.2 $\pm$ 14.6                            |
| 312060 | 37.5 $\pm$ 17.6           | 67.9 $\pm$ 1.7                                  | 67.1 $\pm$ 1.7                              | 67.1 $\pm$ 1.3                              |
| 312066 | 286.1 $\pm$ 126.5         | 369.4 $\pm$ 6.4                                 | 360.5 $\pm$ 6.2                             | 360.6 $\pm$ 3.1                             |
| 312067 | 1092.4 $\pm$ 680.7        | 789.7 $\pm$ 13.5                                | 765.2 $\pm$ 11.4                            | 765.2 $\pm$ 11.7                            |
| 312075 | 314.6 $\pm$ 66.8          | 4527.3 $\pm$ 5.8                                | 527.0 $\pm$ 5.7                             | 527.1 $\pm$ 3.9                             |
| 312080 | 637.3 $\pm$ 133.9         | 913.0 $\pm$ 13.0                                | 910.0 $\pm$ 12.7                            | 910.0 $\pm$ 5.4                             |
| 312090 | 1363.0 $\pm$ 511.9        | 1891.0 $\pm$ 28.6                               | 1818.0 $\pm$ 27.3                           | 1818.0 $\pm$ 8.2                            |

**Table 7.3:** Comparison of the different ABCD estimate methods for the yield in the signal region (A). The actual counts ( $N_A$ ) are the background counts that populate region A for each of the model dependent ABCD planes. The background only arithmetic, signal+background arithmetic, and fit estimates are all given for comparison. All values are from MC scaled to the full Run II luminosity of 139 fb<sup>-1</sup>.

Table 7.3 gives the comparison between the arithmetic and fit estimates, which gives two important facts: the fit properly subtracts the signal contribution from the background estimate, and there is a discrepancy (non-closure) between the estimated counts and the actual counts.

## Correcting The ABCD Estimate

In order to address this non-closure, a correction to the ABCD estimate can be made. This correction is built from the validation ABCD planes given in figure 6.38. The correction factor is given by:

$$\xi_{vr} = \frac{N_{A'}^{vr}}{N_{A',est}^{vr}} = \frac{N_{A'}^{vr} N_{D'}^{vr}}{N_{B'}^{vr} N_{C'}^{vr}} \quad (7.7)$$

where the vr tag could be either the validation region MC or data, and the primed counts refer to the regions after the cut on the BDT response being moved back to  $\mathcal{R} = 0$ . This correction factor would fully correct the corresponding estimate for the given VR plane.

| <b>Model</b> | $T_{est}$ | $T_{corr}^{MC}$ | $T_{corr}^{Data}$ |
|--------------|-----------|-----------------|-------------------|
| 312004       | 1.98      | 1.48            | 0.31              |
| 312008       | 4.06      | 1.55            | 1.43              |
| 312017       | 3.25      | 0.84            | 1.10              |
| 312022       | 7.88      | 0.92            | 1.82              |
| 312028       | 3.43      | 1.03            | 0.08              |
| 312031       | 0.48      | 1.07            | 1.23              |
| 312039       | 1.14      | 0.55            | 0.02              |
| 312046       | 0.88      | 0.08            | 0.18              |
| 312052       | 4.14      | 2.29            | 2.15              |
| 312060       | 1.61      | 1.21            | 0.16              |
| 312066       | 0.58      | 0.46            | 0.24              |
| 312067       | 0.48      | 0.53            | 0.60              |
| 312075       | 3.07      | 0.71            | 0.79              |
| 312080       | 1.95      | 0.86            | 0.17              |
| 312090       | 0.88      | 0.24            | 0.21              |

**Table 7.4:** Comparison of the T-test results for the uncorrected estimate ( $T_{est}$ ), the corrected estimate using validation region MC ( $T_{corr}^{MC}$ ), and the corrected estimate using validation region data ( $T_{corr}^{Data}$ ). The data represents the full set of Data17, and the MC has been scaled to the corresponding luminosity of  $44.3 \text{ fb}^{-1}$ .

$\xi_{vr}$  is then applied to the search region estimate ( $N_{A,est}$ ) to give the corrected estimate:

$$N_{A,corr} = \xi_{vr} N_{A,est}. \quad (7.8)$$

To evaluate the efficacy of the correction, a T-test is used to determine if the estimate (or corrected estimate) is consistent with the actual counts, within uncertainties.

$$T_{est/corr} = \frac{|N_A - N_{A,est/corr}|}{\sqrt{(\sigma_A^{stat})^2 + (\sigma_{A,est/corr}^{stat})^2}} \quad (7.9)$$

such that if  $T_{corr} < T_{est}$ , then the correction brings the estimate closer to the actual value, thus reducing the non-closure.

Table 7.4 gives the results of these T-tests and shows that, in general, the validation region data does a better job of correcting the search region estimate. Thus,  $\xi_{Data}$  can be used, giving a data-driven correction to what will eventually be a data-driven background estimate.

### Systematic Uncertainty on the Corrected Estimate

The correction study must be carried out using only MC in the search region because the analysis is currently blinded. Since there is an obvious (and expected) difference between MC in the search region and the validation region, an additional systematic is applied.

The uncertainty, given by:

$$\sigma_{syst}(\xi) = \frac{|\xi_{MC} - \xi_{Data}|}{\xi_{MC}} \quad (7.10)$$

corresponds to the difference between the corrected estimates from the validation region data and MC.  $\sigma_{syst}(\xi)$  attempts to characterize the difference between data and MC in the validation region, as well as the difference between MC in the search and validation region.

Table 7.5 gives a summary of the correction using this method, as well as the resulting systematic uncertainty on the background estimate.

| Model  | $\xi_{Data} \pm \sigma_\xi$ | $N_{A,corr} \pm \sigma_{stat} \pm \sigma_{syst}$ |
|--------|-----------------------------|--------------------------------------------------|
| 312004 | $0.531 \pm 0.022$           | $54.3 \pm 6.1 \pm 33.1$                          |
| 312008 | $0.615 \pm 0.007$           | $265.2 \pm 13.9 \pm 22.3$                        |
| 312017 | $0.692 \pm 0.004$           | $822.1 \pm 36.2 \pm 37.2$                        |
| 312022 | $0.475 \pm 0.023$           | $65.0 \pm 4.2 \pm 20.3$                          |
| 312028 | $0.475 \pm 0.011$           | $237.7 \pm 12.0 \pm 87.6$                        |
| 312031 | $0.771 \pm 0.004$           | $738.3 \pm 35.9 \pm 41.2$                        |
| 312039 | $0.575 \pm 0.020$           | $42.1 \pm 3.3 \pm 14.5$                          |
| 312046 | $0.659 \pm 0.012$           | $129.6 \pm 9.9 \pm 14.9$                         |
| 312052 | $0.743 \pm 0.005$           | $723.0 \pm 33.9 \pm 46.2$                        |
| 312060 | $0.601 \pm 0.027$           | $40.3 \pm 3.7 \pm 24.3$                          |
| 312066 | $0.709 \pm 0.010$           | $255.5 \pm 14.7 \pm 69.0$                        |
| 312067 | $0.892 \pm 0.005$           | $682.2 \pm 32.1 \pm 44.3$                        |
| 312075 | $0.495 \pm 0.011$           | $261.0 \pm 10.6 \pm 15.3$                        |
| 312080 | $0.675 \pm 0.006$           | $614.7 \pm 27.5 \pm 127.2$                       |
| 312090 | $0.690 \pm 0.004$           | $1254.2 \pm 59.4 \pm 15.6$                       |

**Table 7.5:** Summary table of the ABCD estimate correction factor, the resulting corrected estimate and its associated systematic uncertainty. The MC has been scaled to the full Run II luminosity of  $139 \text{ fb}^{-1}$ .

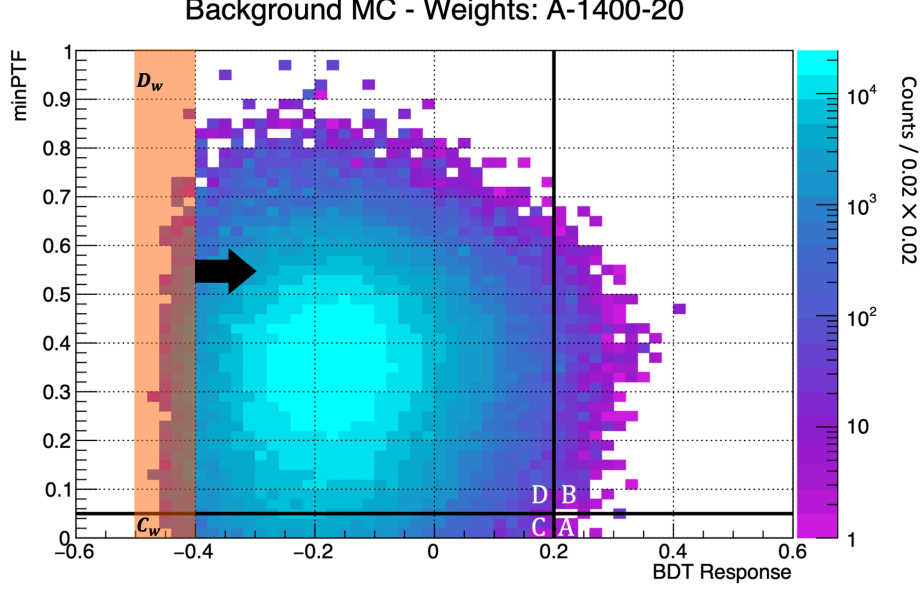
### Control Region Transfer Window

An alternative method to try and improve the ABCD estimate is by changing the control regions. This involves dividing the x-axis range below the cut of  $\mathcal{R} = 0.2$  into windows of fixed width (0.1 in  $\mathcal{R}$ , for a total of 7 windows). Since there is no overlap between the windows, each one is statistically independent.

The y-axis cut then divides each window into two regions, labeled  $C_w$  for events in the window below  $minPTF = 0.05$ , and  $D_w$  for events with  $minPTF > 0.05$ . The transfer factor ( $\frac{N_C}{N_D}$ ) can then be calculated for each window, and these can each be used to give alternative ABCD estimates:

$$N_{A,est}^w = N_B \frac{N_C^w}{N_D^w} \quad (7.11)$$

These window estimates can be compared with both the standard estimate, and the actual counts in region A, to determine whether or not the window estimate performs better than the standard, and if it more accurately represents the actual counts.

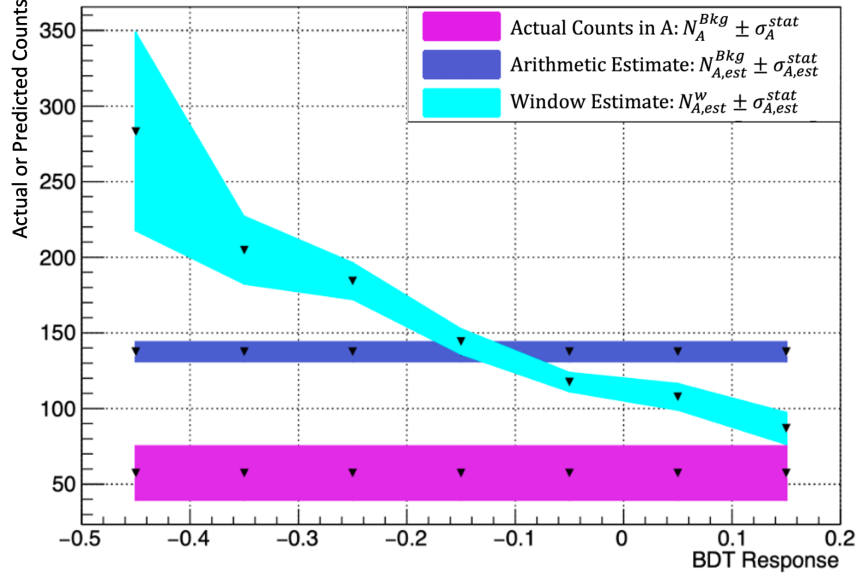


**Figure 7.16:** Background ABCD plane showing the control region transfer window method. The cuts of  $\mathcal{R} = 0.2$  and  $minPTF = 0.05$  (shown in black) define the four regions, which follow the alternate ABCD arrangement. The transfer window is represented by the shaded orange region, with the arrow indicating that the rest of the windows are to the right. The distributions have been scaled to Data17 which corresponds to  $44.3 \text{ fb}^{-1}$ .

As the window moves to the right and approaches the x-axis cut, the window estimate starts to improve and eventually perform better than the original estimate. Figure 7.17 shows how the window estimate evolves as the window moves from left to right in  $\mathcal{R}$ . In most cases, the window whose right edge is along the x-axis cut gives the best estimate, and in some cases it is even consistent (within uncertainties) with the actual counts in region A.

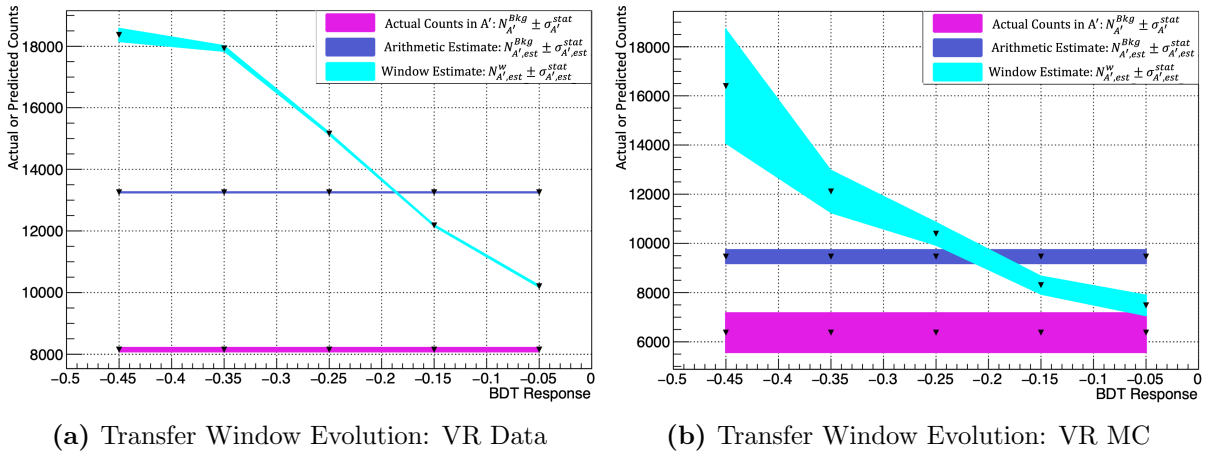
This method can be validated by examining the evolution of the window estimate for the corresponding validation ABCD planes, both in data and background MC. Since the x-axis cut is moved to  $\mathcal{R} = 0$ , there are only 5 transfer windows in the validation region. In general, this study suffers from MC statistics, especially in the validation region. For the data we see that, although





**Figure 7.17:** Transfer window estimate evolution. Each bin represents one of the 7 windows in the control region. The pink band represents the actual counts in region A ( $N_A$ ), the dark blue band represents the standard arithmetic estimate, and the cyan band represents the transfer window estimate. The width of each band is the corresponding statistical uncertainty.

the window closest to the x-axis cut does still improve the estimate, there is still a significant non-closure between the actual counts and the estimate. A similar effect may be seen once the full set of MC statistics is included, thus reducing the efficacy of this method.



**Figure 7.18:** Transfer window estimate evolution for the validation region (a) data and (b) MC. Each bin represents one of the 5 windows in the control region. The pink band represents the actual counts in region A' ( $N_{A'}$ ), the dark blue band represents the standard arithmetic estimate, and the cyan band represents the transfer window estimate. The width of each band is the corresponding statistical uncertainty.

Statistics may also be an issue if the control region is sufficiently small, and close to the x-axis cut. This region is expected to be most similar to the A and B regions, hence the better estimate, but it also has the smallest background contribution and may therefore suffer from statistical uncertainties.

Furthermore, the analysis will have a two-stage unblinding of the search region data: the first where the control regions (B, C, and D) are unblinded so that the data-driven estimate can be performed, and the second where the signal region (A) is unblinded so that it can be compared to the estimate. This means that further validation can be performed using the control region data after the first stage of unblinding.

## 7.2 Sensitivity Estimation

### 7.2.1 Likelihood Fit

After the ABCD plane has been defined, the distributions are passed into the likelihood fit to extract the estimated yield ( $N_{A,est}^{Fit}$ ). Systematic uncertainties can be added to eqn. 6.10 as Gaussian nuisance parameters:

$$L(\mathbf{data}|\tilde{m}, \tilde{N}_B, \tilde{N}_D) \times \prod_{i=1}^3 \mathcal{N}(\chi_i, \alpha_{\chi_i}, \sigma_{\chi_i}) \quad (7.12)$$

where  $\mathcal{N}$  is the Gaussian distribution,  $\alpha_{\chi}$  and  $\sigma_{\chi}$  are the mean and standard deviation of the distribution, respectively, and the  $\chi_i$  correspond to the different nuisance parameters.

$$\mathcal{N}(\chi, \alpha_{\chi}, \sigma_{\chi}) = \frac{1}{\sigma_{\chi}\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\chi-\alpha_{\chi}}{\sigma_{\chi}}\right)^2} \quad (7.13)$$

In the current implementation, there are three nuisance parameters:  $\chi_1 = \mathcal{L}$  which represents the luminosity uncertainty,  $\chi_2 = \mathcal{S}$  which contains all of the systematic uncertainties, and  $\chi_3 = \mathcal{B}$  which represents the uncertainty on the background estimate.

The different sources of  $\sigma_{syst}$  are combined in quadrature to give the width of the systematics nuisance parameter:

$$\sigma_{\mathcal{S}} = \sqrt{\sum_{syst} (\sigma_{syst})^2} \quad (7.14)$$

where  $syst = \{JES, JER, JMS, Pileup, TrkVtx, nJet\}$ . Alternatively, the luminosity nuisance parameter has a single uncertainty given by the luminosity uncertainty, so  $\sigma_{\mathcal{L}} = 1.7\%$  which is the same for all of the signal models. In both cases, the means of the nuisance parameters are set by the likelihood fit.

The prescription for  $\mathcal{B}$  is slightly different; since  $\alpha_{\mathcal{B}}$  is multiplied by the uncorrected background estimate in the fit, the correction factor can be applied as the mean:  $\alpha_{\mathcal{B}} = \xi_{Data}$ . The corresponding systematic uncertainty given by this correction is applied to the width of  $\mathcal{B}$ , so  $\sigma_{\mathcal{B}} = \sigma_{syst}(\xi)$ .

