THE UNIVERSITY OF HULL

Dwarf galaxies with AGN and their environments in observations and simulations

being a Thesis submitted for the Degree of Doctor of Philosophy

in the University of Hull

by

Mikkel Theiss Kristensen

May 2022

Acknowledgements

When I embarked on my PhD journey three and a half years ago, I had no idea where it would take me. I was not sure what was waiting for me moving to a different country and working with research. It was scary.

However, my fears were soon alleviated by the very helpful staff and students of the E.A. Milne Centre. I am grateful to Kevin, who has been an excellent supervisor that helped me carry out interesting research as well as structure my work enabling me to finish degree. Brad has similarly been a good support that has helped with the many intricacies of bureaucracy and understanding cosmological simulations. A lot of the research has not been possible without the enthusiastic help and patience of Samantha, whom I only had the chance to meet once in person, but who contributed with a great wealth of information about dwarf galaxies and AGN.

I was lucky to start in a cohort of three other PhD students, Leah, Iraj, and Tom, with whom it has been possible to share experiences and frustrations about undertaking a PhD degree. They have become some of my closest and most important people outside of the programme as well, and it would not have been possible to be where I am today without them. The other PhD students in the office at the time we started, Tom, James, and Lawrence, provided good company and guidance, and for that I am also grateful.

I found out that doing a PhD is hard, so it has been a huge relief to have an understanding and supportive partner. Roberta has been my rock that gave me the motivation, stability, and support that enabled me to reach the end of my programme and be proud of my work.

Declaration of Originality

Dwarf galaxies with AGN and their environments in observations and simulations is the thesis that is being submitted in fulfilment of my degree of Doctor of Philosophy from the University of Hull. The work undertaken and included in this thesis is my own and carried out under the supervision of Dr Kevin Pimbblet, Dr Brad Gibson, and Dr Samantha Penny.

The thesis consists of a large amount of work that has been submitted to peer-reviewed journals and in collaboration with other researchers. The principle responsibility has been mine, and the collaborating authors have contributed intellectually to the science and writing aspects of the papers.

More specifically, Chapter 2 contains a study on environments of dwarf galaxies with optical AGN characteristics. The results were published in August 2020 in Monthly Notices of the Royal Astronomical Society, Volume 496, Issue 3, pp.2577-2590 and the work was carried out in collaboration with Kevin A. Pimbblet (University of Hull) and Samantha J. Penny (University of Portsmouth) and me as lead author. While I have written the paper and made the plots, collaborating authors have contributed with ideas such as which AGN diagnostics and environmental measures to use leaving me to be 85 per cent responsible for the paper. It has been formatted differently than the published paper to fit in the format of this thesis, and may also differ slightly in typography, but the results and science are unaltered.

Chapter 3 contains a study on merger histories and environments of dwarf galaxies AGN characteristics in simulations. The results were published in December 2021 in The Astrophysical Journal, Volume 922, Issue 2, id.127, 19 pp. and the work was carried out in collaboration with Kevin A. Pimbblet (University of Hull), Brad K. Gibson (University of Hull), Samantha J. Penny (University of Portsmouth), and Sophie Koudmani (University of Cambridge) and me as lead author. While I have written and made all text and plots, collaborating authors have contributed to understanding of the simulation, the physics behind the black hole models, and the limitations of the data leaving me to be 85 per cent responsible for

the paper. It has been formatted differently than the published paper to fit in this thesis, and may also differ slightly in typography, but the results and science are unaltered.

Chapter 4 details the preliminary work in an effort to use environmental and spatially resolved parameters to classify AGN in dwarf galaxies using a machine learning approach. It has not yet been submitted for peer review. It is done in collaboration with Kevin A. Pimbblet (University of Hull) and Samantha Penny (University of Portsmouth). While I have carried out the writing and the making of plots, collaborating authors have contributed to generating the idea of using machine learning (and a random forest approach), understanding of MaNGA data, and guidance on which data to include. Since this project has not been submitted to peer review yet or published in any journals, the work is only outlined in this thesis.

The rest of the work presented was intellectually generated by Kristensen with contributions from K. Pimbblet and S. Penny to the development and writing.

Abstract

This thesis is a study of dwarf galaxies with active galactic nucleis (AGN) characteristics, their environments, and identification of them in both observations and simulations. More specifically, it attempts to answer the questions of what environmental conditions are favourable for AGN activity, if environmental has any influence at all, and to what degree current AGN identification tools are suitable for dwarf galaxies. Using the observational catalogue NASA-Sloan Atlas and the Baldwin-Philips-Terlevich (BPT) and WHAN diagrams as diagnostics, no connection between AGN activity and environment is found based on 62 258 dwarf galaxies, although a weak connection cannot be refuted in a redshift-limited sample of BPT galaxies, while the IllustrisTNG simulation shows an increase in AGN occupation fraction of its 6 771 dwarf galaxies if they have recent mergers. Additionally, dense environments are found to be detrimental for AGN activity, but this finding may be due to numerical reasons. Machine learning does not rank environmental features highly for identifying AGN, but predicted AGN galaxies reside closer to a massive galaxy and denser neighbourhoods. Preliminary results indicate that the best model relies internal features. Other studies find multi-wavelength data provide the best venue to obtain a complete set of AGN in dwarf galaxies, and simulations are now utilising higher resolution and improved black hole (BH) modelling, enabling accurate evolutionary paths of dwarf galaxies. The seemingly contradictory results between different approaches can in part be explained selection bias (e.g BPT favours unobscured AGN), numerical effects (e.g overmassive BH seeding), and statistical framework used to quantify differences. Future work involves constructing a more complete and accurate sample of dwarf AGN, achieved through using multi-wavelength data, higher sensitivity observations like integrated field unit spectroscopy, and simulations with improved dwarf galaxy and BH modelling, tying together the many strings by a fine tuned machine learning approach.

Contents

1 Introduction **1**

1.1 Cosmology . . . 2

1.1.1 The beginning of extragalactic astronomy . . . 2

1.1.2 The Great Debate . . . 4

1.1.3 Leavitt, Hubble, and Lemaître . . . 5

1.1.4 Cosmic microwave background . . . 7

1.1.5 Dark matter . . . 7

1.1.6 Large scale structure . . . 8

1.1.7 Our current cosmological model . . . 9

1.2 Observations of galaxies . . . 10

1.2.1 Basics of observations . . . 10

1.2.2 Photometry and spectroscopy . . . 11

1.2.3 Large scale surveys . . . 12

1.2.4 Deriving galaxy properties . . . 14

1.3 Simulations of galaxies . . . 15

1.3.1 Simulation basics . . . 16

1.3.2 Cosmological simulations . . . 17

1.4 Galaxy evolution . . . 19

1.4.1 Hubble sequence . . . 21

1.4.2 Colour bimodality . . . 21

1.4.3 Tidal interactions and mergers . . . 23

1.4.4 External and internal feedback processes . . . 23

1.4.5 Mass dependent evolution . . . 25

1.5 Active galactic nuclei . . . 27

1.5.1 Discovery of AGN . . . 27

1.5.2 Anatomy of AGN . . . 28
1.5.3 Where are they found . . . 30
1.5.4 How to identify AGN . . . 32
1.5.5 Triggers . . . 35
1.5.6 Co-evolution with galaxies . . . 35
1.6 Dwarf galaxies . . . 36
1.6.1 General dwarf galaxy population . . . 37
1.6.2 Dwarf evolution . . . 38
1.6.3 Dwarfs in observational studies . . . 39
1.6.4 In simulations . . . 40
1.6.5 AGN in dwarf galaxies . . . 41
1.6.6 Research goals of this thesis . . . 43

2 Environments of dwarf galaxies with optical AGN characteristics 45
2.1 Introduction . . . 46
2.2 Data and methods . . . 49
2.2.1 Data and sample selection . . . 49
2.2.2 Classification diagrams . . . 52
2.2.3 Environment estimation . . . 56
2.3 Analysis . . . 58
2.3.1 KS-testing . . . 58
2.3.2 BPT and WHAN comparison . . . 64
2.3.3 Local neighbourhoods of dwarf AGNs, 10NN . . . 65
2.3.4 Immediate neighbourhood of dwarf AGNs, Δv_{NN} . . . 66
2.3.5 Visual inspection . . . 66
2.3.6 Other parameters . . . 69
2.4 Discussion . . . 73
2.4.1 SDSS fiber aperture bias . . . 74
2.4.2 On the environment and nearest neighbours . . . 76
2.4.3 On selection method bias . . . 80

2.5 Conclusions 83

3 Merger Histories and Environments of Dwarf AGN in IllustrisTNG 88

3.1 Introduction 89

3.2 Data and Methods 92

3.2.1 IllustrisTNG and Illustris 93

3.2.2 Dwarf galaxy selection 96

3.2.3 AGN selection 98

3.2.4 Time since last merger 103

3.2.5 Distance to 10th nearest neighbour 105

3.2.6 Kolmogorov-Smirnov testing 108

3.3 Results 109

3.3.1 On time since last merger 110

3.3.2 Current and past environments 111

3.3.3 Sampling size 111

3.3.4 TNG50-1 and Illustris-1 115

3.4 Discussion 117

3.4.1 Mergers as a significant trigger channel 117

3.4.2 Time lag and impact from past environments 118

3.4.3 Black hole requirement 121

3.4.4 Are control galaxies properly constrained? 127

3.4.5 Optimal KS sampling size 128

3.5 Summary 130

4 Dwarf AGN and their environments – a machine learning approach 135

4.1 Introduction 136

4.2 Data 138

4.2.1 NSA and MaNGA 138

4.2.2 Firefly 140

4.2.3 AllWISE 140

4.2.4 4XMM-DR11 . . . 141
4.3 Methods . . . 141
4.3.1 Dwarf galaxy selection and mass splits . . . 141
4.3.2 AGN selection . . . 142
4.3.3 Environment estimations . . . 143
4.3.4 Machine learning classification . . . 144
4.4 Results . . . 147
4.4.1 Cross-validation . . . 148
4.4.2 Feature importance . . . 151
4.4.3 Dwarf predictions - recall rate and precision . . . 151
4.4.4 Characterisation of dwarf galaxy distributions . . . 155
4.5 Discussion . . . 157
4.6 Future work . . . 158
4.7 Conclusions . . . 159

5 Conclusions **161**
5.1 Discussion . . . 161
5.2 Summary . . . 163
5.3 Future work . . . 165

Bibliography **167**

List of Figures

1.1 The Andromeda Nebula, By Isaac Roberts (d. 1904) - A Selection of Photographs of Stars, Star-clusters and Nebulae, Volume II, The Universal Press, London, 1899., Public Domain, https://commons.wikimedia.org/w/index.php?curid=51791 3

1.2 Power spectrum of the CMB. 6

1.3 Large scale structure as seen from SDSS 13

1.4 Large scale structure from two cosmological simulations, TNG100 and TNG300 at redshift zero. Gas density (left), gas temperature (center), and magnetic field amplitude (right) are shown for TNG300 while TNG100 is shown with dark matter density and gas density. 18

1.5 Hubble fork. Credit: NASA & ESA 20

1.6 Colour bimodality, from Baldry et al. (2004b). The left figure shows a colour-magnitude diagram with two distinct population sequences. The right figure shows the deconvolved and parameterized red and blue distributions with a green track showing where a galaxy with a certain specific mass would be located. 22

1.7 From Peng et al. (2010). Dominating quenching mechanism as function of mass and redshift. 26

1.8 AGN anatomy. The different parts are described in Section 1.5.2. (Credit: C.M. Urry and P. Padovani) 29

2.1 Magnitude versus redshift plot. The blue data points are all galaxies in NSA. The orange data points are low-mass galaxies (as defined in Section 2.2.1). There are clear magnitude edges in different redshift intervals, which is due to the completeness selection. 51

2.2 BPT diagram. The solid black lines are follow the Kewley et al. (2001); Kewley et al. (2006) classification diagram. However, no distinction is made between Seyfert and LINERS, and only pure AGNs are included in this sample, thus following the Kewley et al. (2001) classification. Three samples are plotted. The blue dots are all the low-mass galaxies in the NSA catalog. The red dots are the BPT-selected galaxies with $S/N_{ratio} > 3/\sqrt{2}$ on both emission lines ratios. The 'weak' BPT are galaxies with $S/N_{ratio} < 3/\sqrt{2}$. Especially Hβ is responsible for classifying a BPT-selected galaxy as weak ($\approx$ 87.5 per cent of all dwarf BPT galaxies in this sample has $S/N_{H\beta} < 3$). . . 53

2.3 WHAN diagram (For details, see Cid Fernandes et al., 2010, 2011). The solid black lines mark the different regions (from top left, clock-wise); Star-forming, strong AGNs, weak AGNs, and retired galaxies. Both weak and strong AGNs are included in the sample and no distinction is made between them. The blue dots are all the low-mass galaxies in the NSA catalog. The red dots are the WHAN-selected galaxies with $S/N_{ratio} > 3/\sqrt{2}$ on [N II]/Hα. The 'weak' WHAN are galaxies with $S/N_{ratio} < 3/\sqrt{2}$ 55

2.4 NSA Venn diagram showing the different selections. 57

2.5 BPT diagram with WHAN selected galaxies. The dots are colour-coded by their relative point density. The majority of the WHAN selected galaxies would have been classified as star-forming or composite SF/AGN using the BPT classification scheme. 60

2.6 WHAN diagram with BPT selected galaxies. The dots are colour-coded by their relative point density. The majority of BPT selected galaxies (*'AND'*-selected - N = 195) are considered strong AGNs in the WHAN diagram while the non-AGN WHAN-classified galaxies are roughly evenly split between retired galaxies and star forming ones. 61

2.7 Average values of emission line ratios and EW_{H_α} function of mass. The log ratio values are shifted to be in the same area while the EW is log and then scaled by 0.5. The data consists of 32 linear log scale mass bins and the bins with less than 300 galaxies are shaded in grey. 62

2.8 BPT and WHAN diagram showing mass trends. Each data point is the average values in 32 different mass bins and the size of the dot is scaled by the number of galaxies in that bin. Each bin has at least 300 galaxies in them unless surrounded by a black edge and otherwise contains between 312 and 11 581 galaxies. 63

2.9 *Left*: Projected spatial separation from the dwarf AGN galaxies to their 10th nearest neighbour and *right*: The absolute velocity difference between dwarf AGN galaxies and their nearest 2D separated neighbour (within $\pm 1\,000\,\mathrm{km\,s^{-1}}$). Three samples are plotted: Black dotted are BPT-selected galaxies, grey dashed are WHAN-selected galaxies while blue solid are galaxies that appear in neither of the other samples. Generally, there are no discernable differences between the three distributions in either case. The BPT bump near $600\,\mathrm{km\,s^{-1}}$ is not statistically significant. See Section 2.3.1 and Table 2.3 for statistics. . . 67

2.10 Example of cutouts of SDSS data. 6 different observational properties are shown here (only excluding spirals). Detailed information regarding visual inspection can be found in Section 2.3.5 70

2.11 Mass, redshift, and magnitude distributions. The black dash-dotted distribution is BPT-selected galaxies, the grey dashed is WHAN-selected ones while the blue solid is the NOT selection. Regarding mass, AGN galaxies are increasingly common towards higher masses while the NOT galaxies peak around log 9.3 $M_\odot$. For redshift, WHAN and NOT galaxies follow almost the exact same trend, though WHAN has a slight excess at higher redshifts. BPT galaxies are slightly favoured at lower redshifts, but overall follows the same trend. Lastly, on magnitude, AGN galaxies are in general brighter than the NOT galaxies. 71

2.12 AGN fraction as function of mass. The fraction is calculated as the number of galaxies fulfilling the respective AGN criteria divided by the total number of galaxies in that mass bin that also fulfill the S/N criteria outlined in Section 2.2.2. For high masses ($\geq 10^{11}\,M_\odot$), care has to be taken because of incomplete data, which is why there are no mass bins after $\sim 10^{11} M_*/M_\odot$ since a requirement is that there has to be more than 300 galaxies in one bin. . 72

2.13 Plot of fraction of a galaxy covered by the SDSS fiber aperture as function of redshift. The size of a galaxy is taken to be its petrosian 90 per cent light radius, R_{P90}, and is core is defined as $0.1R_{P90}$. The grey dots are all dwarf galaxies overplotted with median of different subsamples. Errorbars show the interquartile range. The subsamples are split into redshift bins with $\Delta z = 0.005$. The solid line at 1.0 equals the R_{P90} while the dashed one at 0.1 is $0.1R_{P90}$. 75

2.14 WHAN and BPT diagrams with weak BPT selected galaxies with dots colour-coded by their relative point density. These galaxies lie primarily in the retired region in the WHAN diagram while their positions in the BPT diagram are uncertain due to low S/N on the y-axis. 81

3.1 Black hole mass versus stellar velocity dispersion, σ. Two relations are plotted (Xiao et al. (2011) dashed line, Kormendy & Ho (2013) solid line) with their intrinsic scatter. Observations of black holes in dwarf galaxies (M_* between 8.5-9.5 log $M_\odot$) from Xiao et al. (2011) are in blue and Baldassare et al. (2020) are in orange and green – *a*: from previous work, *b*: from Baldassare et al. (2020) study. The copper 2D histogram shows TNG100-1 galaxies . . . 95

3.2 Spatial distribution of all subhalos only in top row and with selected dwarf galaxies bottom row projected onto three different planes (XY, XZ, and YZ plane). The gray scale background number density plot includes all subhalos while the coloured distribution is for dwarf galaxies. The data is split into 100 bins on each axis resulting in a bin size of 0.75×0.75 Mpc/h 97

3.3 Colour-magnitude ($u - r$ colour vs r magnitude) diagram showing SDSS dwarfs compared to TNG100 dwarfs with same mass selection criteria. Grey dots and black contour lines are NSA data while blue dots and contour lines are on dwarf subhalos from TNG100. The contour levels are at different levels between the samples since the sample sizes are different. 99

3.4 BH comparison histograms between simulations. Top row shows BH mass, middle row displays BH accretion rate, and the bottom row shows the density of local comoving gas of the BH. The columns display the full sample in the first row, Int AGN in the second column, and lastly non-AGN in the third column. TNG100-1 BH are in blue, TNG50-1 are orange, and Illustris-1 is green. 102

3.5 Time since last merger histogram for three selected subsamples: NOT – low mass galaxies with a black hole but no AGN activity in blue (Section 3.2.2), weak AGN galaxies in orange (Section 3.2.3), and intermediate AGN galaxies in green (also Section 3.2.3). Galaxies with no mergers have a TSLM equal to the age of the universe 104

3.6 Time since last merger (1:10 mass ratio merger) versus merger stellar mass 2D histogram. 106

3.7 Distance to 10th nearest neighbour histogram for selected snapshots. Blue is non-AGN galaxies, orange is weak AGN while green is intermediate. Strong AGNs are not included as this sample size is small. The grey background histogram is observational data from the NASA-Sloan Atlas ($M_* \leq 3 \times 10^9$ $M_\odot$, $z \leq 0.055$). The snapshots (from high to low) roughly corresponds to lookback times of 0.00, 0.48, 1.00, 1.98, 3.97, and 6.01 Gyr. 107

3.8 Distance to 10th nearest neighbour histogram with subject samples and their matched reference sample. Blue is the original subject sample, orange is the resampled subject sample while green is the matched reference sample. Top: Int AGN as subject sample and NOT as reference. Bottom: NOT as subject and Int AGN as reference. Errorbars are calculated as the spread of the averages in each bin of 100 resampling runs with a sampling size of 500. The peak at small distances (between 0.5 to 2 Mpc) for NOT disappears when it is used as a reference sample for AGN samples. 113

3.9 Colour histogram with subject samples and their matched reference sample - like Figure 3.8. Top plot is using Int AGN as subject sample versus non-AGN as reference sample. Bottom plot is in reverse. Errorbars are calculated as the spread of the averages in each bin of 100 resampling runs with a sampling size of 500. Few red ($u - r \geq 2.0$) NOT galaxies are selected when using AGN samples as subject samples. 114

3.10 Distance to 10th nearest neighbour evolution. Each line is the median of the D_{10} distribution of different samples at different snapshots. 119

3.11 Mass distribution of low mass galaxies with (blue) and without (orange) BH. Additionally, the FoF halo mass threshold (7.38×10^{10} $M_{\odot}$) for when a BH is seeded is shown as a black line. Galaxies with (without) a BH, 0.7 per cent (2.0 per cent) have a lower mass than the seed threshold and 99.3 per cent (98.0 per cent) have a higher mass. 122

3.12 Spatial distribution on the XY plane of low mass galaxies with (left) and without (right) BH. There are 100 bins on each axis and the number of subhalos in each bin is then counted. 123

3.13 Spatial distribution on the XY plane of low mass galaxies with (red) and without (blue) BH. There are 100 bins on each axis and summed up in a normalised histogram. Each pixel bin is given a colour corresponding to the ratio between the number of BH to no-BH galaxies with more blue meaning a higher number of no-BH galaxies. The black line in the histograms show the average density, i.e the density distribution if all galaxies were spread out evenly. The departure from the this distribution of the BH and no-BH distributions is calculated and shown next to the histograms. The residual of the no-BH distribution is higher for all axes indicating that they clump together moreso than BH galaxies. 124

3.14 KS-testing results of merger mass ratio. Details on how the p-value and its error is calculated can be found in Section 3.2.6. Colour indicates what the subject sample is with violet being all dwarf galaxies (i.e NOT+Weak+Int+Strong), blue being non-AGN, orange is weak AGN, green is intermediate AGN, and red is strong AGN. On the x-axis is the reference samples with the marker style indicating sample size. Background shading indicates a group of data points with the same subject and reference sample. For example, if you are to look up what the p-value is for non-AGN as subject and weak AGN as reference using a sampling size of 500, it is found as the orange open circle at the 8th tick mark on the x-axis. 132

3.15 Same as Figure 3.14 but for D_{10} for snapshot 99, 96, and 93 133

3.16 Same as Figure 3.14 but for D_{10} for snapshot 87, 75, and 62 134

4.1 Feature importance evolution from CV for [N II] BPT (top 4 figures) and [S II] BPT (bottom 4 figures) galaxies as a function of training mass, inner features only. 150

4.2 Recall rate and precision for dwarf galaxies for different mass bins. Errors are the standard deviation from the sub-classifiers of the CV testing. 153

4.3 Characterisation matrix of dwarf galaxies using intermediate 2 galaxies as training set. Each plot has four box plots, one for each of the different true/false positive/negative populations. 156

List of Tables

2.1 Completeness selection intervals. 50

2.2 Number of galaxies depending on choice of h. While the number of galaxies decreases with decreasing h, the results described in Section 2.3 do not change. 52

2.3 p-values of respectively 10NN and Δv_{NN} 2-sided KS tests. Each row has the subsample in the leftmost column as the subsample to be compared against a control sample from a subsample given by the column name. E.g, the test in row 1, column 2 is found from 152 random galaxies from all low mass galaxies and a matching galaxy (in mass, colour, and redshift) sample is found for each element from the BPT subsample. 'wBPT' is short for 'weak BPT'. 59

2.4 Number (fraction) of galaxies showing visual properties in AGNs (*'AND'* subsample) and a control sample. 69

3.1 Overview of relevant simulation parameters 94

3.2 First column is the simulation name, second one is the corresponding side length given in units of comoving kpc/h. Third column is the mass of a dark matter particle followed by gas cell/particle mass and lastly is the mass of the seeded black hole particle. Masses are in $10^4 M_\odot$ for easy comparison. 94

3.3 Number of subhalos for each AGN selection criteria. 98

3.4 The total number is N and how large a percentage of the total dwarf galaxy population is given in percentage in parenthesis. λ is the Eddington ratio. . . 98

3.5 Summary tables of smallest sampling sizes from KS results. 112

3.6 Summary tables of which smallest sampling sizes KS results are significant at for TSLM for different miniumum merger mass ratios (top table) and D_{10} at different snapshots (bottom table). Open dots, open triangles, filled dots, and filled triangles represent sampling sizes of 500, 1000, 1500, and 2000, respectively. Symbols in parenthesis indicates that the $p \leq 0.05$ is reached within error. For a full plots, see Figures 3.14, 3.15, and 3.16. Strong AGN are not included since no test reaches the threshold. Tests with weak AGN are limited to a maximum sampling size of 988. 112

3.7 Comparison of significant KS results between different simulations. 116

3.8 Columns denote the reference sample while the rows are for subject samples. Each cell is further subdivided into six cells with the columns being the different simulations (TNG100-1, TNG50-1, and Illustris-1 respectively) while the rows are the 10:1 at the top and D_{10} (z=0) is at the bottom. A labelled subtable is shown below (subject Int, reference NOT). A filled circle indicates that the test reached $p \leq 0.05$ while an open circle indicates $p \leq 0.05$ is reached within error. 116

4.1 Overview of AGN selection numbers of MaNGA data 142

4.2 Overview of outer features used for RF 145

4.3 Overview of inner features used for RF 146

4.4 Overview of F1 scores from cross validation for models using different AGN labels and feature sets using three different mass selections for training set. . . 149

4.5 Number of dwarf galaxies in different categories according to pre-existing labels and predicted label. The numbers are the average and standard deviation of the predicted numbers obtained from the 10 CV estimators. 152

1. Introduction

There are two outstanding questions regarding dwarf galaxies with active galactic nuclei (AGN): (1) What conditions are favourable for triggering AGN activity in dwarf galaxies (and are they the same as for massive galaxies), and (2) are the current diagnostic tools suitable for selecting AGN in dwarf galaxies? In order to include the relevant background for this subject, a wide range of data and physical processes are covered in this introduction. The introduction is divided accordingly. The first chapter gives a historical perspective on our understanding of the Universe and the observation of galaxies and concludes with our current understanding and cosmological model.

This is then followed by a section on observations of galaxies and how observations are made. In order to study galaxies, we must first collect light and derive information from them, but this can be done in many ways. Each method has its advantages and limitations, and these are important to keep in mind when interpreting the derived properties.

Simulations are a great way to test our theoretical models and provide a wealth of information that observations are unable to provide due to the fact that it is possible to store all information you are interested in. However, there are limitations and approximations that are similarly important to understand in order to be able to draw robust conclusions.

Following an explanation of simulation basics, a section is devoted to discussing the theoretical models of galaxy evolution that simulations are based on. There are models of both theoretical nature and of empirical nature, and combined they form our current understanding of galaxies, how they evolve, and their relevance.

The next section deals with an important element of galaxy evolution: active galactic nuclei (AGN). AGN is a central subject in this thesis, and after a brief historical rundown of their discovery and how we arrived at our present day understanding, their anatomy is explored. Today, we have a wealth of methods to identify them and several of these methods are explained. Another important question in this thesis is what can actually cause AGN

activity in a galaxy, and a number of processes and current research on this are discussed.

Last in the introduction is a section on dwarf galaxies. This population of galaxies has been less studied in observations and simulations due to their elusive and difficult nature to properly address, but their importance in galaxy evolution, the unique view they offer of the early Universe, and improved observations and simulations mean that they are now a popular research subject. In conclusion, the dwarf-AGN connection is discussed and the research questions that this thesis attempts to answer is given.

1.1 Cosmology

Cosmology is the setting in which galaxies are discussed, so understanding the history and current understanding of cosmology is essential to understand how we view galaxies today. This section includes a brief history section on how we came to the understanding that galaxies are extra-galactic objects, how we discovered what the Universe constitutes of today, and how structure has evolved throughout the history of the Universe.

1.1.1 The beginning of extragalactic astronomy

Around 100 years ago, the nature of *spiral nebulae* was unclear. Were these clouds on the night sky nearby gaseous nebulae or distant unresolved collections of stars (Fath, 1909)? Dr. Edward Arthur Fath found that these nebulae had continua and stellar absorption lines suggesting a collection of stars, although some nebulae had bright spectral lines associated with gaseous nebulae such as the Orion Nebula.

Some years later in 1913, V. M. Slipher published radial velocity observations of M31 showing $-300\,\mathrm{km\,s^{-1}}$ – a rather large radial velocity for a star cluster in the Milky Way. He followed up with observations of 15 spiral nebulae in 1915 which had radial velocities up to $1\,100\,\mathrm{km\,s^{-1}}$ leading him to believe that these nebulae were extragalactic. Contrary positions were held by Adriaan van Maanen and Harlow Shapley, which is showcased by van Maanen (1916) where he argued for observing internal motions in M101 and Shapley (1919) where he comments on the existence of external galaxies, which he calls the average proper motion velocities of nebulae measured by Wirtz (1916, 1917) and Curtis (1915) for *'appaling'*.

Figure 1.1: The Andromeda Nebula, By Isaac Roberts (d. 1904) - A Selection of Photographs of Stars, Star-clusters and Nebulae, Volume II, The Universal Press, London, 1899., Public Domain, https://commons.wikimedia.org/w/index.php?curid=51791

1.1.2 The Great Debate

In 1920, Harlow Shapley and Heber Curtis met at the Smithsonian Museum of Natural history to argue, amongst other things, whether the Andromeda Nebula is *inside* the Milky Way or *outside* [1]. Arguments in favour of Shapleys view of the Milky Way being the entire Universe were the fact that the distance required to the Andromeda nebula if it was a galaxy would be in the order of 500 000 light years. Furthermore, it would require Andromeda to be 50 000 light years in diameter – thus similar to the MW, and the absolute magnitudes of novae would be extremely bright. Such distances, sizes, and absolute luminosities were extraordinary claims and contrary to the current understanding of the universe and as such unlikely to be so. It would also imply that other spiral nebulae were even further away but of comparable size, which implies even more extreme distances and universe sizes. Lastly, Adriaan van Maanen had recently claimed that he recorded rotation in some spiral nebulae, which would indicate extreme rotational velocities if they were galaxies rendering the idea of *island universes* unlikely.

Curtis, while agreeing with Shapley that globular clusters are embedded in our own galaxy, believed spiral nebulae were a class apart (despite having globular cluster-like spectra) and distances of 500 000 to even 10 000 000 light years were correct. He supported these claims by considering the distribution of spiral nebulae are towards the galactic poles and not in the same area as where stars are most numerous (suggesting that spirals do not fit into any coherent scheme of stellar evolution), that most spirals seem to move away from us and at velocities of 1 200 km s^{-1} – a hundred times the velocities of diffuse nebulosities, thirty times that of stars, and even five times more than their spectral intragalactic counterpart, clusters.

Another disagreement between them was of the use and value of Cepheid variable stars for distance measurement. Shapley was fond of them while Curtis did not regard them highly, but ironically, the role of Cepheid in answering the spiral nebula question fell out against Shapley and in favour of Curtis.

[1] https://apod.nasa.gov/debate/1920/cs_nrc.html

1.1.3 Leavitt, Hubble, and Lemaître

A Cepheid variable star is a star that varies in luminosity over days to months. The period of the variability is closely linked to its luminosity – a relation that was discovered by Henrietta Swan Leavitt. This relation was used by Edwin Hubble in 1923 on cepheids in the Andromeda Nebula, and from the distance ladder he inferred that the nebula was actually far outside of the Milky Way at a distance of 285 000 parsec (929 000 lightyears). The universe suddenly became a lot larger.

Arguably, the great distances could also have been inferred from the 'large' redshifts measured by Slipher in 1917. The average line-of-sight velocity of nearby spiral nebulae was $570\,\mathrm{km\,s^{-1}}$ – 30 times larger than the average velocity of stars. However, large radial velocities alone were not compelling enough arguments, but the fact that galaxies appeared to be moving away from us was explored further by Hubble. He found a simple correlation between the distance from the Milky Way to a galaxy and the velocity at which they are receding from us in the form of $v = H_0 \cdot D$, where v is the recession velocity, D is the proper distance to the galaxy, and H_0 is the Hubble constant, which Hubble calculated to be around $500\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$. Modern values of this constant is between $68 - 71\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$ (Planck Collaboration et al., 2020).

This observation that the further away a galaxy is called the the Hubble–Lemaître law. Lemaître had published similar results two years earlier, but it did not receive much attention. He took this observation a step further and suggested that the universe is expanding – a view that is central in the current cosmology paradigm. This suggestion was not without problems, though. If the universe is expanding today, it must have been smaller in the past and in fact have originated from a single point.

Having the Universe starting expanding from a single point gives a mental image of a large explosion, and this was commented on by Fred Hoyle in 1949 where he explained that in this hypothesis *[...] all the matter in the universe was created in one* ***big bang*** *[...]*, and this term caught on in the 1970's when the expanding universe hypothesis was becoming widely accepted due to emerging evidence in support of it. One vital discovery was that of the Cosmic Microwave Background Radiation (CMB).

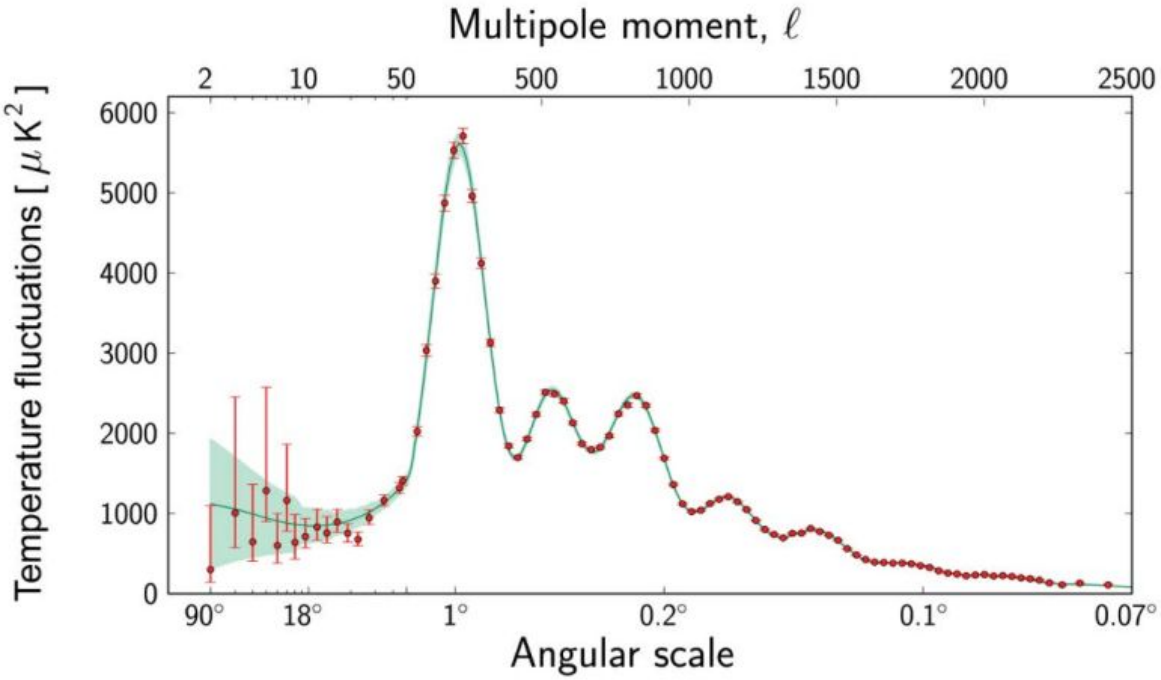

Figure 1.2: Power spectrum of the CMB.

1.1.4 Cosmic microwave background

In short, the CMB is the first light that decoupled from baryons in the early and expanding universe (around $z \sim 1200$,). The decoupling happened due to the universe cooling allowing hydrogen (and helium) and electrons to recombine letting light escape. While the matter of the universe continued to evolve and collapse further, the conditions that the light escaped from were imprinted on it. As such, the CMB provides valuable insights into the initial conditions and cosmological parameters of the Universe such as matter and radiation densities, as well as topology and density pertubation scales.

The peaks in the power spectrum (1.2) corresponds to different physical processes affecting the Universe. Small monopoles ($10 \leq l \leq 100$) corresponds to large spatial scales and by the time of recombination, these anisotropies had not had time to develop much. As such, this part reflects the initial conditions, and the temperature variations of these scales are closely linked to the initial density pertubations.

At larger monopoles ($100 \leq l \leq 1000$), there are several acoustic peaks that are caused by gravity-driven acoustic oscillations happening before recombination. Larger structures oscillate more slow and vice versa, and by the time the photons decoupled to the baryons, the phase of these oscillations were frozen in. The oscillations must be primarily driven by a matter component, but it is inconsistent with a baryonic matter only component and requires an even larger additional matter component that only interacts gravitationally. This type of matter is called dark matter (DM).

From a galaxy evolution perspective, the initial dark matter distribution is important for a number of reasons. Overdense regions are the seeds of future galaxy clusters and groups, and galaxies' environments have impact on parameters such as star formation history and galaxy interactions. While baryonic matter is also important, its lower abundance makes it of less importance for large scale gravitational structures.

1.1.5 Dark matter

One of the first hints towards the existence of an additional gravitational constituent came not long after Hubble established several nebulae to be extragalactic – i.e they are galaxies like the

Milky Way. Galaxies in clusters of galaxies seemed to have orbital velocities far greater than what could be inferred from the visible matter (from a mass-to-light ratio argument, proposed by George Abell and Fritz Zwicky) and orbital velocities of stars in galaxies themselves also did not seem to agree with the apparent mass distribution. Zwicky mentioned that some invisible/dark matter must be present in order to account for this discrepancy in 1933.

Nowadays, the existence of *dark matter* has been confirmed by several other observations and it is required to explain a number of phenomena regarding galaxies such as their rotation curves and cluster/group dynamics. Furthermore, dark matter constitutes around 84 per cent of the matter budget of the Universe (Planck Collaboration et al., 2020) and is thus more responsible for structure formation and evolution than baryonic matter. It is not surprising that the earliest cosmological simulations only focused on modelling dark matter to replicate the large scale structure of the Universe.

As mentioned previously, another type of observation that further supports the existence of dark matter is the temperature anisotropies of the CMB, which was discovered in 1965 by Arno Penzias and Robert Wilson. Small fluctuations in temperature of the CMB are evident of gravitational instabilities, and the strength of the fluctuations are inconsistent with a baryonic matter only component. However, dark matter is able to account for the missing matter.

1.1.6 Large scale structure

While the CMB is the frozen out light from when the Universe became transparent and has not changed since then (except cooled due to expansion), the matter component continued evolving. Gravity works to collapse matter, but in a completely homogeneous Universe, all forces cancel out. However, the minute density pertubations (both initially and from the acoustic oscillations described in previous sections) initiated the collapse to form the structures we see today. Since dark matter constitutes roughly 84 per cent of the matter budget of the universe, the collapse and structure formation is primarily driven by dark matter.

Smaller structures formed first and evolved into more massive structures with time. After the decoupling of photons and baryonic matter, baryonic matter was now also free to collapse further since the radiation pressure from the photons was gone. Baryonic matter was attracted

to the dark matter structure that had been assembling somewhat unimpeded until now, and this clustering is what became the first stars and galaxies. It clustered onto a network of sheets, filaments, and knots and forms the basis of the web-like structure we see today. The knots/intersections are the densest parts and it is here that groups and clusters of galaxies reside today. There are also large empty areas called voids where little to no galaxies reside.

Together, the web like structure with knots and filaments and sheets between them is referred to as the large scale structure of the universe, and while it tells us a lot about the initial conditions and the beginning of the Universe, it also determines the future of galaxies. Galaxies evolve over time through various processes, and several of the processes depend on the environments of the galaxies.

Further collapse of the large scale structure will not happen due to the expansion of the Universe which serves as a counteracting mechanism to gravitational collapse. Furthermore, the expansion is not only constant, it is accelerating (Riess et al., 1998). The component responsible for the accelerating expansion is described by a constant, Λ, in the field equations that describe the expansion of space. This term is an energy term in those equations and is called dark energy since it is invisible and we do not know exactly what it is – only that it is required to fit our current understanding of the Universe.

1.1.7 Our current cosmological model

From the discovery of the expanding Universe in the beginning of the nineteen hundreds to the fine tuning of the cosmological parameters of today and discovery of invisible forces and components, the Big Bang model has been through many iterations. However, it has stood its ground somewhat solidly for almost a century – especially due to its compatibility with general relativity. The current broadly accepted cosmological model is called the *Lambda-cold dark matter* model (or Λ-CDM).

Lambda refers to the dark energy component which dominates the energy budget of the Universe today while the cold dark matter refers to the dark matter component. The 'cold' relates to the velocity of dark matter particles and the cold model (i.e slow moving) is the preferred one due to being able to build structures from a bottom-up approach. The different

components contribution to the energy budget is written as the density parameter Ω, where (for a flat Universe)

$$\Omega = \sum_{components} \Omega_{component} = 1 \tag{1.1}$$

Planck Collaboration et al. (2020) give Ω values for dark energy, dark matter, and baryonic matter as $\Omega_\Lambda = 0.6894$, $\Omega_{DM} = 0.2601$, and $\Omega_b = 0.0489$. For the Hubble constant, H_0, they give a value of $67.70\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$. These values are derived from observations of the microwave and sub-mm sky and thus based of the CMB, but these values also give predictions for the clustering of galaxies and clusters. However, observations of galaxies is not just a simple matter of using a measuring tape and a scale and the intricacies of galaxy observation will be discussed in the next chapter.

1.2 Observations of galaxies

This section will discuss the basics of observations – how to collect light and converting it to a quantifiable signal, and then moves on to highlight the strengths and weaknesses of different methods of observing. While observation methods can be divided into two (photometry and spectroscopy), telescope and instrument design allow for both highly specialised and flexible observations. This section focuses on optical observations, but many discussion points are also valid for infrared and ultraviolet observations. X-ray and radio observations are very different and will only be discussed briefly.

1.2.1 Basics of observations

The basis of modern astronomy is a telescope, the purpose of which is to collect light and either amplifying or magnifying it (or both). Pointing the telescope towards objects that are otherwise too faint or too small become bright enough or large enough to be seen and studied. Once the light is collected, it needs to be recorded.

The earliest astronomical observations were hand-drawn, but this practice was slowly replaced with the advent of photographic plates in the 1800s onto which the light could be recorded. In the 1980's, charged-coupled devices (CCD) had become the staple (Janesick, 2001), and the development since then has led to CCDs with millions of pixels, high efficiency,

and little noise.

The design of the telescope depends on what wavelength is being observed. For optical observations, a main mirror collects the light, and the diameter of the mirror (called the aperture) is the limiting factor in how much light is possible to collect. Another way to manipulate the telescope to suit different needs (other than changing the aperture size) is to change the focal length. Using the same aperture and sensor size, increasing the focal length gives a smaller field of view, which is useful to study single objects, while decreasing the focal length gives a larger field of view, which is useful for multiple source observations.

The collected light is then fed to an instrument which determines what will happen with the light. While instruments are usually designed to answer a specific science question, the types of different instruments can be broken down to two types: spectrographs and imagers. Spectrographs diffract the incoming light onto the CCD and provides a spectrum of the observed object while imaging provides an energy or photon count of the whole observed field, usually in a specific colour/filter.

The resolution of an observation means something different in imaging and spectroscopy. In imaging, the resolution is usually taken to mean the angular resolution – how many arcseconds does a single pixel cover? In spectroscopy, the resolution is given as how many nm or Ångstrom, $\Delta\lambda$, a pixel covers. However, $\Delta\lambda$ is different at different wavelengths, so resolution is often given as the dimensionless resolving power $R = \frac{\lambda}{\Delta\lambda}$. Typically, $R \leq 1\,000$ is referred to as low resolution, $1\,000 \leq R \leq 10\,000$ is considered intermediate resolution while above 10 000 is high resolution.

1.2.2 Photometry and spectroscopy

Photometry is the method of measuring the amount and distribution of light of an area of the sky and the light is usually filtered through bandpasses/filters that only allow a certain range of wavelengths through. While the light is integrated over the interval and thus losing fine details, this type of observation can cover a large area and many sources at once. Additionally, several filters are usually employed simultaneously and the combination of each of these data points can give a general idea of the shape of the spectra, but spectral line information is lost.

Spectroscopy is the measure of energy flux per unit wavelength, and it is done by sending the light through a prism-like object that diffracts it. The diffracted light is then projected onto a CCD with each column corresponding to a different wavelength. Before reaching the prism, the light passes through a spatial limiter (such as a slit) which ensures that only light from the object of interest is collected. Spectroscopy reveals spectral lines since the resolution is high enough to reveal such features. Having spectral lines is useful for galaxies since it provides accurate redshifts and information about dynamics such as rotational velocities.

Although spectroscopy is usually limited to observing single objects at a time, a technique exists that covers multiple objects at once - multifibre spectroscopy. For technique, a bunch of fiber optic cables are connected to holes in a plate, and each fiber/hole observes a different point in the sky. The light from each cable is sent through a prism and projected onto a row on the CCD – thus each row on the CCD represents a single object. This enables building large catalogues of sources with spectroscopy, but the light from each source is integrated over the whole fiber coverage and thus spatial details are lost.

However, spatial details can be recovered by using integrated field units (IFU). The basic idea behind this is that each fiber is divided into further fibers, and each smaller fiber projects onto their own row in the CCD. This method provides even more detail about the dynamics and different parts of a galaxy (e.g difference in metallicity or SFR between the core and outer parts), but it does require a longer exposure time and more CCD rows so less galaxies are observable at the same time.

1.2.3 Large scale surveys

One kind of astronomical observations is large scale surveys. The objective of these surveys is to cover a large area of the sky and characterise many objects simultaneously in that field. This is in contrast to dedicated observations of single objects. Building large catalogues provides a good statistical basis for demographic surveys and details about the distribution of galaxies, which can help constrain cosmological parameters such as initial density fluctuations. Similarly, these surveys provide information about how common and important different processes are. For example, having one galaxy with AGN activity and an old stellar population

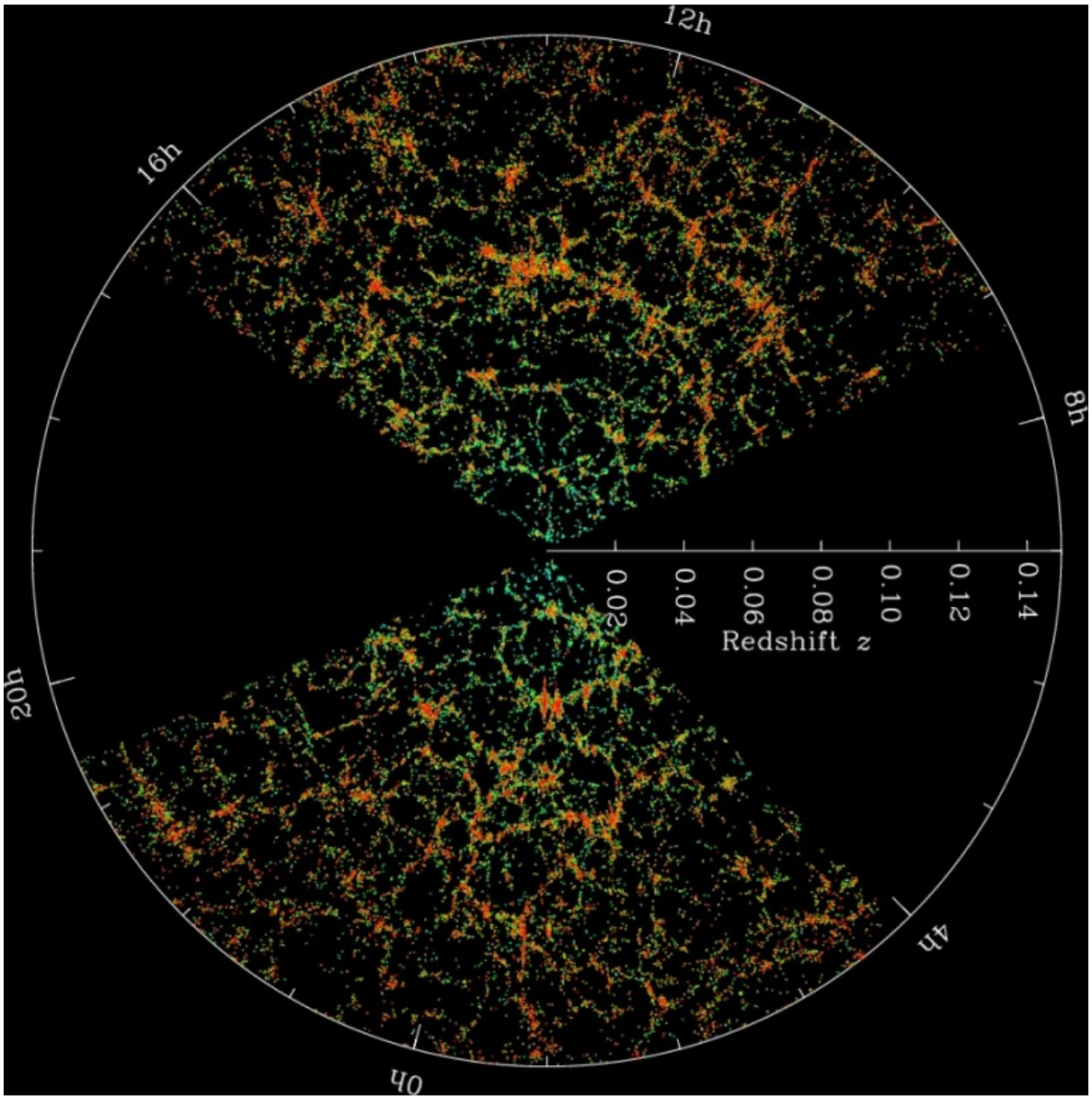

Figure 1.3: Large scale structure as seen from SDSS

may be indicative that red and dead galaxies are favourable for AGN activity, but if it is the only one in a sample of a hundred galaxies, the correlation is more likely spurious.

Examples of recent large scale surveys are the SDSS (York et al., 2000) and UltraVISTA (McCracken et al., 2012). The SDSS is now in its fourth phase (SDSS-IV Blanton et al., 2017) and previous iterations have collected both optical images and optical and near-IR spectroscopy of the northern high Galactic latitude sky. UltraVISTA is a photometric survey focusing on ultra deep observations in optical to infrared bands. SDSS consists of multiple types of surveys. One survey is the Mapping Nearby Galaxies at APO (MaNGA) is an IFU survey taking spatial spectral measurement of around 10,000 nearby galaxies.

When observing galaxies, the only property that is directly measured is the light output. Other properties such as stellar content, star formation rate, and gas content are derived from interpretations of these measurements. One method of inferring galaxy properties is fitting its spectral energy distribution with combinations of stellar distributions.

1.2.4 Deriving galaxy properties

While spectroscopy provides fine details about a galaxy, it is generally more time intensive than photometry/imaging. A less demanding way of obtaining galaxy properties is spectral energy distributions. This method exploits the fact that the slope between different points in a galaxy spectra depends on parameters such as stellar composition, age, gas fraction, and redshift. Spectroscopy, however, can give details about the kinematics of the components of the galaxy and information about the central black hole by measuring emission lines and their widths.

To derive properties from photometry, the points in the spectra are measured in different filters and data from multiple telescopes can even be combined. For example, visible light telescopes can measure the red and blue light and from the slope between these points, the ratio of young to old stars can be inferred. However, the dust content of the galaxy can not as easily be inferred from visible observations in which case infrared telescopes like WISE can be utilised.

Then, a fitting algorithm combines galaxy templates with different stellar populations with

various metallicities and ages. The galaxy templates are made from models that synthesise a range of different stellar populations. The core of these stellar synthesis models is the mass distribution of newly formed stars (also called *initial mass function* (IMF)) and how these populations evolve (stellar evolution models). The fitted SED can then give information about stellar mass of the galaxy, its stellar population age, and even redshifts.

The derived properties are thus reliant on the theoretical assumptions behind the fitting templates. For example, using an IMF that is bottom-heavy (i.e produces more low mass stars) for a dusty galaxy may yield a higher stellar mass because the infrared radiation from the dust may be confused to be stellar radiation instead. A top-heavy IMF could easily overshoot the blue and UV radiation parts of the SED and thus require a lower stellar mass.

By measuring spectral lines with spectroscopy, several properties can be measured to a higher accuracy. For example, redshift can be accurately determined by comparing the peak wavelength of known emission line to their laboratory measured peak. A measured peak value of Hα ($\lambda = 6563\,\text{Å}$) at e.g $\lambda = 7000\,\text{Å}$ yields a redshift of $z = \frac{\lambda_{obs} - \lambda_{emit}}{\lambda_{emit}} = 0.067$. Such accuracy in redshift is hard to obtain with photometry only although it is possible. Furthermore, the abundance of different elements can be inferred from the strength of the emission lines (e.g Pilyugin et al., 2012), and ratios between different emission lines can reveal details about their excitation mechanism (e.g Baldwin et al., 1981). Absorption lines are suitable for measuring column densities amongst other things and is thus a good tool for measuring abundances.

The theoretical assumptions behind the interpretation of galaxy observations and properties can be put to the test by doing a numerical time evolution of a galaxy and see if it ends up reproducing or matching observations. It works the other way as well – simulations are able to predict observable parameters that then can be searched for. It is therefore important to know how galaxies are simulated, which will be discussed in the next section.

1.3 Simulations of galaxies

Simulations of galaxies involve a numerical time evolution of a galaxys's constituents. For example, one of the earliest simulations was by Holmberg (1941) who investigated the tidal

disturbances of two stellar systems passing each other. Each system was composed of $N = 37$ light bulbs, each of which represented a single mass element. The light emitted from the light bulbs represented the gravitational force, so at each light bulb/mass element, the light intensity was measured in all directions to estimate the overall 'gravitational' acceleration. The model was then moved forward in time and each mass element moved to their new positions.

A decade later, these N-body simulations were made on programmable computers which increased the processing time significantly. Even then, Lindblad (1960) mentioned that the choice of $N = 160$ is not limited by storage but *keeping the computing time within reasonable bounds*. He was similarly trying to replicate the structure of a galaxy – specifically a barred spiral structure.

Nowadays with increased computing power, not only has the number of simulated bodies increased but the size, scale, and detail have as well. Furthermore, modern simulations model not only gravity but also dark matter, dark energy, ordinary matter, and central black holes over several Gyrs. Still, the scope of present day cosmological simulations is limited in order to keep computing time within reasonable bounds.

1.3.1 Simulation basics

Given the way the different constituents (dark matter, baryonic matter, and dark energy etc) interact with other parts, each part is usually modelled somewhat differently. Dark matter is usually modelled through an N-body approach while baryonic matter are usually modelled in a Lagrangian smoothed particle hydrodynamics (SPH) fashion or a Eulerian mesh-based hydrodynamics with adaptive mesh refinement (AMR) one.

Regardless of what is being modelled, a single mesh cell or particle represent a collection of what they model. For example, masses of DM particles are in the order of $\sim 10^3 - 10^9\,\mathrm{M}_\odot$, and stellar particles (particles that represent one or more stellar populations) can be of similar sizes.

The physical size of the simulations are referred to as the side length or box length. This is the size of a single side of the 3D space being simulated, so the total volume being simulated is the box length cubed. Similarly, the particle count is often given as N^3. For example, the

IllustrisTNG100-1 has a box length of 75 Mpc/h with 2180^3 (DM/gas) particles.

For this thesis, central black holes are also of interest. They can be included in simulations in various of ways. Either, they exist at the beginning of a simulation or be spawned, or *seeded*, when their host galaxy is sufficiently massive. The next question is then how they should interact with their environment. In large scale simulations, the scale on which they operate is too small to be resolved, so they are modelled in a semi-analytical way. For example, gas accretion can be modelled by an Eddington-limited Bondi-Hoyle accretion[2]. The Eddington accretion rate is the point where the gravitational force is equal to the radiative pressure caused by accretion, and accretion rates higher than this is assumed to be unfeasible in simulations.

1.3.2 Cosmological simulations

Cosmological simulations involve simulation boxes that range from tens of Mpc to hundreds of Mpc depending on what science questions they attempt to answer. One of the earliest ones, the Millennium Simulation (Springel et al., 2005b) aimed to replicate the large scale structure observed by surveys such as SDSS and thus had a large box side ($500\,h^{-1}$ Mpc) but large particles sizes ($8.6 \times 10^8\,h^{-1}\,M_\odot$).

Attempting to replicate the large scale distribution of galaxies means that galaxies need to be properly identified in simulations. Several methods exist for this, and a common tool is the Friends-of-Friends (FoF) algorithm. FoF identifies groups of particles by using a linking length to check if they are connected. Linked particles are then assigned to a unique halo.

A large linking length is useful for identifying clusters and groups while a smaller linking length can be used to identify substructures/subhalos such as galaxies, although much more complicated subhalo finders exist (see Onions et al., 2012, for a comparison). The identified halos and subhalos can then be stored in a group catalogue, which is simulation equivalent of an observational galaxy catalog like the NASA-Sloan Atlas.

An example of a recent cosmological simulation is the IllustrisTNG project. 3 different runs exist: TNG300, TNG100, and TNG50, with the number indicating the box length in Mpc. There are $2500^3, 1820^3, 2160^3$ particles of each type with $m_{DM}/10^5 = 590, 75, 4.5$

[2]A standard accretion estimate that relies on the sound speed and pressure of surrounding gas as well as mass of the black hole

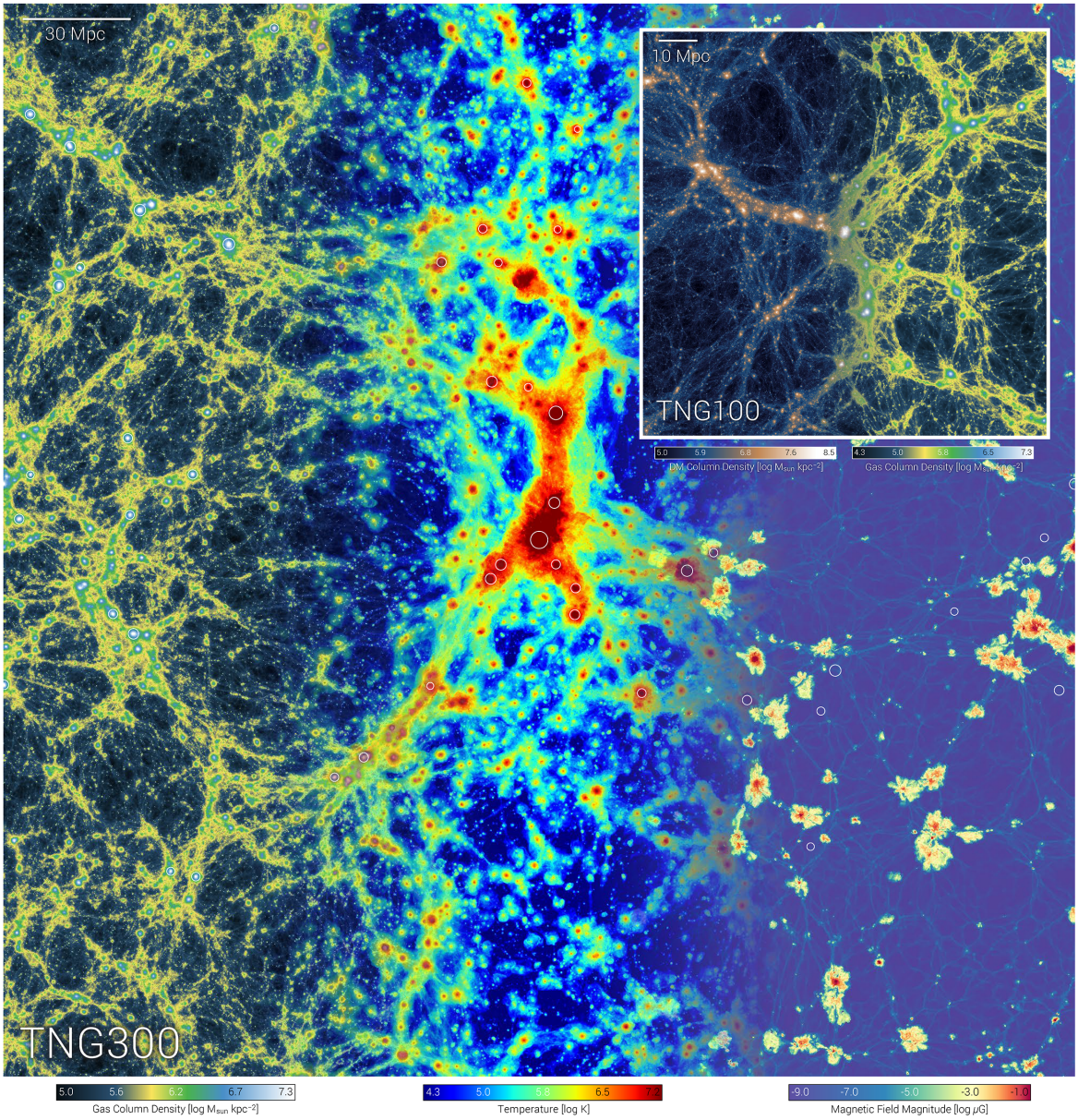

Figure 1.4: Large scale structure from two cosmological simulations, TNG100 and TNG300 at redshift zero. Gas density (left), gas temperature (center), and magnetic field amplitude (right) are shown for TNG300 while TNG100 is shown with dark matter density and gas density.

and $m_{gas}/10^4 = 1100, 140, 8.5\,M_\odot$, respectively. Particles and gas are evolved using the AREPO code, in which dark matter is modelled in an N-body fashion while gas is evolved in smoothed particle hydrodynamics in a unstructured Voronoi-mesh that moves the fluid in a quasi-Lagrangian way. They employ a FoF algorithm to identify halos and use the SubFind algorithm (Springel et al., 2001) to identify subhalos/galaxies.

Blacks holes do not exist at the beginning of the simulations, but when halos reach a certain mass threshold ($M_{halo} \geq 7.38\times10^{10}\,M_\odot$), they are seeded a BH of roughly $M_{BH} \sim 10^6$ $M_\odot$. Accretion onto the SMBH is Eddington-limited Bondi-Hoyle accretion, and the energy release is inputted into the surrounding gas as either thermal or kinetic energy depending on the efficiency of accretion. At high efficiencies ($\dot{M}_{Bondi}/\dot{M}_{Edd} \geq 0.1$), the SMBH is in quasar-mode and inputs thermal energy while below is wind-mode and the energy output is kinetic energy.

As mentioned in Section 1.2.4, the way that galaxies evolve over time is based on a theoretical framework. This framework involves many different mechanism over a wide range of scales and certain processes are not worth it or too intrinsic to simulate in detail so approximations or semi-analytical models are used instead. Nevertheless, these models are based on an elaborate set of theories which will be described in the next section.