### 7.2.2 Confidence Limits

After extracting the signal estimate from the fit, the signal efficiency ( $\mu$ ) can be calculated:

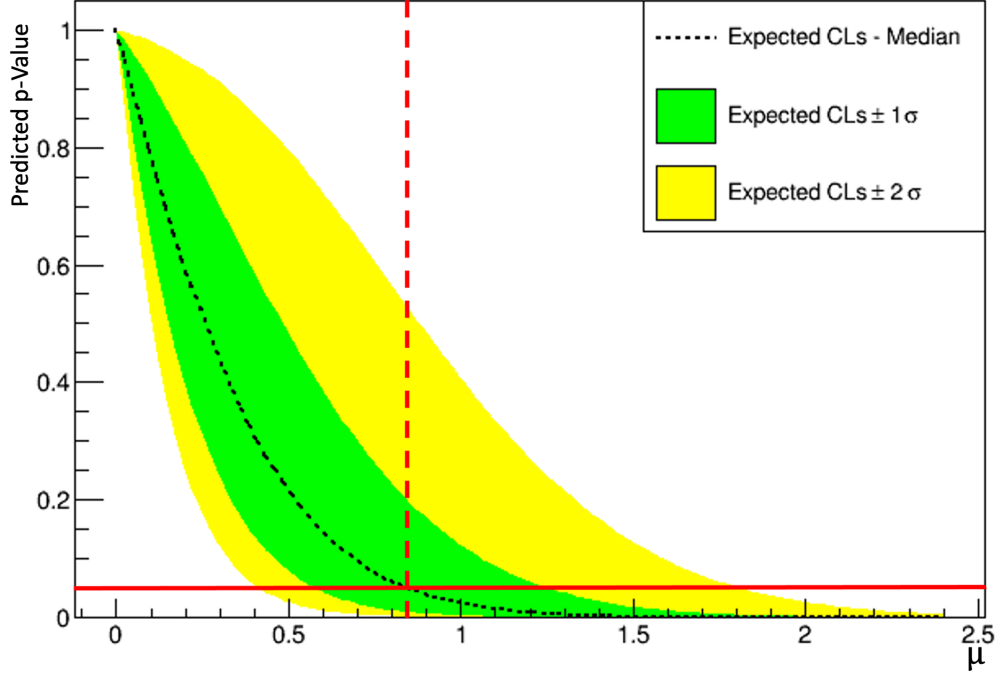
$$\mu = \frac{N_{A,fit}^{Sig}}{N_{A,MC}^{Sig}} \quad (7.15)$$

where  $N_{A,fit}^{Sig}$  is the estimated signal counts from the fit, and  $N_{A,MC}^{Sig}$  is the corresponding counts from the signal MC. This is then multiplied by the theoretical cross-section to give the expected upper limit:

$$\sigma_{UL} = \mu \times \sigma_{theoretical}. \quad (7.16)$$

The statistical machinery which handles the CLs will scan through a range of  $\mu$  values, and calculate a p-value for each one. The machinery then determines which  $\mu$  gives a p-value of 0.05, corresponding to a 95% confidence limit. If  $\mu \leq 1$  then the upper limit on the estimated cross-section will be lower than the theoretical.

The uncertainties on  $\mu$  are given by a combination of the propagated statistical uncertainties, and the applied systematic uncertainties. The resulting uncertainty on the CL is determined by where the uncertainty bands in figure 7.19 cross the p-values line at 0.05.



**Figure 7.19:** Asymptotic confidence limit scan of the signal efficiency and the corresponding p-value. The dashed black curve represents the central value of  $\mu$ , while the green and yellow bands are the  $1\sigma$  and  $2\sigma$  uncertainties, respectively. The horizontal red line corresponds to the p-value of 0.05, and the dashed, vertical red line is the resulting  $\mu$  value. Figure from B. Death.

### 7.3 Results

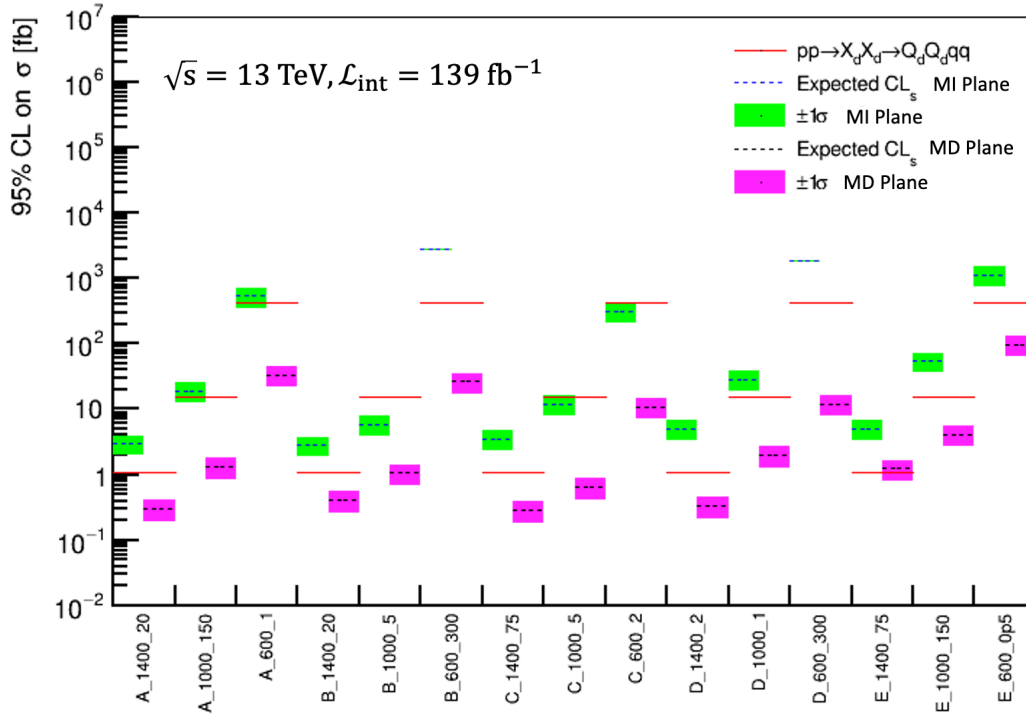
Results for this thesis come in the form of expected cross-sections for each of the models in the canonical set, given by the CL machinery. In order to demonstrate the benefit and effectiveness of including the BDT into the analysis, three main results will be shown:

1. Comparison of the Model Independent method to the Model Dependent method
2. Comparison of the different BDT training scenarios (individual vs. mediator mass mix)
3. Sensitivity of the Model Dependent method with the full set of systematics included

Since at this stage the analysis is still blinded, all results are for Monte-Carlo estimates using both the signal MC and the QCD background MC, both scaled to the full Run II luminosity of  $139 \text{ fb}^{-1}$ .

### 7.3.1 Comparison of Methods

Figure 7.20 shows the comparison of the model independent method to the Model Dependent method. In the plot, the dashed lines represent the upper limit on the cross-section given by the result of the ABCD fit, while the size of the coloured bands represents the uncertainty on the expected value. The red lines represent the theoretical cross-sections; if the expected cross-section and its uncertainty are below the theoretical, then the analysis is sensitive to that model and is said to *test* the theoretical limit. If a dashed line is present without the uncertainty band, it means that the fit did not converge properly, likely due to large associated uncertainties, so no CL can be set.



**Figure 7.20:** Sensitivity plot showing the expected cross-sections (dashed lines) for each model in the canonical set. A comparison is shown for the Model Independent (MI) method in green, and the Model Dependent (MD) in magenta. Note that systematics have not been included in the uncertainties for this comparison. Figure from B. Death.

Three main conclusions can be drawn from the comparison. The first is that the model independent method, at least in its current state, does not provide sensitivity for the majority of the signal points, since many of the expected cross-sections are above the theoretical. The second is that the

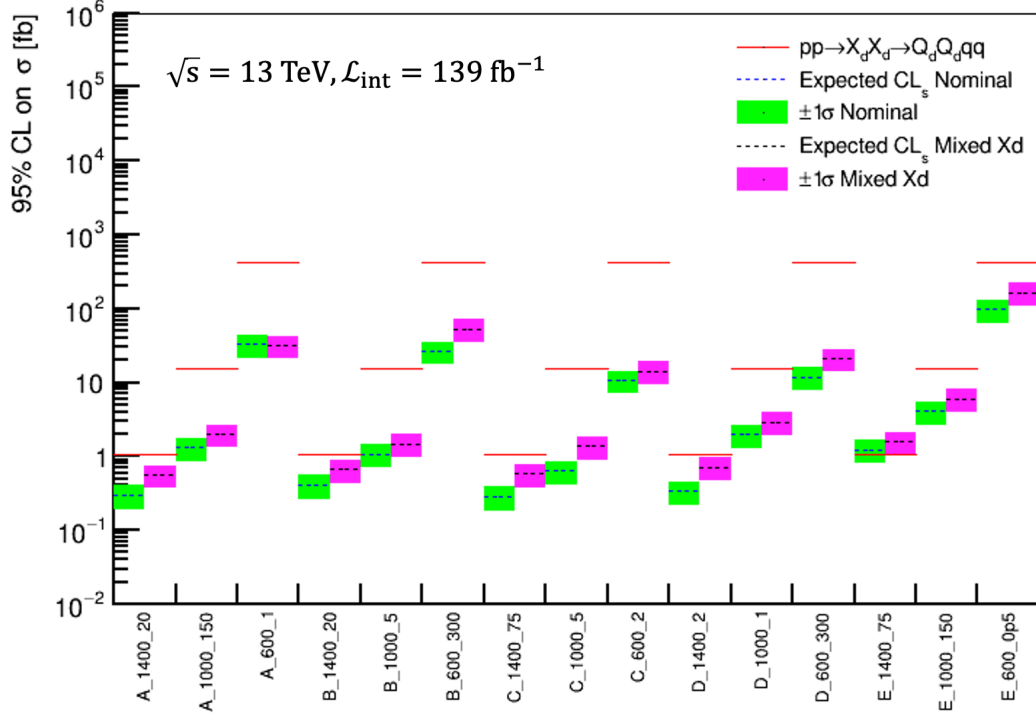
model dependent method performs significantly better for each model in the canonical set; this is to be expected since the BDT makes use of more of the available information, and provides better background suppression. The third conclusion is that, for all but one signal point (312075: Model E,  $m_{X_d} = 1400$  GeV,  $c\tau_{\pi_d} = 75mm$ ), the model dependent method provides sufficient sensitivity for each model to test the theoretical cross-section.

### 7.3.2 Comparison of Training Scenarios

Figure 7.21 shows the comparison of the mediator mass mixed training and the nominal (individual) training for the BDT. The mixed training was introduced in an attempt to reduce the model dependence of the BDT as well as the technical load of training on each signal point individually. The plot shows that there is a general reduction in the sensitivity for almost all signal points, with the exception of 312017 (Model A,  $m_{X_d} = 600$  GeV,  $c\tau_{\pi_d} = 1mm$ ), which shows slight improvement with the mixed training. However, this reduction still leaves most signal points sufficient sensitivity. The models with  $m_{X_d} = 1400$  GeV tend to lose the most sensitivity with the mixed scenario, with 312060 (Model D,  $m_{X_d} = 1400$  GeV,  $c\tau_{\pi_d} = 1mm$ ) likely becoming insensitive after the inclusion of systematics.

It should be noted that the mixed training is done using an additional set of 21 signal points, which gives 12 models per mediator mass. This means that the BDT learns from a variety of  $(m_{\pi_d}, c\tau_{\pi_d})$  combinations, and should not become too dependent on any one model in particular. However, in order to truly reduce the model dependence, the full signal grid of 90 signal points must eventually be used to create the training groups, giving 30 combinations of  $(m_{\pi_d}, c\tau_{\pi_d})$  per  $m_{X_d}$ . The inclusion of the full signal grid may, however, degrade the signal efficiency, and therefore the sensitivity, to a point where it is no longer beneficial to train the BDT in the mixed scenario.

For the sake of this thesis, the analysis will proceed with the nominal training of the BDT. A comparison is not shown for the Canon15 mixed training since this degraded the signal efficiency to a point where it was no longer a viable option.

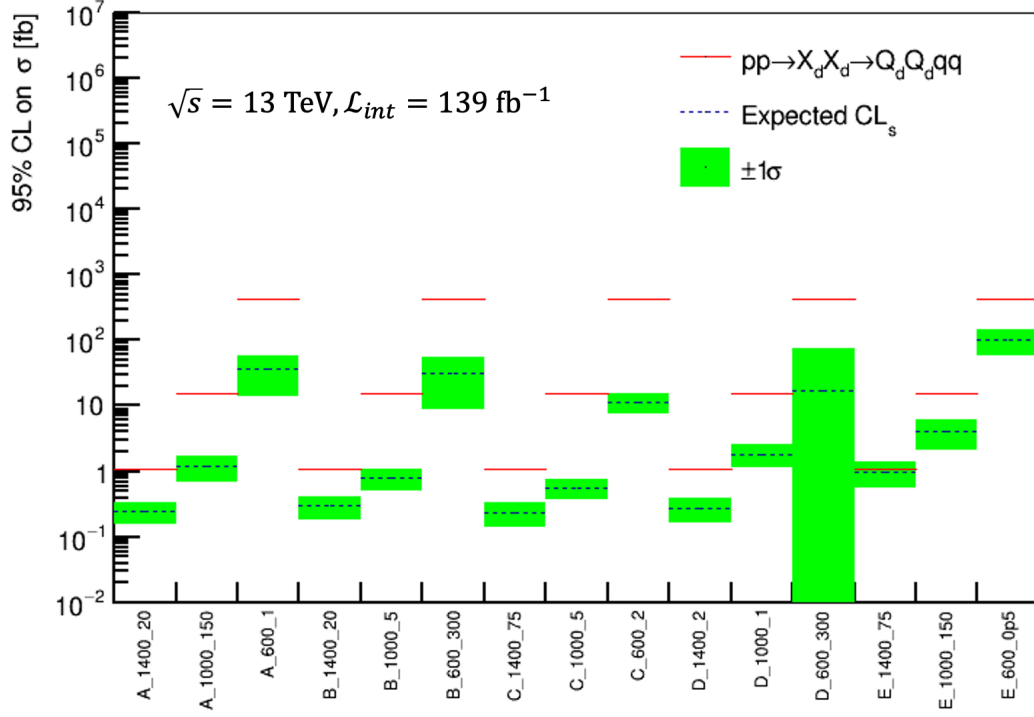


**Figure 7.21:** Sensitivity plot showing the expected cross-sections (dashed lines) for each model in the canonical set. A comparison is shown for the Nominal (individual) BDT training in green, and the Mixed BDT training in groups of mediator mass in magenta. Note that systematics have not been included in the uncertainties for this comparison. Figure from B. Death.

### 7.3.3 Including Systematics

In general, including systematic uncertainties should not change the central value of the expected cross-section, however it does change the size of the uncertainty band, and therefore can reduce the sensitivity of a given model.

Figure 7.22 shows the nominal BDT sensitivity with the full set of systematic uncertainties, including both signal systematics and the background estimate correction ( $\xi_{Data}$ ) with the corresponding systematic ( $\sigma_{syst}(\xi)$ ). With the exception of model 312075 (Model E,  $m_{X_d} = 1400$  GeV,  $c\tau_{\pi_d} = 75mm$ ), which struggled with sensitivity before the inclusion of the systematics, the canonical set remains sensitive, with some signal points excluding the theoretical cross-section by an order of magnitude or more.



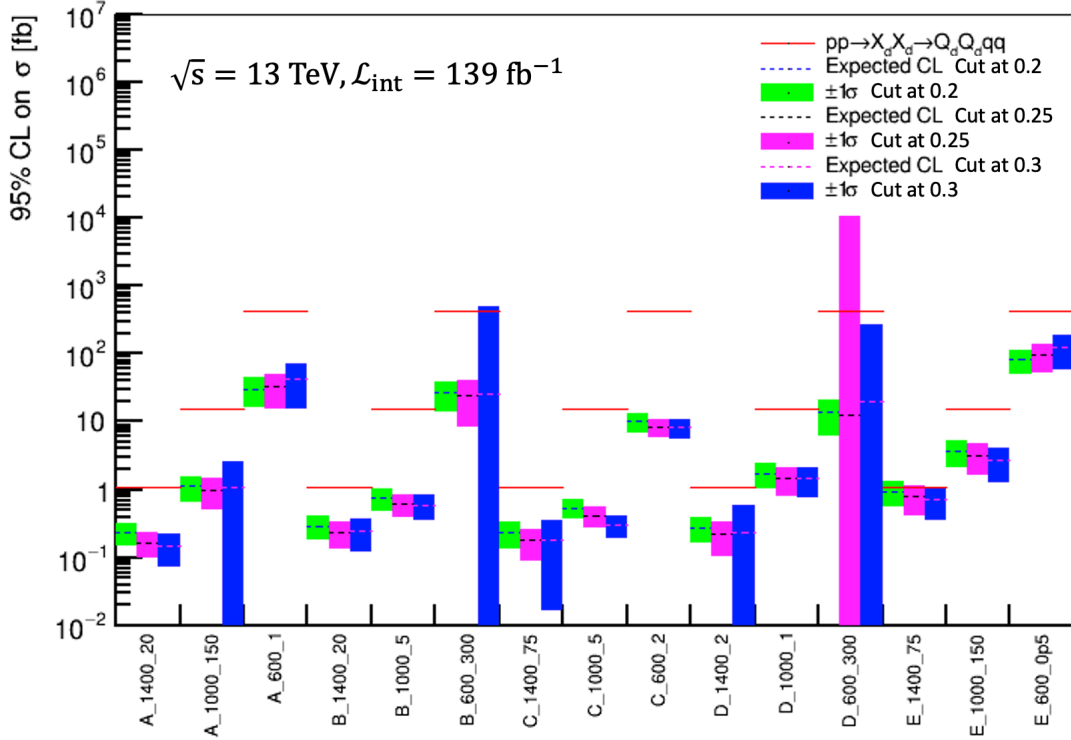
**Figure 7.22:** Sensitivity plot showing the expected cross-sections (dashed lines) for each model in the canonical set. The sensitivity is given for the Nominal BDT training with the full set of systematic uncertainties included. The asymmetric uncertainties are due to the logarithmic scale of the y-axis. Figure from B. Death.

### BDT Cut Optimization

Although the cut at  $\mathcal{R} = 0.2$  is shown throughout the analysis, several different cuts on the BDT response have been tested for each situation. Since the cut changes both the signal and background yield in region A, it can seriously impact the sensitivity of a given model. The cut on  $\mathcal{R}$  must therefore be optimized to give the best sensitivity for the majority of the models in the canonical set.

Figure 7.23 shows the expected sensitivities of the nominal BDT at three cuts on BDT response, of increasing severity. Three general trends are visible. Firstly, as the cut moves farther to the right the total uncertainty increases, and in some cases it becomes large enough to render the model insensitive, even if the central value is similar to the cut at  $\mathcal{R} = 0.2$ . This is because the stricter cut reduces the yield in the signal region, this increases the relative statistical uncertainty, while also increasing the impact of the systematics. The second trend is for the high mediator





**Figure 7.23:** Sensitivity plot showing the expected cross-sections (dashed lines) for each model in the canonical set. The sensitivity is given for the Nominal BDT training with the full set of systematic uncertainties included. The comparison is given for three cut:  $\mathcal{R} = 0.2$  (green),  $\mathcal{R} = 0.25$  (magenta), and  $\mathcal{R} = 0.3$  (blue). Figure from B. Death.

mass models ( $m_{X_d} = 1400$  GeV); these tend to be the better separated models, and therefore the stricter cuts tend to improve the CLs. In these cases, the background distribution falls off faster at larger  $\mathcal{R}$  than the signal, so the background reduction has a larger impact on the sensitivity than the loss of signal events. The third trend shows the opposite effect for the low mediator mass cases ( $m_{X_d} = 600$  GeV), where they perform worse as the cut moves farther to the right in BDT response. This is because the signal and background fall off at more similar rates, so a significant reduction in the background has a corresponding reduction in the signal. The result of this study finds that the optimal cut, for the majority of the models in the canonical set, is at  $\mathcal{R} = 0.2$ .

Table 7.6 gives the final summary for the expected yields in the signal region, for both signal and background, based on the cut at  $\mathcal{R} = 0.2$ . The background counts represent the corrected ABCD estimate, with the corresponding systematic uncertainty.

| <b>Model</b> | $N_A^{sig} \pm \sigma_{stat}^{sig} \pm \sigma_{syst}^{sig}$ | $N_{A,corr}^{bkg} \pm \sigma_{stat}^{bkg} \pm \sigma_{syst}^{bkg}$ |
|--------------|-------------------------------------------------------------|--------------------------------------------------------------------|
| 312004       | $75.3 \pm 0.4 \pm 22.3$                                     | $54.4 \pm 6.1 \pm 33.1$                                            |
| 312008       | $497.7 \pm 3.5 \pm 184.8$                                   | $265.2 \pm 13.9 \pm 23.1$                                          |
| 312017       | $986.5 \pm 12.6 \pm 316.1$                                  | $822.1 \pm 36.2 \pm 37.3$                                          |
| 312022       | $64.4 \pm 0.4 \pm 23.1$                                     | $65.0 \pm 4.2 \pm 20.3$                                            |
| 312028       | $664.2 \pm 4.3 \pm 266.5$                                   | $237.7 \pm 12.0 \pm 87.6$                                          |
| 312031       | $1039.8 \pm 23.7 \pm 344.3$                                 | $738.3 \pm 35.9 \pm 41.2$                                          |
| 312039       | $67.4 \pm 0.4 \pm 22.8$                                     | $42.1 \pm 3.3 \pm 14.5$                                            |
| 312046       | $700.8 \pm 3.5 \pm 228.0$                                   | $129.7 \pm 9.9 \pm 14.9$                                           |
| 312052       | $2616.3 \pm 16.4 \pm 749.8$                                 | $723.0 \pm 33.9 \pm 46.2$                                          |
| 312060       | $57.1 \pm 0.3 \pm 14.5$                                     | $40.4 \pm 3.7 \pm 24.3$                                            |
| 312066       | $308.4 \pm 2.1 \pm 57.2$                                    | $255.5 \pm 14.7 \pm 69.0$                                          |
| 312067       | $2058.3 \pm 13.9 \pm 723.5$                                 | $682.2 \pm 32.1 \pm 44.3$                                          |
| 312075       | $41.0 \pm 0.4 \pm 14.1$                                     | $261.0 \pm 10.6 \pm 15.3$                                          |
| 312080       | $232.1 \pm 3.0 \pm 60.9$                                    | $614.7 \pm 27.5 \pm 127.2$                                         |
| 312090       | $456.1 \pm 13.6 \pm 125.3$                                  | $1254.2 \pm 59.4 \pm 15.6$                                         |

**Table 7.6:** Summary table of the region A yields for both signal and background. The background counts are based on the corrected ABCD estimate. The MC has been scaled to the full Run II luminosity of  $139 \text{ fb}^{-1}$ .

## Chapter 8

# Conclusions

The focus of this thesis has been to demonstrate that, by implementing a machine learning component in the form of a BDT, the Emerging Jets analysis can recover sufficient sensitivity to exclude the theoretical cross-section for 14 of the canonical 15 signal points, and give partial sensitivity to the 15<sup>th</sup>. This was shown using a Monte-Carlo study which represents  $pp$  collisions at  $\sqrt{s} = 13$  TeV, and is scaled to the full ATLAS Run II integrated luminosity of  $139 \text{ fb}^{-1}$ .

A two-step structure to the BDT was developed to first try and classify emerging-like jets, and then try and separate signal-like events from the QCD background, using the information from the first step. The jet-level BDT was trained on calorimeter information such as the jet's mass, energy, width and electromagnetic fraction. The event-level BDT then uses the jet-level BDT response as well as other calorimeter information such as the  $H_T$  of the four leading jets, and combines it with both tracker information (from the displaced vertex multiplicity) and event-shape information to build an event-level discriminator. Although the BDT makes this component of the analysis model dependent, it is designed to be complimentary to a model independent result, obtained using a cut-based approach. Both components will use the ABCD method to obtain a data-driven estimate of the background.

An attempt was made to reduce the model dependence of the BDT. This involved training the BDT with three groups of signal points determined by their mediator mass. While it was determined that this mixed training maintains a similar level of sensitivity (before the inclusion of systematics), the viability of this method cannot be determined until the full grid of 90 signal points can be used for the training.

Since the BDT is more susceptible to the modelling of the input variables, a thorough evaluation of the systematic uncertainties was conducted. This involved the creation of a new validation region with a higher jet multiplicity, so that the modelling of each variable could be studied and the BDT response could be validated. This also prompted the development of two new systematic uncertainties, in addition to the standard Jet uncertainties and pileup reweighting, which attempt to address the mismodelling seen in the jet and displaced vertex multiplicities.

| Model  | Upper Limit ( $\sigma_{UL}$ ) [ fb <sup>-1</sup> ] | Theoretical ( $\sigma_{Theory}$ ) [ fb <sup>-1</sup> ] |
|--------|----------------------------------------------------|--------------------------------------------------------|
| 312004 | $0.24 \pm 0.09$                                    | 1.08                                                   |
| 312008 | $1.19 \pm 0.52$                                    | 15.20                                                  |
| 312017 | $36.40 \pm 22.64$                                  | 430.09                                                 |
| 312022 | $0.30 \pm 0.11$                                    | 1.08                                                   |
| 312028 | $0.79 \pm 0.28$                                    | 15.22                                                  |
| 312031 | $32.03 \pm 23.07$                                  | 430.17                                                 |
| 312039 | $0.24 \pm 0.10$                                    | 1.08                                                   |
| 312046 | $0.56 \pm 0.19$                                    | 15.18                                                  |
| 312052 | $11.20 \pm 3.84$                                   | 429.97                                                 |
| 312060 | $0.28 \pm 0.11$                                    | 1.08                                                   |
| 312066 | $1.84 \pm 0.72$                                    | 15.19                                                  |
| 312067 | $17.40 \pm 56.05$                                  | 430.01                                                 |
| 312075 | $0.98 \pm 0.42$                                    | 1.08                                                   |
| 312080 | $4.12 \pm 2.02$                                    | 15.18                                                  |
| 312090 | $103.35 \pm 44.92$                                 | 430.26                                                 |

**Table 8.1:** Summary table of the current expected cross-section upper limits for the canonical set of 15 signal points. The corresponding theoretical cross-section is given in the last column. The limits represent the Model Dependent yields given by the Nominal BDT training with the cut at  $\mathcal{R} = 0.2$ , and are set using the full Run II luminosity of 139 fb<sup>-1</sup>.

In addition to the signal systematics, a data-driven correction to the ABCD background estimate was also developed to try and reduce the non-closure that occurs from the basic ABCD assumption given by equation 6.1. This correction required the use of a validation region ABCD plane for

the model dependent component of the analysis. Furthermore, a systematic uncertainty on the corrected background estimate was also developed, which characterizes the difference between the data and Monte-Carlo background distributions in this validation ABCD plane.

The work presented in this thesis is now also a major component of an ongoing ATLAS analysis which is looking for Emerging Jets in the full Run II dataset, and aims for publication in early 2022.

Some components of the analysis suffer from limited MC statistics; in particular, the ABCD non-closure studies and correction methods could benefit from additional MC events to try and reduce the statistical uncertainties, which may be skewing the effectiveness of these methods in the current iteration. The other area affected by low statistics is in the training of the BDT; where the presence of overtraining may be degrading the signal efficiency of some models. This becomes a concern since the models which are more difficult to distinguish from the QCB background, also tend to be more statistically limited, and are therefore more prone to overtraining. Presently, only one of the three MC campaigns (mc16d) is being used to conduct the analysis studies. Work is ongoing to include the other two MC campaigns (mc16a and mc16e) and their corresponding datasets (Data15/16 and Data18, respectively), which should help reduce the statistical uncertainties, and the effects of overtraining.

Part of the ongoing work in the larger analysis is the improvement of the model independent method, in order to recover some sensitivity to the EJs models. This has involved experimenting with different ABCD axis variable combinations and preselections to try and separate the signal points from the QCD background. Additionally, the analysis has begun work on including additional signal points, namely the same 21 signal points that were used in the mediator mass mixed BDT training. Increasing the set of signal points makes it more difficult to find a model independent method which is effective for the majority of the set. Furthermore, the analysis is also working on getting the remaining signal points in the full grid processed.

Similarly, in order to keep the BDT component sensitive to the expanded set of signal points improvements to the preselection have been implemented, including tighter selection on displaced vertices to reduce the effect of fakes. This improved BDT has already demonstrated nominal sensitivity to the set of 36 signal points, before the inclusion of systematic uncertainties.

In conjunction with the improvements to the Run II analysis, prospects for Run III, which is set

to begin soon, bring potential improvements which may increase the sensitivity to the Emerging Jets signature. The high mediator mass models, whose cross-sections are relatively small, will benefit from the increase in luminosity that is expected. Run III also entails improvement to the track and displaced vertex reconstruction, which should help improve the sensitivity for the models whose dark-sectors correspond to the limits of ATLAS' current reconstruction capabilities, specifically those with low mediator and dark pion masses, and very long or very short dark pion lifetimes. In particular, the  $(m_{X_d} = 600 \text{ GeV}, c\tau_{\pi_d} = 300mm)$  combination models tend to test the limits of the model space; the dark component tends to live beyond the limits of the pixel detector, making displaced vertex reconstruction difficult, and the low mass visible component tends to be more difficult to distinguish from the QCD background. The expected improvements for Run III may help to recover some additional sensitivity for these edge-cases.

# Appendices

# Appendix A

## Physical Constants

| Symbol      | Name                     | Value [SI Units]                           | Value [Natural Units]                        |
|-------------|--------------------------|--------------------------------------------|----------------------------------------------|
| $e$         | Elementary Charge        | $1.602 \times 10^{-19}$ [C]                | 1 [–]                                        |
| $\hbar$     | Reduced Plancks Constant | $1.054 \times 10^{-34}$ [ $J s$ ]          | 1 [–]                                        |
| $c$         | Speed of Light in Vacuum | $2.998 \times 10^8$ [ $\frac{m}{s}$ ]      | 1 [–]                                        |
| $v$         | Vacuum Expectation Value | $4.38 \times 10^{-25}$ [kg]                | 246 [GeV]                                    |
| $G_F$       | Fermi Coupling Constant  | $4.544 \times 10^{14}$ [ $\frac{1}{J^2}$ ] | $1.166 \times 10^{-5}$ [ $\frac{1}{GeV^2}$ ] |
| $m_{W^\pm}$ | $W^\pm$ Boson Mass       | $1.433 \times 10^{-25}$ [kg]               | 80.379 [GeV]                                 |
| $m_Z$       | $Z$ Boson Mass           | $1.626 \times 10^{-25}$ [kg]               | 91.188 [GeV]                                 |
| $m_H$       | Higgs Boson Mass         | $2.23 \times 10^{-25}$ [kg]                | 125.10 [GeV]                                 |
| $m_e$       | Electron Mass            | $9.109 \times 10^{-31}$ [kg]               | 0.511 [MeV]                                  |
| $m_p$       | Proton Mass              | $1.673 \times 10^{-27}$ [kg]               | 938.272 [MeV]                                |

**Table A.1:** Table of Physical Constants.

Table A.1 gives all of the physical constants (above) and particle masses (below) that are use or referenced in the thesis, predominantly in Chapter 2. The values are given both in standard SI units, and in *natural units*, which takes  $c = \hbar = e = 1$ .



## Appendix B

# PYTHIA Dijet Background Slices

| PYTHIA Dijet Slice  | $\sigma$ [fb]          | $N^{MC}$ | $N^{exp}$              |
|---------------------|------------------------|----------|------------------------|
| JZ2withSW           | $2.433 \times 10^{12}$ | 19982000 | $1.078 \times 10^{14}$ |
| JZ3withSW           | $2.645 \times 10^{10}$ | 19965400 | $1.172 \times 10^{12}$ |
| JZ4withSW           | $2.546 \times 10^8$    | 19972600 | $1.128 \times 10^{10}$ |
| JZ5withSW           | $4.553 \times 10^6$    | 9958500  | $2.017 \times 10^8$    |
| JZ6withSW           | $2.575 \times 10^5$    | 9988500  | $1.141 \times 10^7$    |
| JZ7withSW           | $1.622 \times 10^4$    | 4994000  | $7.184 \times 10^5$    |
| JZ8withSW           | 625.060                | 4995000  | $2.769 \times 10^4$    |
| JZ9withSW           | 19.639                 | 2498000  | 870.145                |
| JZ10withSW          | 1.196                  | 660000   | 53.000                 |
| JZ11withSW          | $4.226 \times 10^{-2}$ | 660000   | 1.872                  |
| JZ12withSW          | $1.038 \times 10^{-3}$ | 660000   | $4.593 \times 10^{-2}$ |
| <b>QCD MultiJet</b> | 0.144                  | 94334000 | $1.090 \times 10^{14}$ |

**Table B.1:** Table of Pythia Dijet Slices. For each slice, the cross-section ( $\sigma$ ) is given, as well as the total number of events generated for the slice ( $N^{MC}$ ). The expected number of events ( $N^{exp}$ ) is given by the slice cross-sections multiplied by the corresponding Data Luminosity.

Table B.1 gives the relevant information for the PYTHIA dijet slices for mc16d (which corresponds to Data17), and make up the full QCD multijet background for the analysis. Note that  $N^{MC}$  is not the number of events in the analysis ntuples, which is just a subset of the total events.

The expected number of events is given by:

$$N_i^{exp} = \mathcal{L}^{17} \times \sigma_i$$

which is calculated using the Data17 luminosity ( $\mathcal{L}^{17} = 44.3 \text{ fb}^{-1}$ ) and the cross-section for each given slice ( $\sigma_i$ ).

The final row gives the information for the full QCD multijet background, which is the combination of slices 2 to 12. Here

$$\sigma^{QCD} = \sum_{i=2}^{12} \sigma_i$$

is the effective cross-section for the full background; similarly

$$N^{QCD} = \sum_{i=2}^{12} N_i^{MC}$$

is the total number of generated events across the full background, while

$$N^{Exp} = \sum_{i=2}^{12} N_i^{exp} = \mathcal{L}^{17} \times \sum_{i=2}^{12} \sigma_i$$

is the expected number of events for the full background corresponding to the Data17 luminosity. This indicates that the number of generated background events for this campaign (mc16d) underestimates the corresponding Data17 luminosity.

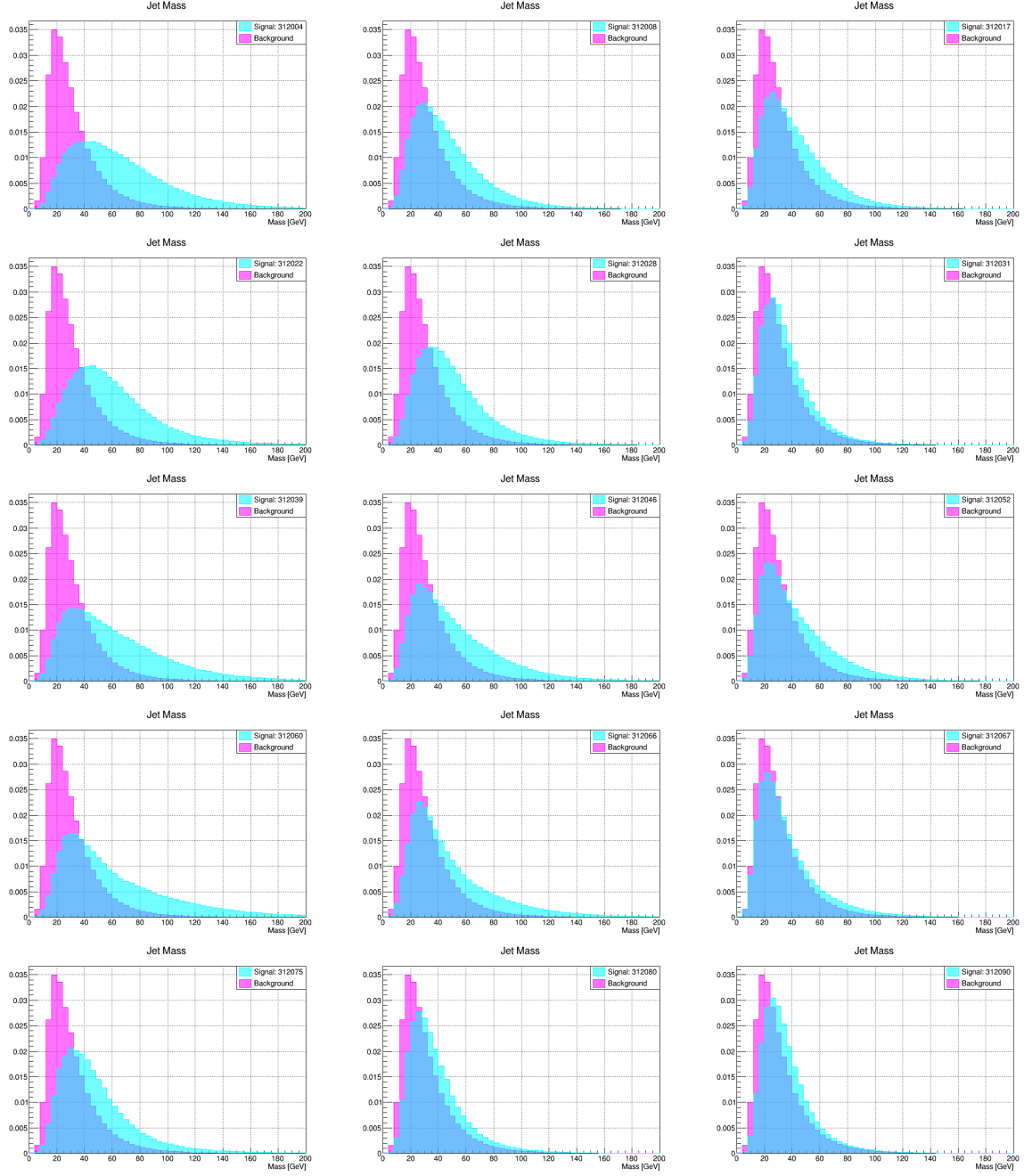
## Appendix C

# Jet-Level BDT Input Variables

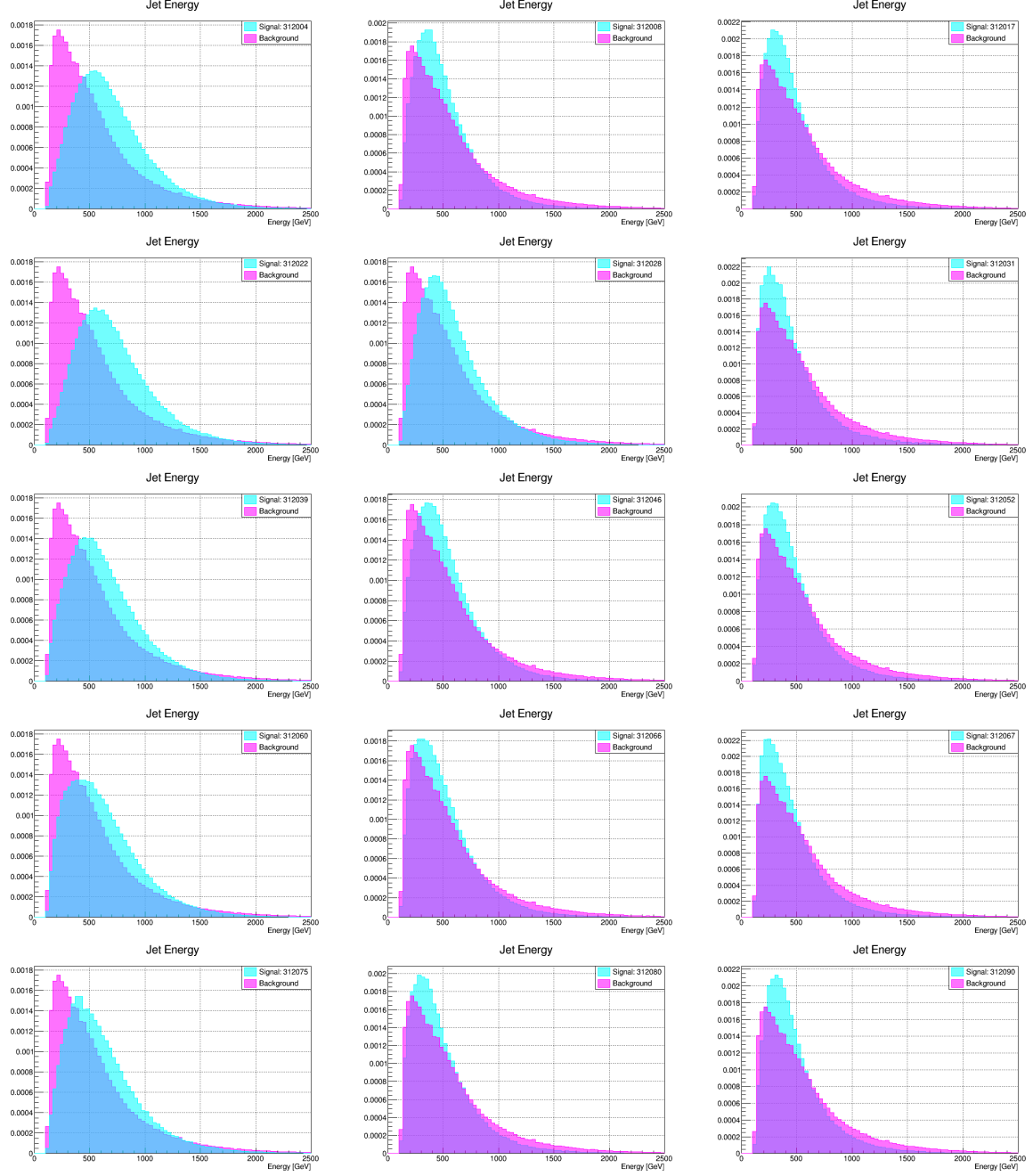
Figures C.1 - C.6 give the input distributions for the jet-level BDT for the 15 signal points of the canonical set. The DSID for each signal point is given in the legend, but the plots are organized as follows:

- The rows are arranged alphabetically by model, with model A at the top and model E at the bottom.
- The columns are arranged by mediator mass, with  $m_{X_d} = 1400$  GeV on the left and  $m_{X_d} = 600$  GeV on the right.