1.4 Galaxy evolution

Though galaxies appear static in the sky, they are quite dynamic and evolving entities, but most changes in a galaxy are on such large timescales that they are not observable to us. However, it is possible to piece together the whole sequence of a galaxy's life by assuming different galaxies are in different stages of their evolution. One of the earliest suggestions of an evolutionary track of galaxies was by Edwin Hubble in 1926, where galaxies start out as blobs and develop structure such as spirals arms over time. Today, the picture of galaxy evolution is much more complicated where parameters such as environment, merger history, feedback processes, and redshift are taken into account.

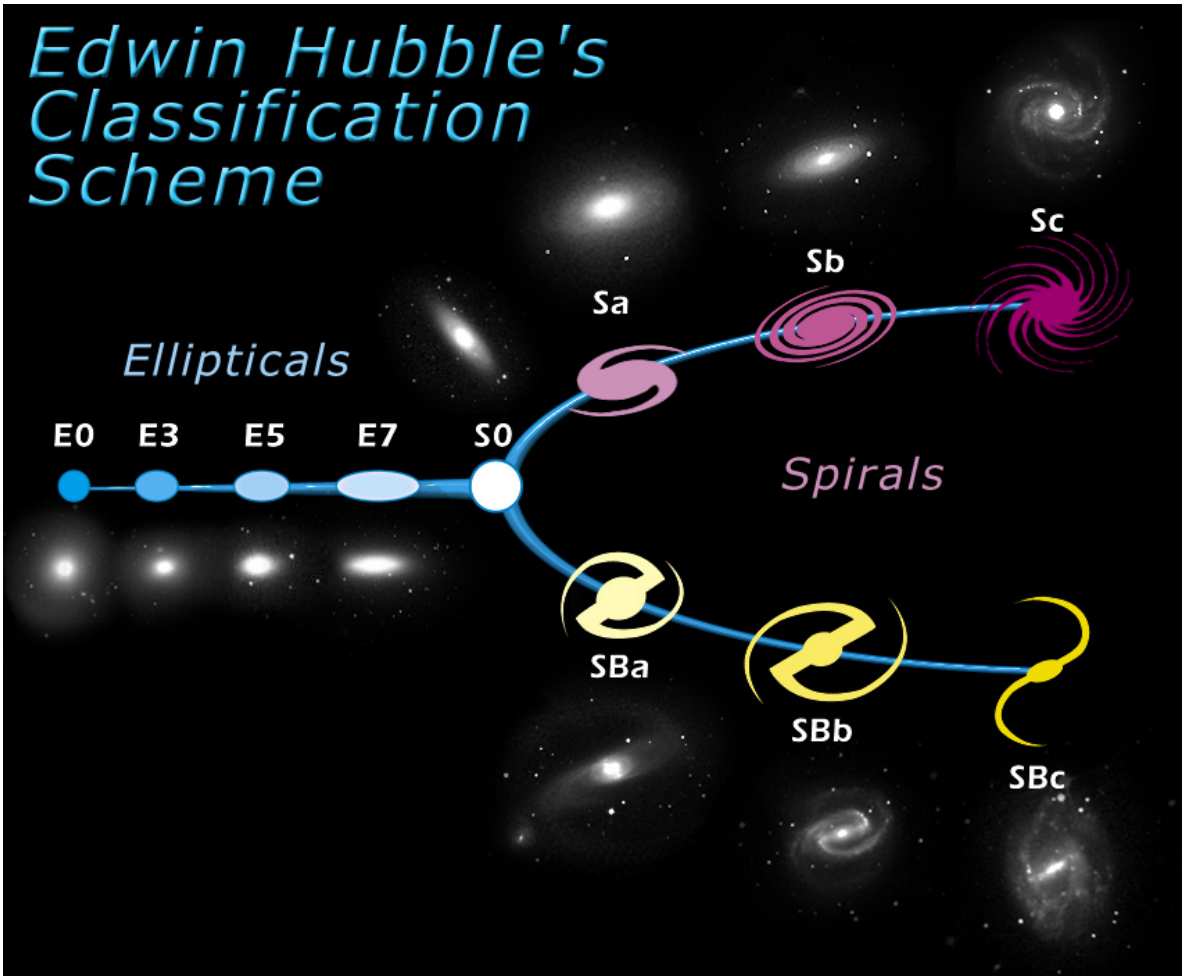

Figure 1.5: Hubble fork. Credit: NASA & ESA

1.4.1 Hubble sequence

An easily identifiable property of galaxies are their appearance or *morphologies*. The perhaps most well-known morphological classification diagram was invented in 1926 by Edwin Hubble in what is today known as the Hubble sequence or Hubble tuning fork diagram. It describes the apparent evolution from spherically elliptical E0 galaxies to increasingly elliptical E{1-7} ones before becoming lenticular S0 galaxies forking out to become barred/unbarred spiral galaxies that are decreasingly less wound/have more spiral arms. This progression towards spiral galaxies has given rise the labelling elliptical galaxies as early type (since they are early in their evolution) and late type galaxies are spiral galaxies (since the are later in their evolution).

1.4.2 Colour bimodality

Another way of categorising galaxies is by inserting them into a colour-magnitude diagram. This reveals two distinct populations in a u-r versus r diagram – a red and blue population (Holmberg, 1958; Roberts & Haynes, 1994; Baldry et al., 2004a). Generally, the red population consists of elliptical galaxies that have little to no star formation and only an old population of stars remains. Furthermore, their gas content is often low and thus the light output is dominated by old red stars. Blue galaxies tend to be spiral galaxies with active star formation which means there are plenty of young stars present (Larson & Tinsley, 1978). These young stars give rise to the blueness of these galaxies. In between the two populations is a valley that very few galaxies inhabit. This is called the green valley and is thought to be galaxies in transit to become red galaxies. This transformation is short lived (roughly equal to the life time of massive stars) which explains why so few galaxies are found there.

Similarly to the Hubble tuning fork, the colour-magnitude diagram also separates elliptical and spiral galaxies, but it is not based on morphology but rather on stellar populations and gas content. Furthermore, there is an environmental difference between these populations with red galaxies being more common in high density environments such as galaxy clusters while blue spiral galaxies are found in groups. These observations lead to a different theory of galaxy evolution compared to Hubble: Galaxies start out as irregular or spiral galaxies (if they

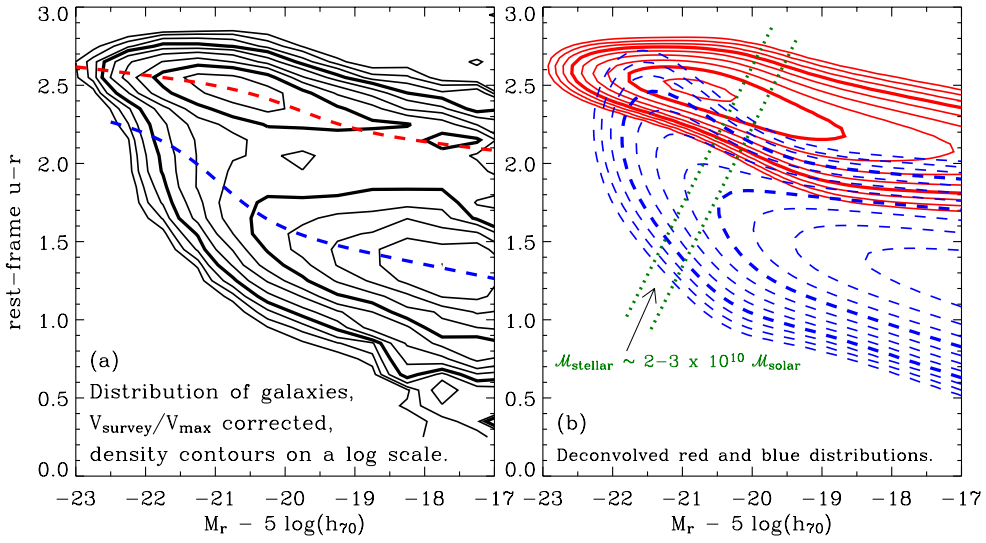

Figure 1.6: Colour bimodality, from Baldry et al. (2004b). The left figure shows a colour-magnitude diagram with two distinct population sequences. The right figure shows the deconvolved and parameterized red and blue distributions with a green track showing where a galaxy with a certain specific mass would be located.

are massive and undisturbed enough), but their environment can cause their gas reservoirs to deplete (either through stripping them or triggering rapid star formation) and randomise stellar orbits. The resultant effects are elliptical stellar orbits, gas loss, and stellar population of primarily old stars – an archetypal elliptical galaxy (Moore et al., 1998).

1.4.3 Tidal interactions and mergers

An important way that galaxies evolve over time is through the interaction with other galaxies. Galaxies passing each other exert a tidal force on one another, distorting their morphologies. These disturbances can cause otherwise stable configurations to become unstable leading to major changes in the galaxy (Toomre & Toomre, 1972; Moore et al., 1998; Kormendy et al., 2009).

One such change is the star formation rate (SFR), which can increase by one or two orders of magnitude (Larson & Tinsley, 1978). Galaxies with a highly increased SFR are referred to as *starburst galaxies*, and they often show highly irregular morphologies – indicative of a merger being in progress. The starburst process deplete the cold gas reservoirs both through formation of new stars and energy input from supernovae. After the starburst phase, stars in the galaxies have highly random orbits and new stars struggle to form.

Tidal interactions do not necessarily mean merging, and one effect on the gas is that it just loses angular momentum. Tidally interacting systems are seen having more frequently active galactic nuclei, and a possible explanation is that the gas needs to lose angular momentum to sink to the centre of the galaxy where a supermassive black hole (SMBH) resides. The frequency of encounters depend on the environment that a galaxy resides in.

1.4.4 External and internal feedback processes

Given that galaxies are affected by other galaxies, it is not surprising that the galactic neighbourhood that they reside in also affects them. As a rule of thumb, groups usually refer to collections of 10-50 galaxies while the number of members in clusters range between 50-1000+ galaxies. Indirectly, being in a cluster rather than a group or in the field increases the number of tidal interactions and mergers, and the processes described in Section 1.4.3 occur more frequently. Directly, there are several components of a galaxy group or cluster

that influence the evolution of a galaxy.

One component is the intergalactic medium (IGM). This is the catch-all name for the material and gas that is between galaxies in clusters and can make up to 90 per cent of all the baryonic matter in clusters. As galaxies move through the cluster, they encounter the IGM, which increases in density towards the centre of the group or cluster. The IGM exerts a force on the gas contents of the galaxies moving through it while the stellar contents are left unaffected. This effect is called ram-pressure stripping (Gunn & Gott, 1972).

Other effects from the environment can be inferred from studying the galaxy populations differences between the two environments. Galaxy clusters have more red and elliptical galaxies while blue spiral galaxies dominate group environments, which suggest that the star formation histories for the average galaxy depends highly on the environment. However, not only external processes affect galaxies, internal processes and feedback mechanisms do as well (Kauffmann et al., 2004; Kormendy et al., 2009).

Generally, the two internal mechanisms that can impact the evolution of galaxies are stellar feedback consiting of stellar winds and supernova (SN) feedback (Larson, 1974; Dekel & Silk, 1986; Klypin et al., 1999; Kormendy et al., 2009) and AGN feedback (Fabian, 2012). SN feedback is thought to be especially important for the formation and evolution of dwarf galaxies (Larson, 1974; Dekel & Silk, 1986) since their low gravitational potential makes it difficult for them to prevent their gas reservoirs from being blown out of the galaxy by SN winds. AGN winds and the energy input can similarly push out gas and heat cold gas, but compared to supernovae, this process dominates in high mass galaxies (Schawinski et al., 2007; Fabian, 2012; Kormendy & Ho, 2013) and thought of to be too weak in dwarf galaxies to be important in their evolution.

Feedback are important processes in regards to the gas content of galaxies – both when it comes to the temperature of the IGM but also the existence of it. When galaxies are left with little to no cold gas reservoirs, their star formation is suppressed and they turn into *passive galaxies*. Processes that either strip or heat these reservoirs are collectively called quenching processes, and they can be either external ones like ram-pressure stripping or internal ones like SN feedback. While these processes are described individually, the act of quenching is

often a complex interplay between several of the processes happening either simultaneously or consequently.

Nevertheless, Peng et al. (2010) argue that environmental (i.e external) and feedback (i.e internal) processes produce unique signatures that make it easy to associate observations with what processes are affecting a given galaxy. Another factor that is a strong indicator how a galaxy will evolve is its mass.

1.4.5 Mass dependent evolution

A parameter that strongly decides how galaxies evolve is their mass. Peng et al. (2010) found that the effects of mass evolution is separable. They focus particularly on quenching and find that mass quenching (i.e quenching that is directly related to galaxy mass through some physical mechanisms that they do not identify) is relevant for galaxies with stellar masses above $10^{10.2}$ $M_{\odot}$while satellite galaxies (usually dwarfs and especially galaxies below $10^{10.2}$ $M_{\odot}$) tend to be quenched by environmental effects more so than their higher mass counterparts.

Direct observational evidence for quenching is hard to come by since it is difficult establishing a causal link between the proposed quenching processes and the quenching itself – often because the quenching itself takes place over typical timescales of 10^7 years. Some environmental effects like ram-pressure stripping and interactions can be observed directly in the form of jellyfish galaxies and irregular galaxies, respectively. Such galaxies often show enhanced SF (Peng et al., 2010; Vulcani et al., 2018) and subsequent quenching has to be inferred from evolution models.

Regarding AGN feedback, Feruglio et al. (2010) found outflows from a nearby AGN (Mrk 231) expelling more gas than what is being used for star formation, and suggested this could lead to a depletion of gas reservoirs in $\sim 10^7$ years. In simulations, Croton et al. (2006) have similarly found that both outflows and ionisation winds from AGN can quench star formation in especially the most massive galaxies. However, Penny et al. (2018) found that several isolated dwarf galaxies showed signs of quenching while hosting an AGN at the same time highlighting the fact that AGN feedback is also relevant for dwarf galaxies. To explore this further, understanding AGN and what drives them will be explored in the next section.

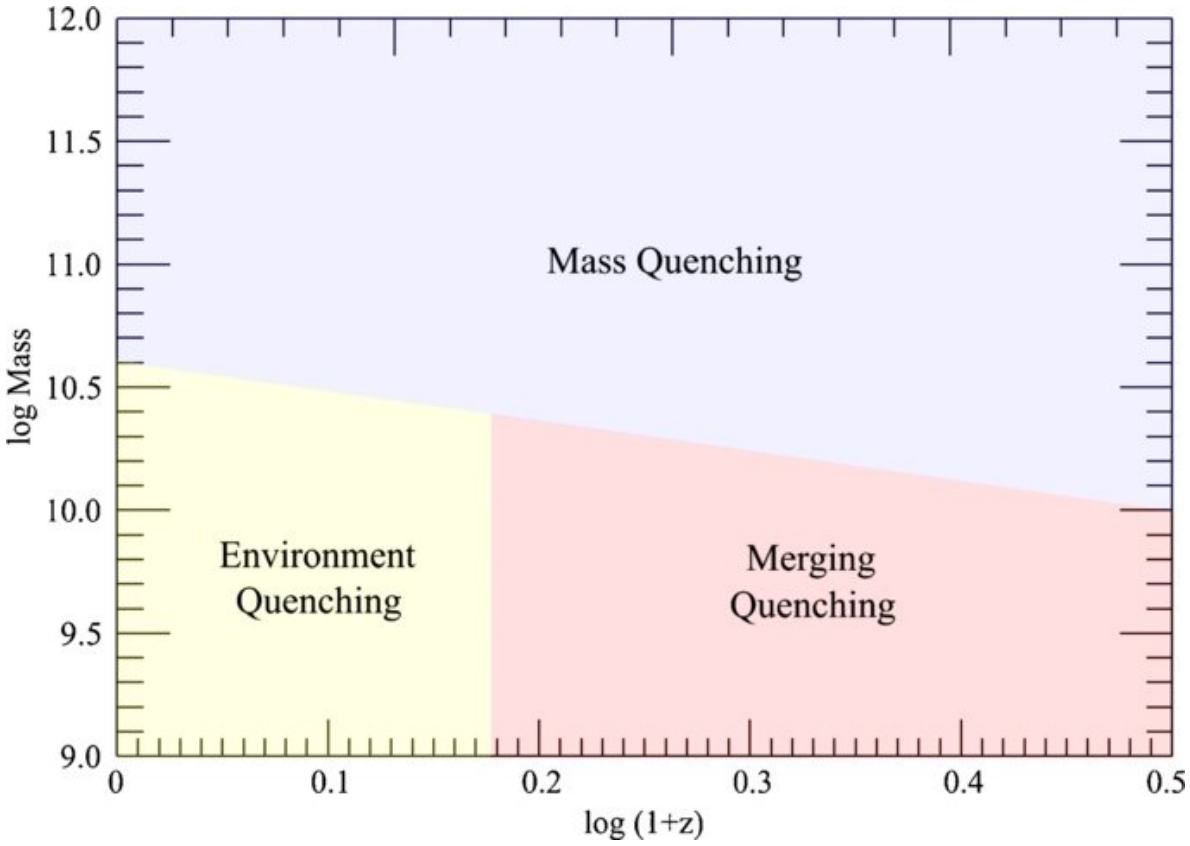

Figure 1.7: From Peng et al. (2010). Dominating quenching mechanism as function of mass and redshift.

1.5 Active galactic nuclei

Active galactic nuclei (AGN) have been mentioned several times already and are at the core of this thesis. In short, they are black holes residing in the centre of galaxies that are several hundred thousand to millions – even billions solar masses heavy. Surrounding them is gas and material, some of which is being accreted and heated. The energy output from this process can be comparable to that of the host galaxy which is a testament to the impact they can have on host galaxy.

However, their existence was not recognised until the late 1960's – decades after that galaxies were found to be extragalactic collection of stars. Today, they play an important part in galaxy evolution, especially for high mass galaxies.

1.5.1 Discovery of AGN

Two of the first ones to notice that certain galaxies had unusual spectral lines in the nucleus were Slipher (1917) and Hubble (1926). It was noted that certain galactic nuclei showed spectra similar to planetary nebulae rather than composite stellar spectra. In 1943, Carl Seyfert conducted a systematic study on the nuclear emission of 6 spiral nebulae/galaxies (Seyfert, 1943) and found that their emission could be described as G-type (Sun-like) stellar like with superimposed emission lines. He remarked that the widths of these emission lines corresponded velocities upwards of $8\,500 kms^{-}1$ from a doppler shift interpretation. Furthermore, certain forbidden transitions were also present in the nuclear spectra which hint towards different excitation environments that are not associated with regular stellar nebulae.

Despite these extreme velocities and unusual emission lines, focused efforts into figuring out the nature of these mysterious conditions was were not carried out until the advent of radio observations. However, in the infancy of radio observations of AGN, the link to the unusual nucleus emission features was not clear. In fact, the naming reveals as much: quasi-stellar objects (QSO, or quasars), named so because they were first confused with and resembled stellar sources, and their strange spectra earned them the prefix of *quasi*.

The first QSO was discovered in 1960 by Allan Sandage, but the breakthrough in the study of quasars came in 1963 when Schmidt (1963) obtained a redshift of $z = 0.16$ for object 3C

273 challenging the notion that quasars were stellar objects in the Milky Way. Accepting the origin of quasars to be extragalactic was also difficult because it would imply that the nuclear region of the galaxy to be 100 times optically brighter than the luminous radio galaxies discovered so far.

1.5.2 Anatomy of AGN

The different parts of an AGN will be described in this section with the names of the parts bolded. The engine of an AGN is the **supermassive black hole (SMBH)** in the centre. These black holes have masses in the range of millions to billions of solar masses ($\sim 10^6 - 10^9\,M_\odot$) with an accretion disk of hot, ionised gas surrounding it. The accretion disk feeds the black hole a few $M_\odot\,yr^{-1}$ of material and the infall of this gas is the main source of energy in the form of release of gravitational potential energy.

In simple models, the **accretion disk** as assumed to be axisymmetric and thin and described as a Keplerian disk. The innermost part of the disk is the hottest and decreases outward and it is subsequently where most emission emanates from. The thermal emission from the inner part of an accretion disk surrounding a $10^8\,M_\odot$ SMBH will peak near a wavelength of $\sim 100\,$Å, which is in the extreme UV or soft X-ray regime. Further out, the temperature decrease and give rise to thermal emission peaking near UV and optical wavelengths, and thus the combined emission from the inner and outer part of the disk resembles not thermal emission but rather that of a powerlaw ($F_\nu \propto \nu^{1/3}$).

Surrounding the accretion disk is a region where gas clouds reprocess the ionising UV radiation from the accretion disk into broad emission lines. This region is called the **Broad-Line Region (BLR)**. Emission lines here show Doppler-broadening of typically $5\,000\,km\,s^{-1}$, which is inconsistent with thermal velocity dispersions. Rather, these clouds are thought to orbit the accretion disk and SMBH in distances of a few light days. The emission lines from this region differ from HII regions, planetary nebulae, and another emission line region in AGN primarily due to the fact that electron densities collisionally suppress forbidden emission lines.

Further out is the **Narrow-Line Region (NLR)**, which are similarly ionised by the central

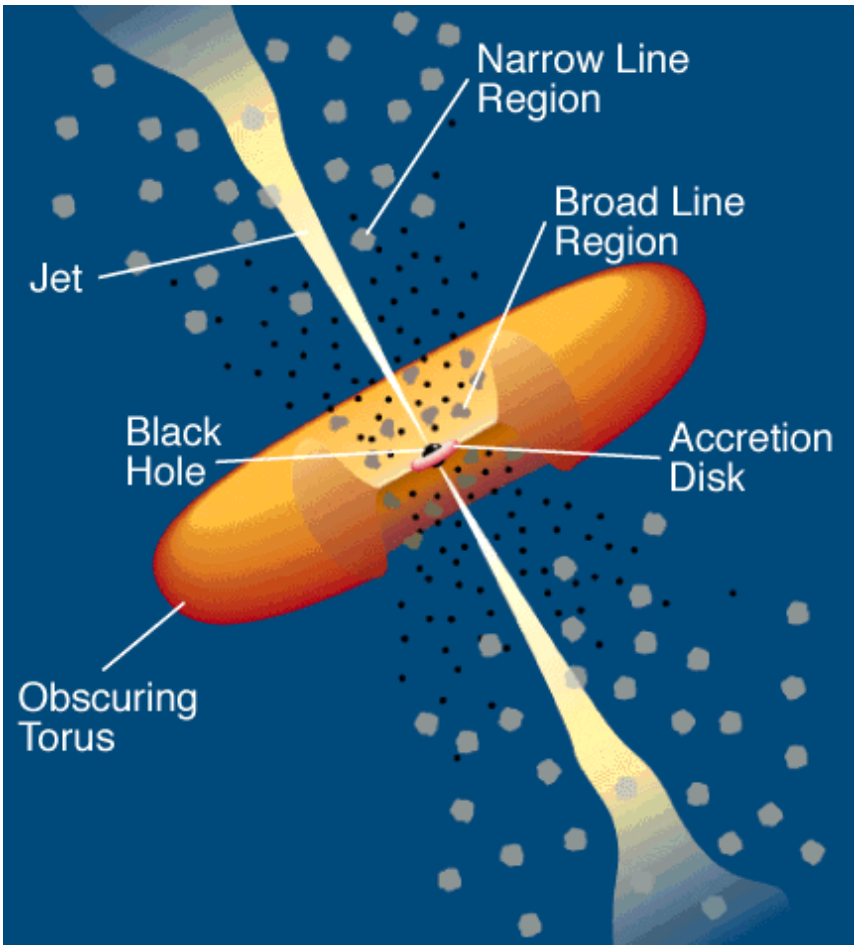

Figure 1.8: AGN anatomy. The different parts are described in Section 1.5.2. (Credit: C.M. Urry and P. Padovani)

source although in a non-isotropic matter and where the electron densities are low enough to allow for forbidden transitions. The orbital velocities are low (few hundreds of km s^{-1}) making the emission lines narrow from which this region gets its name from.

Surrounding the accretion disk is a dusty and optically thick **torus** that obscures radiation coming from the accretion disk and the BLR. The torus is believed to be one of the main contributors to the infrared radiation as the dust and grains in the torus absorbs UV and visible light and re-emits it as infrared radiation.

Extending out from the black hole along its axis are **jets** of ionised particles moving at relativistic velocities and can extend to even hundreds of thousands of light years. The particles move around the magnetic field lines and give rise to large amounts of synchrotron radiation. Not all AGN are associated with a jet since some are lacking or have little radio emission.

In the current black hole paradigm of AGN, depending on which components are observable and the components strength, the observational signature is different. For example, Seyfert galaxies are the 'vanilla' AGN galaxy, but some of them do not have broad lines because the obscuring torus is blocking line-of-sight of the BLR. This leads to a different classification; Seyfert galaxies with broad emission lines are called Seyfert 1 and galaxies without broad emission lines are Seyfert 2 – despite the fact that they are believed to be the same type of object. Similarly, radio galaxies and blazars are two classes of AGN with radio emission and are thought to be the same type of AGN but where the jets in a blazar is pointing directly towards the observer. This idea that the underlying nature of the different types of AGN is the same is called the unified model of AGN.

1.5.3 Where are they found

The host mass of AGN is not randomly distributed, and the likelihood of hosting an AGN increases with host mass (Kauffmann et al., 2003; Dunlop et al., 2003; Aird et al., 2012; Pimbblet et al., 2013)), but the AGN fraction also depends on selection method. Kauffmann et al. (2003) found that the peak AGN fraction is near $\log M_*/M_\odot = 11.0$ while e.g Pimbblet et al. (2013) found it to be at least increasing to $\log M_*/M_\odot = 11$.

While the exact position of the AGN fraction peak is still debated, there is a consensus that it is near $\log M_*/M_\odot = 11.0$ and that the low mass end ($\log M_*/M_\odot \leq 9.5$) has an almost negligible number of AGN. However, this picture is changing with more and more AGN found in dwarf galaxies in large scale surveys (e.g Reines et al., 2013) and more sensitive studies finding hard-to-identify AGN (e.g Baldassare et al., 2015). As such, the AGN fraction for dwarf galaxies may be underestimated due to the typical diagnostic tools for AGN are calibrated towards high mass galaxies.

Regarding morphology, the AGN fraction is highest in galaxies with a strong bulge component (e.g Hubble type E-Sb) with fractions of between 50-70 per cent while later types only have a fraction of 10 per cent or less. It is worth noting, though, that morphology also overlaps with host mass and environment, which are other deciding parameters in AGN activity. Furthermore, several studies find that bar galaxies have increased AGN activity (Oh et al., 2012; Chown et al., 2019), but this relation is disputed by e.g Cheung et al. (2015). Galloway et al. (2015) note that while they find an increased AGN fraction in barred galaxies, it is only a minor enhancement.

Highly distorted galaxies that are in the process of merging or strongly tidally interacting with another galaxy are strongly associated with increased AGN activity (Hernquist, 1989; Ellison et al., 2019; Kristensen et al., 2021), although some find that mergers and tidal interactions are associated with highly luminous AGN (Kocevski et al., 2012; Marian et al., 2020) or the AGN activity happens at a later stage (Ellison et al., 2008; Hopkins, 2012; Satyapal et al., 2014; Shabala et al., 2017).

A more disputed connection to AGN activity is the environment. Some find no connection (Miller et al., 2003; Kristensen et al., 2020), while others do (Monaco et al., 1994; Deng et al., 2012), although some suggest only a weak connection (Xin & Deng, 2021; Wethers et al., 2022). For example, Pimbblet et al. (2013) found that towards larger virial radii from cluster centres the AGN fraction increases. Man et al. (2019) found the contrary – AGN are more often found in denser environments, but they note that the evidence for this is marginal. If anything, this may even be due to more frequent interactions and mergers in groups.

Kauffmann et al. (2004) found that AGN with strong [O III] emission reside in low density

environment twice as often as in dense environments, and Silverman et al. (2009) examined X-ray selected AGN and found them also preferring under-dense regions. This can be taken to mean that the gas which needs to be driven to the core is unlikely to have been stripped away and thus being able to feed and maintain an AGN.

One explanation of the disagreements in literature is that they arise due to differences in selection method, environmental measures, and bias corrections. Indeed, Amiri et al. (2019) note that the AGN fraction increases with mass faster in clusters than in voids suggesting that failing to correct for mass properly can lead to wrong correlations. Man et al. (2019) suggests to leave out passive and dead galaxies in comparison studies since their cosmological histories cannot provide insight into environmental effects, and in doing so can eliminate the strong correlation between passive galaxies (i.e non-star forming) and environment.

Depending on which selection method is used thus has an impact of the population being examined. The next section will go over the different selection methods in different wavelengths that currently are being used as well as a brief characterisation of the AGN they select.

1.5.4 How to identify AGN

Given that different types of AGN have different observable components, finding and identifying AGN is not straight-forward. While a galaxy may appear to harbour an AGN in optical observations, it may not show up in infrared, or vice versa. Therefore, different identification methods are employed depending on which wavelength regime is being looked at. For example, a flux limit can be used for X-ray identification, but association to an optical source is usually required. Optical identification can be done by identifying broad emission lines, but often emission line ratios are used as a more robust method.

X-ray

Very few astrophysical processes produce a constant stream of X-ray photons, and the soft X-ray (0.5-2 keV) background is believed to be dominated by Seyfert 1 and quasars (Schmidt et al., 1998). As such, any constant soft X-ray source is a strong indication of AGN activity.

Seyfert galaxies, however, do not dominate the hard X-ray background (2-10 keV), but there is some indication that other types of AGN dominate this region (e.g Jackson et al., 2012).

Some contamination exist from processes such as X-ray binaries (Aird et al., 2017; Birchall et al., 2020) and star formation (Lehmer et al., 2016; Aird et al., 2017) so low X-ray fluxes can be ambiguous, which means that SMBHs accreting at low rates are more difficult to identify. Latimer et al. (2021) suggest that up to 1 350 dwarf AGN are expected at low redshifts, most of which are not known currently.

UV and optical

In optical wavelengths, spectroscopic confirmation of nuclear broad emission lines (such as Lyα and C IV) is often used to infer AGN activity. However, the BLR can be obscured and thus the broad emission lines are not visible. In such cases, narrow emission lines can also be used, but high SFR in a galaxy can give similar signals and is therefore a more uncertain identification technique.

To overcome this difficulty, emission line ratios are commonly used. There are two commonly used emission line ratio diagrams used to identify AGN: The Baldwin, Phillips, and Terlevich (BPT) diagram (Baldwin et al., 1981) and the $W_{\mathrm{H}\alpha}$ versus [N II]λ6584/Hα (WHAN) diagram (Cid Fernandes et al., 2010). The BPT diagram comes in several flavours and has been modified by e.g Kewley et al. (2001) and Kauffmann et al. (2003) to further categorise galaxies into Seyfert/Low ionisation, nuclear emission-line region (LINER) and composite galaxies, respectively. The physical reasoning behind the diagrams is the fact that different excitation mechanisms (e.g shock-excitation, photoionisation, H II regions, or planetary nebulae) affect certain emission lines differently, and using the ratio between emission lines reveal the underlying excitation mechanism.