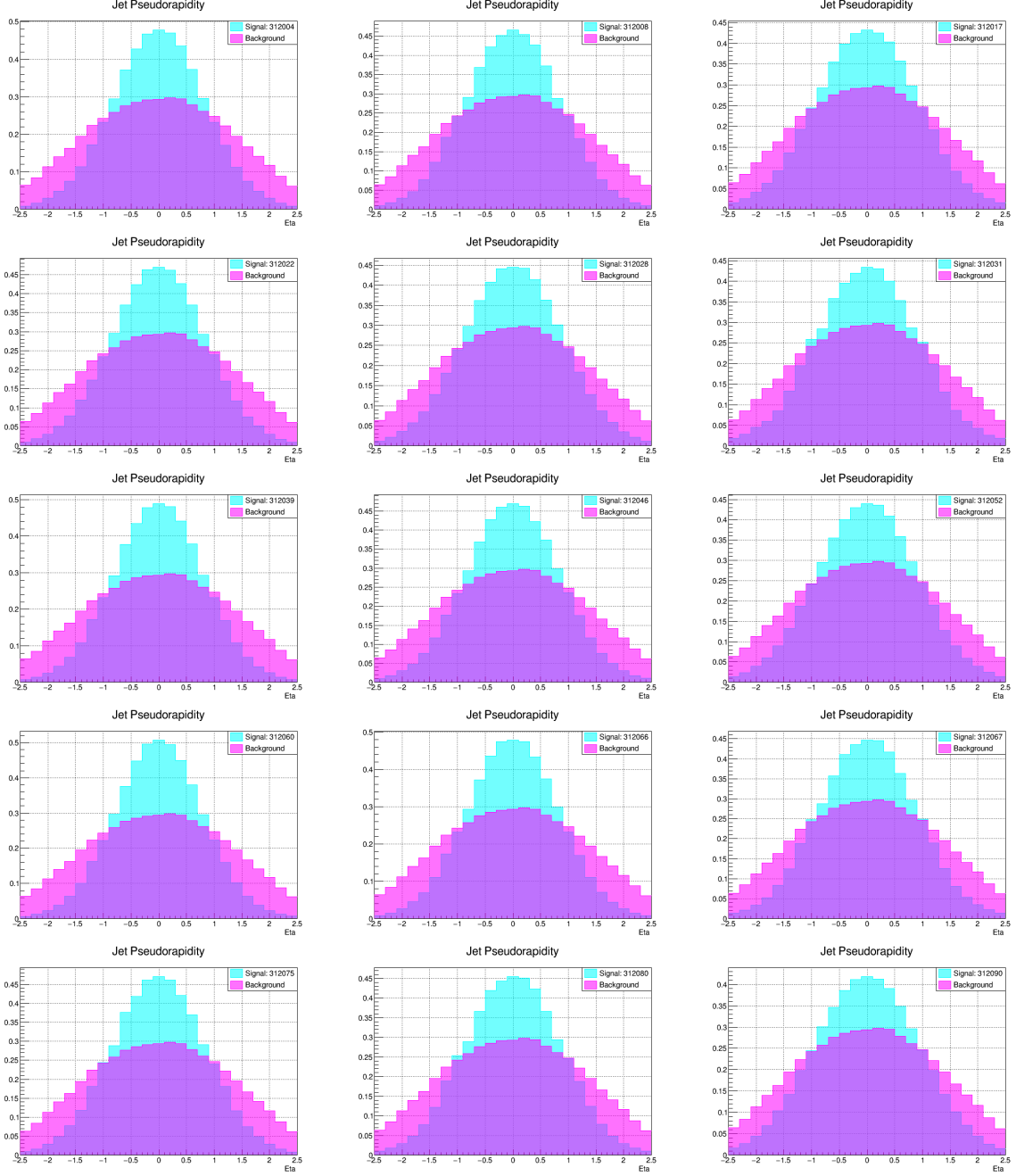
The plots are arranged in the same way for each of the 6 input variables.



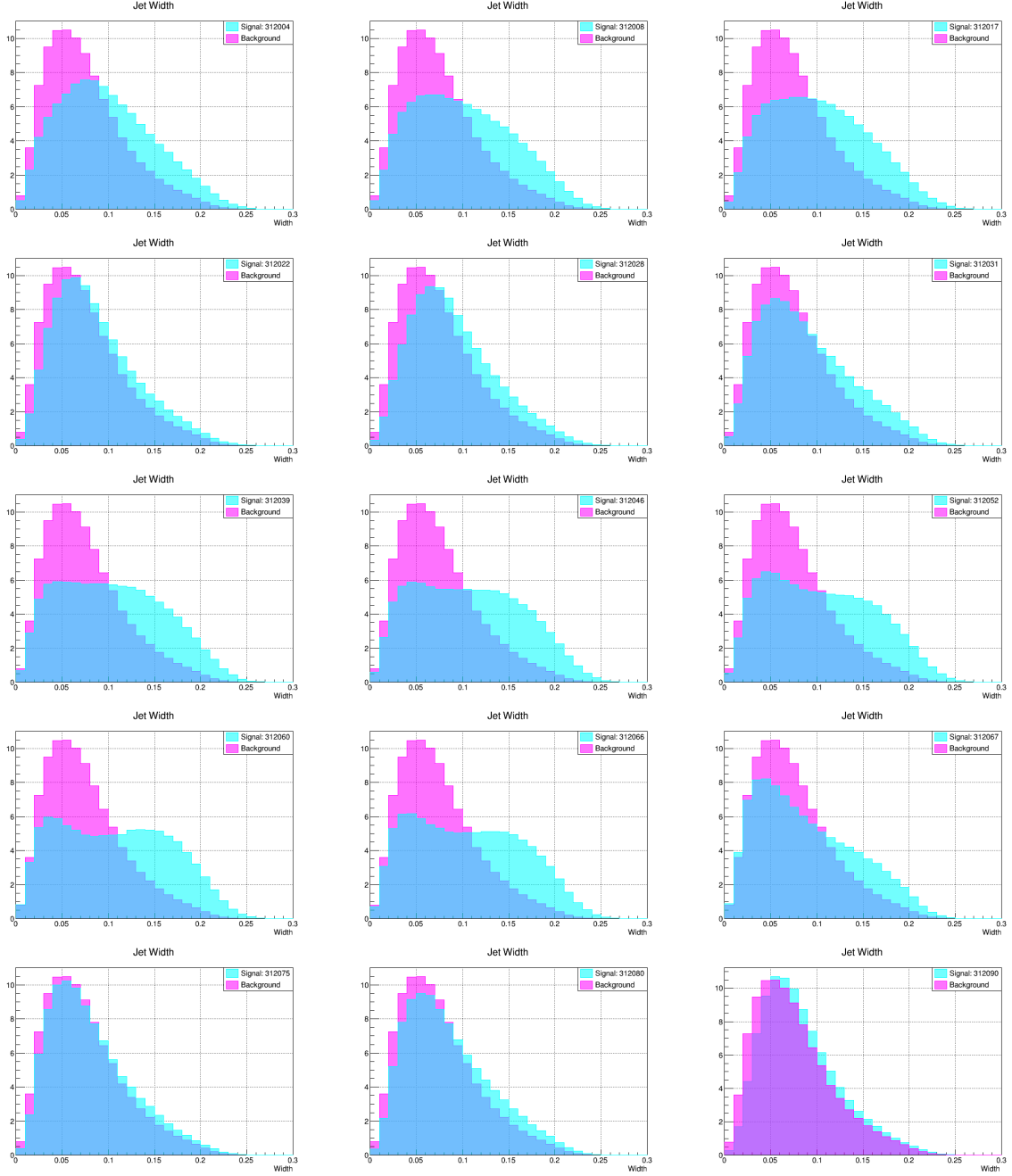
**Figure C.1:** Jet Mass Distributions (Signal and Background) for the Canonical 15 Signal Points.



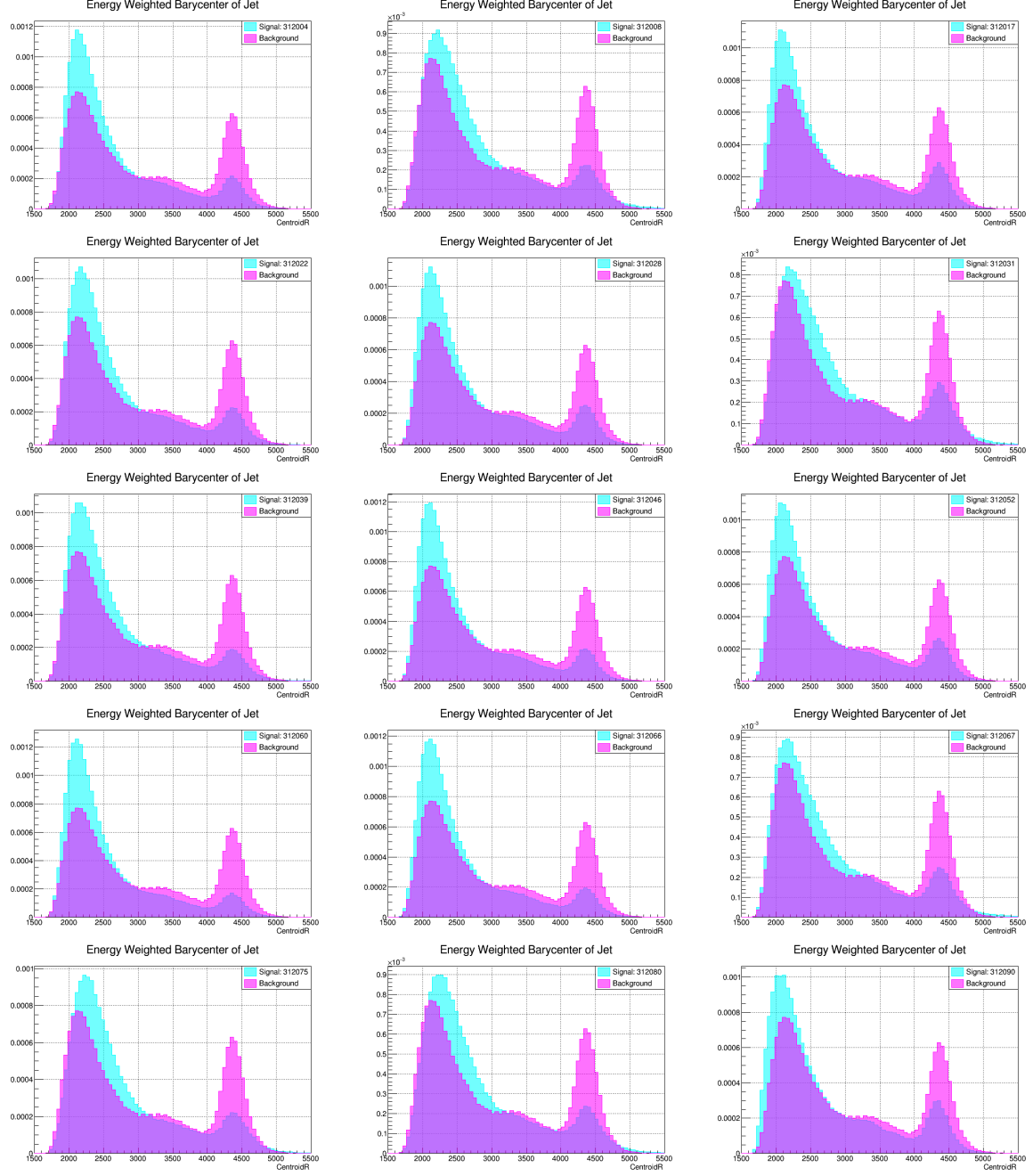
**Figure C.2:** Jet Energy Distributions (Signal and Background) for the Canonical 15 Signal Points.



**Figure C.3:** Jet Eta Distributions (Signal and Background) for the Canonical 15 Signal Points.

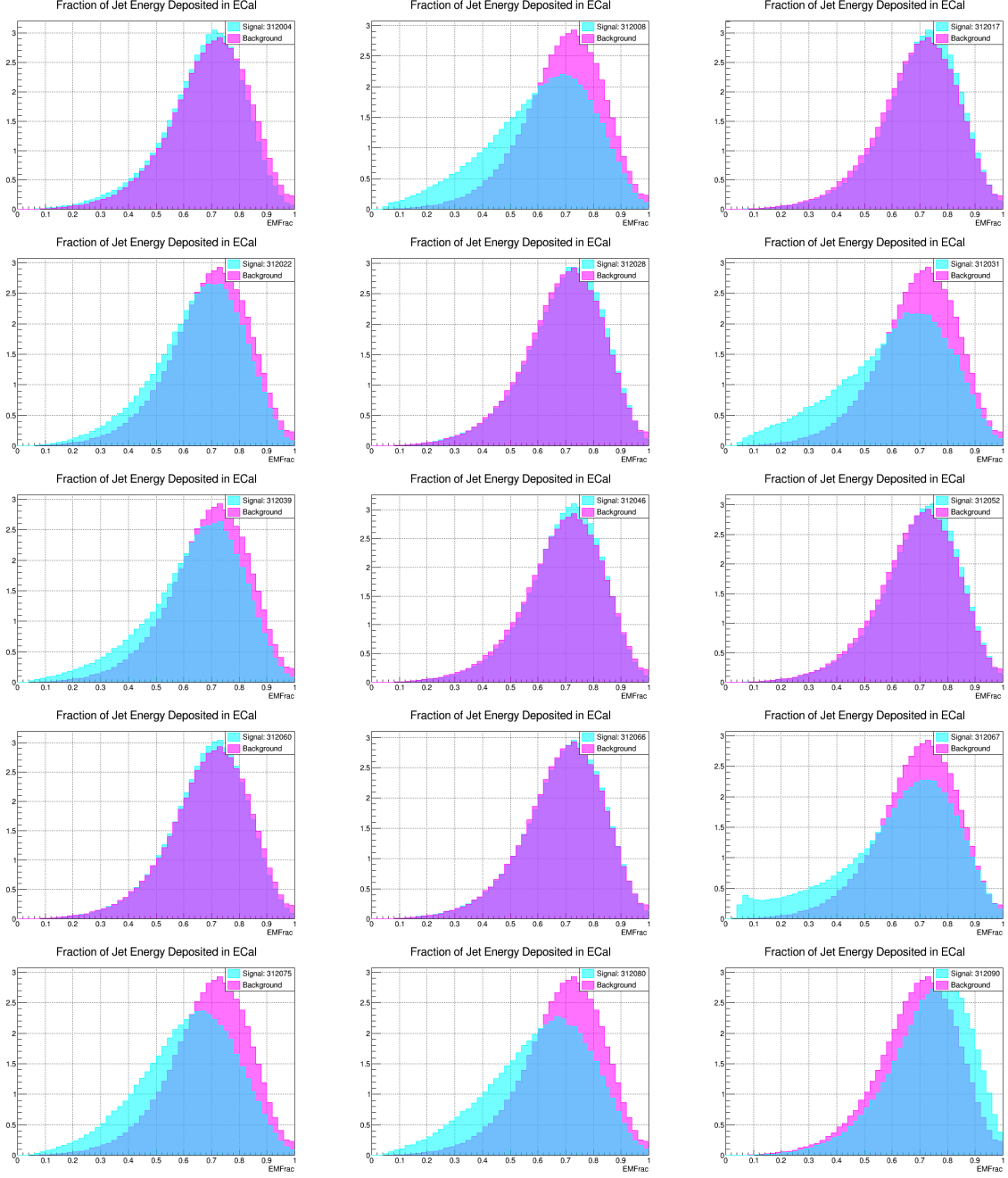


**Figure C.4:** Jet Width Distributions (Signal and Background) for the Canonical 15 Signal Points.



**Figure C.5:** Jet Centroid-R Distributions (Signal and Background) for the Canonical 15 Signal Points.





**Figure C.6:** Jet EM Fraction Distributions (Signal and Background) for the Canonical 15 Signal Points.

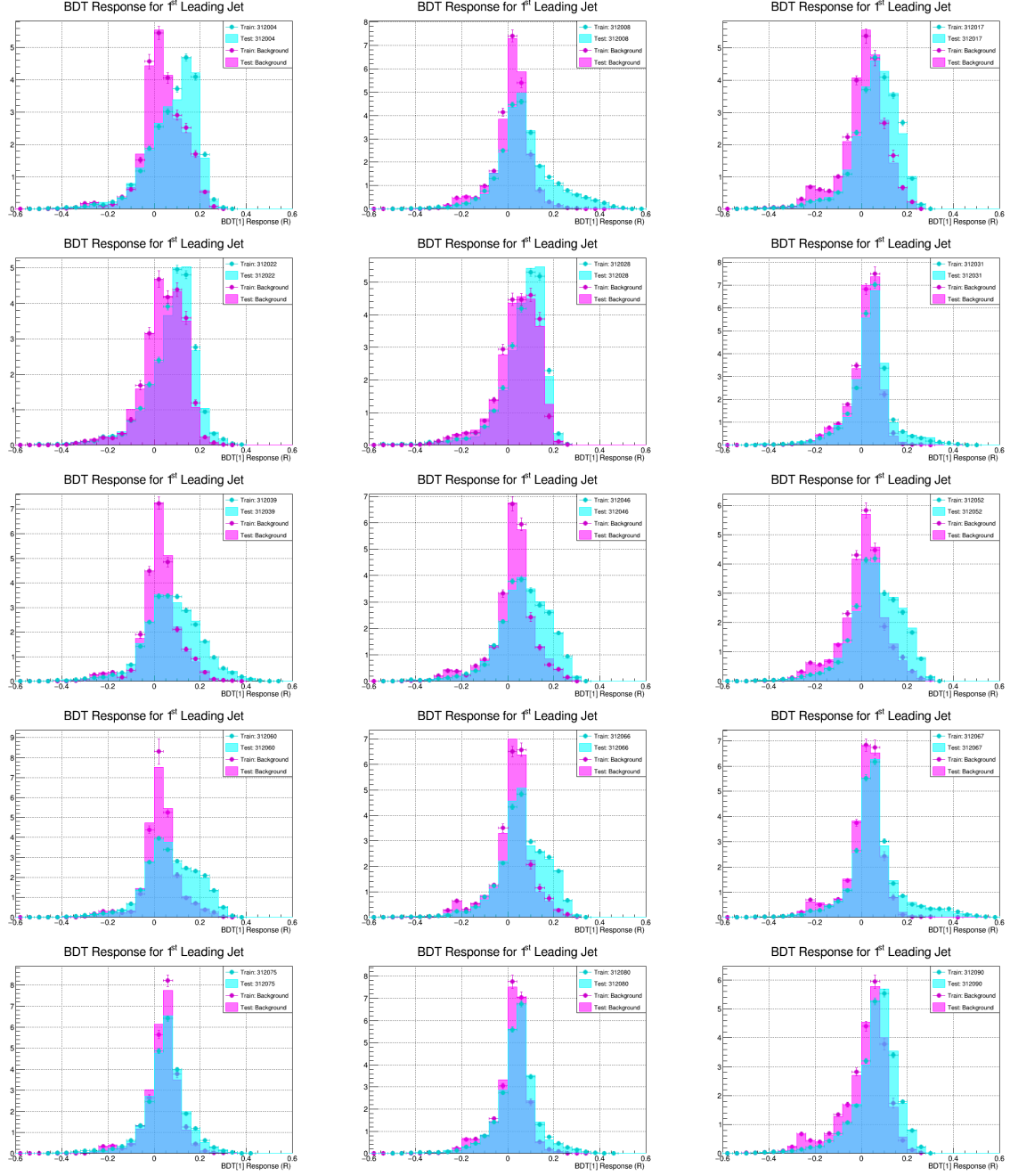
## Appendix D

# Event-Level BDT Input Variables

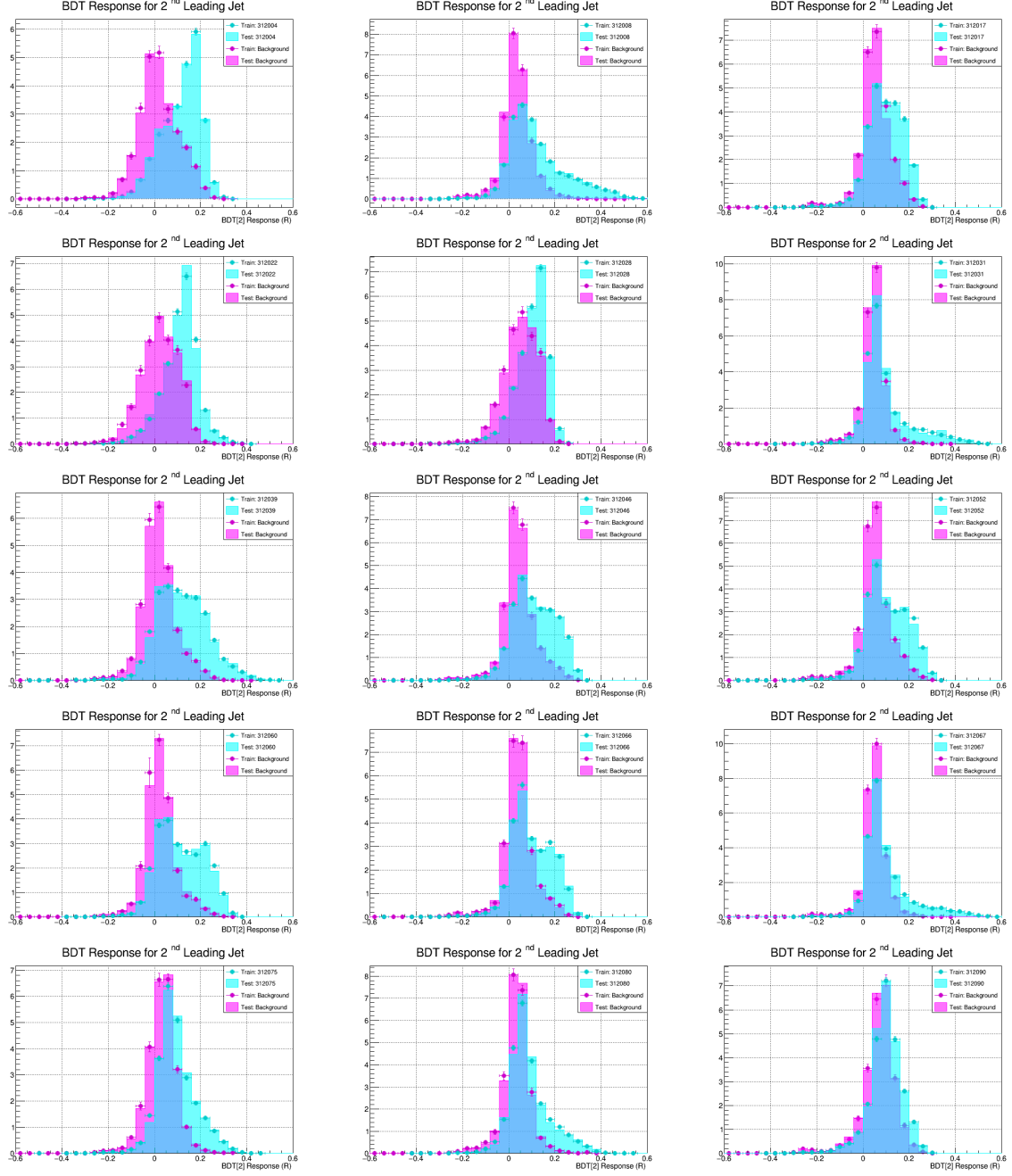
Figures D.1 - D.9 give the input distributions for the event-level BDT for the 15 signal points of the canonical set. The DSID for each signal point is given in the legend, but the plots are organized as follows:

- The rows are arranged alphabetically by model, with model A at the top and model E at the bottom.
- The columns are arranged by mediator mass, with  $m_{X_d} = 1400$  GeV on the left and  $m_{X_d} = 600$  GeV on the right.

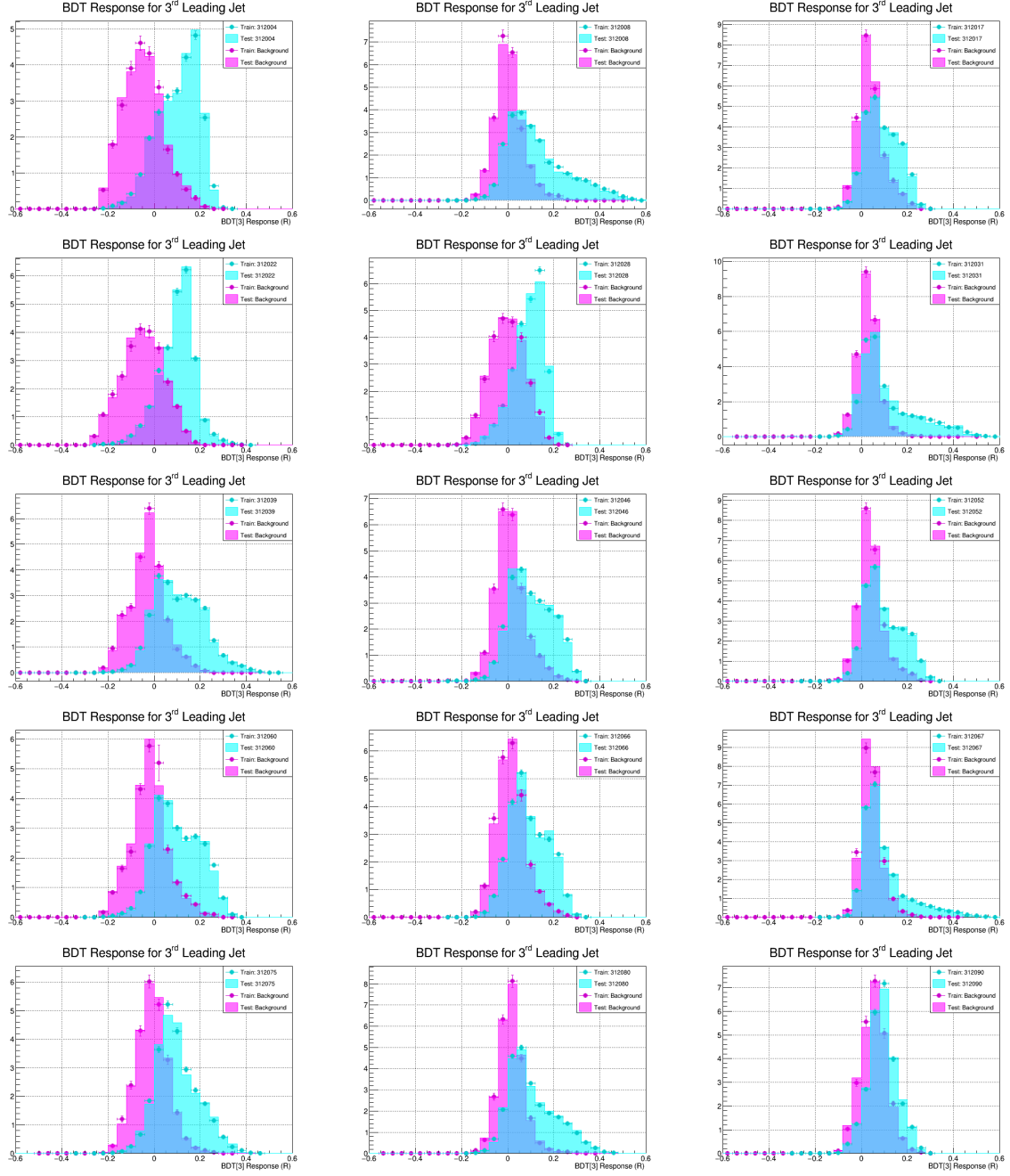
The plots are arranged in the same way for each of the 9 input variables.



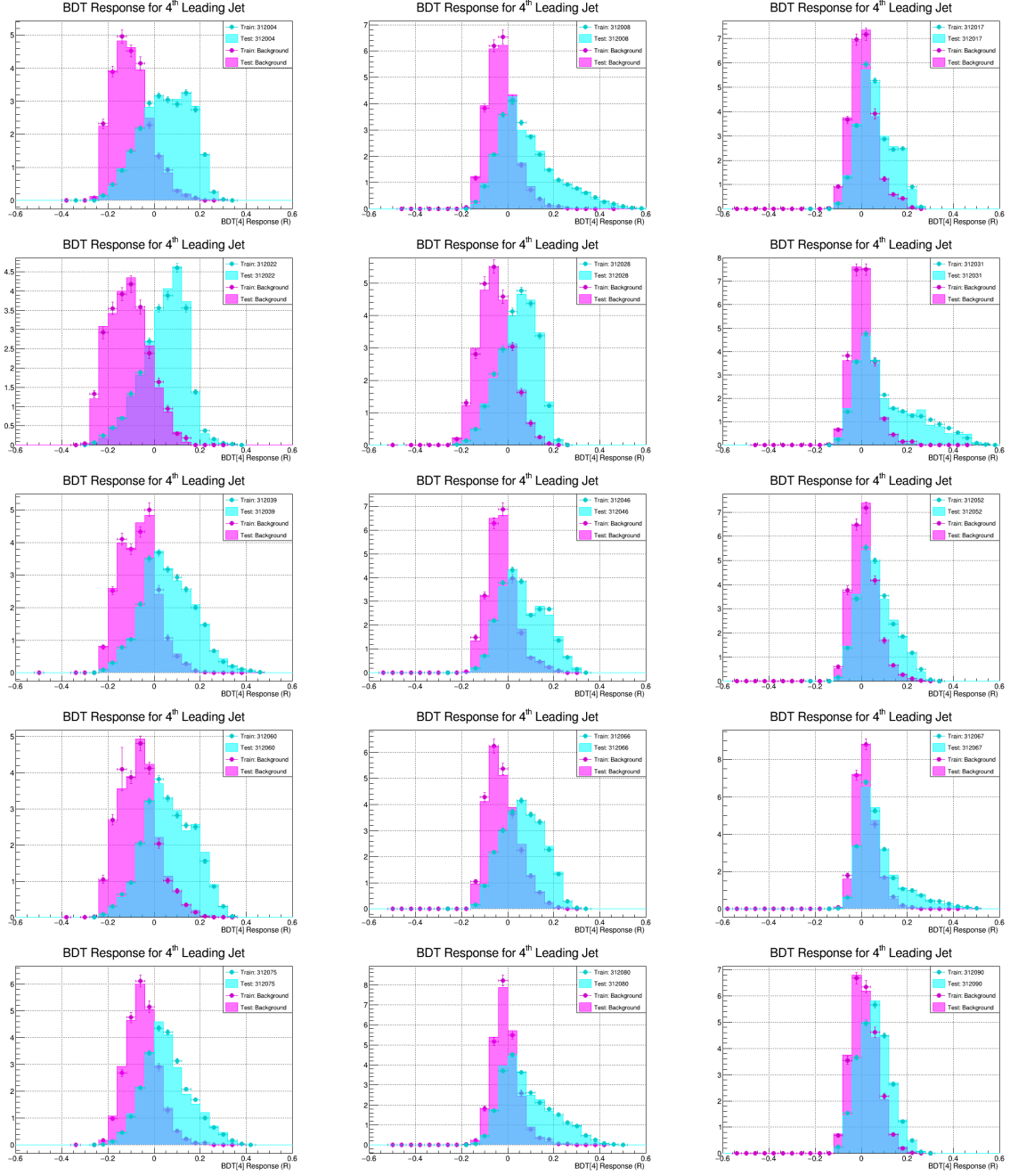
**Figure D.1:** BDT Response of the 1<sup>st</sup> Leading Jet Distributions (Signal and Background) for the Canonical 15 Signal Points.



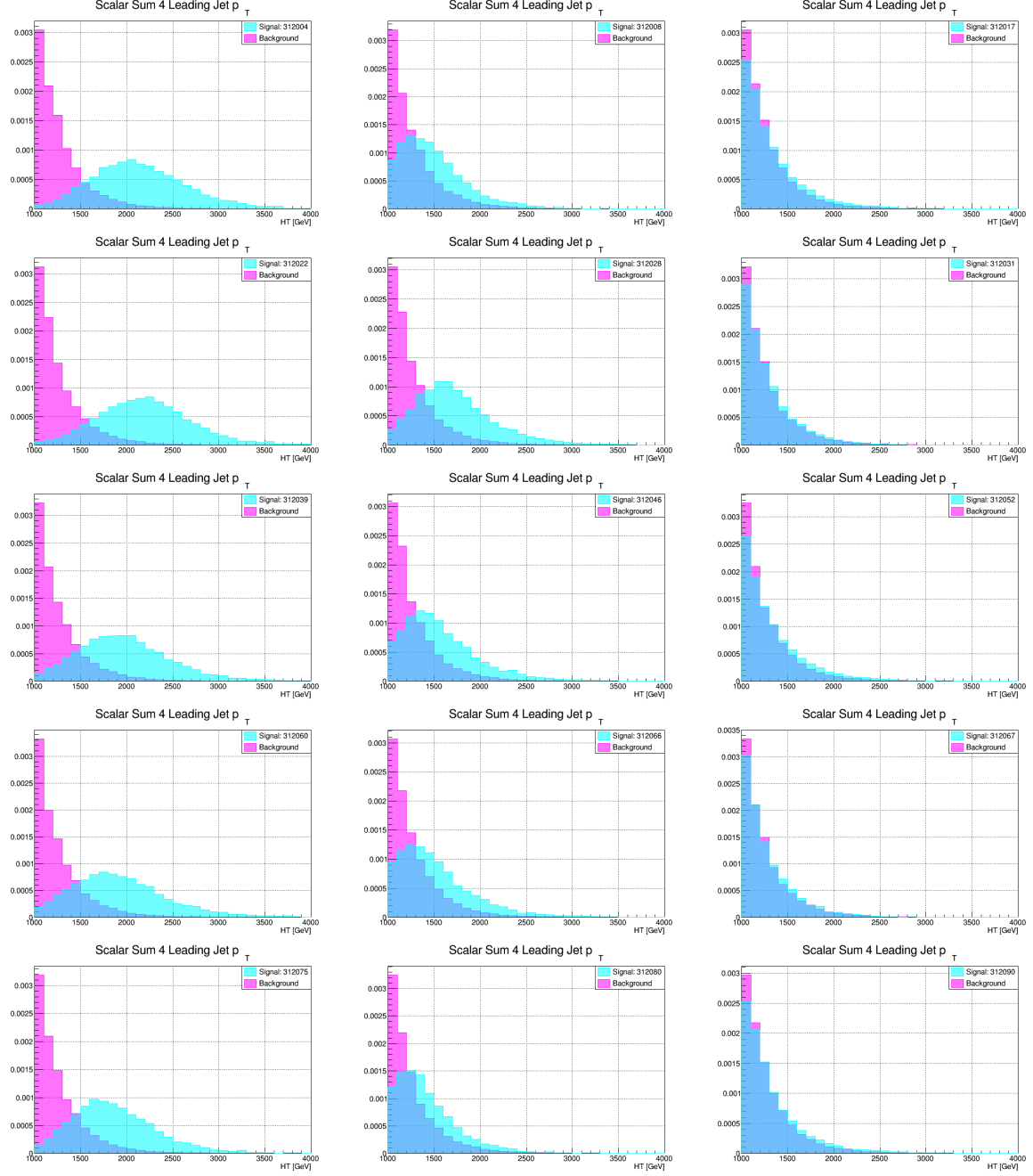
**Figure D.2:** BDT Response of the 2<sup>nd</sup> Leading Jet Distributions (Signal and Background) for the Canonical 15 Signal Points.