The regular BPT diagram makes use of four different emission lines (Hα, Hβ, [O III]λ5007, and [N II]λ6584), which makes classification robust since it takes more underlying physics into account, but also misses weak sources since not all emission lines may have been measured at a high signal-to-noise (SN) ratio. The WHAN diagram only makes use of Hα and [N II]λ6584, both of which usually have high SN-ratios and as such can classify more

galaxies with certainty. However, it is prone to contamination from star formation and may misclassify high SF galaxies – especially dwarf galaxies.

Infrared

There are several parts of an AGN that emit infrared radiation with the accretion disc being responsible for a large part of the continuum from UV to near-IR energies. The dusty torus further out absorbs the radiation, is heated, and re-emits it as infrared radiation at wavelengths longer than $\sim 1\,\mu$m. At energies lower than $\sim 50\,\mu$m cold dust usually associated with star formation can dominate the IR radiation, and this difference between accretion disk induced and SF induced IR can be utilised in AGN identification. Since the light originates from the torus, mid- and far IR identification is useful for identifying Type 2 galaxies where the optical and the X-ray identification features often are obscured.

This selection technique is thus not based on emission lines but rather on the colour of the infrared radiation. Since colour is calculated from the difference of two photometric measurements in different bandpasses, infrared selection diagrams differ from from telescope to telescope. For example, selecting AGN using the Wide-field Infrared Survey Explorer (WISE; Wright et al., 2010; Jarrett et al., 2011) is different from that of using the Spitzer telescope (Stern et al., 2005) since their observation channels have different central wavelengths and widths (Jarrett et al., 2011; Yan et al., 2013). Therefore, defining the boundaries for AGN in colour-colour diagrams for mid-IR is still an ongoing effort (Blecha et al., 2018; Satyapal et al., 2018).

Radio

The main radio emission in AGN is from synchrotron emission originating from the jets. The jets themselves consists of plasma with electrons moving at relativistic speeds, and although jets are a fundamental part of the unified AGN picture, not all AGN have radio emission. One explanation is that these 'radio-quiet' AGN do not have a jet (Padovani, 2017; Radcliffe et al., 2021) unlike their 'radio-loud' counterparts. Nevertheless, high luminosity extended radio emission is a clear identification fingerprint, but at lower luminosities, the radio signal can be

significantly contaminated by SF processes in massive SF galaxies (Radcliffe et al., 2021).

1.5.5 Triggers

The nature of what causes AGN activity is a hotly debated topic. Many suggested mechanisms can both be used to explain the presence and lack of AGN activity, and it is difficult to resolve the physical space to accurately determine what is happening. Zhang et al. (2021) provide an overview of current research, but the locus of all these processes is explaining how to drive gas to the central part of the galaxy.

The triggers can be divided into two processes: internal and external. Internal triggers arise from asymmetries in galactic structure that induce a torque which forces gas inwards and feeds the central regions. One such asymmtric structure is a galactic bar (Oh et al., 2012; Chown et al., 2019), but the effect of bars on AGN activity is heavily debated (Arsenault, 1989; Mulchaey & Regan, 1997; Oh et al., 2012; Galloway et al., 2015; Goulding et al., 2017; Alonso et al., 2018).

External triggers are processes such as mergers and tidal interactions. In fact, many of the same processes affecting SF and galaxy evolution are also invoked in regards to AGN trigger mechanisms. There is an increase in AGN activity seen in merging systems in both observations and simulations (Kristensen et al., 2021), but establishing a connection to environment is more elusive. This may be due to not properly considering the evolutionary stage of AGNs as well as biases in both selection, observation, and environmental measures.

1.5.6 Co-evolution with galaxies

The simple observation that the more massive a galaxy is, the more massive its black hole is, poses an interesting question: Do galaxies and their black holes evolve together and by the same processes? Clues to this question can be found by narrowing down that the black hole mass scales more tightly with the bulge velocity dispersion (Ferrarese & Merritt, 2000; Gebhardt et al., 2000; Kormendy & Ho, 2013). Kormendy & Ho (2013) remark that this tight relation is still found when replaced the velocity dispersion with the bulge mass. This tight relationship suggests that galaxies (or at least their bulges) and their black holes co-evolve.

Deviations from this relationship also exist. Galaxies in the process of merging have

'abnormally' small black holes compared to their velocity dispersion, and a possible explanation of this is that the feeding gas is scrambled and becomes unavailable for feeding the BH. Overmassive black holes also exist, and these generally reside in old galaxies that turned elliptical a long time ago (Kormendy & Ho, 2013). As for the mechanism that enabled this early feeding frenzy, it remains an open question. These deviations hint towards a more nuanced evolution of both galaxy and black hole where they affect and regulate each other while also grow through entirely separate channels – at least for regular galaxies.

Historically, the impact of AGN on dwarf galaxies has been disputed. Other processes such as environmental effects and supernova feedback have a more dominating role in the evolution of dwarf galaxies, but recently, more observations find that many dwarf galaxies host AGN and some even being strongly affected by their AGN (by either being quenched, highly star forming, or little baryionic content). Thus, simply ignoring the role of AGN for dwarf galaxies is not reasonable. In order to explore this further, the next section will cover our current understanding of dwarf galaxies.

1.6 Dwarf galaxies

All of the processes mentioned in the previous sections are accepted as regular parts of galaxies' lives, but it is more uncertain how important and what role these effects have in early parts of galaxies' lives. In order to study this, dwarf galaxies constitute good laboratories since they are pristine relics of the early universe and are the building blocks of the large galaxies of today.

However, dwarf galaxies are difficult to work with. Their lower surface brightness compared to regular galaxies means that they require deeper or longer observations in order to be seen, so they have not received much attention in surveys. In large scale simulations, they are computationally expensive to resolve so their cosmological histories and their role in galaxy evolution are not well-explored.

Therefore, studying dwarf galaxies is an interesting undertaking – both because of how they themselves evolve and are affected by astrophysical processes, but also because of how they are tied to galaxy formation and evolution. This section ties together the physical

processes discussed in previous sections and puts them in a dwarf galaxy perspective.

1.6.1 General dwarf galaxy population

The highest concentration of dwarf galaxies are in our local galactic neighbourhood, the Local Group (LG). Especially low luminosity ones are mostly found in our neighbourhood (Mateo, 1998), but this is more a testament to the difficulty finding and observing ones that a further away. Clusters and groups are also environments rich with dwarf galaxies, and numerically, they outnumber spiral and elliptical galaxies. However, their meagre size and mass do not measure up to that of the massive galaxies in these environments.

Not all dwarf galaxies are the same and the term covers a wide range parameter space. The Magellanic Clouds are the closest dwarf galaxies to us, but the Large Magellanic Cloud (LMC) is similar to low-luminosity spiral galaxies in terms of mass, luminosity, and size. The smallest know dwarf galaxies are similar to some stellar clusters, so the definition is not strict. Historically, a magnitude cut could be used, e.g $M_{b,v} \geq -18$ (Grebel, 1999; Gallagher & Wyse, 1994), but this does not take into account the e.g their internal dynamics, morphology, nor environment.

This lack of clear definition of what a dwarf galaxy is also shows in in the many subclassifications that exist; early-type dwarf spheroidals (dSphs), dwarf ellipticals (dEs), late-type starforming dwarf irregulars (dIs), very-low surface brightness, ultrafaint dwarfs (uFd), centrally concentrated actively star-forming BCDs, and ultracompact dwarfs (UCDs) that are globular cluster sized but with 'regular' dwarf galaxy spectra.

One of the clearest separation of dwarf galaxies is between star forming systems with a large gas reservoir (usually dIs) and those without (usually dEs and dSphs). This demarcation has almost existed since the first reviews of dwarf galaxies (Hodge, 1971), and even before the advanced models of today, stellar content and age were inferred from the red colours of dSphs versus the bluer colours of dIs suggesting that dSphs are made up of much older stars, if not entirely made up of them, and shares many properties with the brightest and most massive globular clusters of the MW.

1.6.2 Dwarf evolution

In order to explain where the large galaxies we see today come from, the best explanation invoked is the concept of hierarchical structure formation. This theory states that smaller structures/galaxies collapse first and merge other similarly small objects, which then merge with another object and so on. This is in contrast to a top-down approach to structure formation where the largest objects are formed first and then the small details settle later.

An example of the role of dwarf galaxies comes from out own galaxy, the Milky Way, through galactic archeology. Galactic archeology is a field that is about understanding the origins of stellar populations with different chemical and dynamic properties. From this, it can be inferred that the MW has had two infall periods that has given rise to two separate components (Chiappini et al., 1997).

What this means for dwarf galaxies is that they constitute the earliest building blocks of galaxies we see today and understanding them – how they evolve and change – will help in understanding regular galaxies today. A good laboratory for studying dwarf galaxies are the Local Group dwarf galaxies, which bears resemblance to the larger population of dwarf galaxies (Weisz et al., 2011), but their proximity to us makes them easier to observe in detail.

While it seems reasonable to assume that all galaxies affected by the same processes behave similarly, the non-linearity of e.g the stellar mass/halo mass relation suggests otherwise (see Section 1.4.5). Especially the smaller potential wells of dwarf galaxies make them more susceptible to energy feedback processes and environmental effects (e.g Woo et al., 2013; Fillingham et al., 2016).

Hodge (1971) differentiates between dwarf elliptical galaxies and dwarf irregular galaxies, noting that the two populations generally show significant different characteristics when it comes to shape, stellar content, environment, and gas content. These differences hint towards a different cosmological history. Indeed, this is a notion that still exists today and a split still being investigated (Tolstoy et al., 2009; McConnachie, 2012). One explanation of this split is similar to the one between more massive galaxies – that there is an evolution from one type to another. Tolstoy et al. (2009) noted that there are transition type galaxies between dIs and dSphs suggesting an evolutionary pathway, and that the transition galaxies mark the average

mass or point in a dwarf galaxy's life where it start losing its gas reservoirs.

Regardless of type or the exact nature of the evolution history of different dwarfs, the majority of all dwarf galaxies are very metal poor. This is somewhat counterintuitive since they have a fairly high SFR throughout their life (Tolstoy et al., 2009), but there are a couple of mechanisms invoked to explain this such as metal rich outflows (Fujita et al., 2003), variations in IMF, or accretion of metal poor ISM. This metal poor gas content of the dwarf galaxies make them pristine systems in the sense that they have not changed much since their formation, or they are similar to the first galaxies formed in the Universe (Mateo, 1998). However, the local dwarf galaxies will have had a high degree disruption and interference due to their proximity and interactions with massive galaxies, and to find even more pristine dwarf galaxies, all sky surveys or deep surveys are required.

1.6.3 Dwarfs in observational studies

One of the earliest studies of nearby dwarf galaxies is the review of Hodge (1971). Some of the earliest observations of dwarf galaxies (barring the Magellanic Clouds) were done in the mid 20th century by e.g Shapley (1938); Baade & Hubble (1939); Zwicky (1957). These studies mostly focused on single targets or ones belonging to the same cluster, but Baum et al. (1959) showed a general progression of colour between dwarf galaxies and massive ones. More recent studies similarly suggest a continuity of structural properties from dwarf galaxies to larger ones (Tolstoy et al., 2009).

Most studies of dwarf galaxies have been focused on dwarf galaxies in the Local Group and for good reason. They are the closest and most easily observable dwarf galaxies, and in some cases, even individual stars are resolvable. This means that detailed star formation histories are inferable and kinematics, dynamics, and stellar content can be measured and resolved. Mateo (1998) offer the perspective that the LG dwarf galaxies provide a window into detailed properties of dwarf galaxies as a whole.

With the advent of deeper and deeper large scale surveys, the number of dwarf galaxies were found to dominate by numbers (Ferguson & Binggeli, 1994; Gallagher & Wyse, 1994). Furthermore, the multiobject spectroscopy (such as the Sloan Digital Sky Survey; York et al.,

2000) increased the sample size of dwarf galaxies enough to enable quantitative analyses of the dynamics, radial velocities, and stellar content.

The Local Group is still the target for intense studies, one of the reason being that new dwarf galaxies continue to be found (Cerny et al., 2021; Martínez-Delgado et al., 2021; Mutlu-Pakdil et al., 2022) and that the local system provides an excellent benchmark for theoretical models. An early contention between observations and theory were the fact that the number of observed dwarf galaxies was far too low compared to what was expected from simulations. Conversely, theoretical models have helped guide observational efforts to find missing pieces in out understanding of galaxy evolution, for example highlighting the fact that there may be many dwarf galaxies still to be found in our neighbourhood.

1.6.4 In simulations

Similarly as with observational studies of dwarf galaxies, simulations have not historically dealt with dwarf galaxies on a large scale. However, newer cosmological simulations such as Illustris do include dwarf galaxies down to stellar masses of $10^8\,M_\odot$, so they can now examine some of the discrepancies found between early simulations and observations such as the missing satellite problem and the core-cusp problem.

The missing satellite problem is the discrepancy between the number of observed dwarf/satellite galaxies in the Local Group versus the number of the predicted galaxies from theoretical models and cosmological simulations. Klypin et al. (1999) were some of the first to formulate this issue with hierarchical structure formation models and early cosmological simulations such as Moore et al. (1999) predict about 500 dwarf galaxies in the MW halo, but recent observational surveys put this number at around 100 (McConnachie, 2012)

However, the problem has been alleviated slightly but not solved entirely with the discovery of ultra faint dwarf galaxies suggesting that very faint – almost unobservable – galaxies exist and that various mechanism (such as tidal stripping and feedback processes) might exist to suppress the visible parts of these galaxies. Furthermore, with cosmological simulations including more complex baryonic physics, feedback processes, and higher resolution having become more commonplace, more and more studies are finding a reduced discrepancy – even

agreement – between observations and simulations (Engler et al., 2021; Fattahi et al., 2020).

Being able to reproduce the diversity in dwarf galaxies as observed is still at the limit of what cosmological simulations can do today, but the inclusion of baryonic processes and feedback has been a step forward. For example, while AGN feedback has mostly been ignored for dwarf galaxies, the discovery of active nuclei in many dwarf galaxies (up to 10 per cent depending on selection method) warrants attention. This is further emphasised by recent observations and simulations (Penny et al., 2018; Koudmani et al., 2021) that consider AGN feedback central to explain star formation suppression – a view contrary to the paradigm of only invoking SN feedback in dwarf galaxies.

1.6.5 AGN in dwarf galaxies

Historically, AGN in dwarf galaxies has been an unexplored subject – in part due to the weak impact the AGN seem to have on their host galaxy, and in part due to the difficulty observing them. Large surveys like SDSS are biased towards luminous galaxies and luminous AGN. An early effort into looking for low luminosity AGN was that of Filippenko & Sargent (1985). However, these effort focused primarily on the AGN/Seyfert aspect of galaxies rather than the host properties, and only a couple of the 75 observed 'dwarf' Seyfert' galaxies can be classified as a dwarf galaxy (Ho et al., 1997). One galaxy that seemed to fit in both categories is NGC 4395, which consequently received a lot of interest since it constituted an an excellent to study the quasar phenomenon at low intrinsic luminosity, and that it is close enough to estimate the mass of the central black hole (Filippenko & Sargent, 1989; Filippenko & Ho, 2003).

A brief review a decade and a half ago (Greene et al., 2006) gave the current status on AGNs in dwarf galaxies – by then, still a relatively unexplored frontier with only a two prominent examples. Greene & Ho (2007) expanded the sample of known dwarf galaxies with intermediate mass BHs (IMBH) by an order of magnitude to 174 thanks to the availability of new large scale surveys such as SDSS. An even larger trove of dwarf galaxies with AGN was found in Reines et al. (2013). This sample has only increased in size with other large scale surveys in other wavelengths such as mid-IR (Sartori et al., 2015), UV, and X-ray (Baldassare

et al., 2017; Birchall et al., 2020).

The notion that AGN are irrelevant for dwarf galaxy evolution (Haines et al., 2007) has been brought into question by recent studies that found dwarf galaxies with strong BH feedback. For example, AGN-driven outflows were found by Manzano-King et al. (2019); Liu et al. (2020) in optically selected AGNs and radio-selected ones, too (Mezcua & Domínguez Sánchez, 2020; Schutte & Reines, 2022). Typically, AGN feedback is thought of as detrimental to star formation rates (e.g Penny et al., 2018), but some work show the contrary (Schutte & Reines, 2022). Supernovae (SNe) have been the preferred internal feedback mechanism that regulates the growth of dwarf galaxies, but it is strong enough alone to account for some dwarf galaxies (Garrison-Kimmel et al., 2013), and the neutral hydrogen content in some isolated dwarf galaxies (Bradford et al., 2018).

This renewed interest in the role of AGN feedback in dwarf galaxies has carried over into simulations and modelling. Dashyan et al. (2018) suggested that AGN feedback can indeed drive negative feedback in dwarf galaxies to an even greater degree than SNe. Koudmani et al. (2021) similarly found that dwarf galaxies in the FABLE simulations with overmassive black holes can lead to significantly reduced gas fraction in the host galaxy.

Acknowledging the importance of AGN then raises the question of under what conditions AGN activity occurs – what are the triggers? Ultimately, the mechanisms is the same as for regular galaxies: Gas needs to be driven to the IMBH or SMBH in the centre (or, in some cases for dwarf galaxies, off-centre (Reines et al., 2020)). The same processes described in Section 1.5.5 are valid for dwarf galaxies, but the difference in mass (both halo and black hole), stellar content, and so on between dwarf galaxies and 'regular' galaxies make them differently susceptible to the various processes.

For example, SN feedback has been linked to decreased BH growth in dwarf galaxies (Habouzit et al., 2017; Anglés-Alcázar et al., 2017) whereas environmental effects on AGN activity have been both affirmed (e.g Deng et al., 2012; Pimbblet et al., 2013; Sabater et al., 2013; Satyapal et al., 2014) and rejected (e.g Miller et al., 2003; Man et al., 2019; Kristensen et al., 2020). However, there are many nuances to this correlation such as AGN identification method, environment estimation, and statistical threshold. For example, obscured AGN are

not often picked up in optical searches **but** more often in infrared ones. Infrared galaxies are more often in a stage of merger (Satyapal et al., 2014) and thus more likely to find a link to environment.

Lastly, common selection techniques are tuned towards regular mass galaxies (see Section 1.5.4 which may not be directly transferable to the dwarf mass regime. E.g Mackay Dickey et al. (2019) found extended AGN emission in isolated dwarf galaxies with optically selected AGN characteristics, which suggests that the emission is a false positive of actual AGN activity. Hainline et al. (2016) note that young starbursts mimick AGN in dwarf galaxies, and Lupi et al. (2020) remark a large number of contaminants in mid-IR selection.

1.6.6 Research goals of this thesis

These concerns and considerations regarding dwarf galaxies and AGN pose two research avenues: (1) What conditions are favourable for triggering AGN activity in dwarf galaxies (and are they the same as for massive galaxies), and (2) are the current diagnostic tools suitable for selecting AGN in dwarf galaxies?

This thesis attempts to answer point (1) focusing on the environmental impact AGN activity in dwarf galaxies using both large scale observational surveys and cosmological simulations. More specifically, Chapter 2 uses the NASA-Sloan Atlas (NSA) and selects dwarf galaxies with optical AGN characteristics. Their environments are measured by the distance to their 10th nearest neighbour and velocity difference to their nearest one and compares them to a control group. Chapter 3 takes a similar approach but using simulation data from the IllustrisTNG project. Further measures are included as well such as merger history since simulation data retains the full cosmological histories of the dwarf galaxies. Both of these chapters are published in peer-reviewed journals.

Point (2) is also mentioned in these two chapters, but it is taken further in Chapter 4. This chapter similarly attempts to shed light on whether there is an environmental impact on AGN activity in dwarf galaxies. The chapter uses further selection diagnostics, environmental measures, and integrated field unit (IFU) spectroscopy in an attempt to build a more complete set of dwarf AGN. The high number of parameters and their impact are then characterised

using a machine learning approach, which may be used to improve AGN selection in dwarf galaxies thus expanding on point (2).

2. Environments of dwarf galaxies with optical AGN characteristics

This chapter contains a study on environments of dwarf galaxies with optical AGN characteristics. The results were published in August 2020 in Monthly Notices of the Royal Astronomical Society, Volume 496, Issue 3, pp.2577-2590 and the work was carried out in collaboration with Kevin A. Pimbblet (University of Hull) and Samantha J. Penny (University of Portsmouth) and me as lead author. I have written and made 100 per cent of the text and plots. It has been formatted differently than the published paper to fit in this thesis, and may also differ slightly in typography, but the results and science are unaltered.

Abstract

This study aims to explore the relation between dwarf galaxies ($M_* \leq 5 \times 10^9 M_\odot$) with AGNs and their environment by comparing neighbourhood parameters of AGN and non-AGN samples. Using the NASA-Sloan Atlas, both the local environment and the immediate environment of dwarf galaxies with $z \leq 0.055$ are analysed. Of the 145 155 galaxies in the catalogue, 62 258 of them are classified as dwarf galaxies, and by employing two AGN selection methods based on emission line fluxes (BPT and WHAN), 4 476 are found to have AGN characteristics in their optical spectra. Regardless of selection method, this study finds no discernible differences in environment between AGN and non-AGN host dwarf galaxies and these results indicate that environment is not an important factor in triggering AGN activity in dwarf galaxies. This is in line with existing literature on environments of regular galaxies with AGNs and suggests universality in terms of reaction to environment across the mass regime. The biases of AGN selection in low-mass galaxies, and the biases of different measures of environment are also considered. It is found that there are several mass-trends in emission line ratios and that the SDSS fiber covers galaxies non-uniformly with redshift.

These biases should be accounted for in future work by possibly including other wavelength regimes or mass-weighting of emission line ratios. Lastly, a discussion of the environment estimation methods is included since they may not gauge the desired properties due to factors such as time delay or using loosely constrained proxy parameters.

2.1 Introduction

Galaxies are dynamical objects that evolve and mature over time. Internal processes such as star formation, supernovae, and nuclear activity and external ones such as galaxy interactions (Moore et al., 1996), ram-pressure stripping (Gunn & Gott, 1972), and intergalactic medium accretion can change the composition and structure of galaxies and decide their futures. Many of these processes are strongly correlated with stellar mass or environment (Kauffmann et al., 2003; Miller et al., 2003; Baldry et al., 2006; Peng et al., 2010, 2012).

Multiple processes can affect galaxies simultaneously. Dwarf galaxies can be used to isolate a single evolutionary process due to their low masses and relatively low frequency of mergers. These properties potentially give a single process a huge impact on the evolution of them. For example, field dwarfs are very much shaped only by internal processes while environmental effects dominate low-mass galaxies in clusters and groups (Haines et al., 2007; Peng et al., 2010).

Observing and analysing dwarf galaxies is observationally expensive and time consuming since their low surface brightness require long exposures. For example, in a survey similar to the Sloan Digital Sky Survey (SDSS, York et al., 2000), a galaxy such as the Large Magellanic Cloud (LMC) will only be observable out the $z \sim 0.35 - 0.45$ in the r-magnitude[1]. However, more and more large scale surveys (such as SDSS) are now reaching these depths and include more dwarf galaxies, which means that the statistical basis for studying dwarf galaxies is becoming better. Furthermore, since dwarf galaxies constitute the first link in the chain of hierarchical structure formation theory, they constitute an invaluable source in figuring out the full galaxy formation and evolution puzzle.

[1]Assuming $M_r = -18.5$ and SDSS depth $m_r = 22.70$, https://www.sdss.org/dr14/imaging/other_info/

Furthering their importance, most dwarf galaxies are believed to host intermediate-mass black holes (e.g Moran et al., 2014; Silk, 2017, IMBHs; $M_{BH} \sim 10^2 - 10^6\,M_\odot$) – a characteristic that has been studied in more detail in a number of papers; Barth et al. (2004) examined the host galaxy properties and the IMBH properties in the POX 52 galaxy. Reines et al. (2013) examined dwarf galaxies with optical signatures of active massive black holes. Sartori et al. (2015) searched for IMBHs using mid-IR and optical data while Baldassare et al. (2015) looked at the core region of RGG 118 and could infer an IMBH from the kinematics. Since IMBHs are the root of the super-massive black holes (SMBH) either through acting as a seed of gas accretion or merging of several IMBHs (e.g Micic et al., 2007), observing IMBHs during these phases (i.e AGN phase) can shed light on conditions required for IMBH growth.

There are several mechanisms thought to trigger AGN activity. Merging or harassing galaxies are effective ways of accreting inter-stellar medium (ISM) or removing angular momentum from native gas reservoirs (Miller et al., 2003; Sabater et al., 2013; Gordon et al., 2018; Ellison et al., 2019), and the influx of material to the central regions of galaxies can then trigger AGN activity.

Other AGN triggers include environmental effects (Kauffmann et al., 2004), where for example cooling gas from cluster cores accrete onto the central galaxies or the intergalactic medium compressing and shocking gas within a galaxy and driving the gas towards the core. Complicating this picture are observations that there might be a time delay between interactions and the onset of AGN activity (e.g Pimbblet et al., 2013), which means that the current environment of a galaxy may not represent the environment that triggered the AGN activity.

The effect of the environment on a single galaxy can be analysed from detailed and focused observations, but such an undertaking is not feasible for a large scale survey containing thousands of objects. However, several methods exist to quantify the environment of galaxies (for a review, see Muldrew et al., 2012), which are more suitable for a study like this. For example, Miller et al. (2003) calculated a galactic density using the 10th nearest neighbour as the shell edge while Baldry et al. (2006) used the 4th and 5th nearest neighbour. Sabater et al. (2015) calculated a tidal estimator that traced the relation between tidal forces exerted

by companions and the internal binding force of a galaxy.

These methods all attempt to quantify the environment, but they all have different strengths and weaknesses. Some are better at describing the local galactic environment, others are better for the group/cluster environment while some are better for the immediate environment (i.e whether a close neighbour exerts strong influence or not).

Even the task of identifying AGNs is not straightforward since they have different signatures in different wavelength regimes. In this work, spectroscopic data from the Sloan Digital Sky Survey SDSS will be used and two different AGN selection methods are utilised. Since this work is based on SDSS data, optical diagnostic diagrams are used. The first one is the common Baldwin, Phillips, & Terlevich (BPT) diagram (Baldwin et al., 1981) with the Kewley et al. (2001); Kauffmann et al. (2004) criteria for AGN. BPT takes advantage of the fact that different excitation mechanisms have different emission line fingerprints.

The second diagnostic is the less common WHAN diagram (Cid Fernandes et al., 2010, 2011). WHAN utilises the equivalent width, W_λ, of Hα and the [N II]/Hα line ratio and thus covers the same wavelength regime as BPT. The WHAN diagram was developed as a response to the BPT since BPT leaves a large population of emission line galaxies (ELG) unclassified in SDSS data. The advantage is that it recovers most things that the BPT does, but it also gains the weaker AGNs. Both methods will be discussed further in section 2.2.2

Whether environment quenches AGN, triggers AGN, or has no effect, is unclear when it comes to dwarf galaxies. The broad goals of this work are to therefore determine the environment of dwarf galaxies with AGN characteristics and construct arguments based on these environmental measures on how such dwarf galaxies with AGN trigger and evolve. The environmental analysis consists of the 10th nearest neighbour (10NN) method and the velocity difference to nearest neighbours (Δv_{NN}), and the distributions for each sample is then compared to non-AGN galaxies using two-sample Kolgomorov-Smirnov tests.

This paper is structured as follows: Section 2.2 contains details about the data and methods used. Section 2.3 includes the analysis and interpretation of the results and Section 2.4 has discussions on the findings. Conclusions and a summary is found in Section 2.5. This study assumes a Λ-CDM Universe with $H_0 = 70$ and $\Omega_{m_0} = 0.3$.

2.2 Data and methods

This section describes the data used and the cuts made to classify dwarf galaxies and the diagnostics used to select AGN in that sample.

The selection criteria can be summarised as the following: Low mass galaxies: $M_* \leq 5\times10^9\,\mathrm{M}_\odot$, $\sigma \leq 100\,\mathrm{km\,s^{-1}}$, and completeness corrections. BPT galaxies follow the classification in Kewley et al. (2001) while the WHAN AGN selection requires $\log([\mathrm{N\,II}]/\mathrm{H}_\alpha) \geq -0.4$ and $W_{H_\alpha} \geq 3$ Å. The environment analysis involves two methods; distance to the 10th nearest neighbour and velocity difference to the nearest angular separated galaxy.

2.2.1 Data and sample selection

The data used for identifying dwarf galaxies and AGN is from the NASA-Sloan Atlas (NSA) catalogue. This catalogue is constructed by using several catalogues; Sources are found from a combination of SDSS DR8 (York et al., 2000; Aihara et al., 2011), NASA/IPAC Extragalactic Database[2], Six-degree Field Galaxy Redshift Survey, Two-degree Field Galaxy Redshift Survey, ZCAT and ALFALFA catalogues. Spectroscopic measurements (e.g line fluxes) are performed on SDSS spectra while all catalogues are used to determine redshifts. The final NSA catalogue contains extragalactic sources to a high completeness to $z < 0.05$, which there are 145 155 of.

The detection and de-blending technique for the photometry analysis is described in Blanton et al. (2011). It is in spirit based on the SDSS photometric pipeline (Lupton et al., 2001), but there are differences in the way objects are deblended and use r-band templates for all bands[3]. Furthermore, the sources in NSA are only included if they are matched to a spectroscopy survey. Not all sources have SDSS spectroscopy, but the ones that do, have had their spectra remeasured by Yan (2011) using an improved calibration, which affects small equivalent width lines making this catalogue well suited for classification diagrams based on emission line ratios and equivalent widths.

[2]The NASA/IPAC Extragalactic Database (NED) is funded by the National Aeronautics and Space Administration and operated by the California Institute of Technology.

[3]For a more in-depth summary, see http://nsatlas.org/documentation

Table 2.1: Completeness selection intervals.

z	$\leq M_r$
$0.00 \leq z < 0.01$	-15.0
$0.01 \leq z < 0.02$	-16.0
$0.02 \leq z < 0.03$	-17.0
$0.03 \leq z < 0.04$	-17.5
$z \geq 0.04$	-18.0

Furthermore, since dwarf galaxies tend to have weaker emission, the better measurements (i.e higher signal-to-noise) of spectroscopic data make this catalogue preferable to others for this study. Another argument for this catalogue is the stricter significance in SDSS in r-band images on splitting 'child' objects from 'parent' objects – basically when an algorithm decides that a source is two objects rather than one. For dwarf galaxies, it means fewer false positives making the dwarf galaxy sample more robust, although there is a risk of large galaxies 'absorbing' small and weak ones.

Low mass galaxies are selected by imposing a stellar mass limit of $M_* \leq 5 \times 10^9\, M_\odot$ and velocity dispersion $\sigma \leq 100\,\mathrm{km\,s^{-1}}$. This follows similar limitations as other work in the field (e.g Reines et al. 2013 and Penny et al. 2016, 2018) and corresponds roughly to the stellar mass of the LMC. The masses in NSA is given in units of $M_\odot\, h^{-2}$, and while other studies assume $h \approx 0.70$ (Reines et al., 2013; Hainline et al., 2016; Baldassare et al., 2018), we have assumed $h = 1$ for galaxy masses despite the cosmology assumed. The analysis and results in Section 2.3 and Table 2.3 have been analysed using both values, and no significant difference is found. Therefore, the choice of $h = 1$ remains unchanged. The effect on sample sizes can be seen in Table 2.2.

From inspection of z vs r-magnitude (see Figure 2.1), upper limits for several redshifts bins are imposed for the sake of completeness. The specific redshift bins and their corresponding magnitude-cuts can be seen in Table 2.1. Using both the low mass galaxy criteria and the completeness restrictions, the sample size is reduced to 62 258 objects. This constitutes the parent sample from which further analysis is carried out.

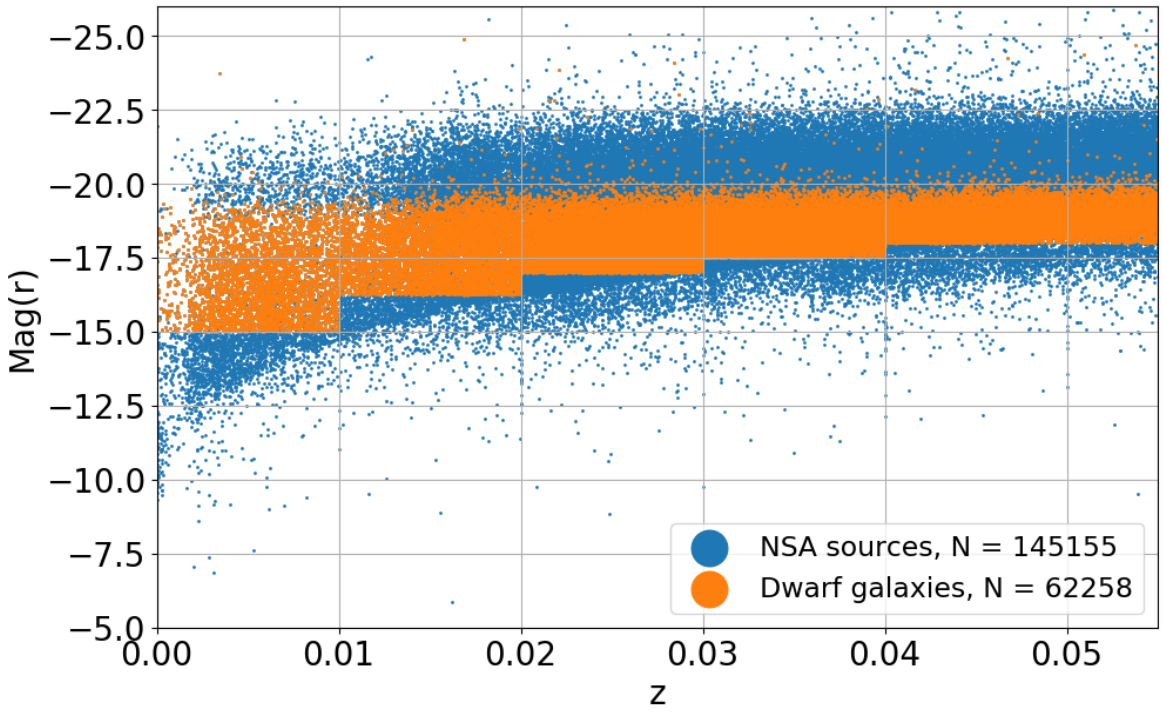

Figure 2.1: Magnitude versus redshift plot. The blue data points are all galaxies in NSA. The orange data points are low-mass galaxies (as defined in Section 2.2.1). There are clear magnitude edges in different redshift intervals, which is due to the completeness selection.

Table 2.2: Number of galaxies depending on choice of h. While the number of galaxies decreases with decreasing h, the results described in Section 2.3 do not change.

h	Dwarfs	NOT	BPT	WHAN	AND	OR
1.00	62,258	55,643	387	4,323	228	4,476
0.73	43,774	41,341	124	1,399	62	1,461
0.70	41,289	39,189	102	1,182	47	1,237

For the environmental analysis, the NSA catalogue is also used. The only interesting properties of the neighbour galaxies are their positions and redshift. The full number of sources is then 145 155 and all objects contain coordinates and redshifts from spectroscopy. The environmental analysis will be described in detail in Section 2.2.3.

2.2.2 Classification diagrams

Two AGN selection methods are employed: The familiar BPT diagram (Baldwin et al., 1981; Kewley et al., 2001; Kauffmann et al., 2003, 2004) and the lesser-used WHAN (Cid Fernandes et al., 2010, 2011). They are used both in conjunction and in parallel since they both have different strength and weaknesses. The diagnostics are used on the dwarf galaxy sample consisting of 62 258 objects. Below is a more detailed description of each classification scheme and an overview of the numbers can be found in Table 2.2.

BPT diagram

The BPT diagram is used as one of the diagnostics to identify AGN. More specifically, the [N II] λ6584/Hα vs. [O III] λ5007/Hβ line ratios are used and follow the Kewley et al. (2001) distinction between composite star-forming galaxies and pure AGNs. While massive composite galaxies do include AGNs, too, we are uncertain of the intepretation in the low mass regime. This division yields 2 644 objects. However, requiring a S/N $\geq$ 3 on the 4 emission lines reduces this number to 296 – a ~88.8 per cent rejection rate.

The primary reason for rejection is due to the low S/N on Hβ. 95.5 per cent of the 2,348 rejected BPT galaxies have a $S/N_{H_\beta} < 3$. As noted by Cid Fernandes et al. (2010), AGN

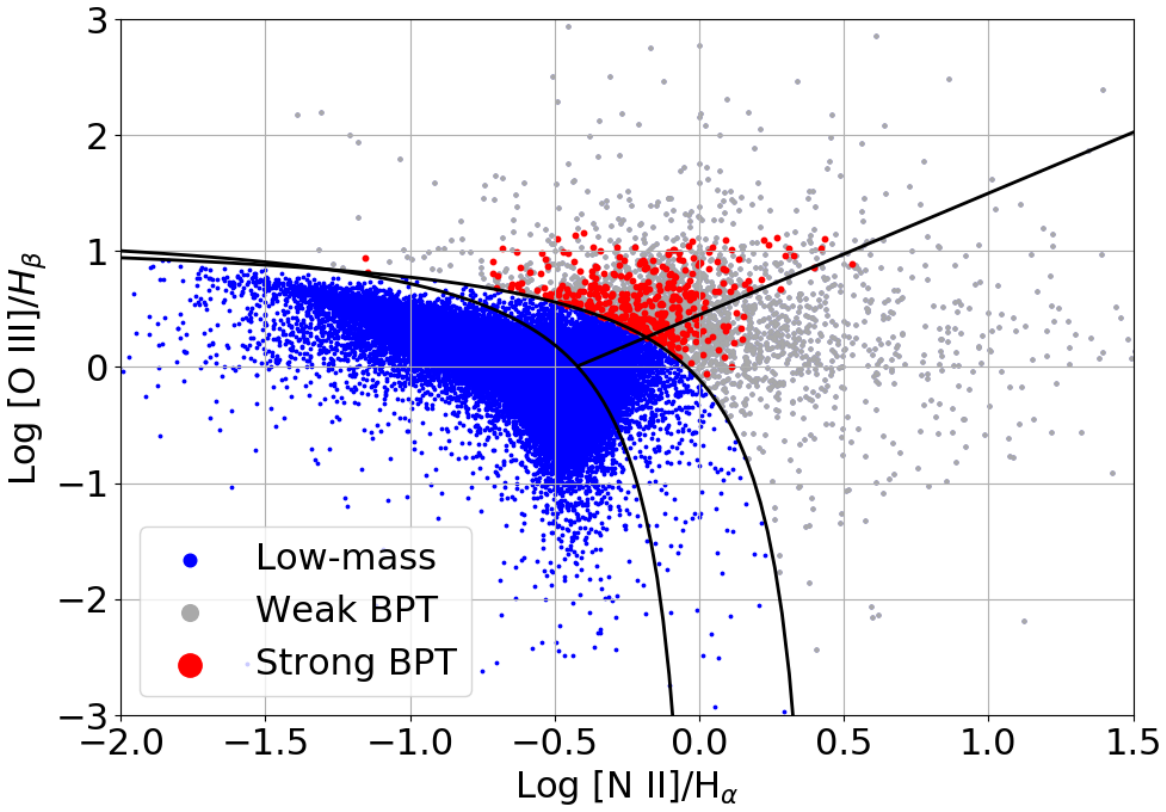

Figure 2.2: BPT diagram. The solid black lines are follow the Kewley et al. (2001); Kewley et al. (2006) classification diagram. However, no distinction is made between Seyfert and LINERS, and only pure AGNs are included in this sample, thus following the Kewley et al. (2001) classification. Three samples are plotted. The blue dots are all the low-mass galaxies in the NSA catalog. The red dots are the BPT-selected galaxies with $S/N_{ratio} > 3/\sqrt{2}$ on both emission lines ratios. The 'weak' BPT are galaxies with $S/N_{ratio} < 3/\sqrt{2}$. Especially Hβ is responsible for classifying a BPT-selected galaxy as weak ($\approx$ 87.5 per cent of all dwarf BPT galaxies in this sample has $S/N_{H\beta} < 3$).

galaxies have intrinsically low Hβ emission, which gives rise to low S/N measurements – a problem that is exaggerated in dwarf galaxies because of their already weak signal. In Cid Fernandes et al. (2010), 53 per cent of their sample of emission line galaxies (ELGs) had weak measurements of Hβ, which supports the notion that dwarf galaxies are particularly vulnerable to this effect.

Another approach is to require use the $S/N_{ratio} > 3/\sqrt{2}$ instead. This means that if one emission line is well-determined but the other is not, it is not automatically rejected. This follows the same approach as e.g Juneau et al. (2014) and Trump et al. (2015). Using this S/N cut yields 387 BPT galaxies and a $\sim$ 85.4per cent rejection rate, and this is the sample used going forward.

BPT classification can also be performed using other line pairs such as[S II] $\lambda\lambda$6717, 6731/Hα and [OI]λ6300/Hα. However, they are also compared against [O III] λ5007/Hβ and thus do not provide a way to bypass the low SNR on Hβ. Therefore, these BPT diagrams are not chosen for further analysis in this work.

WHAN diagram

The criteria for being classified as an AGN in the WHAN diagram follow Cid Fernandes et al. (2011); $\log([\mathrm{N\,II}]/\mathrm{H}_\alpha) \geq -0.4$ and $W_{H_\alpha} \geq 3$Å. In the WHAN classification, there is a distinction between strong and weak AGN (*weak* here meaning to be an indicator of energy output of the AGN and not low S/N like for the weak BPT classification). The used limit on W_{H_α} is such that both weak and strong AGN are included and no further distinction are made between them. This yields 4 323 objects. Using a S/N $\geq$ 3 requirement of Hα and a $S/N_{ratio} > 3/\sqrt{2}$ recovers 4,317 sources. This is the WHAN sample in onwards analysis.

Cid Fernandes et al. (2010) suggest that the WHAN diagram is more suitable for selecting weaker AGNs – especially ELGs – compared to the BPT diagram. The BPT diagram is a very strict selection technique since it requires 4 emission lines of high quality. In fact, they argue that the choice of a strong (here meaning S/N$\geq$ 3) Hβ biases against objects with low W_λ and thus leaving out weaker AGN galaxies. As the goal of this paper is to quantify the environment of dwarf galaxies hosting AGN, this makes the WHAN diagram an ideal selection method for

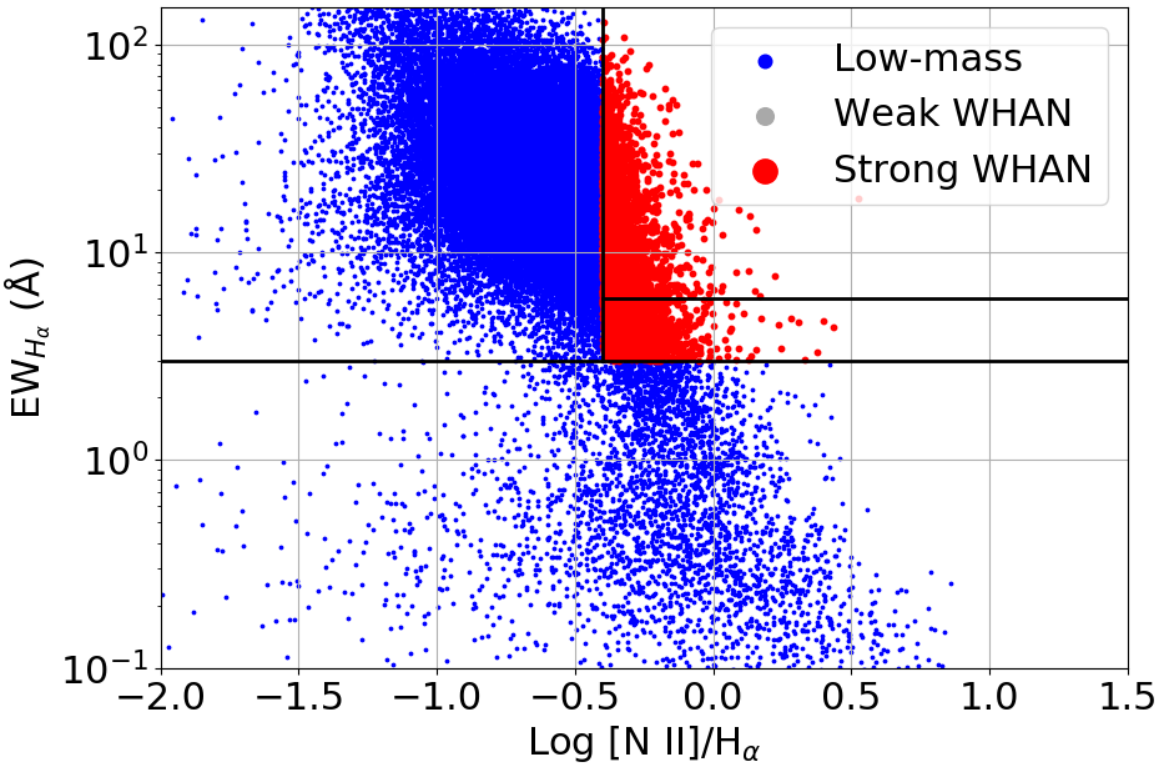

Figure 2.3: WHAN diagram (For details, see Cid Fernandes et al., 2010, 2011). The solid black lines mark the different regions (from top left, clock-wise); Star-forming, strong AGNs, weak AGNs, and retired galaxies. Both weak and strong AGNs are included in the sample and no distinction is made between them. The blue dots are all the low-mass galaxies in the NSA catalog. The red dots are the WHAN-selected galaxies with $S/N_{ratio} > 3/\sqrt{2}$ on [N II]/Hα. The 'weak' WHAN are galaxies with $S/N_{ratio} < 3/\sqrt{2}$

our sample selection.

'AND', 'OR' & 'NOT' samples

As mentioned in Section 2.1, the two diagrams are used both separately and in conjunction with each other. Two further samples are made from the BPT and WHAN samples: *'AND'* and *'OR'*. The *'AND'* sample is comprised of galaxies that fulfil *both* the BPT and WHAN selection criteria while *'OR'*-selected galaxies fulfil *either*. The size of the samples are 228 and 4 476, respectively.

Additionally, dwarf galaxies that does not appear in either non-S/N-corrected sample are labelled *'NOT'*. This is the largest subsample and comprises 55 643 objects. In *'NOT'*, galaxies with either a BPT or WHAN AGN classification (before correcting for low S/N and are thus not considered AGNs in this study). They are not included because the sample size is sufficiently large without them – even if they were included, it would only change the sample size by less than ~4 per cent and excluding them makes the *'NOT'* sample more robust because only onjects with clear classification is included.

2.2.3 Environment estimation

There are two different properties of the environment that this study attempts to examine: The density of the local environment and the recent interaction history of AGN galaxies where the local environment is to be understood as the area of the group or cluster that the dwarf galaxy is situated in. Though both properties are not straight-forward to quantify, there are a number of methods to infer them (for a discussion of these, see Muldrew et al., 2012).

One method to infer the density of the environment is the projected distance to the 10th nearest neighbour (10NN) while the recent interaction history can be inferred from the velocity difference to the nearest angular separated neighbour (Δv_{NN}). Throughout this study, the environment inferred from 10NN is often referred to as the local environment, and from Δv_{NN}, the environment is referred to as the immediate environment. While other studies describe the local environment by the galaxy surface density, translating r_{10} to galaxy surface density is straightforward through the equation $\Sigma_{10} = \frac{N}{\pi r_{10}}$. Therefore, the use of r_{10} is as

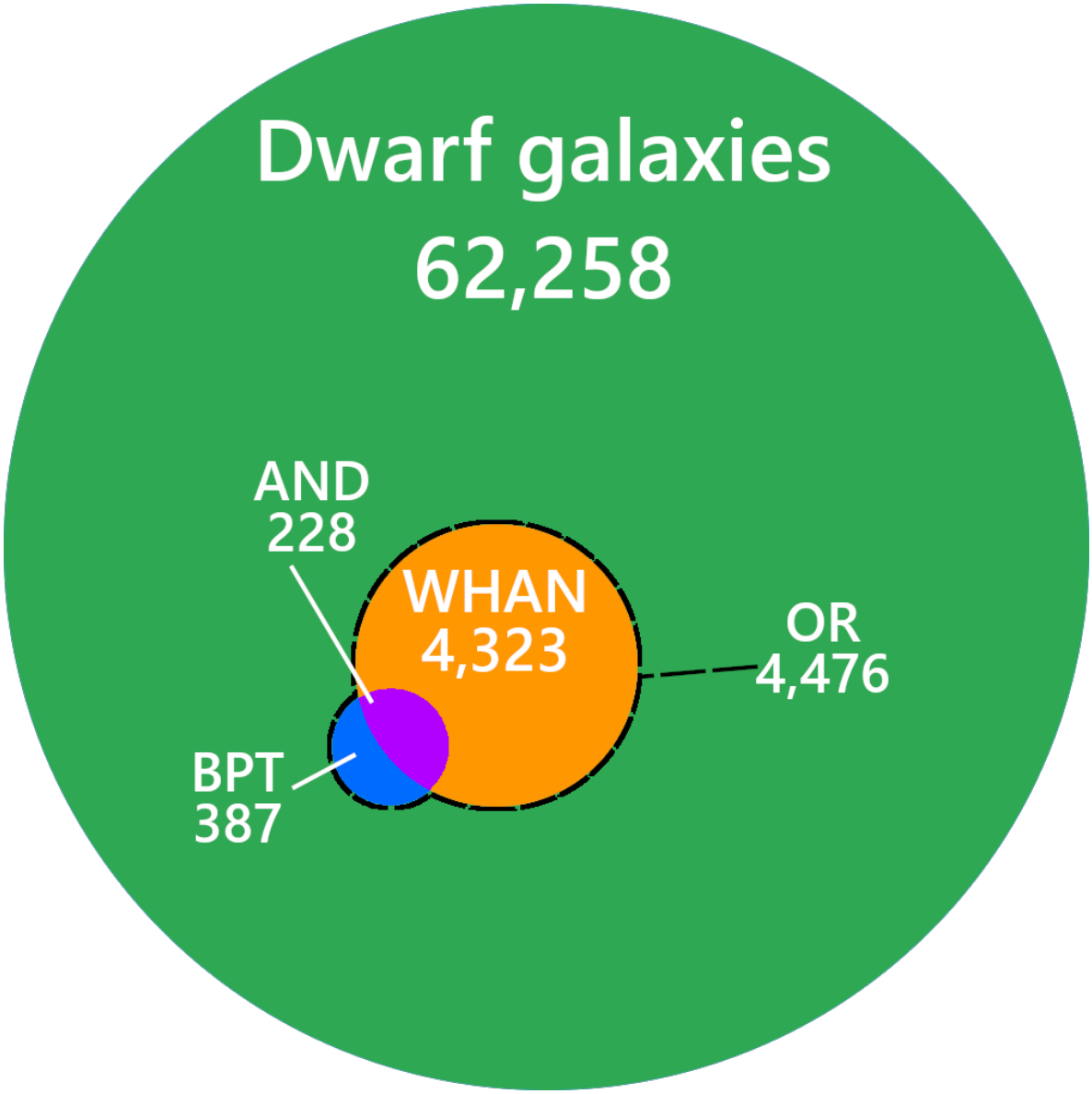

Figure 2.4: NSA Venn diagram showing the different selections.

good as Σ_{10}.

Both methods only consider galaxies within $\pm 1\,000\,\mathrm{km\,s^{-1}}$, which is slightly higher than the average galaxy cluster velocity dispersions (see e.g Bilton & Pimbblet, 2018). This is to ensure that galaxies are only neighbours if they are close spatially (i.e member of the same group or cluster) and not just angularly close. Muldrew et al. (2012) remarks that a nearest neighbour approach is a better measure of the local density compared to cluster density, and a higher *n*th separated neighbour smooths out local variances. Smoothing out local variances is desirable for estimating the local environment in general, but local variances are exactly what is important for immediate environments.

2.3 Analysis

This section contains the statistical analysis of differences between the subsamples. The neighbourhood parameters (10NN and Δv_{NN}) will be looked at with a Monte Carlo Kolmogorov-Smirnov (KS) test procedure while other properties such as stellar mass and redshift will receive a short statistical rundown. A visual inspection is also carried out on the galaxies fulfilling both the WHAN and BPT criteria (i.e the *'AND'* subsample) and compared to a similar sized subsample from the *'NOT'* subsample.

2.3.1 KS-testing

To quantify the difference of environment between different subsamples, two-sample Kolmogorov Smirnov (KS) tests are carried out. It is a test of whether or not two samples come from the same parent distribution – for example whether the distribution of the distance to the 10th nearest neighbour of the 'BPT' sample is the same as the distribution of the 'NOT' sample. Though two samples of different sizes can be used, the input sample sizes are scaled to 152 elements. 152 is the number of objects in the smallest subsample (WHAN AGNs that are rejected in BPT because of low S/N).

Each KS test is iterated 1 000 times, each time with 152 different random elements from the subsamples listed in Table 2.3. Next, a comparison sample is found from another subsample (although each subsample is also tested against itself) where a matching galaxy is found for

Table 2.3: p-values of respectively 10NN and Δv_{NN} 2-sided KS tests. Each row has the subsample in the leftmost column as the subsample to be compared against a control sample from a subsample given by the column name. E.g, the test in row 1, column 2 is found from 152 random galaxies from all low mass galaxies and a matching galaxy (in mass, colour, and redshift) sample is found for each element from the BPT subsample. 'wBPT' is short for 'weak BPT'.

10NN	**All**	**BPT**	**WHAN**	**AND**	**OR**	**NOT**	**wBPT**
All	0.52	0.26	0.47	0.14	0.46	0.54	0.00
BPT	0.20	0.57	0.40	0.43	0.40	0.21	0.00
WHAN	0.52	0.27	0.55	0.12	0.53	0.50	0.00
AND	0.14	0.51	0.25	0.53	0.26	0.13	0.00
OR	0.50	0.27	0.54	0.13	0.56	0.49	0.00
NOT	0.55	0.28	0.44	0.14	0.44	0.54	0.00
wBPT	0.17	0.00	0.01	0.00	0.01	0.20	0.56

Δv_{NN}	**All**	**BPT**	**WHAN**	**AND**	**OR**	**NOT**	**wBPT**
All	0.53	0.31	0.46	0.40	0.46	0.53	0.39
BPT	0.44	0.56	0.41	0.46	0.43	0.44	0.29
WHAN	0.49	0.33	0.52	0.25	0.53	0.48	0.38
AND	0.32	0.49	0.28	0.55	0.29	0.31	0.21
OR	0.51	0.34	0.54	0.26	0.52	0.50	0.36
NOT	0.51	0.28	0.43	0.39	0.45	0.53	0.37
wBPT	0.45	0.34	0.37	0.22	0.38	0.47	0.53

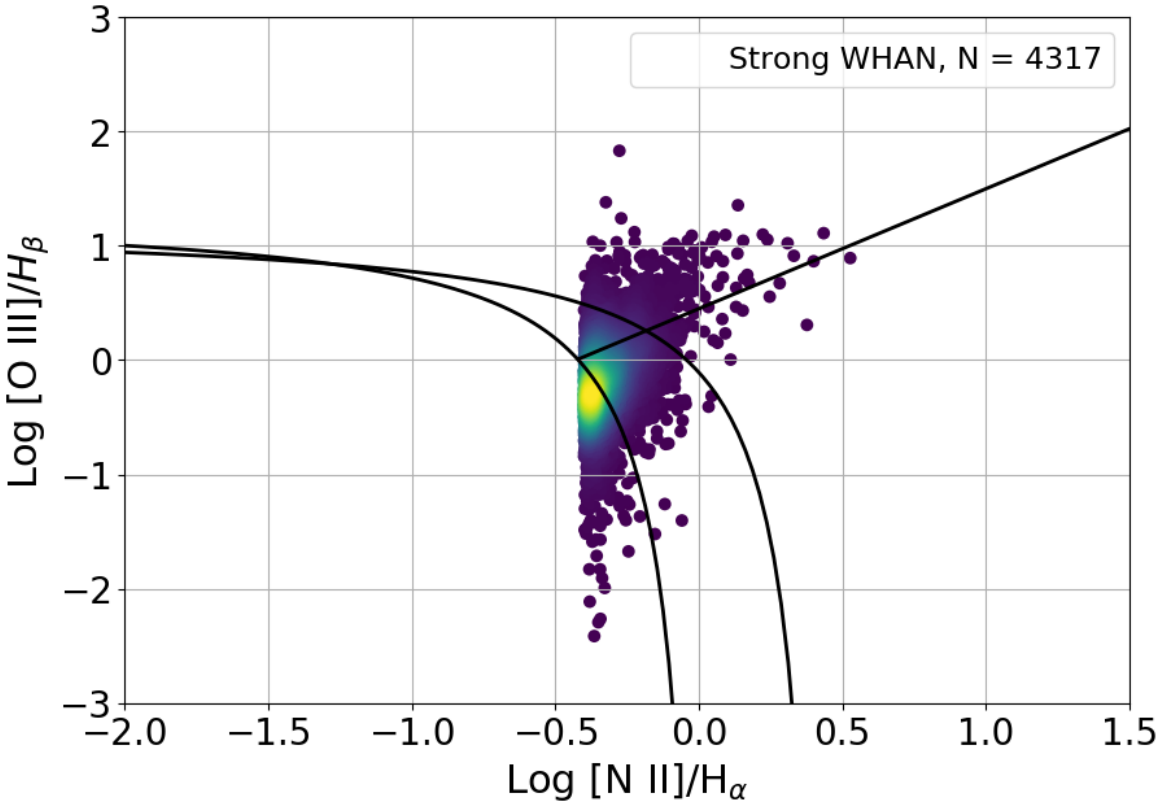

Figure 2.5: BPT diagram with WHAN selected galaxies. The dots are colour-coded by their relative point density. The majority of the WHAN selected galaxies would have been classified as star-forming or composite SF/AGN using the BPT classification scheme.

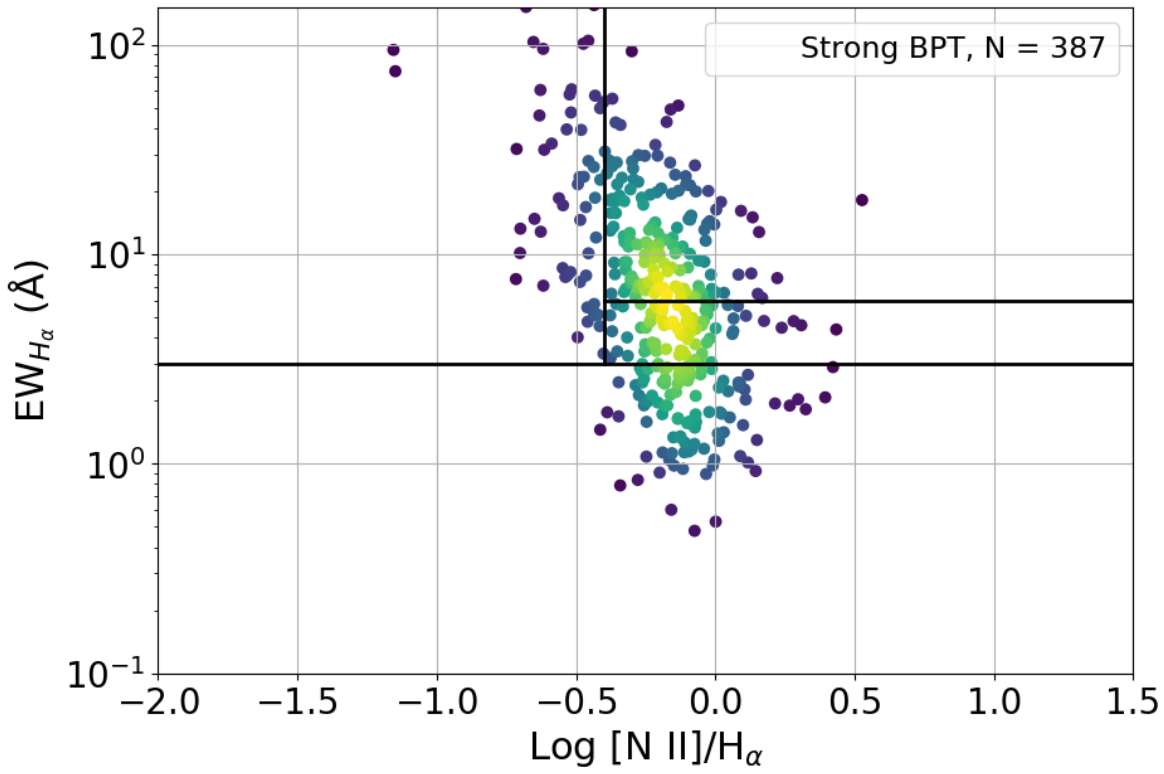

Figure 2.6: WHAN diagram with BPT selected galaxies. The dots are colour-coded by their relative point density. The majority of BPT selected galaxies (*'AND'*-selected - N = 195) are considered strong AGNs in the WHAN diagram while the non-AGN WHAN-classified galaxies are roughly evenly split between retired galaxies and star forming ones.