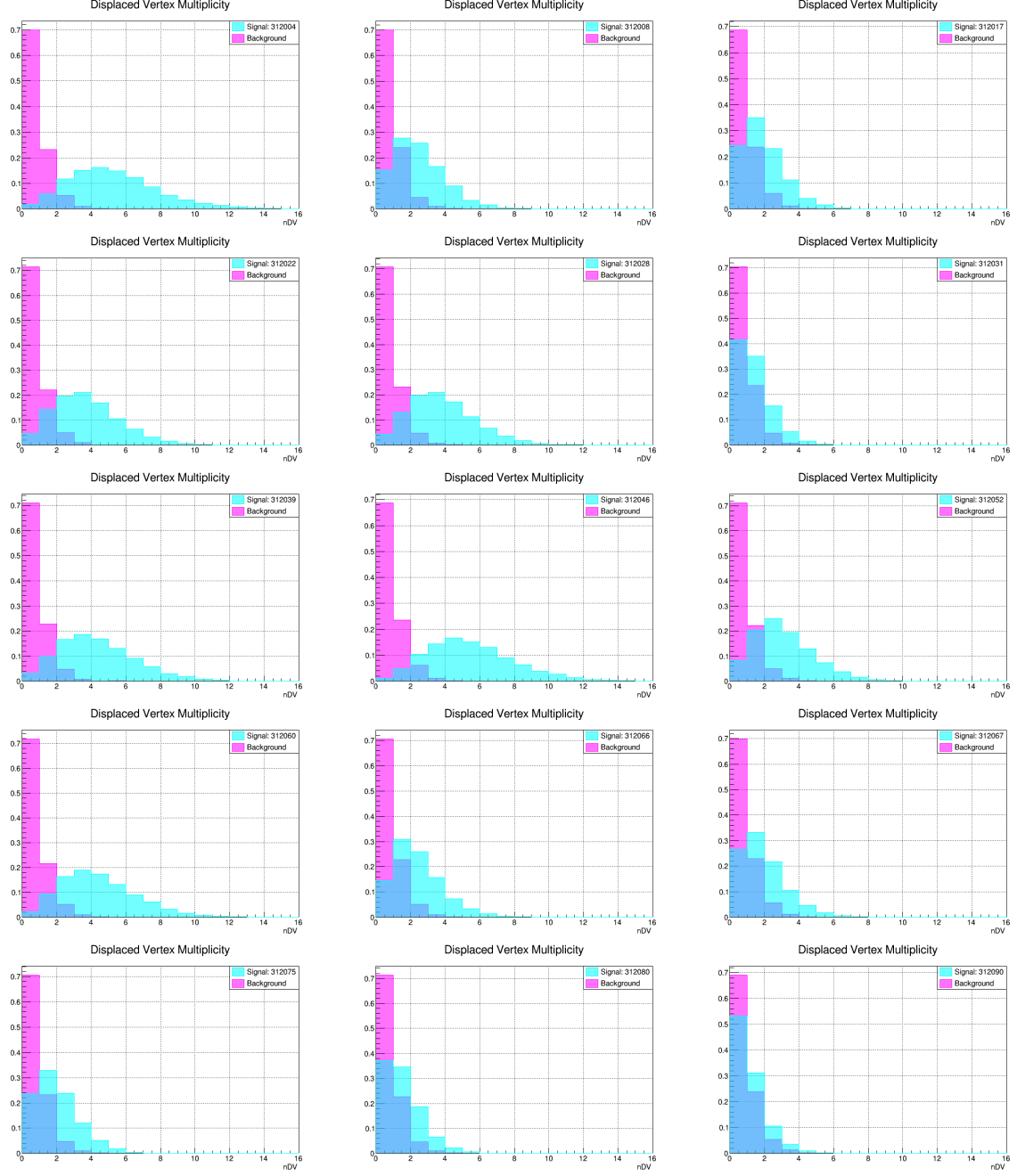
**Figure D.3:** BDT Response of the 3<sup>rd</sup> Leading Jet Distributions (Signal and Background) for the Canonical 15 Signal Points.



**Figure D.4:** BDT Response of the 4<sup>th</sup> Leading Jet Distributions (Signal and Background) for the Canonical 15 Signal Points.

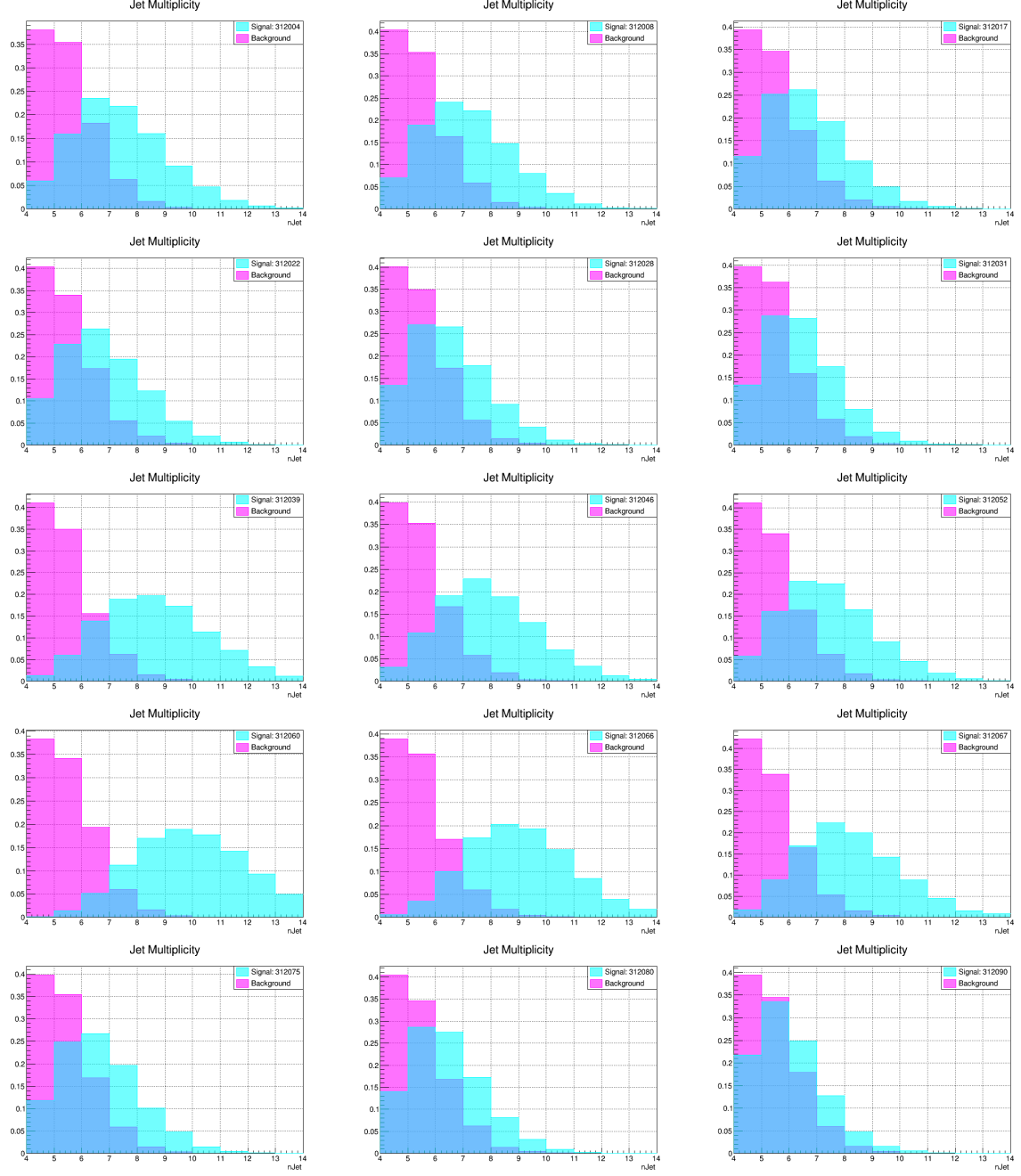


**Figure D.5:**  $H_T$  Distributions (Signal and Background) for the Canonical 15 Signal Points.

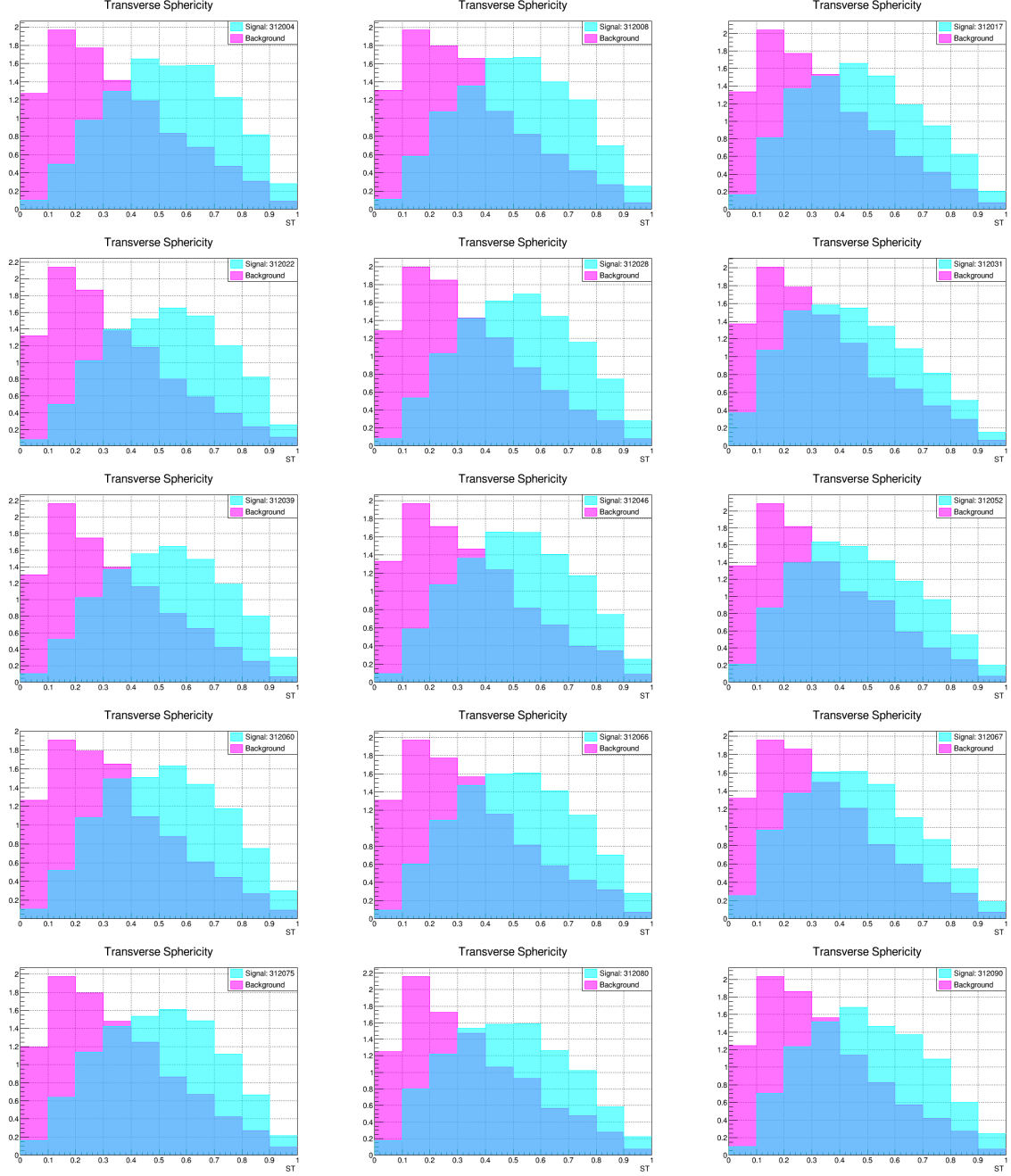


**Figure D.6:**  $nDV$  Distributions (Signal and Background) for the Canonical 15 Signal Points.

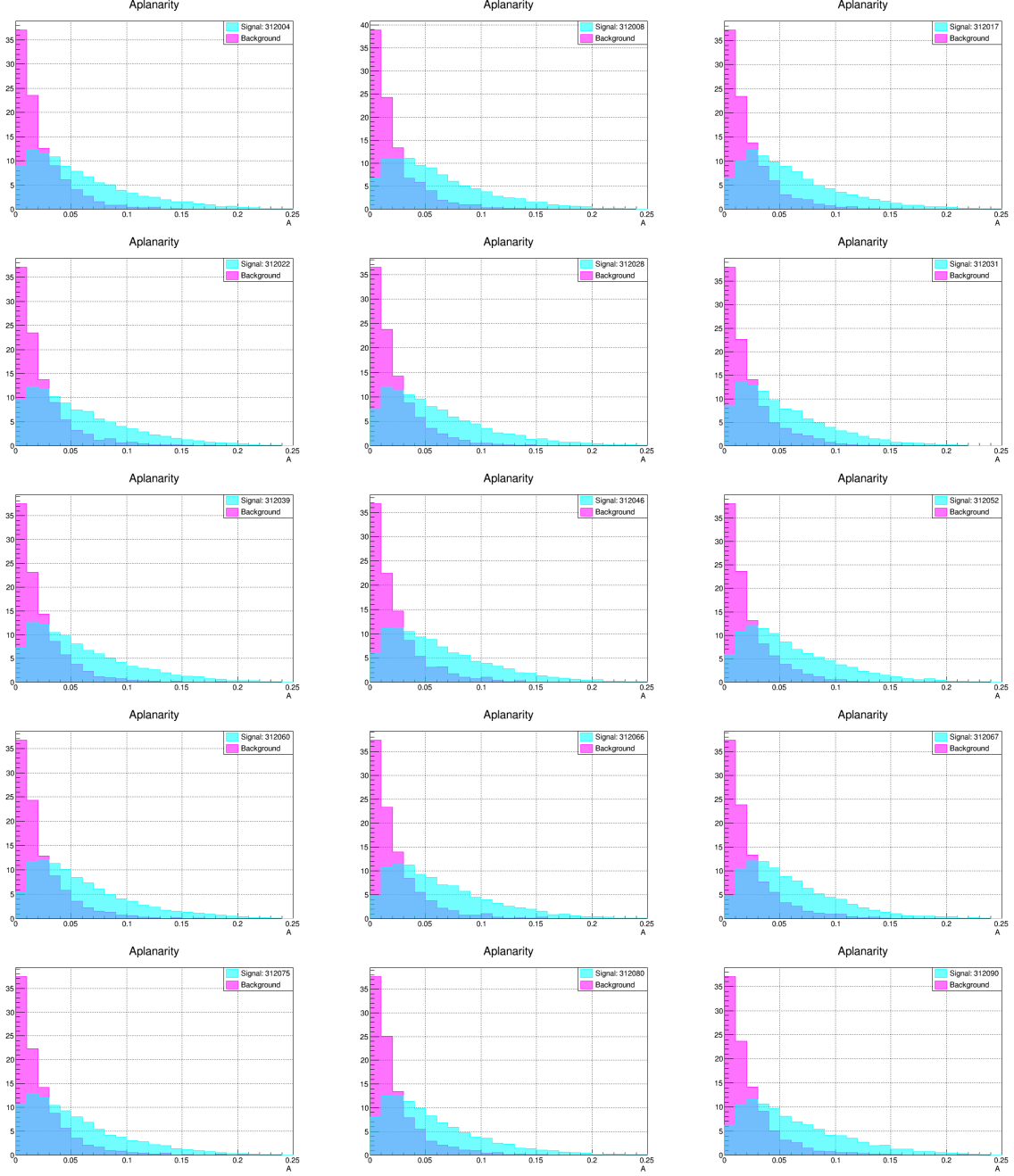




**Figure D.7:**  $n_{Jet}$  Distributions (Signal and Background) for the Canonical 15 Signal Points.



**Figure D.8:**  $S_T$  Distributions (Signal and Background) for the Canonical 15 Signal Points.



**Figure D.9:**  $\mathcal{A}$  Distributions (Signal and Background) for the Canonical 15 Signal Points.

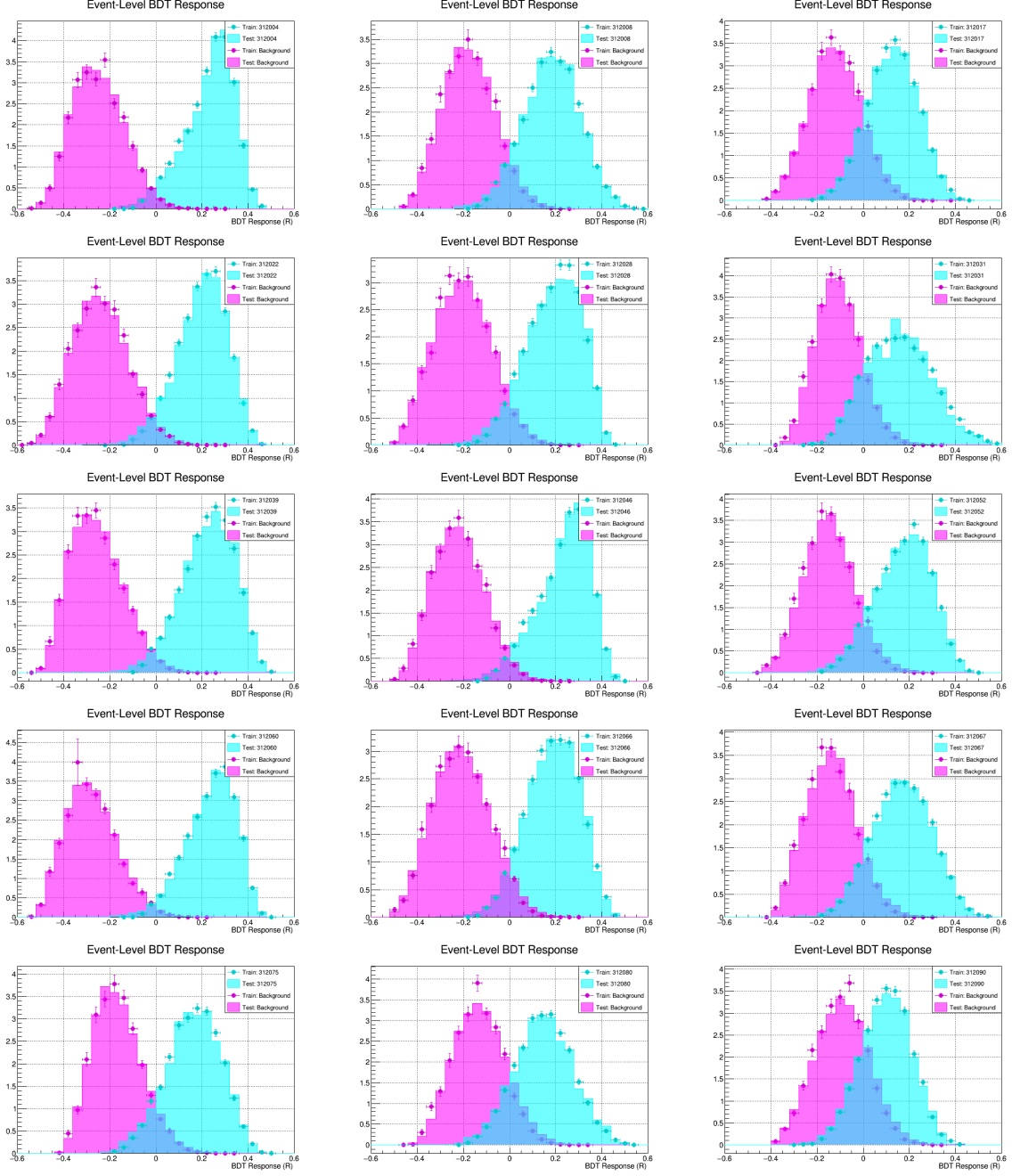
## Appendix E

# Model Dependent ABCD Variables

Figures E.1 and E.2 give the distributions for the variables which define the model dependent ABCD plane for the 15 signal points of the canonical set. The DSID for each signal point is given in the legend, but the plots are organized as follows:

- The rows are arranged alphabetically by model, with model A at the top and model E at the bottom.
- The columns are arranged by mediator mass, with  $m_{X_d} = 1400$  GeV on the left and  $m_{X_d} = 600$  GeV on the right.

The plots are arranged in the same way for both of the defining variables.