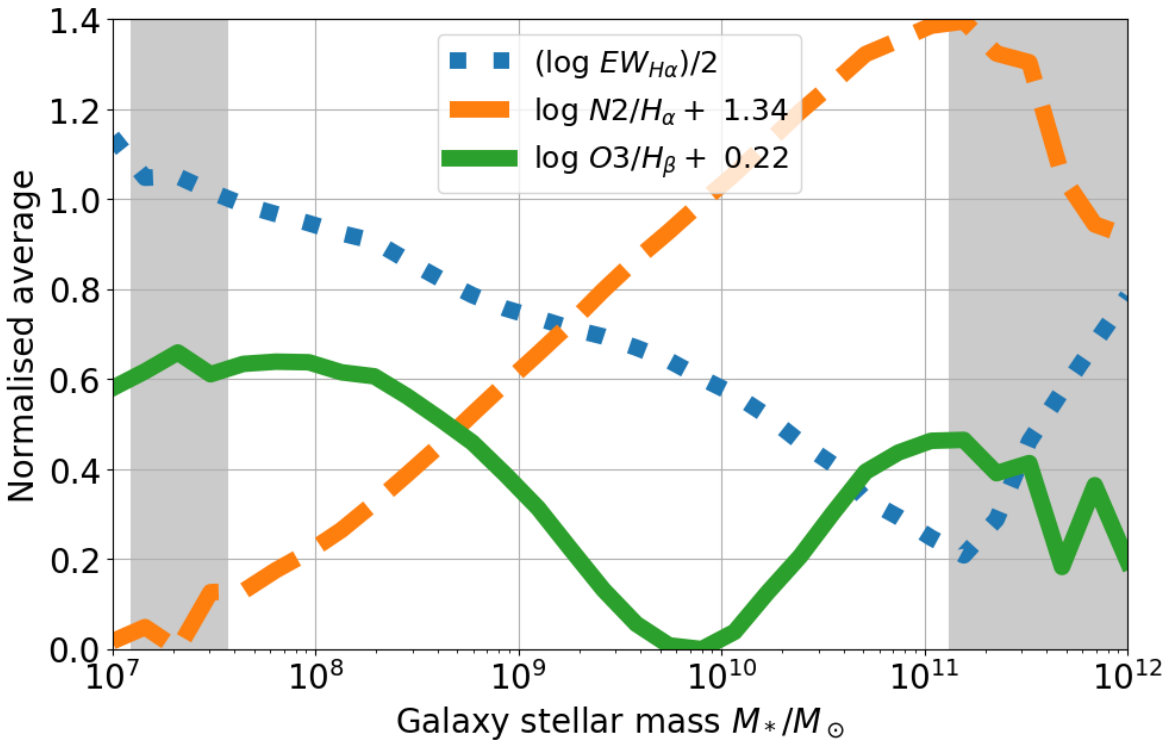

Figure 2.7: Average values of emission line ratios and EW_{H_α} function of mass. The log ratio values are shifted to be in the same area while the EW is log and then scaled by 0.5. The data consists of 32 linear log scale mass bins and the bins with less than 300 galaxies are shaded in grey.

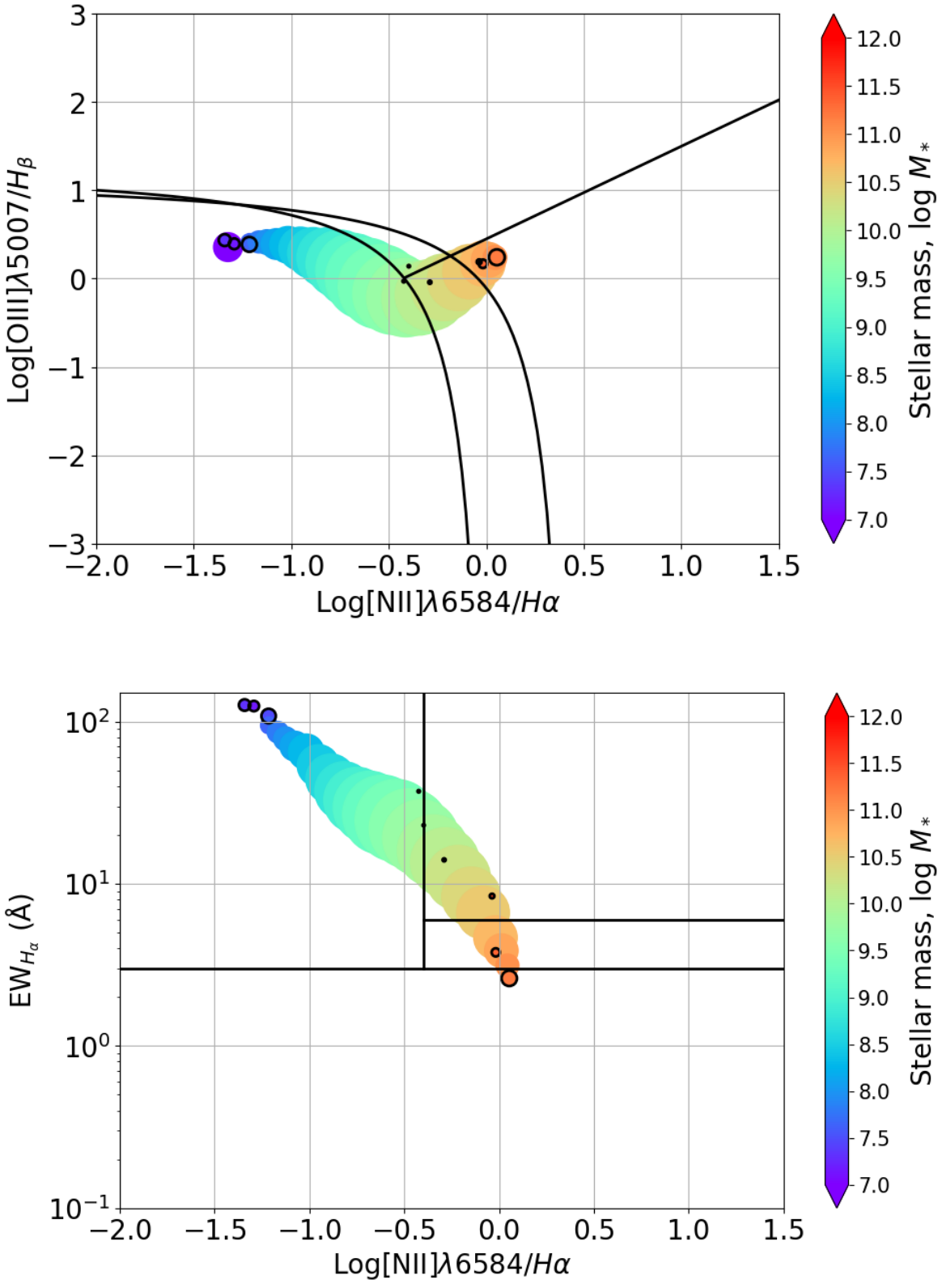

Figure 2.8: BPT and WHAN diagram showing mass trends. Each data point is the average values in 32 different mass bins and the size of the dot is scaled by the number of galaxies in that bin. Each bin has at least 300 galaxies in them unless surrounded by a black edge and otherwise contains between 312 and 11 581 galaxies.

each of the 152 in the original subsample. The matching critera involves mass, redshift, and colour. The critera are:

1. $\|1 - M_*/M_{*,\mathrm{AGN}}\| \leq 0.20$
2. $\|z_{\mathrm{AGN}} - z\| \leq 0.01$
3. $\|(u_{\mathrm{AGN}} - r_{\mathrm{AGN}}) - (u - r)\| \leq 0.4$

The matching criteria are similar to that of Cheung et al. (2015) but with stricter limits on mass and redshift. The masses of the galaxies in this sample are more similar than the galaxies in of Cheung et al. (2015) and are more numerous which allows for more strict criteria without eliminating all possible matches.

A stricter redshift interval is necessary due to that fact that the SDSS fiber covers different fraction of galaxies at different redshift – an effect that is very pronounced at lower redshifts. E.g, at $z = 0.005$, the 3" fiber covers 0.3 kpc while at $z = 0.055$, it covers 3.3 kpc. Thus, it is not the same reagion of each galaxy that is examined with redshift. A discussion of this effect can be found in Section 2.4.1.

This follows the same methodology as in other papers such as Penny et al. (2016). That study's sample size is smaller (39) since it is drawn from the smaller MaNGA survey. Therefore, the statistics are not directly comparable because the p-value from KS-testing changes with sample size (decreases with larger sample size). They are sufficiently similar to allow for adaptation of the method. The values shown in Table 2.3 shows the average p-values of these iterations.

2.3.2 BPT and WHAN comparison

To compare the two selection methods, the AGN-subsamples are classified in the other's diagnostic diagram (see Figure 2.5 and 2.6). While BPT galaxies tend to be in the AGN part of the WHAN diagram, WHAN galaxies are mostly in the star-forming or composite region in the BPT diagram. Interestingly, ~53 per cent of WHAN selected galaxies have $\mathrm{S/N}_{H\beta,[\mathrm{O\,III}]} \geq 3$, which is the same fraction as Cid Fernandes et al. (2010) found for all galaxies. This means that they are more robustly classified in the BPT diagram than the initial

(i.e before SNR rejection) BPT galaxies. The BPT diagram classifies the majority of WHAN galaxies as only star-forming but with a significant number of composite galaxies.

An interesting finding is that the emission line ratios and equivalent width of Hα all have clear mass trends. Towards lower stellar mass galaxies, [N II]/Hα decreases while EW_{H_α} increases. This means that galaxies move towards the upper left corner of the WHAN diagram – deep in the star formation region. This is in agreement with the literature on dwarf galaxies that they are very star forming (Kauffmann et al., 2004; Yang et al., 2007; Geha et al., 2012). Towards higher masses, the average EW_{H_α} drops to below 3 Å, which helps explain why the WHAN AGN fraction peaks around $M_* = 10^{10}\,M_\odot$ (see Figure 2.12 for a visualisation).

In the BPT diagram towards lower masses, the trend of [N II]/Hα moves the galaxies away from the vertical cut-off for AGN/LINER classification, and the [O III]/H_β trend for $M_* \leq 10^9\,M_\odot$ is declining while it increases afterwards. In BPT, whenever [N II]/$H_\alpha \gtrsim -0.1$, galaxies are classified as either composite or pure AGNs. This condition is met for the average [N II]/Hα for galaxies $2 \times 10^{10}\,M_\odot \geq M_* \geq 2 \times 10^{11}\,M_\odot$ possibly explaining the BPT AGN fraction peak around $M_* \sim 10^{11}\,M_\odot$.

2.3.3 Local neighbourhoods of dwarf AGNs, 10NN

From the KS-testing, it appears that there are no discernible differences between the distances to the 10th nearest neighbours of any of the subsamples. This means that the density of the environment does not seem to affect AGN activity in dwarf galaxies, and the implications will be discussed further in 2.4.2. Figure 2.9 shows the 10NN distribution BPT, WHAN and NOT and similar figures for the *AND*, *OR*, and weak BPT and *NOT* samples can be found in the appendix. The statistics can be found in Table 2.3.

The average projected separations are between $d_p = 3.7 - 4.3$ Mpc with $\sigma = 2.0 - 2.2$ Mpc, which further shows that the distributions are indiscernible. The BPT and WHAN distributions tend to lie at the lower end of both intervals (respectively, 3.7 ± 2.0 Mpc and 4.1 ± 2.2 Mpc) suggesting they do prefer denser environments compared to NOT galaxies (4.3 ± 2.2 Mpc) though the KS statistics make this inconclusive.

Weak galaxies

The only subsample that shows a significant difference in distribution is BPT selected galaxies with low S/N and thus rejected as AGNs. Though it is uncertain whether this subsample has AGN characteristics due to low S/N, this subsample will be referred to as 'weak BPT'. These galaxies will be discussed further in Section 2.4.2.

2.3.4 Immediate neighbourhood of dwarf AGNs, Δv_{NN}

Similarly to 10NN, this measure shows no discernible between the any of the subsamples – even weak emission line galaxies. This seems to suggest that the velocity difference to a dwarf AGN galaxy's nearest neighbour is not deciding factor in its AGN activity. The distributions can be seen in Figure 2.9. A notable anomaly/feature is an excess at around 600 km s^{-1} in the BPT distribution, but this 'bump' does not significantly affect the KS statistics. However, the bump does seem to make the BPT distribution have the highest average Δv.

Overall, most galaxies tend to have a very small velocity differences to its nearest neighbour. There is no adjustment for the fact that the velocities are only in the line of sight, which partially explains the shape of the distribution.

2.3.5 Visual inspection

A visual inspection is carried out to look for any morphological disruptions. Such tidal interactions are not necessarily quantified by the two primary environment estimation methods and thus serves as a complementary qualitative method. Two subsamples are used: The *'AND'* subsample and a simlar sized control sample from the *'NOT'* subsample. The control subsample is comprised of galaxies that are matched in stellar mass, colour, and redshift to the AGN galaxies. The matching criteria are the same as in Section 2.3.1. The purpose of this is to look for any obvious asymmetries or tidal interactions with neighbours.

The images are 40" by 40" and from SDSS. They are characterised by a number of properties which will be explained below. Figure 2.10 showcases 4 of these properties in the different subsamples. A number of galaxies are rejected due to either appearing as a massive galaxy or observational artefacts.

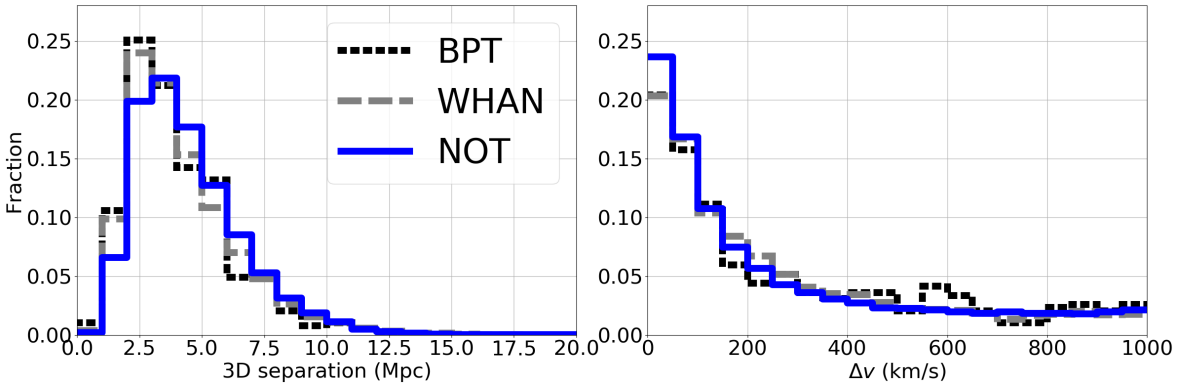

Figure 2.9: *Left*: Projected spatial separation from the dwarf AGN galaxies to their 10th nearest neighbour and *right*: The absolute velocity difference between dwarf AGN galaxies and their nearest 2D separated neighbour (within $\pm 1\,000\,\mathrm{km\,s^{-1}}$). Three samples are plotted: Black dotted are BPT-selected galaxies, grey dashed are WHAN-selected galaxies while blue solid are galaxies that appear in neither of the other samples. Generally, there are no discernable differences between the three distributions in either case. The BPT bump near $600\,\mathrm{km\,s^{-1}}$ is not statistically significant. See Section 2.3.1 and Table 2.3 for statistics.

1. Unstructured. Does the galaxy lack any morphology or have any discernible structure, e.g spiral arms or dust lanes?

2. Bright core. Does the galaxy have a concentrated peak in brightness at the centre?

3. Elongated. Is the galaxy flatter than roughly an ~E6 galaxy?

4. Compact. The appearance of the galaxy is that only of a core and confined within 4″

5. Spiral. Does the galaxy show clear spiral arms from either an angle or face-on?

6. Neighbour. Does the galaxy have a neighbour in the image? A neighbour is defined as a source of roughly the same colour and brightness.

7. Asymmetric. Does the galaxy have asymmetric features such as a tidal tail, a warped appearance or unevenly distributed light.

Item (i)-(v) are descriptive of the intrinsic properties of the galaxies whereas (vi) and (vii) can be used to infer properties about their environments. The numbers between the two samples are similar (within $\sim$ 6 %-points) in most aspects except frequency of bright cores and being compact.

The higher frequency of bright cores and compactness of AGNs can be explained by the intrinsic properties of AGNs: They are defined as having a high degree of radiation originating from the nucleus, which explains the bright cores. Furthermore, a galaxy with weak galactic emission or being at a high redshift with a relatively strong AGN results in only the core region being visible thus appearing compact on the sky.

The neighbour numbers are somewhat comparable to Ellison et al. (2019) that found roughly 78 percent of non-AGN galaxies and 64 per cent of AGN galaxies to be isolated and whereas the numbers for this study is 78 per cent and 75 per cent respectively. Tidal features (equivalent to asymmetries in this study) for AGNs in Ellison et al. (2019) are higher than the fraction in this study, but their numbers for non-AGNs are comparable to the control sample here. The sample size in this study is lower by a factor of $\sim$ 6, though, and the approach here is not as meticulous as Ellison et al. (2019), which may partially account for the difference.

Table 2.4: Number (fraction) of galaxies showing visual properties in AGNs (*'AND'* subsample) and a control sample.

Parameter	**AGN (181)**		**Control (192)**	
Structureless	156	(86%)	177	(92%)
Bright core	163	(90%)	81	(42%)
Elongated	61	(34%)	62	(32%)
Compact	76	(42%)	59	(31%)
Spiral	17	(9%)	18	(9%)
Neighbour (N)	45	(25%)	43	(22%)
Asymmetric (AS)	21	(12%)	19	(10%)
N+AS	7	(4%)	5	(3%)

Furthermore, it should be noted that mass and redshift distributions are different in Ellison et al. (2019), so some differences are expected. More specifically, their sample goes to $z \simeq 0.25$ and only 3 out of 1 124 optical galaxies are low mass galaxies and 7 out of 254 mid-IR galaxies. Also, more massive galaxies are larger and will have a larger angular size on average than the dwarf galaxies, which would make tidal features easier to spot[4].

2.3.6 Other parameters

The other parameters that are compared are: Stellar mass, M_*, redshift, z, and r-magnitude. The conclusions from them are mostly used to check whether or not the samples behave as expected – e.g higher AGN fraction at higher masses. The distributions are in Figure 2.11.

The mass distribution is not surprising. AGNs are primarily found in the higher mass systems which agrees with e.g Miller et al. (2003); Kauffmann et al. (2004); Sabater et al. (2013), and these distributions follow the same trend. BPT and WHAN are indiscernible with a very rapid rise from $\log(M_*/M_\odot) \approx 9.2$ while *'NOT'* plateaus around that mass and falls

[4]A galaxy the size of the Milky Way (MW, $d \simeq 30$ kpc) at $z = 0.25$ will have the same angular size as an LMC-like galaxy ($d \simeq 4$ kpc) at $z = 0.03$. Thus, most galaxies in Ellison et al. (2019) are more resolved than even the largest galaxies in this sample at $z = 0.03$ – yet the sample of this study even goes to $z = 0.055$. For a MW-like to have the same angular size as an LMC-like galaxy, its redshift would have to be $z = 0.48$

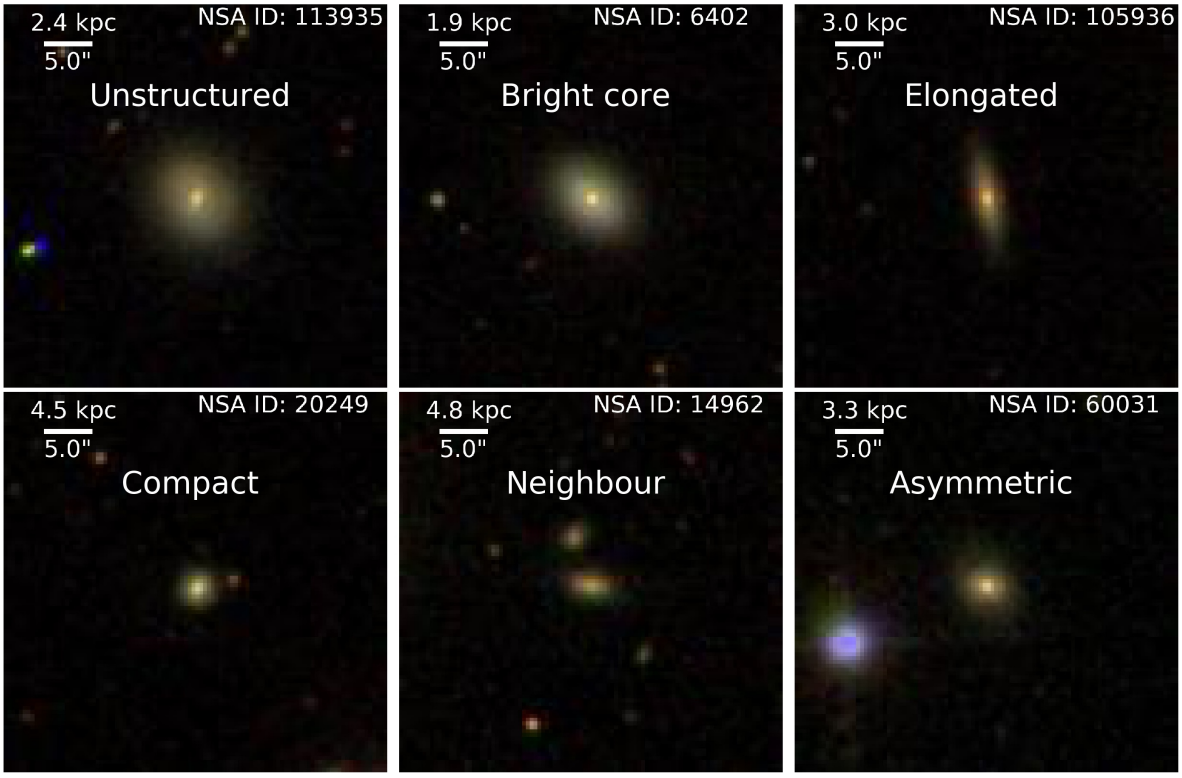

Figure 2.10: Example of cutouts of SDSS data. 6 different observational properties are shown here (only excluding spirals). Detailed information regarding visual inspection can be found in Section 2.3.5

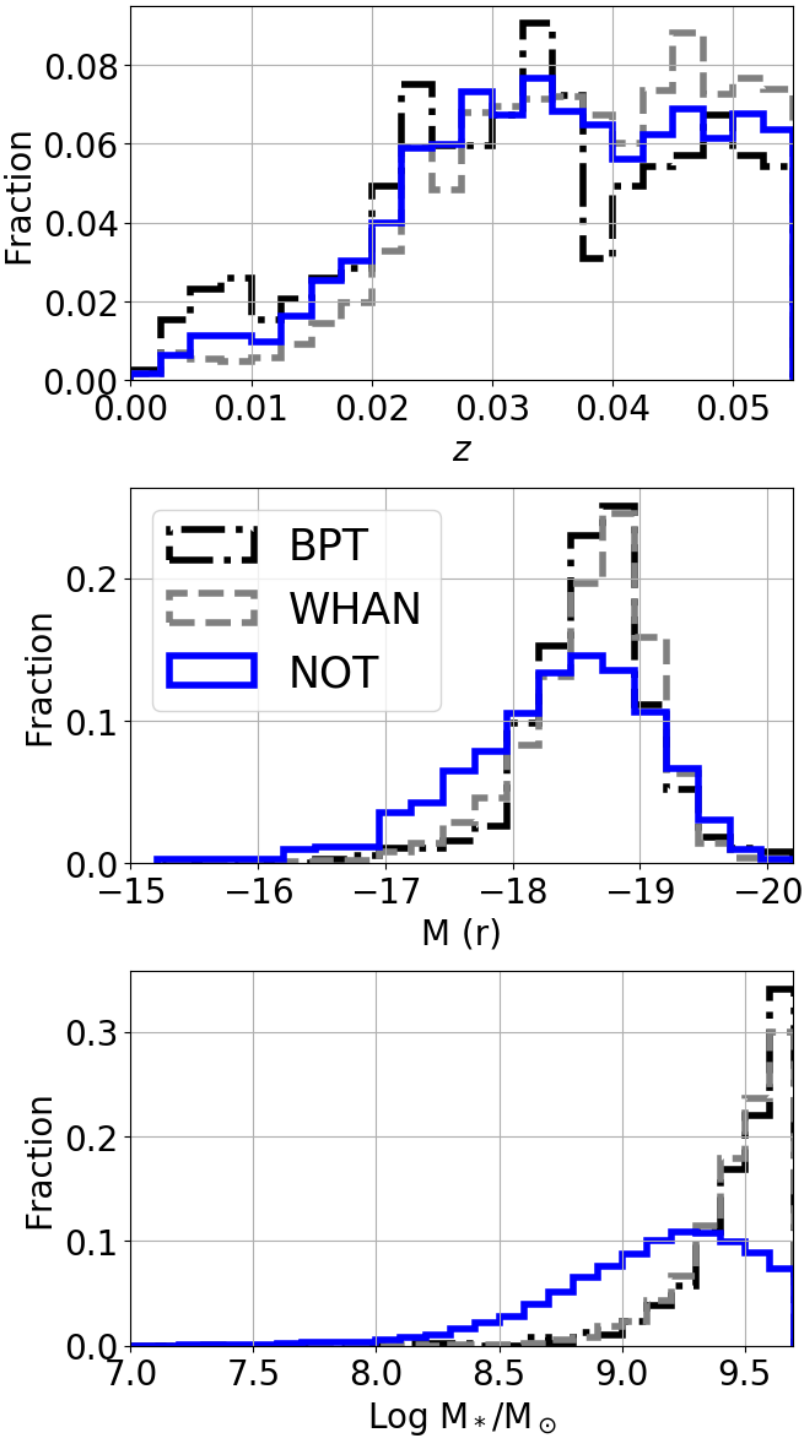

Figure 2.11: Mass, redshift, and magnitude distributions. The black dash-dotted distribution is BPT-selected galaxies, the grey dashed is WHAN-selected ones while the blue solid is the NOT selection. Regarding mass, AGN galaxies are increasingly common towards higher masses while the NOT galaxies peak around log 9.3 $M_{\odot}$. For redshift, WHAN and NOT galaxies follow almost the exact same trend, though WHAN has a slight excess at higher redshifts. BPT galaxies are slightly favoured at lower redshifts, but overall follows the same trend. Lastly, on magnitude, AGN galaxies are in general brighter than the NOT galaxies.

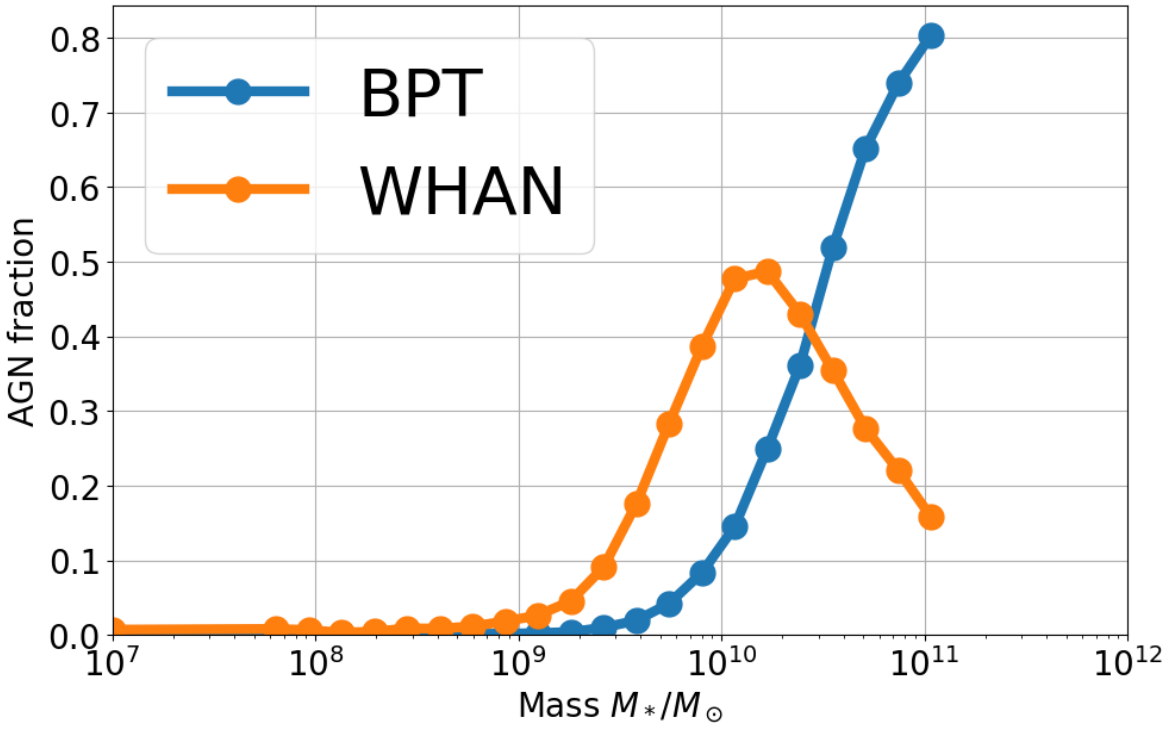

Figure 2.12: AGN fraction as function of mass. The fraction is calculated as the number of galaxies fulfilling the respective AGN criteria divided by the total number of galaxies in that mass bin that also fulfill the S/N criteria outlined in Section 2.2.2. For high masses ($\geq 10^{11}\,M_\odot$), care has to be taken because of incomplete data, which is why there are no mass bins after $\sim 10^{11} M_*/M_\odot$ since a requirement is that there has to be more than 300 galaxies in one bin.

off steadily – a trend that is partially caused by AGN distributions 'stealing' galaxies at these masses. The AGN fraction as a function of mass bin (Figure 2.12) does reveal that the mass cut of $M_* \leq 5 \times 10^9\,M_\odot$ is where the AGN fraction starts changing the most in WHAN

The majority of the low mass galaxies are found at redshifts above $z = 0.02$. Generally, all distributions follow the same pattern, although BPT galaxies seem to be found and lower redshifts whereas WHAN has more objects at higher redshifts.

The magnitudes of the subsamples show that AGN galaxies tend to be brighter than non-AGN galaxies. There are quite a few parameters to untangle, though. Since the mass distribution of AGNs in this sample are shifted towards higher masses, it seems natural that the magnitudes are shifted accordingly. However, the masses are derived from the r-magnitudes (Blanton et al., 2011), which may mean that some of the luminosity from the AGN contributes to the stellar mass estimate. What it shows at least is that the subsamples behave as expected with active galaxies being more luminous than regular dwarf ones.

2.4 Discussion

Overall, the apparent non-dependence on environment of dwarf galaxy AGN hosts found in this study is in line with existing literature (e.g Miller et al., 2003; Kauffmann et al., 2004; Padilla et al., 2010; Man et al., 2019, although several of these studies find other properties that trend with environment like AGN colour or [O III] strength). It suggests that AGNs in dwarf galaxies react similarly to environment as regular galaxies.

This non-uniqueness of dwarf galaxies is surprising since the gravitational potential, cold gas content (e.g Bradford et al., 2018), and morphology of dwarf galaxies are different to regular galaxies. Sabater et al. (2013) suggest the most important factor in fueling AGN activity is having a supply of gas to feed the core, and the cold gas content is more vulnerable in dwarf galaxies due to their shallow gravitational potentials. Dense local environments have a detrimental effect on the cold gas reservoirs by stripping and heating it while strong galaxy interactions can enhance AGN activity by perturbing the otherwise stable structures, though neither effects can be inferred from the results of this study..

The implications are that; a) environment plays an insignificant role on AGN activity

regardless of host mass, or b) the environmental effect on AGN activity is either delayed or obfuscated such that the environment measurement methods do not probe the desired properties, or c) the selection methods for AGN cannot be applied directly to the low mass regime due to biases, or d) a mix of the above.

2.4.1 SDSS fiber aperture bias

The SDSS fiber aperture does not cover the same fraction of a galaxy as a function of redshift, which can be seen from Figure 2.13. In the lowest redshift bin, the fiber does not even cover the whole core. This means that measured emission lines are affected by the redshift of the galaxy. Emission line flux will be left out if the aperture size covers less than the core while AGN signatures may be drowned out by SF emission for larger fractions (e.g Trump et al., 2015).

Although the galaxy cores are not resolved, if a core region is smaller than the fiber diameter, the total AGN flux will not be recovered which will lead to inaccurate emission line measurements thus making equivalent width methods such as WHAN inaccurate. The emission line ratios may not be affected though making the BPT diagnostic robust.