**Figure E.1:** Event  $\mathcal{R}$  Distributions (Signal and Background) for the Canonical 15 Signal Points.



**Figure E.2:** Minimum Prompt Track Fraction Distributions (Signal and Background) for the Canonical 15 Signal Points.

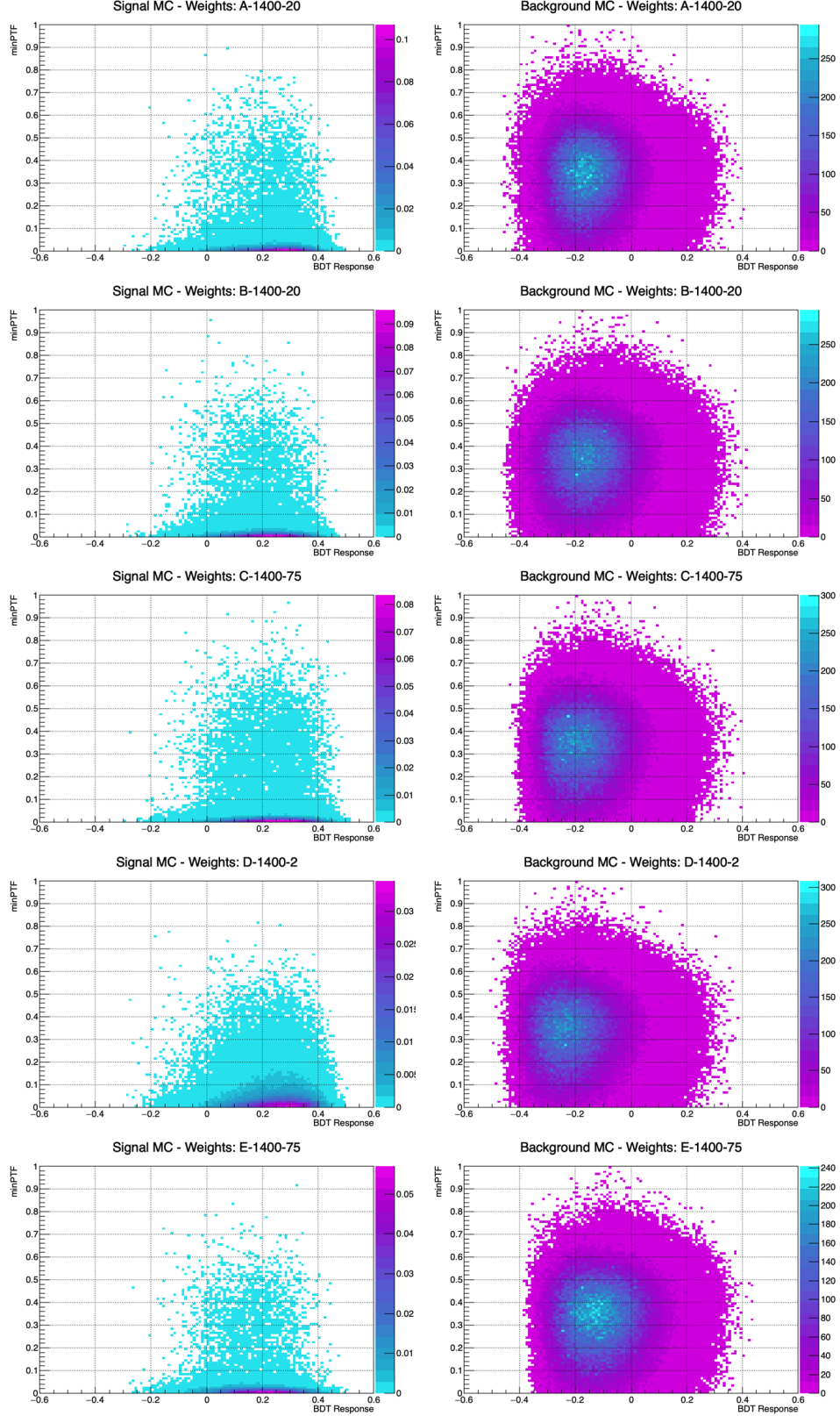
## Appendix F

# Search Region Model Dependent ABCD Planes

Figures F.1 - F.3 give the search region signal and background distributions for model dependent ABCD plane. All 15 signal points of the canonical set are shown, but the plots are organized in blocks depending on the corresponding mediator mass. The model is indicated in the title, but the plots are organized as follows:

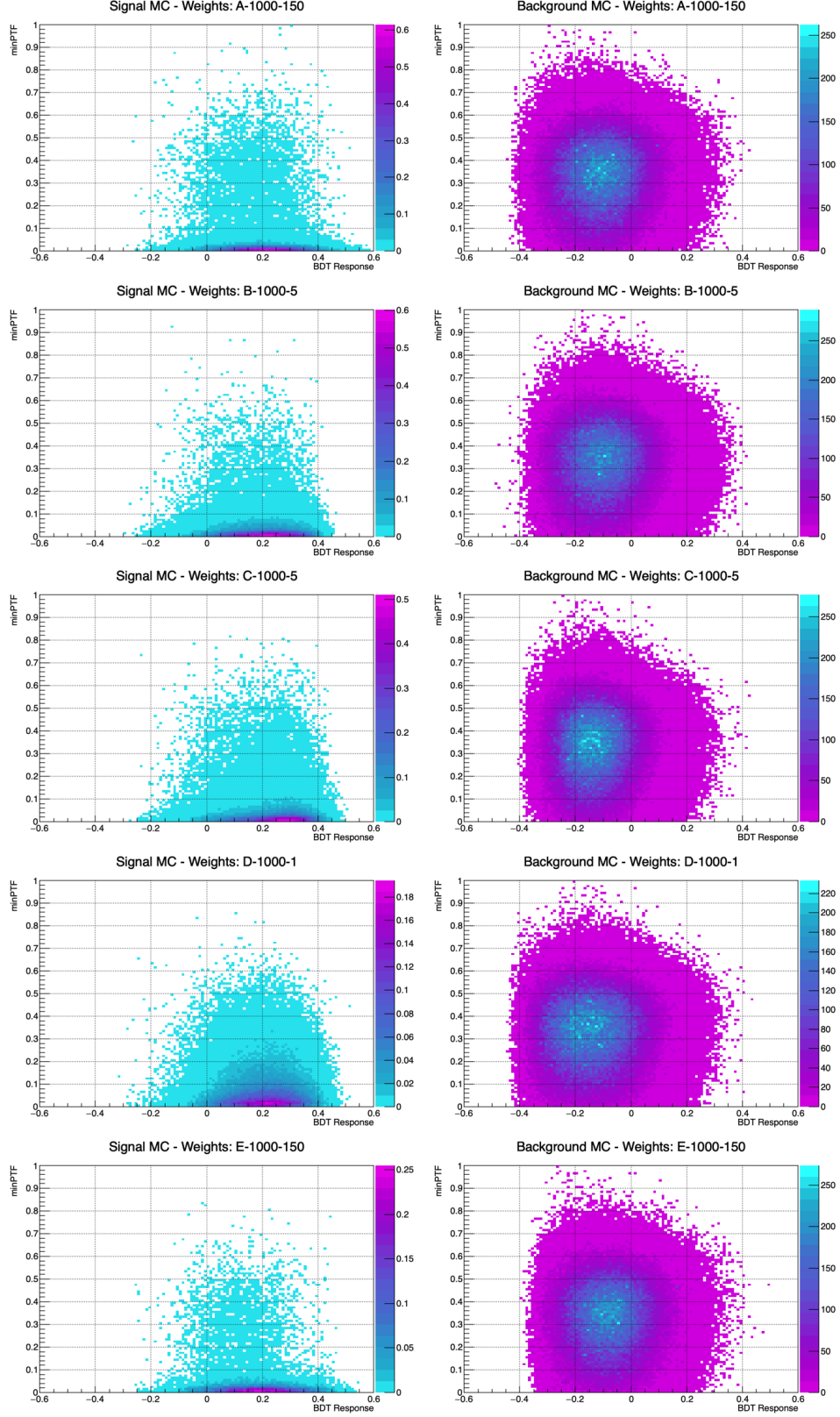
- The rows are arranged alphabetically by model, with model A at the top and model E at the bottom.
- The signal distributions are shown in the left column, and the background are on the right.
- The blocks are arranged by mediator mass, with  $m_{X_d} = 1400$  GeV appearing first and  $m_{X_d} = 600$  GeV shown last.

The plots are arranged in the same way for all three blocks.

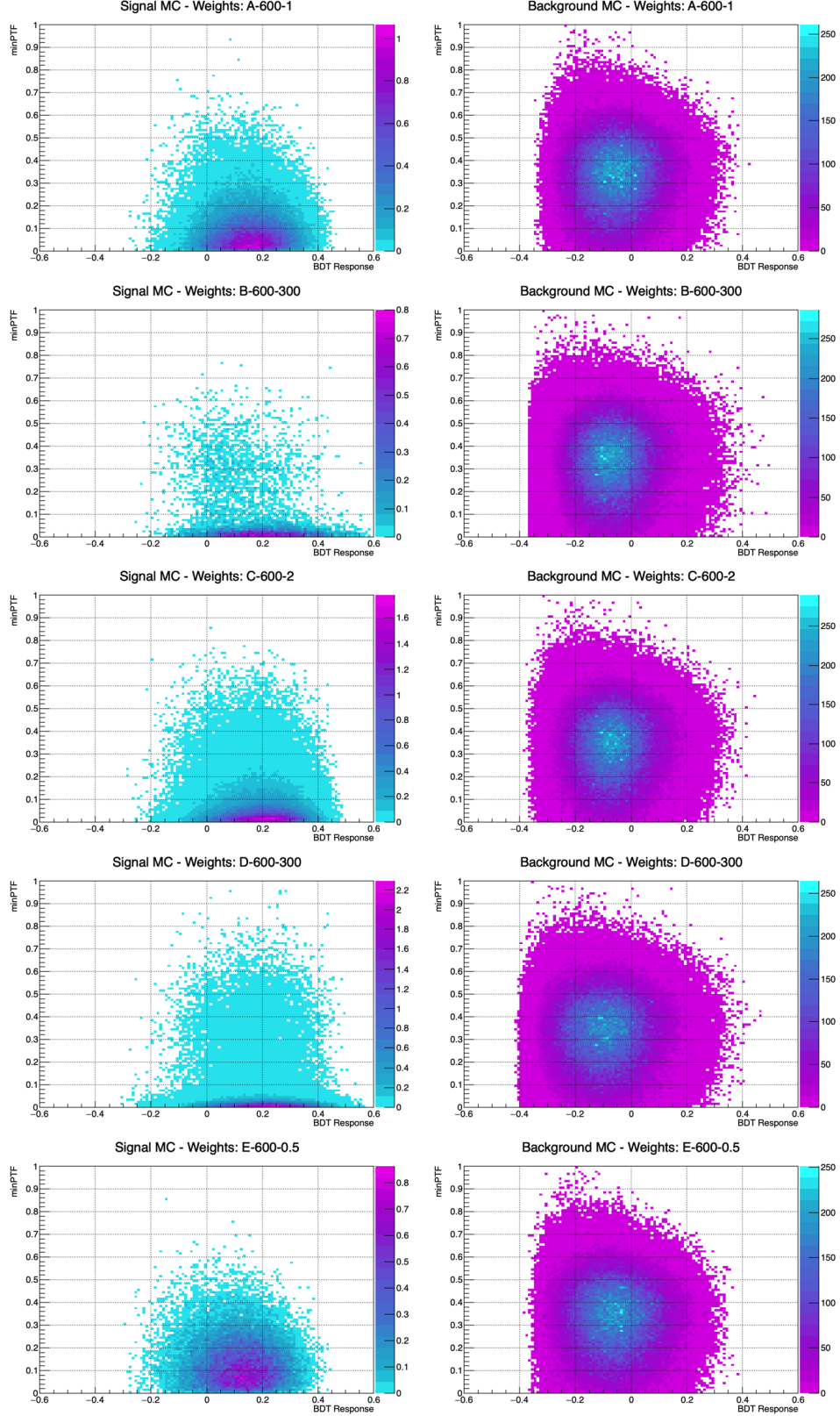


**Figure F.1:** Model Dependent ABCD Planes (Signal and Background) for the  $m_{X_d} = 1400$  GeV Models in the Canonaical Set, in the Search Region.





**Figure F.2:** Model Dependent ABCD Planes (Signal and Background) for the  $m_{X_d} = 1000$  GeV Models in the Canonaical Set, in the Search Region.



**Figure F.3:** Model Dependent ABCD Planes (Signal and Background) for the  $m_{X_d} = 600$  GeV Models in the Canonaical Set, in the Search Region.

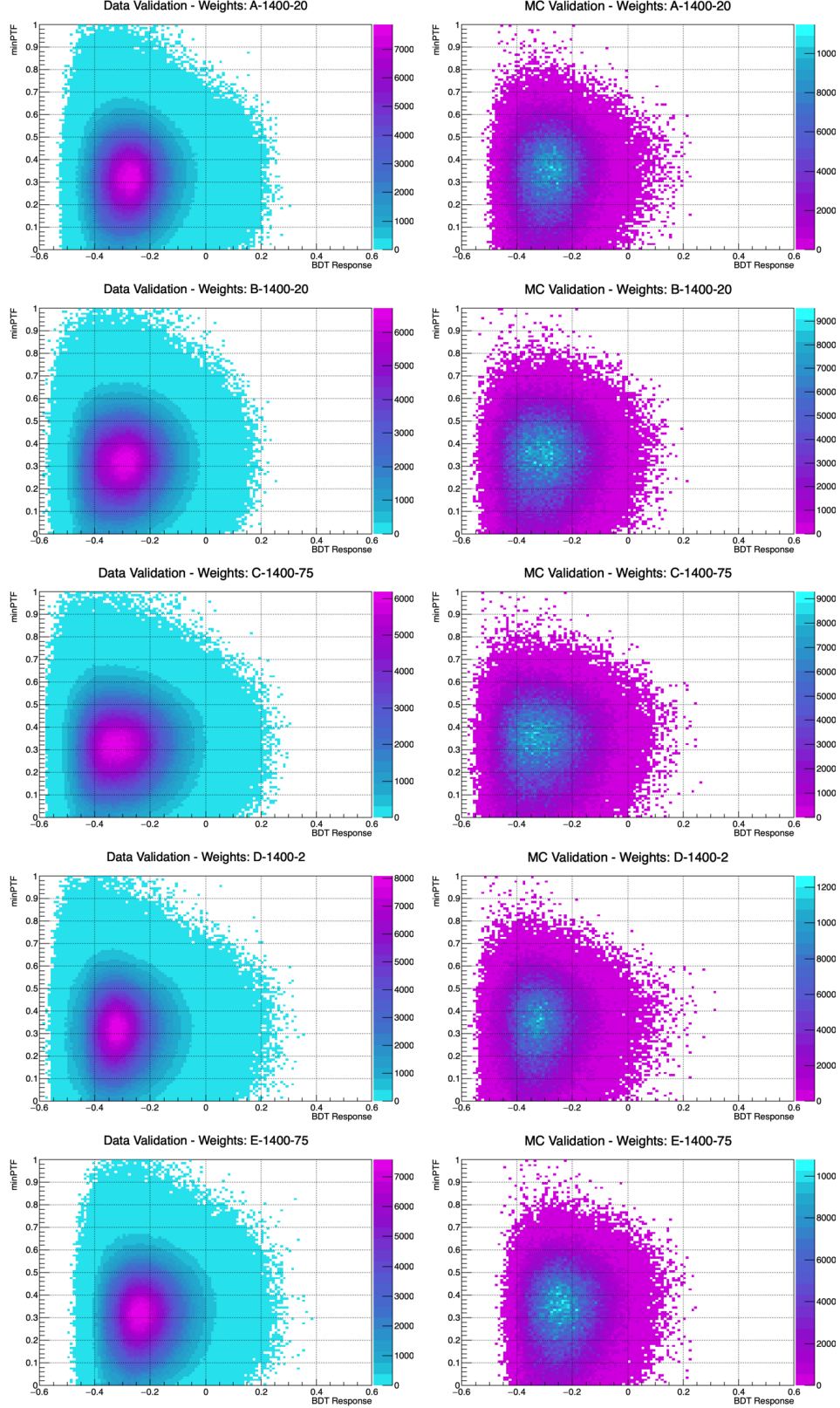
## Appendix G

# Validation Region Model Dependent ABCD Planes

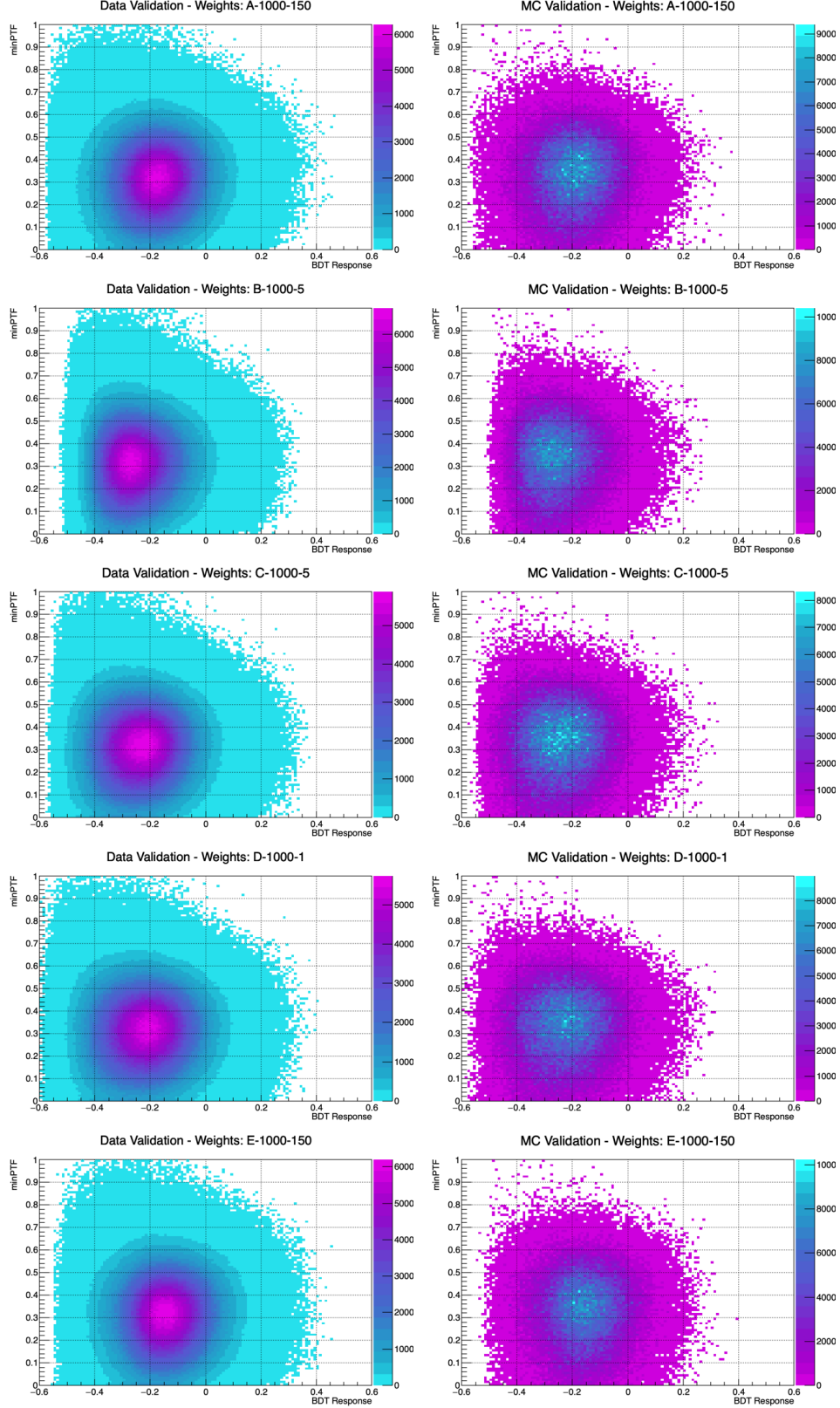
Figures G.1 - G.3 give the 4-jet validation region data and background MC distributions for model dependent ABCD plane. The plots for all 15 signal points of the canonical set are shown, but the plots are organized in blocks depending on the corresponding mediator mass. The model is indicated in the title, but the plots are organized as follows:

- The rows are arranged alphabetically by model, with model A at the top and model E at the bottom.
- The data distributions are shown in the left column, and the background MC are on the right.
- The blocks are arranged by mediator mass, with  $m_{X_d} = 1400$  GeV appearing first and  $m_{X_d} = 600$  GeV shown last.

The plots are arranged in the same way for all three blocks.

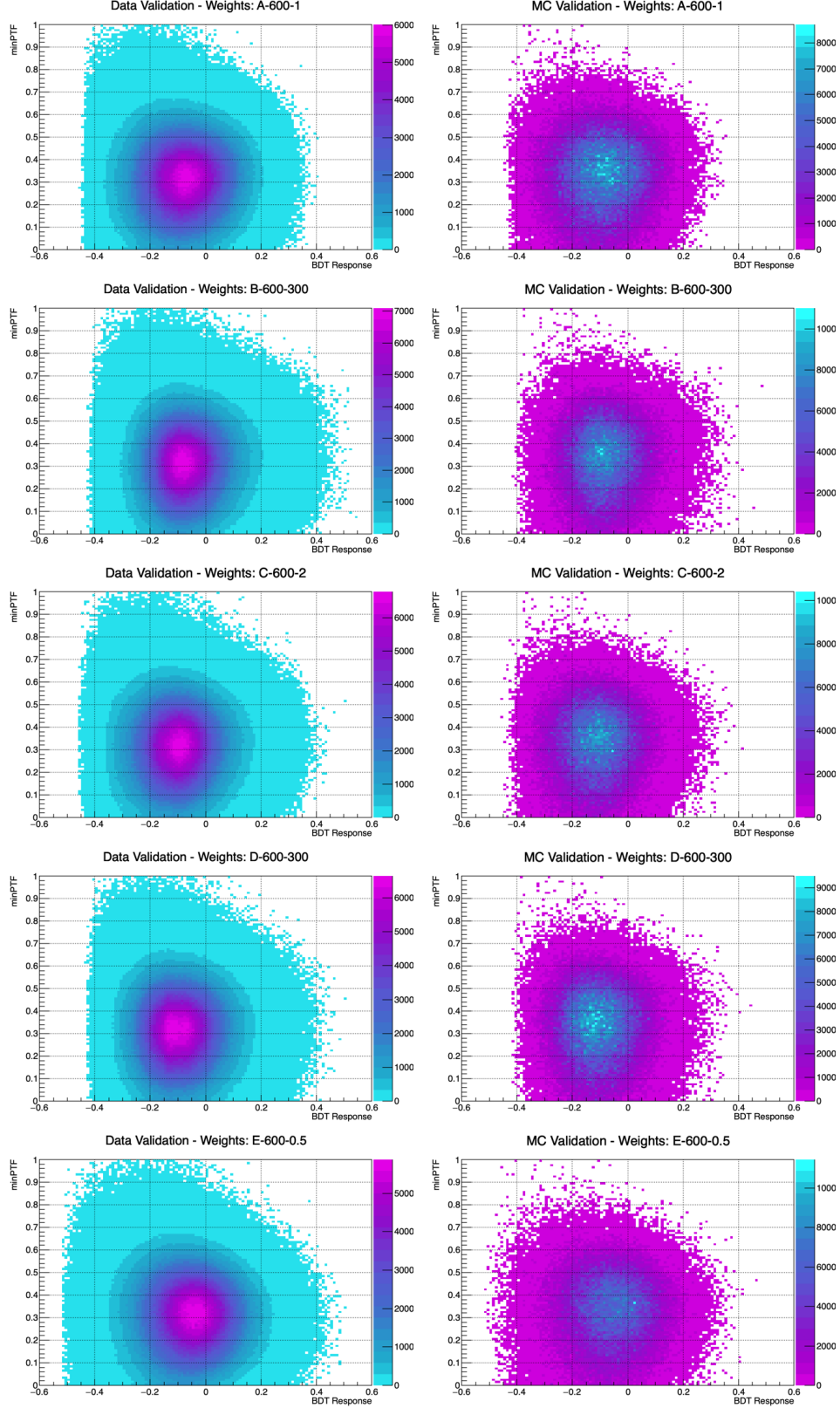


**Figure G.1:** Model Dependent ABCD Planes (Data and MC) for the  $m_{X_d} = 1400$  GeV Models in the Canonaical Set, in the 4-Jet Valdation Region.



**Figure G.2:** Model Dependent ABCD Planes (Data and MC) for the  $m_{X_d} = 1000$  GeV Models in the Canonaical Set, in the 4-Jet Validation Region.





**Figure G.3:** Model Dependent ABCD Planes (Data and MC) for the  $m_{X_d} = 600$  GeV Models in the Canonaical Set, in the 4-Jet Validation Region.

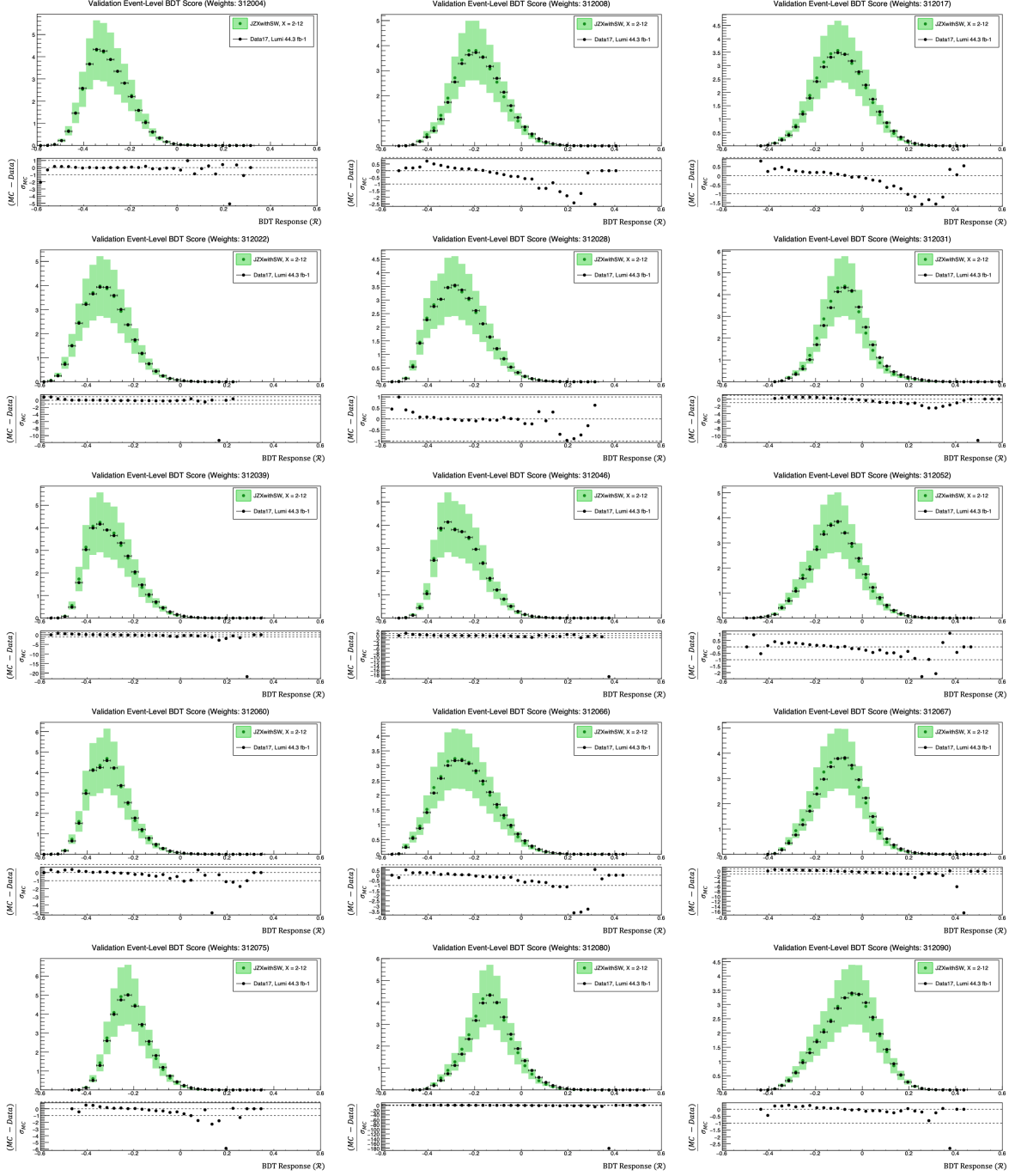
## Appendix H

# Evaluation of Systematics

Figure H.1 gives the evaluation of the full set of systematics for the 15 signal points of the canonical set in the 4-Jet validation region. The DSID for each signal point is given in the title, but the plots are organized as follows:

- The rows are arranged alphabetically by model, with model A at the top and model E at the bottom.
- The columns are arranged by mediator mass, with  $m_{X_d} = 1400$  GeV on the left and  $m_{X_d} = 600$  GeV on the right.

The green bands give the total combined uncertainty,  $\sigma_{MC} = \sqrt{\sigma_{stat,MC}^2 + \sigma_{syst}^2}$ , where the dominant effect is from the systematic uncertainty. The comparison in the bottom panel gives the significance of the difference between data and MC, and the error bars correspond to the data statistical uncertainty, which in some are too small to see.



**Figure H.1:** Evaluation of the effect of the full set of systematics on the event-level  $\mathcal{R}'$ , in the 4-Jet validation region for the Canonical 15 signal points.



# References

- [1] M. E. Peskin and D. V. Schroeder. *An Introduction To Quantum Field Theory*, CRC Press, Taylor and Francis Group, 1995.
- [2] D. Griffiths. *Introduction to Elementary Particles*, John Wiley and Sons, Inc., 1987.
- [3] M. Tanabashi et al. (Particle Data Group). *Review of Particle Physics*, Physical Review D and the PDG, 2018.
- [4] Chinese Physical Society (Particle Data Group). *Chinese Physics C (Vol. 38, No. 9): Review of Particle Physics*, Science Press and IOP Publishing, 2014.
- [5] ATLAS Collaboration. *The ATLAS Experiment at the CERN Large Hadron Collider*, Journal of Instrumentation, 2008.
- [6] The CMS Collaboration. *The CMS Experiment at the CERN Large Hadron Collider*, Journal of Instrumentation, 2008.
- [7] L. Evans and P. Bryant. *LHC Machine*, Journal of Instrumentation, 2008.
- [8] G. Duda and K. Gattett. *Dark Matter: A Primer*, Advances in Astronomy, 2010.
- [9] P. Schwaller, D. Stolarski, and A. Weiler. *Emerging Jets*, arXiv:1502.05409 [hep-ph], 2015.
- [10] J. D. Jackson. *Classical Electrodynamics*, 3<sup>rd</sup> eds. John Wiley and Sons, Inc., 1999.
- [11] B. M. Peterson and B. Ryden. *Foundations of Astrophysics*, Pearson Education Inc., 2011.
- [12] J. Beringer et al. (Particle Data Group). *Higgs Boson: Theory and Searches*, Physical Review D and the PDG, 2012.

- [13] J. C. Peng and J. W. Qiu. *The Drell-Yan Process*, Popular Science Article Vol. 4, No. 3, 2016.
- [14] M. D’Onofrio and K. Rummukainen. *Standard Model Cross-Over on the Lattice*, arXiv:1508.07161 [hep-ph], 2015.
- [15] M. Thomson. *Modern Particle Physics*, Cambridge University Press, 2013.
- [16] The European Organization for Nuclear Research (CERN). *Dark Matter*, CERN Physics, 2012.
- [17] F. Zwicky. *On the Masses of Nebulae and of Clusters of Nebulae*, Astrophysical Journal, Volume 86, 1937.
- [18] V. C. Rubin and W. K. Ford Jr. *Rotation of the Andromeda Nebula from a Spectroscopic Survey of Emission Regions*, Astrophysical Journal, Volume 159, 1970.
- [19] E. Corbelli and P. Salucci. *The Extended Rotation Curve and the Dark Matter Halo of M33*, Monthly Notices of the Royal Astronomical Society, Volume 311, Issue 2, 2000.
- [20] J. B. Hartle. *Gravity: An Introduction to Einstein’s General Relativity*, Pearson Education Inc., 2003.
- [21] National Aeronautics and Space Administration (NASA). *Shining a Light on Dark Matter*, NASA Discoveries - Highlights, 2019.
- [22] M. Schumann. *Direct Detection of WIMP Dark Matter: Concepts and Status*, arXiv:1903.03026 [astro-ph.CO], 2019.
- [23] The European Organization for Nuclear Research (CERN). *Supersymmetry*, CERN Physics, n.d.
- [24] L. Roszkowski, E. M. Sessolo, S. Trojanowski. *WIMP Dark Matter Candidates and Searches - Current Status and Future Prospects*, arXiv:1707.06277 [hep-ph], 2018.
- [25] M. J. Strassler and K. M Zurek. *Echoes of a Hidden Valley at Hadron Colliders*, arXiv:hep-ph/0604261[hep-ph], 2006.
- [26] H. Mies, C. Scherb, and P. Schwaller. *Collider Constraints on Dark Mediators*, arXiv:2011.13990 [hep-ph], 2020.

- [27] S. Giagu. *WIMP Dark Matter Searches With the ATLAS Detector at the LHC*, Frontiers in Physics, 2019.
- [28] J. Alimena et al. *Searching for Long-Lived Particles Beyond the Standard Model at the Large Hadron Collider*, arXiv:1903.04497 [hep-ex], 2019.
- [29] B. Shuve. *Theory Overview of Long-Lived Particles*, LHC-LLP Workshop, CERN, 2017.
- [30] L. Lee et al. *Collider Searches for Long-Lived Particles Beyond the Standard Model*, arXiv:1810.12602 [hep-ph], 2018.
- [31] D. Curtin et al. *Long-Lived Particles at the Energy Frontier: The MATHUSLA Physics Case*, arXiv:1806.07396 [hep-ph], 2019.
- [32] The ATLAS Collaboration. *Search for Heavy Charged Long-Lived Particles in the ATLAS Detector in  $36.1 \text{ fb}^{-1}$  of Proton-Proton Collision Data at  $\sqrt{s} = 13 \text{ TeV}$* , arXiv:1902.01636 [hep-ex], 2019.
- [33] The Planck Collaboration. *Planck 2018 Results. VI. Cosmological Parameters*, arXiv:1807.06209 [astro-ph.CO], 2020.
- [34] C. J. Treado. *A Search for Emerging Jets at  $\sqrt{s} = 13 \text{ TeV}$  at ATLAS Run 2*, CERN-THESIS-2020-242, 2020.
- [35] The European Organization for Nuclear Research (CERN). *The Large Hadron Collider*, CERN Science - Accelerators, n.d.
- [36] The European Organization for Nuclear Research (CERN). *The LHC Design Report - Volume I The LHC Main Ring*, CERN-2004-003, 2004.
- [37] The ATLAS Collaboration. *ATLAS Fact Sheet*, 2011.
- [38] The ATLAS Collaboration. *Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC*, arXiv:1207.7214v2 [hep-ex], 2012.
- [39] W. Herr and B. Muratori. *Concept of Luminosity*, CERN-2006-002, 2006.
- [40] The ATLAS Collaboration. *Luminosity Public Result Run 2*, ATLAS Public Results, n.d.

- [41] The European Organization for Nuclear Research (CERN). *LHC Report: Collisions and spring cleaning at the LHC*, CERN News - Accelerators, 2018.
- [42] The ATLAS Collaboration. *Detector and Technology*, ATLAS Discover - Detector, n.d.
- [43] The ATLAS Collaboration. *ATLAS Inner Detector: Technical Design Report - Volume 1*, ATLAS-TDR-4-Volume-I, 1997.
- [44] The ATLAS Collaboration. *ATLAS Pixel Detector Electronics and Sensors*, Journal of Instrumentation, 2008.
- [45] The ATLAS Collaboration. *Operation and Performance of the ATLAS Semiconductor Tracker*, arXiv:1404.7473v2 [hep-ex], 2014.
- [46] The ATLAS Collaboration. *The ATLAS Transition Radiation Tracker (TRT) proportional drift tube: Design and performance*, Journal of Instrumentation, 2008.
- [47] V. Lacuesta. *Track and Vertex Reconstruction in the ATLAS Experiment*, Journal of Instrumentation, 2013.
- [48] The ATLAS Collaboration. *ATLAS Liquid-Argon Calorimeter: Technical Design Report*, ATLAS-TDR-2, 1996.
- [49] The ATLAS Collaboration. *ATLAS Tile Calorimeter: Technical Design Report*, ATLAS-TDR-3, 1996.
- [50] The ATLAS Collaboration. *ATLAS Muon Spectrometer: Technical Design Report*, ATLAS-TDR-10, 1997.
- [51] J. Pequeno and P. Schaffner. *How ATLAS Detects Particles: Diagram of Particle Paths in the Detector*, 2013.
- [52] L. Carloni and T. Sjöstrand. *Visible Effects of Invisible Hidden Valley Radiation*, arXiv:1006.2911 [hep-ph], 2010.
- [53] The ATLAS Collaboration. *Performance of the reconstruction of large impact parameter tracks in the inner detector of ATLAS*, ATL-PHYS-PUB-2017-014, 2017.

- [54] S. Schramm. *ATLAS Jet Reconstruction, Calibration, and Tagging of Lorentz-boosted Objects*, ATL-PHYS-PROC-2017-236, 2017.
- [55] M. Cacciari, G. P. Salam, and G. Soyez. *The Anti- $k_t$  Jet Clustering Algorithm*, arXiv:0802.1189 [hep-ph], 2008.
- [56] The ATLAS Collaboration. *Performance of Vertex Reconstruction Algorithms for Detection of New Long-Lived Particle Decays Within the ATLAS Inner Detector*, ATL-PHYS-PUB-2019-013, 2019.
- [57] The ATLAS Collaboration. *Performance of the ATLAS trigger system in 2015*, The European Physical Journal. C, Particles and Fields, 2017.
- [58] W. Buttinger. *Background Estimation with the ABCD Method*, 2017.
- [59] The ATLAS Collaboration. *Search for long-lived, massive particles in events with displaced vertices and missing transverse momentum in  $\sqrt{s} = 13$  TeV pp collisions with the ATLAS detector*, arXiv:1710.04901v1 [hep-ex], 2017.
- [60] K. Albertsson et al. *TMVA 4 - Toolkit for Multivariate Data Analysis with ROOT: User Guide*, arXiv:physics/0703039 [physics.data-an], 2009.
- [61] Y. Freund and R. E. Schapire. *A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting*, Journal of Computer and System Sciences, 1997.
- [62] The ATLAS Collaboration. *Measurement of event shapes at large momentum transfer with the ATLAS detector in pp collisions at  $\sqrt{s} = 7$  TeV*, arXiv:1206.2135v1 [hep-ex], 2012.
- [63] S. Schramm. *Small- $R$  ( $R=0.4$ ) Jet Energy Scale and Jet Energy Resolution Uncertainties for Full Run 2 Analyses*, ATLAS Twiki, 2021. [<https://twiki.cern.ch/twiki/bin/view/AtlasProtected/JetUncertaintiesRel21Summer2018SmallR>]
- [64] C. M. Delitzsch. *Small- $R$  ( $R=0.4$ ) Jet Mass Scale Uncertainties for Full Run 2 Analyses*, ATLAS Twiki, 2021. [<https://twiki.cern.ch/twiki/bin/view/AtlasProtected/JetUncertaintiesRel21Spring2021SmallRJMS>]

- [65] W. Buttinger. *Extended Pileup Reweighting*, ATLAS Twiki, 2021. [<https://twiki.cern.ch/twiki/bin/viewauth/AtlasProtected/ExtendedPileupReweighting>]