A test is performed to see if excluding galaxies whose core regions are not fully covered changes any results. From Figure 2.13, galaxies with $z \leq 0.02$ are excluded and the KS-testing described in Section 2.3.1 is run over this new sample of galaxies. This exclusion reduces the number of dwarf galaxies from 62 258 to 51 971, BPT galaxies from 387 to 326 and WHAN from 4 323 to 4 029.

The results do not differ from those in Table 2.3 with the exception of BPT galaxies in 10NN. The p-value drops to 0.05 when comparing BPT galaxies versus non-AGN galaxies (*'NOT'* galaxies). Raising the sampling size to 326 decreases the p-value even further suggesting that the distributions are *not* similar. However, excluding nearby galaxies does not necessarily mean that it is an aperture effect. Only more energetic dwarf AGNs are visible at higher redshifts, so this may be a luminosity bias. Clearly, a more in-depth analysis is required to disentangle this result, but this is outside the scope of this study. This result does mean that it is not possible to conclusively rule out an environmental connection in dwarf

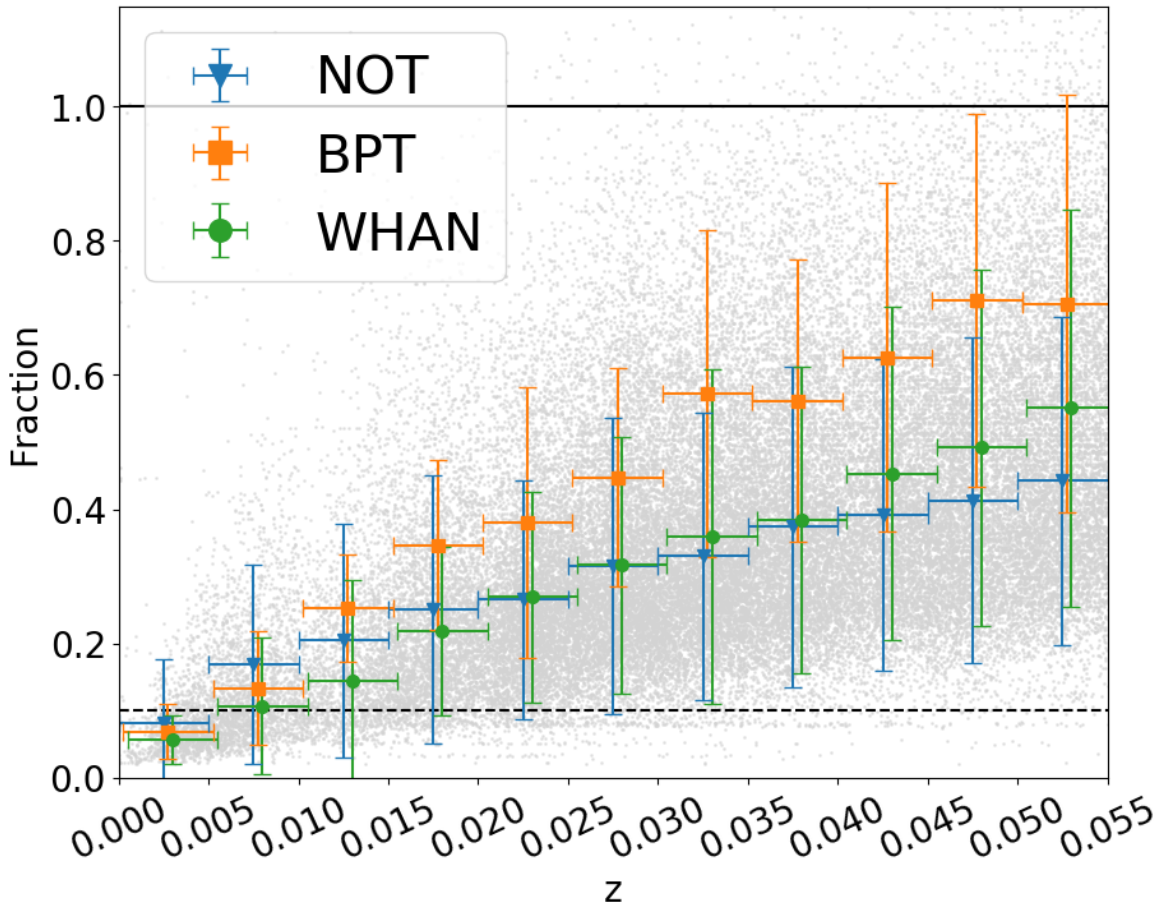

Figure 2.13: Plot of fraction of a galaxy covered by the SDSS fiber aperture as function of redshift. The size of a galaxy is taken to be its petrosian 90 per cent light radius, R_{P90}, and is core is defined as $0.1R_{P90}$. The grey dots are all dwarf galaxies overplotted with median of different subsamples. Errorbars show the interquartile range. The subsamples are split into redshift bins with $\Delta z = 0.005$. The solid line at 1.0 equals the R_{P90} while the dashed one at 0.1 is $0.1R_{P90}$.

BPT AGN galaxies.

For further analysis, low mass galaxies remain in the sample, but this presents another concern, namely whether the offset between the center of the galaxy and the fiber position is large enough for the fiber to not cover the core at all. The fiber will be fully offset if $\Delta\text{pos} > R_{\text{core}} + R_{\text{fiber}}$ which is the case for 8 103 of the 62 258 dwarf galaxies. Assuming the positions of the galaxies are equal to their core region, the spectra of these galaxies do not include their nuclei. However, most of these (8 100) are not linked to any spectroscopy runs and thus have no emission line fluxes. Of the 3 with emission lines, none of these galaxies appear in the BPT subsample but 2 of them appear in the WHAN subsample. Inspection of images of these WHAN galaxies with overplotted apertures reveals that their R_{P90} are underestimated, but even then, the fiber position is not covering the core (by visual inspection). The NSA does list off-center SDSS spectroscopy as one of its caveats. However, only 2 out of all 4 323 WHAN galaxies are affected by this so this effect is ignored.

Regarding galaxies towards higher redshift, SF dilution (i.e weakening of the AGN signal due to an increasing ratio of emission by SF processes to AGN emission) increases since the fiber encloses an increasing fraction of the galaxies' total light. Moran et al. (2002) demonstrated that this effect biases against narrow line AGNs since that emission is drowned out by the host galaxy light. While no AGN detection limits have been imposed in this study like the one used in Trump et al. (2015), the importance of such methods appears crucial in studies focused on dwarf AGN selection improvement.

2.4.2 On the environment and nearest neighbours

Accepting the fact that environments do not affect AGN activity should not be done unconditionally as multiple studies have found connections (although sometimes weak) between AGN activity and environment (Miller et al., 2003; Kauffmann et al., 2004; Sabater et al., 2013; Amiri et al., 2019). Some find that specific types of AGNs (e.g strong [O III] emitters or redder AGNs) are dependent on the local environment, so further subclassification of AGNs in dwarf galaxies may show a connection. However, classifying AGNs in the dwarf mass regime has a number of challenges. Emission line ratios and equivalent widths follow a mass

trend (see Figure 2.7) and AGN characteristics in dwarf galaxies are hard to distinguish from e.g SF (Trump et al., 2015), which suggests that AGNs in the low mass-regime have to be treated differently. As noted by Mackay Dickey et al. (2019), active dwarf galaxy samples can include many false positives. This results in non-AGN galaxies being included in the AGN samples and the statistics would be biased towards regular galaxy distributions and thus not representative of the AGN population - an issue also raised by Hainline et al. (2016)

Furthermore, while the environment estimation methods used in this study are tried and tested in other work for other purposes, there is a risk that they do not properly gauge the desired properties or the properties are not showing up in the statistics due to obfuscating factors such as trigger time lag (the time delay it takes for AGN activity to begin after an interaction or harassment event, see e.g Schawinski et al., 2007; Pimbblet et al., 2013; Shabala et al., 2012, 2017) or SF contamination (Trump et al., 2015). Other environment estimation methods may reveal a connection to AGN activity, but the findings in this paper are in line with existing literature (e.g Miller et al., 2003; Kauffmann et al., 2004; Padilla et al., 2010; Sabater et al., 2013; Sabater et al., 2015; Man et al., 2019). Therefore, using other relatively simple environment estimates will likely show similar results but more complicated ones such as the so-called tidal force estimator may find a difference in immediate environment (Sabater et al., 2013). Other options for improving the environment estimation method involves higher resolution and better spatially resolved observations such as IFU surveys. They enable methods that more correctly gauge e.g recent merger history (e.g Penny et al., 2018, who inferred recent mergers from kinematically offset cores).

Regarding galaxy tidal interactions, this study found no dependence in Δv to the nearest neighbour, although it is established that mergers can trigger AGN activity (see e.g Miller et al., 2003; Sabater et al., 2013; Ellison et al., 2019). Treister et al. (2012) suggest that they are not necessary – only for the brightest AGNs. As mentioned previously, using different methods such as distance to nearest bright neighbour (Penny et al., 2016) or tidal force estimator (Sabater et al., 2013), this sample may show an excess of galaxy interactions. However, Kaviraj et al. (2019) found no excess of dwarf merger rate compared to regular merger rate for non-AGN dwarf galaxies suggesting that mergers are not important for AGN

activity in the low-mass regime.

Ellison et al. (2019) found that mergers can trigger AGN activity, though it may not be the dominant trigger. Furthermore, the fraction of disturbed galaxies are different depending on AGN selection method with mid-IR candidates being more often disturbed ($\sim$ 60 per cent) than optical ones ($\sim$ 30 per cent). This excess in mid-IR selected AGNs was also found by Satyapal et al. (2014). Furthermore, Ellison et al. (2019) note that the excess of morphologically disturbed galaxies with AGN activity compared to disturbed non-AGN galaxies does increase with host mass and AGN luminosity. Conversely, the excess decreases towards lower mass galaxies giving credence to the notion that mergers are of lesser importance to AGNs in dwarf galaxies.

However, the majority of if the galaxies in Ellison et al. (2019) are galaxies with $\log M_*/M_\odot > 9.5$, and therefore extending the findings into the low mass regime should be done with care. The luminosity on [O III] are also several orders of magnitude brighter. If the arguments from Ellison et al. (2019) are extended to dwarf galaxies, it is based on the assumption that AGNs in both low- and high mass galaxies are similar and can simply be scaled accordingly. Searching for mid-IR AGN dwarf galaxies has proven difficult as remarked by Lupi et al. (2020). This means that a comparative study with Ellison et al. (2019) with dwarf galaxies is a difficult task.

Furthermore, the findings of Lupi et al. (2020) may point towards that AGNs in dwarf galaxies may be different from regular AGNs. It is remarked by e.g Mendez et al. (2013) and Azadi et al. (2017) that different wavelength diagnostics probe different AGN populations. It is therefore not possible to to conclude whether findings from Ellison et al. (2019) can be extended into the dwarf mass regime or not.

In this work, no restrictions were put on the neighbouring galaxy. Other work on this area such as Penny et al. (2016, 2018) required $M_k < -23$ of the neighbour since the mass of the neighbour decides how strong the tidal interactions are, and may be what is required to drive gas to the central region. Conversely, a strongly disturbed dwarf galaxy may not have sufficient gas reservoirs to feed an AGN. From the method and results of this study alone, neither scenario is favoured.

To explore this further, manual or automated visual inspection of the AGN sample may be required to give clues to properties such as morphology, nearby neighbours, and immediate environment. Morphology such as large-scale bars have found be correlated to AGN activity in e.g Galloway et al. (2015) (who suggested that a bar increases the probability of an actively accreting central black hole), while other studies such as Cheung et al. (2015) did not find this connection. While dwarf galaxies can be well-structured, they are often irregular (due to their low gravitational potential) and thus do not have a morphology that triggers AGN activity (e.g a bar).

However, visual inspection of SDDS images of the active dwarf galaxies (specifically the majority of the *'AND'* subsample, $N = 195$, see Section 2.3.5) in this study has revealed no excess of morphology disturbances compared to a similar sized control sample (from *'NOT*), which can be seen in Table 2.4. This is in line with what Kaviraj et al. (2019) found and Satyapal et al. (2014) noted that that optically selected AGNs do not tend to show an excess of mergers whereas mid-IR ones did. Goulding et al. (2017) found a similar excess in mid-IR data.

Complicating the morphology discourse further is the fact that Kruk et al. (2017) found that dwarf galaxies can be morphologically disturbed when found in isolation. What it means is that morphological disruptions of dwarf galaxies does not necessarily mean that they have been tidally affected or harassed by companion galaxies and as such, morphology is not an indicator of environment.

Another important complication to consider is delayed triggering times of AGNs, which is suggested by e.g Pimbblet et al. (2013); Shabala et al. (2017). The idea is that AGN activity does not start during an encounter or disturbance but rather 0.2-0.3 Gyr later. This timescale is the same order of magnitude as crossing time in rich galaxy clusters, which means that any present day AGN activity would be difficult to pin on a past event. Penny et al. (2018) also found dwarf galaxies with post-starburst spectra which supports the hypothesis that AGN activity can be delayed from an interaction since star bursts tend to be found in actively merging or harassed systems (Hopkins et al., 2006).

This would mean that the methods used in this study are not suitable to examine these

properties. Other research such as Penny et al. (2018) found kinematically offset cores which could indicate accretion of IGM or merger, and the analysis of spatially resolved spectroscopy may be required to gauge the galaxy's past interaction history.

Weak galaxies

The only subsample to show a different environment is *weak BPT galaxies*, although it is misleading to label them as BPT galaxies since their AGN activity is not definitive. As seen in Figure 2.14, the are primarily retired galaxies – a classification defined in Stasińska et al. (2008). Stasińska et al. (2008); Stasińska et al. (2015) discusses the implication that ionising radiation from evolved stellar population can place a galaxy in the LINER-region (and in the Seyfert-region, to a lesser extend) of the BPT diagram. They argue for using the equivalent width of Hα as a method to break the degeneracy of the BPT classified LINERs and retired galaxies.

Although most of the weak BPT galaxies are in the BPT-Seyfert region (Figure 2.14), their exact positions are uncertain as their S/N is too low for proper line ratio measurements. Their horizontal positions, though, are mostly correct as measurements of Hα and [N II] are of good quality. This would put most of them in the composite or LINER region of the BPT diagram.

It is not surprising to find retired galaxies in denser environments as the connection between dense environments and SF quenching is well-established (Balogh et al., 1998; Lewis et al., 2002; Peng et al., 2010, 2012; Penny et al., 2016; Spindler et al., 2018). A number of processes such as ram-pressure stripping, heating of cold gas reservoirs, and galaxy harassment have been found theorised to quench star formation and these processes are more likely to happen in dense environments. Therefore, this subsample behaves as one would expect.

2.4.3 On selection method bias

A concern in this study is the extension of regular optical AGN diagnostic in to dwarf mass regime – a concern explored in detail in e.g Cann et al. (2019) – and studies like Trump et al. (2015), Penny et al. (2018), and Mackay Dickey et al. (2019) does bring into question whether

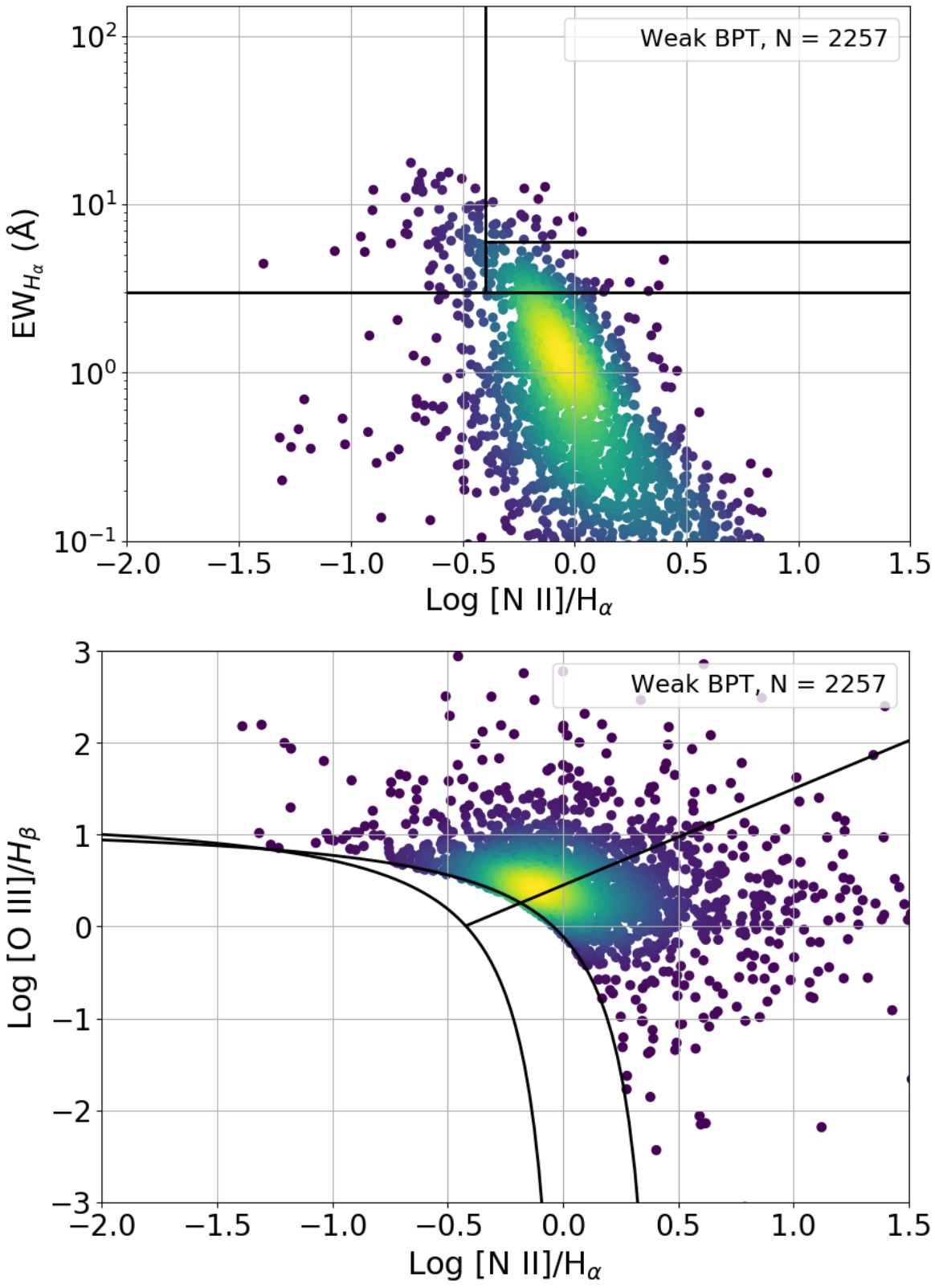

Figure 2.14: WHAN and BPT diagrams with weak BPT selected galaxies with dots colour-coded by their relative point density. These galaxies lie primarily in the retired region in the WHAN diagram while their positions in the BPT diagram are uncertain due to low S/N on the y-axis.

AGNs in dwarf galaxies are robustly identified in regular diagnostic tools.

One observation from this study is that there a clear mass trends in both BPT and WHAN, and extending these diagnostic into the low-mass regime may carry biases, which are often not corrected for. As shown in e.g Reines et al. (2013); Sartori et al. (2015); Stasińska et al. (2015); Baldassare et al. (2018); Mackay Dickey et al. (2019); Cann et al. (2019), a number of AGNs will not be identified with standard optical AGN diagrams even though they clearly show AGN characteristics in other diagnostics (e.g X-ray or mid-IR selecttion), or numerous non-AGN will be included in a sample if the selection criteria are too lenient – something that is exaggerated in low-mass galaxies (Stasińska et al., 2015; Trump et al., 2015; Hainline et al., 2016).

Trump et al. (2015) suggest that SF dilution biases against AGN in low mass galaxies. The emission lines are drowned out by SF radiation and also suggest that AGNs are fueled by the same gas as SF resulting in SF 'stealing' available gas from the AGN and preventing it from reaching very high energy outputs. The consequence is that low mass galaxies have weaker relative AGN emission compared to high mass galaxies. In order to correct for this, one solution could be to mass-weigh emission line ratios (and equivalent widths). However, it is well-established that AGN fraction is strongly correlated with host mass, and finding a correction factor on this parameter may prove difficult. However, this mass bias may be caused by selection biases due to aperture (see Section 2.4.1) – something that Moran et al. (2002) argue.

Studies in other wavelengths may reveal further AGNs in dwarf galaxies, and this has been explored in e.g Lupi et al. (2020); Birchall et al. (2020) in mid-infrared and X-ray respectively. However, Lupi et al. (2020) remark that the poor resolution of mid-infrared surveys and contamination makes this wavelength regime a bad choice for identifying AGNs in dwarfs. Birchall et al. (2020) found that out of 4 331 dwarf galaxies, 61 show AGN activity, and 85 per cent of these identified AGNs did not show up in optical wavelengths. This suggests that X-ray data is suitable to complement optical data in search of dwarf AGNs. Unfortunately, the X-ray data and coverage of the sky is limited, but it may be used in conjunction with optical data sets to find a potential correction function for optical diagnostics. Baldassare et al. (2018) used

long-term optical variability to identify dwarf AGN and noted that star formation dilution and low metallicity may be likely reasons why AGNs are missed in dwarf galaxies. Using nuclear variability to identify AGNs would circumvent problems with correlating observations across wavelength regimes or mass-weighing emission lines –two methods which carry limitations discussed earlier in this section. The obvious downside to this approach is the requirement of observations at different times over several years (data from Baldassare et al., 2018, spans over ∼5 years) and the need to spatially resolve cores of dwarf galaxies, which are very small.

Another finding in this study is that the WHAN diagram tends to classify low mass galaxies with low S/N on Hβ and/or [O III] as retired (Figure 2.14). This suggests that using the WHAN diagram on BPT selected AGNs is a fast way of identifying contaminating retired galaxies in AGN samples – at least in low-mass samples. This is assuming that WHAN classified AGNs are indeed AGNs. The AGN fraction as function of stellar mass (Figure 2.12) shows that the two methods find different AGN fractions and follow different mass trends. As noted by Cid Fernandes et al. (2010), WHAN tends to probe weaker AGNs, which obviously are more common in lower stellar mass galaxies.

2.5 Conclusions

The main discussion points will be summarised in this section. For a short summary, skip to the end. This study finds that the environments of AGN dwarf galaxies are no different than the environments of regular dwarf galaxies regardless of AGN selection method. There is neither a difference in local galactic density nor a velocity difference to the nearest neighbours suggesting that the main AGN trigger is an internal process. However, the non-dependence found can be a result of the method – either from biases in the AGN selection methods or from not taking various factors such as time delay, mass trends, or SDSS fiber aperture bias into account.

For example, using only galaxies with $z \geq 0.02$ – basically galaxies whose whole core region is covered by the SDSS fiber – the environments of BPT galaxies are distinguishable enough from a matched control sample of non-AGN galaxies to show up as statistically significant. However, this effect may be due to other reasons than just fiber aperture such

as e.g luminosity. Without the redshift restriction, the only subsample to show a difference in environment is weak BPT galaxies that show up as retired galaxies in WHAN. This subsample prefers denser environments, which makes sense because this subsample most likely consists of quenched star formation and dense environments are known to cause star formation quenching.

The analysis also looked at other galactic parameters. The distributions of stellar mass, redshift, and r-magnitudes were used as a test to see how they compare to existing literature. The stellar masses and magnitudes behave as expected with AGNs tending to be brighter and more frequent in higher mass galaxies. The redshift distributions between the samples are slightly different with WHAN tending to be found at higher redshifts and BPT at lower redshifts compared to the regular galaxies. This might be due to observational effects – BPT requires high quality measurements of weak lines and thus favours brighter and closer galaxies while WHAN probes weaker AGNs that are harder to detect at higher redshift.

The environment description methods only probe the galaxies in their current environment, though, and does not take their past into account. Some research suggests a time delay of $\sim 0.2 - 0.3$ Gyr (Pimbblet et al., 2013; Shabala et al., 2017). Strong encounters in the past might have left an impression on the galaxies' morphologies, but visual inspection did not reveal any significant disturbances. This is in line with e.g Kaviraj et al. (2019), who did not find an excess of disturbed or merging dwarf AGN galaxies compared to a control sample. Other indicators might exist of past encounters or significant disturbances such as kinematically offset cores, so other diagnostics may be needed for better analysis and understandings.

As an attempt to avoid bias from the selection method, two AGN selection methods were used. However, the samples from both methods were indiscernible from each other and from regular dwarf galaxies regarding environmental analysis, which means that optical AGN features are not affected by the environment. The two methods themselves seem to probe slightly different galaxy populations. While most BPT galaxies are also identified as AGNs in WHAN ($195/296 \sim 66$ per cent), the majority of WHAN galaxies are classified as star-forming galaxies or composite ones in BPT ($4099/4294 \sim 95$ per cent). However, this

does not mean the WHAN AGN classification is untrustworthy. The advantage of WHAN is that it aims to probe weaker AGNs and as several studies have found, AGNs in dwarf galaxies can be diluted by star formation (Trump et al., 2015), be rejected because of intrinsic weak emission lines (Cid Fernandes et al., 2010), or have higher sensitivity to environment (Wetzel et al., 2013) – effects that all may weaken AGN signatures.

Regardless of AGN selection method, neither local nor immediate environment seem to play a role in triggering AGNs in dwarf galaxies judging from the similarity of environment between AGNs and non-AGNs. While this is in agreement with existing literature, there are a number of factors that weakens this conclusion. Firstly, whether the environment estimates actually gauge the desired properties can be called into question. It could also be that the observable environment parameters have changed since they first triggered the AGN activity. Secondly, the sample of AGNs are found using diagnostic tools developed for regular galaxies, and extending these to the low-mass regime carries biases that results in an unpure sample consisting of many non-AGNs.

A solution to the first issue would be to develop more complicated environment description tools as already seen in e.g Sabater et al. (2013), Baldassare et al. (2018), and Penny et al. (2018). Generally, these methods can be thought of as having a longer lookback time compared to simpler methods. Regarding the second issue, other wavelength regimes can help identify a large fraction of optically undiscovered AGNs (in Birchall et al., 2020, , optical diagrams failed to find 85 per cent of AGNs) or variability surveys (e.g Baldassare et al., 2018) also help. A third option is to mass-weigh emission line ratios – an option motivated by the fact that there are mass trends in optical diagnostic diagrams.

These findings can be summarised as follows:

- There is no difference in neither local or immediate environment between AGN dwarf galaxies and non-AGN dwarf galaxies suggesting that the environment does not play a role in triggering AGN activity.

- This non-dependence on environment was found regardless of selection method (BPT and WHAN), although a redshift-limited (due to SDSS fiber coverage) sample of BPT galaxies did show a difference in environment. Thus it is not possible to conclusively

rule out an environmental dependance of BPT galaxies.

- Concerns were raised regarding both the AGN selection methods and the environment selection methods. AGN diagnostics are calibrated to regular galaxies, and extending these into the low-mass galaxy regime may produce samples either including many non-AGN galaxies or excluding AGNs.

- Regarding the environment, the utilised methods probe the current environment, but the current environments of galaxies may not be the environments that triggered AGN activity – there may be a time delay before the onset of AGN activity

- For future work on the subject of AGNs in dwarf galaxies, involves calibration of AGN diagnostics such as the BPT and WHAN diagrams to low-mass galaxies. This may include using other wavelength regimes in the identification of dwarf AGN or develop a mass-weighing factor on emission line ratios since they trend with stellar mass.

Acknowledgements

We appreciate the thorough and insightful comments by the reviewers that helped improve the paper, especially in regards to aperture bias and constructing control samples. Furthermore, KAP acknowledges the support of the Science and Technology Facilities Council (STFC) through the University of Hull's Consolidated Grant ST/R000840/1. MTK acknowledges the support of University of Hull Astrophysical Data Science Cluster.

Funding for SDSS-III has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Science Foundation, and the U.S. Department of Energy. The SDSS-III web site is http://www.sdss3.org. SDSS-III is managed by the Astrophysical Research Consortium for the Participating Institutions of the SDSS-III Collaboration including the University of Arizona, the Brazilian Participation Group, Brookhaven National Laboratory, University of Cambridge, University of Florida, the French Participation Group, the German Participation Group, the Instituto de Astrofisica de Canarias, the Michigan State/Notre Dame/JINA Participation Group, Johns Hopkins University, Lawrence Berkeley National Laboratory, Max Planck Institute for Astrophysics, New Mexico State University,

New York University, Ohio State University, Pennsylvania State University, University of Portsmouth, Princeton University, the Spanish Participation Group, University of Tokyo, University of Utah, Vanderbilt University, University of Virginia, University of Washington, and Yale University.

Data availability

The data underlying this article is the NASA-Sloan Atlas (`v0_1_2`) and can be accessed at `nsatlas.org`.

3. Merger Histories and Environments of Dwarf AGN in IllustrisTNG

This chapter contains a study on merger histories and environments of dwarf galaxies AGN characteristics in simulations. The results were published in December 2021 in The Astrophysical Journal, Volume 922, Issue 2, id.127, 19 pp. and the work was carried out in collaboration with Kevin A. Pimbblet (University of Hull), Brad K. Gibson (University of Hull), Samantha J. Penny (University of Portsmouth), and Sophie Koudmani (University of Cambridge) and me as lead author. I have written and made 100 percent of the text and plots. It has been formatted differently than the published paper to fit in this thesis, and may also differ slightly in typography, but the results and science are unaltered.

Abstract

The relationship between active galactic nuclei activity and environment has been long discussed, but it is unclear if these relations extend into the dwarf galaxy mass regime – in part due to the limits in both observations and simulations. We aim to investigate if the merger histories and environments are significantly different between AGN and non-AGN dwarf galaxies in cosmological simulations, which may be indicative of the importance of these for AGN activity in dwarf galaxies, and whether these results are in line with observations. Using the IllustrisTNG flagship TNG100-1 run, 6 771 dwarf galaxies are found with 3 863 (~57 per cent) having some level of AGN activity. In order to quantify 'environment', two measures are used: 1) the distance to a galaxy's 10th nearest neighbour at 6 redshifts and 2) the time since last merger for three different minimum merger mass ratios. A similar analysis is run on TNG50-1 and Illustris-1 to test for the robustness of the findings. Both measures yield significantly different distributions between AGN and non-AGN galaxies; more non-AGN than AGN galaxies have long term residence in dense environments while recent ($\leq$ 4 Gyr)

minor mergers are more common for intermediate AGN activity. While no statements are made about the micro- or macrophysics from these results, it is nevertheless indicative of a non-neglible role of mergers and environments.

3.1 Introduction

An important part of galaxy evolution is the co-evolution of the central black hole and the central bulge. The black hole mass and the luminosity and mass of the bulge follow a tight correlation for classical bulges and elliptical galaxies (Marconi & Hunt, 2003; Gültekin et al., 2009; Alexander & Hickox, 2012; Volonteri & Bellovary, 2012; Kormendy & Ho, 2013). Although the evolution of the two components are closely linked, divergences from the trend suggest that the bulge and supermassive black hole do not follow the exact same channels of evolution (e.g Simmons et al., 2017).

Conventionally, the growth and even formation of elliptical galaxies and classical bulges are believed to be mainly through mergers (Kormendy & Ho, 2013). Galaxies in the process of merging often show undermassive super massive black holes (SMBHs) and several studies do not find a significant link between mergers and the so-called active galactic nucleus phase (AGN; a phase where the SMBH grows through gas accretion) both observationally (Villforth et al., 2016; Simmons et al., 2017; Kaviraj et al., 2019; Smethurst et al., 2019) and in simulations (Steinborn et al., 2018; Martin et al., 2018; Ricarte et al., 2019) lending credit to the fact that the two components evolve somewhat independently – although some studies find merger activity and type of merger linked to the type and strength of AGN (e.g Satyapal et al., 2014; Simmons et al., 2017; Donley et al., 2018; Shah et al., 2020).

The abovementioned growth phase, the AGN phase, is when a SMBH is accreting gas and material. In this phase, matter is being driven to the central region of the galaxy and is being deposited onto the black hole. There are a number of mechanisms thought to be able to drive gas and dust to the center, either through internal processes such supernova feedback or dynamic friction (commonly referred to as secular evolution) or through external ones such as ram pressure stripping (Gunn & Gott, 1972), galaxy interactions (Moore et al., 1996) or accretion from the intergalactic medium (i.e environmental effects).

However, studies examining black hole growth have primarily been focused on intermediate mass galaxies or high mass galaxies (understood as $M_* \geq 10^{10}\,M_\odot$, see e.g Di Matteo et al., 2005; Bower et al., 2006; Sijacki et al., 2009; Amiri et al., 2019). While these galaxies are easier to study with their greater brightness and size, they are not necessarily representative of lower mass galaxies and their evolution, which is evidenced by the difference in susceptibility to different quenching mechanisms between mass regimes of galaxies (Peng et al., 2010, 2012; Geha et al., 2012).

In order to decrypt the potential differences between populations, a good statistical basis is required. Obtaining a large number of sources requires large scale surveys. However, observations have traditionally had to choose between either large field of view or faint magnitude limits where deep surveys do not yield a high number of sources, but it would include the faint ones (i.e the low mass galaxies) while wide surveys would include many sources but little to no low mass galaxies. As prominent examples of this is the Sloan Digital Sky Survey (SDSS York et al., 2000) DR16 which covers 14 555 square degrees to a limiting magnitude of around 22 (*ugriz* bands) while the UltraVISTA (McCracken et al., 2012) covers 1.5 square degrees to a limiting magnitude of around 25 (Y band).

Similarly, in large-scale cosmological simulations, resolving dwarf galaxies requires low particle masses and/or low mass cells which results in a high particle/cell count, which is computationally expensive and thus not feasible to pursue. As an example, one of the largest and earliest cosmological scale simulations, the Millenium Simulation (Springel et al., 2005b), used $2\,160^3$ dark matter particles with an individual particle mass of $8.6\times10^8\,h^{-1}\,M_\odot$ – roughly half the dark matter mass of the Small Magellanic Cloud (Di Teodoro et al., 2019). However, progress being made with high resolution boxes such as TNG50 (Pillepich et al., 2019; Nelson et al., 2019) and zoom-in simulations of large regions such as NewHorizon (Dubois et al., 2021).

Evidently, the low mass regime of galaxies is a relatively under explored field, and even more so when considering dwarf galaxies with AGN characteristics. However, with advances in both observations and simulations, this undertaking has become more feasable. Observationally, Greene & Ho (2004, 2007); Reines et al. (2013); Sartori et al. (2015); Mezcua

et al. (2016) were some of the first ones to look at dwarf AGNs and central black holes on a large scale. Baldassare et al. (2017) examined the X-ray and UV properties of AGNs in nearby dwarf galaxies, again expanding dwarf AGNs into a new realm. Detailed studies on the impact of AGN in dwarf galaxies is also feasible now – from outflows (Manzano-King et al., 2019; Liu et al., 2020) to feedback and gas kinematics (Dashyan et al., 2018; Penny et al., 2018; Kaviraj et al., 2019; Reines et al., 2020)

Cosmological simulations, too, now reach somewhat resolved dwarf galaxies ($M_* \leq 5 \times 10^9$ $M_\odot$) such as IllustrisTNG with baryonic particle masses between $5.7{\times}10^4{-}7.6{\times}10^6$ $M_\odot/h$ (in TNG50-1 and TNG300-1, respectively). Although smaller scale simulations (local group size) with dwarf galaxies have been around for longer (e.g Wadepuhl & Springel, 2011) and are still being refined today (Trebitsch et al., 2018; Barai & de Gouveia Dal Pino, 2019; Koudmani et al., 2019; Bellovary et al., 2019; Sharma et al., 2020), only now is the emphasis on the effect of AGNs and black hole growth.

One of the keystone subjects of AGNs is under what circumstances they are found and what triggers their activity. Examples of such questions are whether field galaxies more frequently host AGNs, whether mergers and tidal interactions are the main culprit of triggering AGN activity, or what the effect of a dense environment is. While a connection between density of a galaxy's environment and its star formation rate has been established (Baldry et al., 2006; Peng et al., 2010; Davidzon, I. et al., 2016; Penny et al., 2016), the environment-AGN connection is disputed (Yang et al., 2018; Smethurst et al., 2019; Kristensen et al., 2020).

However, a connection between strong AGN and merger activity has been found (e.g Steinborn et al., 2018; Ellison et al., 2019; Kaviraj et al., 2019; Marian et al., 2020) suggesting that external factors are not without a say. The lack of an apparent connection to environment may be due to a time delay between the conditions that triggered AGN activity and when the AGN activity turned on (e.g Hopkins, 2012; Pimbblet et al., 2013; Kristensen et al., 2020)

Observationally, past environments and events are hard – if not impossible – to find unless morphological disturbances are still present. Some promise has been found using integrated field unit (IFU) spectroscopy where Penny et al. (2018) found kinematically offset cores in a sample dwarf AGN galaxies. However, simulations retain the complete environmental history

and past mergers – tracers of which are erased over time in real galaxies.

This study aims to test whether or not the current and past environments of a sample low-z dwarf AGN galaxies are different from those of a matched control sample with no AGN activity. The environment is examined in the IllustrisTNG simulation (more specifically, the TNG100-1 run), and observational data from the NASA-Sloan Atlas (NSA) is also included for comparative purposes. The AGN samples and the control samples are compared against each other using a Monte Carlo Kolmogorov-Smirnov (KS) test suite following a similar procedure as Kristensen et al. (2020).

This paper is organised as follow: Section 3.2 describes the data used, the sample selection criteria, and the environmental measures used. Section 3.3 contains the different distributions of the parameters of the the different samples along with the results from the KS-testing. Caveats and discussion of the results follow in Section 3.4 and the findings are summarised in Section 3.5. This study assumes the same cosmology as IllustrisTNG, namely a Λ-CDM Universe with $\Omega_{\Lambda,0} = 0.6911$, $\Omega_{m,0} = 0.3089$, $\Omega_{b,0} = 0.0486$, and $h = 0.6774$

3.2 Data and Methods

This section will describe the data used and details about the analysis carried out on said data. The data used is mostly simulation data from the IllustrisTNG project using their 75 Mpc/$h \sim 106.5$ Mpc simulation (TNG100-1) with some observational data from the NASA-Sloan Atlas (NSA, details of this data set can be found in Kristensen et al. (2020), but can be summarised as SDSS dwarf galaxies with $M_* \leq 3 \times 10^9$ $M_\odot$, $z \leq 0.055$) included for comparison purposes. The samples are first found using dwarf galaxy selection criteria similar to Kristensen et al. (2020) combined with simulation specific requirements, and that sample is then subdivided into subsamples according to AGN selection criteria based on Eddington ratios. Finally, a number of environmental measures are found for all dwarf galaxies, and those properties are then compared between the different subsamples using a Monte Carlo Kolmogorov-Smirnov testing suite.

3.2.1 IllustrisTNG and Illustris

The Illustris 'The Next Generation' (IllustrisTNG) simulation is the successor to the original Illustris simulation (Vogelsberger et al., 2014; Genel et al., 2014; Sijacki et al., 2015) with updated and new physics and refinements over the original. The simulations are evolved with the AREPO code (Springel, 2010), and consists of three different runs (TNG50, TNG100, and TNG300), though only TNG100 is used for this analysis. The number indicates the physical box size, and for TNG100, side lengths of the box are 75 Mpc/h $\sim$ 106.5 Mpc, with $h = 0.6774$. More specifically, the TNG100-1 run is used which has $1\,820^3$ dark matter particles with a mass of $7.5 \times 10^6\,\mathrm{M}_\odot$ and a baryonic mass of $1.4 \times 10^6\,\mathrm{M}_\odot$.

Of particular interest is the evolution and modelling of supermassive black holes. As described in Weinberger et al. (2018), friends of friends (FoF) groups are identified on the fly on dark matter particles and a SMBH of mass $1.18 \times 10^6\,\mathrm{M}_\odot$ is seeded whenever a FoF halo exceeds a total mass threshold of $7.38 \times 10^{10}\,\mathrm{M}_\odot$ and does not yet contain a SMBH. This does mean that some low mass subhalos in very dense environments are not seeded a black hole. The subhalos without black holes are not included in the analysis. A thorough discussion of this bias can be found in Section 3.4.3.

The mass accretion of the black holes are reliant on the local environment – more specifically, it is a Bondi-based accretion prescription, which relies on the ambient sound speed and ambient density. It is not boosted as in other simulations (e.g Springel et al., 2005a), which gives more validity to the environmental analysis since the accretion rate is based only on the physical space and processes surrounding the black hole. There are, however, caveats with the BH modelling. Therefore, to check the validity of the results obtained from TNG100-1, the TNG50-1 and Illustris-1 runs are used for comparison purposes.

The main differences in resolution and particle masses between the three simulations can be found in Table 3.1, but to summarise: TNG100-1 and Illustris-1 have roughly the same box size and particle masses, though BH seeds are lighter in Illustris but BH accretion is boosted. BH feedback mechanisms are also slightly different and a more detailed discussion can be found in Pillepich et al. (2018) and references therein. TNG100-1 and TNG50-1 differ only in box size and particle masses – their BH prescriptions are the same.

Table 3.1: Overview of relevant simulation parameters

Simulation	L_{box} [ckpc/h]	m_{DM}	m_{gas}	$m_{BH\ seed}$
TNG100-1	75000	750	140	80/h
TNG50-1	35000	45	8.5	80/h
Illustris-1	75000	630	130	10/h

Table 3.2: First column is the simulation name, second one is the corresponding side length given in units of comoving kpc/h. Third column is the mass of a dark matter particle followed by gas cell/particle mass and lastly is the mass of the seeded black hole particle. Masses are in $10^4 M_\odot$ for easy comparison.

Both TNG and Illustris have caveats and limitations in the BH modelling (Weinberger et al., 2018; Li et al., 2020; Terrazas et al., 2020; Habouzit et al., 2021). Habouzit et al. (2021) remarks that current cosmological simulations – Illustris(TNG) and Horizon-AGN, EAGLE, and SIMBA – struggle to produce the diversity of BHs observed in the local Universe. A strong caveat of TNG BH modelling is the seeding mass of $\sim 10^6\ M_\odot$, which is ~ 0.5 dex more massive than currently observed BH in dwarf galaxies (e.g Xiao et al., 2011; Kormendy & Ho, 2013; Reines et al., 2013; Moran et al., 2014; Reines & Volonteri, 2015; Manzano-King & Canalizo, 2020; Baldassare et al., 2020). This effect can be seen in Figure 3.1, which shows the black hole-velocity dispersion relationship of both TNG100-1 subhalos and observed black holes in dwarf galaxies in Xiao et al. (2011) and Baldassare et al. (2020). However, our AGN selection relies on the Eddington ratios (see Section 3.2.3, which scales with M_{BH} and not accretion rates, which scales with M_{BH}^2, lessening the importance of the bias of overmassive SMBHs. Therefore, the results should also be seen as a comment on the simulation physics. Nevertheless, the impact of seed mass is checked by including Illustris-1, which has a lower seed mass than the TNG runs.

Merger trees and thus merger data are from the SubLink algorithm (Rodriguez-Gomez et al., 2015). Colours are from Nelson et al. (2018) which are calculated from the stellar particles in a subhalo by summing the luminosities of the particles and applying a dust attenuation model. The response function is modelled to SDSS photometry. For Illustris-1 and TNG50-1, dust corrected measurements are not available and thus just the sum of

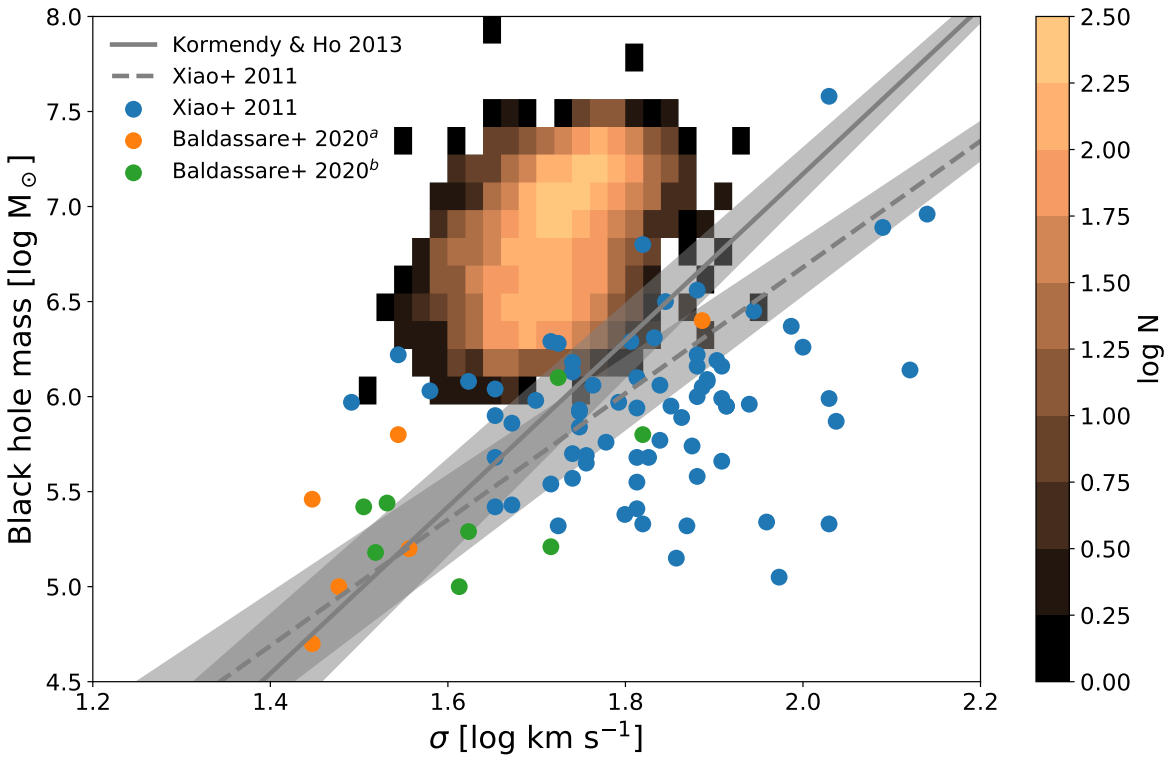

Figure 3.1: Black hole mass versus stellar velocity dispersion, σ. Two relations are plotted (Xiao et al. (2011) dashed line, Kormendy & Ho (2013) solid line) with their intrinsic scatter. Observations of black holes in dwarf galaxies (M_* between 8.5-9.5 log $M_\odot$) from Xiao et al. (2011) are in blue and Baldassare et al. (2020) are in orange and green – *a*: from previous work, *b*: from Baldassare et al. (2020) study. The copper 2D histogram shows TNG100-1 galaxies

luminosities of the stellar particles are used.

3.2.2 Dwarf galaxy selection

During the simulation run, group catalogues are computed using friends-of-friends (FoF) and Subfind algorithms at each snapshot on dark matter particles. For this study, dwarf galaxies are selected from the $z = 0$ (snapshot 99) group catalogue, and a number of requirements are imposed to ensure the selected dwarf galaxies are of acceptable quality. Working with dwarf galaxies means working near the resolution limit of the simulation, and extra care has to be taken.

Simulation specific requirements are that subhalos are required to have a dark matter component and a black hole. An upper stellar mass limit $M_* = 3 \times 10^9\,M_\odot$ is also imposed as this follows observational definitions. To resemble observations further, the stellar mass used is the mass enclosed in 2 times the stellar half mass radius, `PartType4` of `SubhaloMassInRadType`. Similarly, a lower stellar mass limit of $10^9\,M_\odot$ is also used to resemble large scale observational surveys where this the effective mass limit (see e.g Kristensen et al., 2020, Figure 11 for a mass distribution of local SDSS dwarf galaxies) and to ensure well resolved galaxies. This is in line with other studies on dwarf galaxies in simulations (Sharma et al., 2020; Fattahi et al., 2020; Dickey et al., 2021; Koudmani et al., 2021; Reddish et al., 2022; Jahn et al., 2022).

These requirements can be summarised as follows:

1. $1 \times 10^9\,M_\odot \leq M_{*,2HM} \leq 3 \times 10^9\,M_\odot$
2. $M_{DM} > 0\,M_\odot$
3. $M_{BH} > 0\,M_\odot$

A projected number density of dwarf galaxies on three planes can be seen in Figure 3.2. These requirements yield 6 771 dwarf galaxies.

The reasoning behind requiring a black hole is due to large number of dwarf galaxies with no black holes ($N = 1\,001$, assuming same requirements for regular dwarf galaxies except $M_{BH} = 0\,M_\odot$). This requirement is also used in other studies (e.g Koudmani et al., 2021).

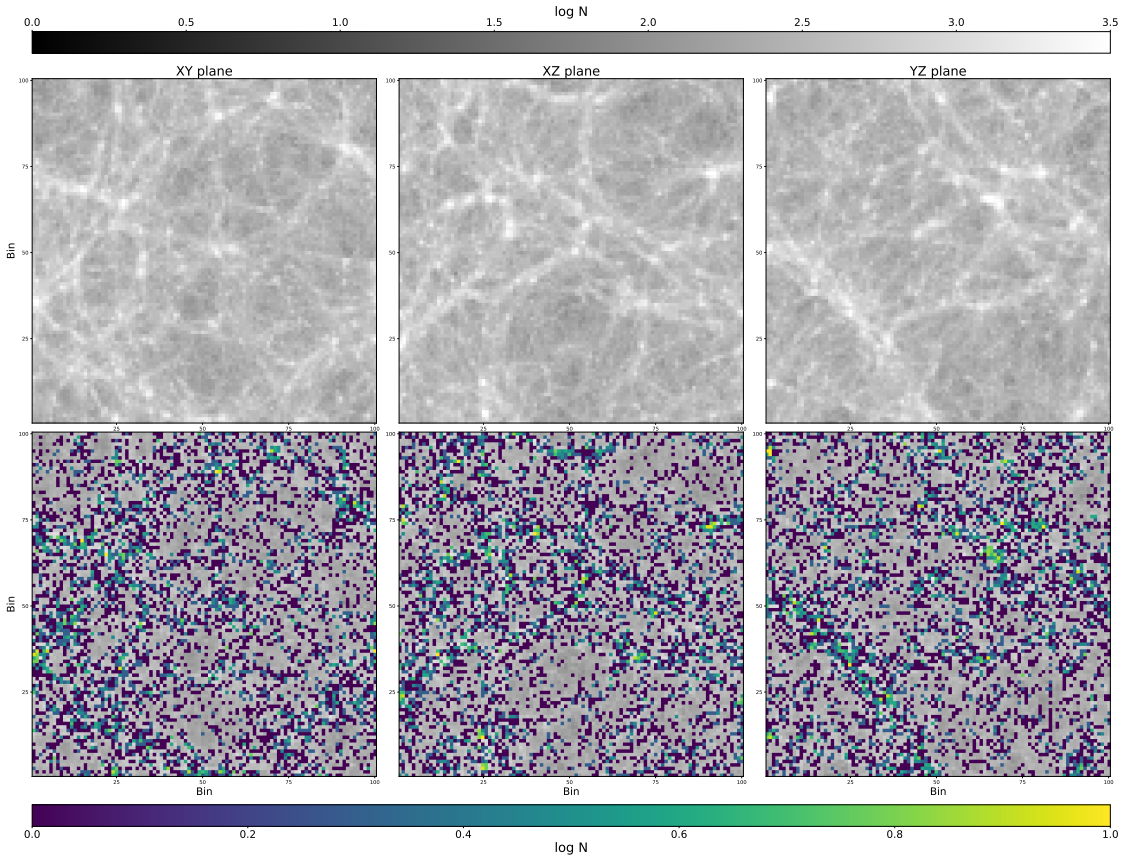

Figure 3.2: Spatial distribution of all subhalos only in top row and with selected dwarf galaxies bottom row projected onto three different planes (XY, XZ, and YZ plane). The gray scale background number density plot includes all subhalos while the coloured distribution is for dwarf galaxies. The data is split into 100 bins on each axis resulting in a bin size of 0.75×0.75 Mpc/h

Table 3.3: Number of subhalos for each AGN selection criteria.

	Dwarfs	NOT	Weak	Intermediate	Strong
Simulation	Eddington ratio	$\lambda < 0.005$	$0.005 \leq \lambda < 0.01$	$0.01 \leq \lambda < 0.1$	$\lambda \geq 0.1$
TNG100-1	6 771 (100 %)	2 908 (42.95 %)	988 (14.59 %)	2 821 (41.66 %)	54 (0.80 %)
TNG50-1	1 003 (100 %)	417 (41.58 %)	247 (24.63 %)	337 (33.60 %)	2 (0.20 %)
Illustris-1	10 914 (100 %)	10 402 (95.31 %)	226 (2.07 %)	254 (2.33 %)	32 (0.30 %)

Table 3.4: The total number is N and how large a percentage of the total dwarf galaxy population is given in percentage in parenthesis. λ is the Eddington ratio.

For these galaxies' real life counterparts, there is no reason to assume they should not have a black hole as seeding mechanisms are still very unconstrained. It comes down to the BH seeding mechanism in IllustrisTNG, which leaves the simuated galaxies without a BH. Thus, they are unable to host AGN activity and will skew the AGN to non-AGN distributions that will be described in Section 3.2.3. A discussion of this effect can be found in Section 3.4.3.

Divergence from observational dwarf galaxies

A concern when using observational data is bias towards low surface brightness galaxies not picked up due to observational constraints. This concern is not relevant when selecting subhalos from simulations since even the lowest surface brightness galaxies will be selected. Figure 3.3 shows a colour-magnitude diagram of TNG100-1 and NSA dwarfs. While the majority of dwarf galaxies and subhalos are centered around roughly the same values ($u - r = 1.3$, $r = -18.5$), NSA sources are more spread out – especially towards bluer galaxies. Consequently, some bias exists in observational data but we do not attempt to correct for it since there are also uncertainties of the colour modelling of TNG100. The details of the colour model and subsequent bias of TNG are described in Nelson et al. (2018).

3.2.3 AGN selection

AGNs are selected from their Eddington ratios. The Eddington ratio is given as:

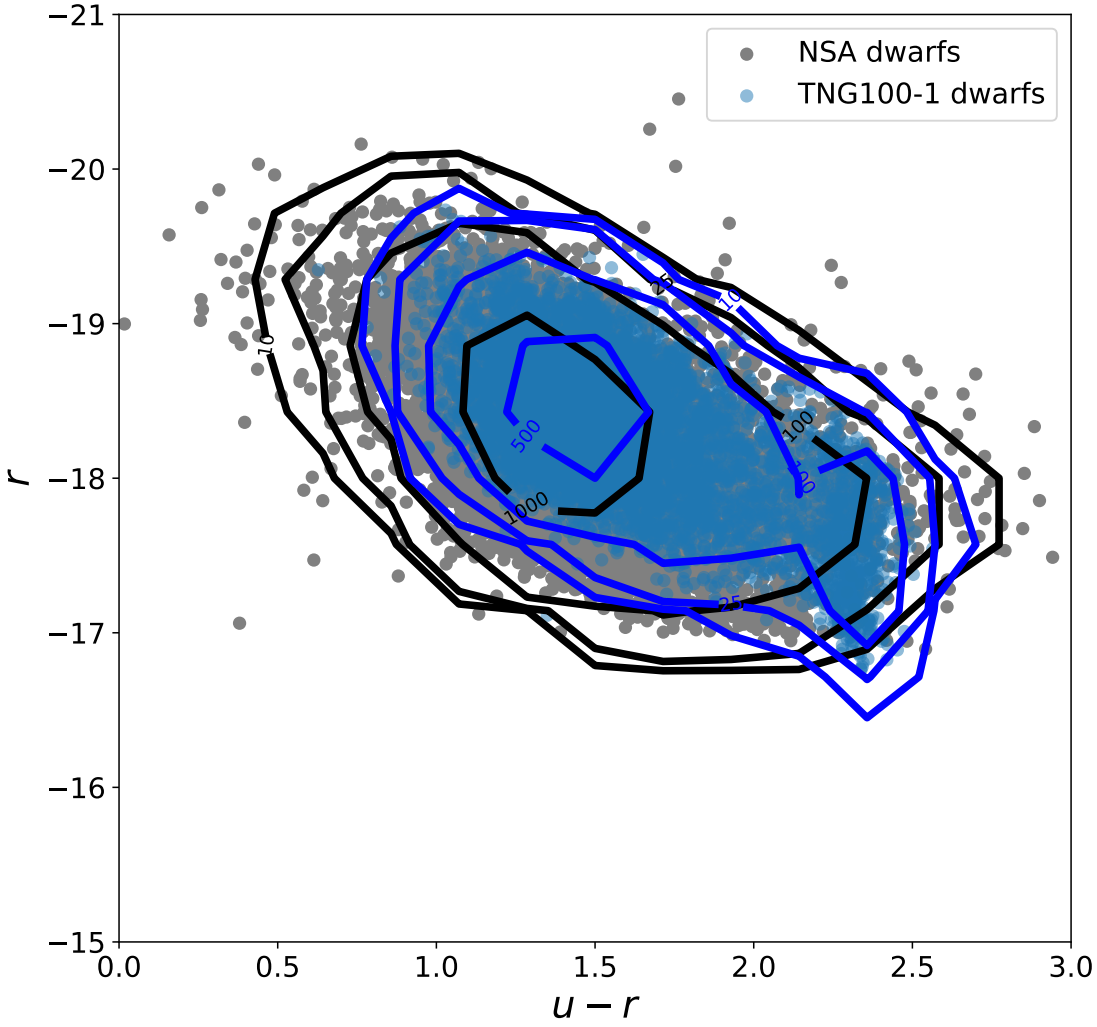

Figure 3.3: Colour-magnitude ($u - r$ colour vs r magnitude) diagram showing SDSS dwarfs compared to TNG100 dwarfs with same mass selection criteria. Grey dots and black contour lines are NSA data while blue dots and contour lines are on dwarf subhalos from TNG100. The contour levels are at different levels between the samples since the sample sizes are different.

$$\begin{aligned} \lambda &= \dot{M}/\dot{M}_{\rm Edd}, \\ \dot{M}_{\rm Edd} &= \frac{4\pi G M_{\rm BH} m_p}{\varepsilon_r \sigma_T c}, \\ \dot{M} &= \alpha 4\pi G^2 M_{\rm BH}^2 \rho / c_s^3 \end{aligned} \tag{3.1}$$

where $\dot{M}$ is the black hole mass accretion (Eddington limited Bondi accretion), G is the gravitational constant, M_{BH} is the mass of the black hole, m_p is the proton mass, $\varepsilon_r = 0.2$ is the black hole radiative effeciency, σ_T is the Thompson cross section, c is the speed of light, $\alpha = 1$, ρ is the local comoving gas density, and c_s is the speed of sound in the local gas cells.The black hole mass and its accretion rate are based on the prescription in Springel et al. (2005a) and are both available from the group catalogues and described in detail in Section 2.3 of Weinberger et al. (2018).

Three Eddington ratios are used for AGN selection splits: $\lambda = 0.005, 0.01, 0.1$. Table 3.3 contains the size of the samples along with their classification names; weak, intermediate, and strong. This follows similar selection as e.g Bhowmick et al. (2020); McAlpine et al. (2020). Since the $\lambda \geq 0.1$ selection yields so few objects, it is difficult to draw convincing statistical conclusions for this sample. However, they are still kept for some analysis parts but are excluded in some plots and conclusions. Non-AGN (also referred to as the 'NOT' sample) are defined as dwarf galaxies with $\lambda < 0.005$ and consists of 2 908 (42.95 per cent) sources. An overview can be found in Table 3.3 which also contains an overview for TNG50-1 and Illustris-1. The discrepency in AGN fraction between TNG and Illustris is ascribed to the difference in BH-modelling and overmassive SMBH in TNG, which yields overly efficient black hole accretion (see Section 3.2.1).

The nature of Eddington ratio selected AGN are expected to be different than observationally chosen ones such as those in Kristensen et al. (2020), which relies on optical emission line ratios. In fact, even AGN selected by different observational methods are expected to be different in nature from each other (e.g Ji et al., 2022). AGNs selected by optical emission lines tend to not pick up obscured or low luminosity AGN – attributes that do not matter when using an Eddington ratio selection. Following the classification scheme above on an

observational study by Baldassare et al. (2017), out of 12 dwarf AGN galaxies, 2 would be low intensity, 4 would be intermediate intensity, and 4 would be considered high intensity AGN. As such, observations appear to be biased towards high intensity/luminosity AGN.

Nevertheless, while simulations may include more low intensity AGN compared to observations, this bias is not important for the most part since comparisons between subsamples are within the same simulation runs, which means that compared galaxies have the overmassive black holes and boost factors. However, it does mean that direct comparisons to observations and between simulations need to be done cautiously.

BH comparisons between simulations

Since TNG100 and TNG50 have the same BH modelling while Illustris runs follow slightly different prescriptions (see Pillepich et al., 2018, Table 1 for a complete overview of the differences), some differences are expected between the AGN populations of the different simulations. Table 3.3 shows a much lower AGN fraction for Illustris simulations ($\sim$ 4.7 per cent) compared to TNG runs ($\sim$ 58 per cent).

There are also differences in BH properties such as BH mass, accretion rate, and density of local comoving gas. Figure 3.4 shows that, as expected, the Illustris-1 BH mass is lower (by 1.5-2.0 dex). The BH accretion rate is also lower – in fact by even more orders of magnitude than the black hole mass despite being boosted by a factor of $\alpha = 100$. This is not too surprising since the Bondi accretion rate is proportional to M_{BH}^2. Nevertheless, this is indicative of that AGN populations between simulations are different; TNG has more dwarf subhalos appearing as AGN because of a more efficient accretion, while subhalos sharing similar characteristics in Illustris may not have reached the $\lambda \geq 0.005$ threshold to be considered AGN. Further differentiating the accretion rates is the BH density – the gas density from which accretion is calculated. In Illustris, it is only the parent gas cell while TNG calculates it from nearby gas cells (evaluated over a sphere enclosing certain number of neighbours (where the neighbout number is scaled with the mass resolution of the simulation, see Weinberger et al., 2018; Pillepich et al., 2018, for details)). Furthermore, TNG simulations have magnetohydrodynamic modelling unlike Illustris (Pillepich et al., 2018), which further

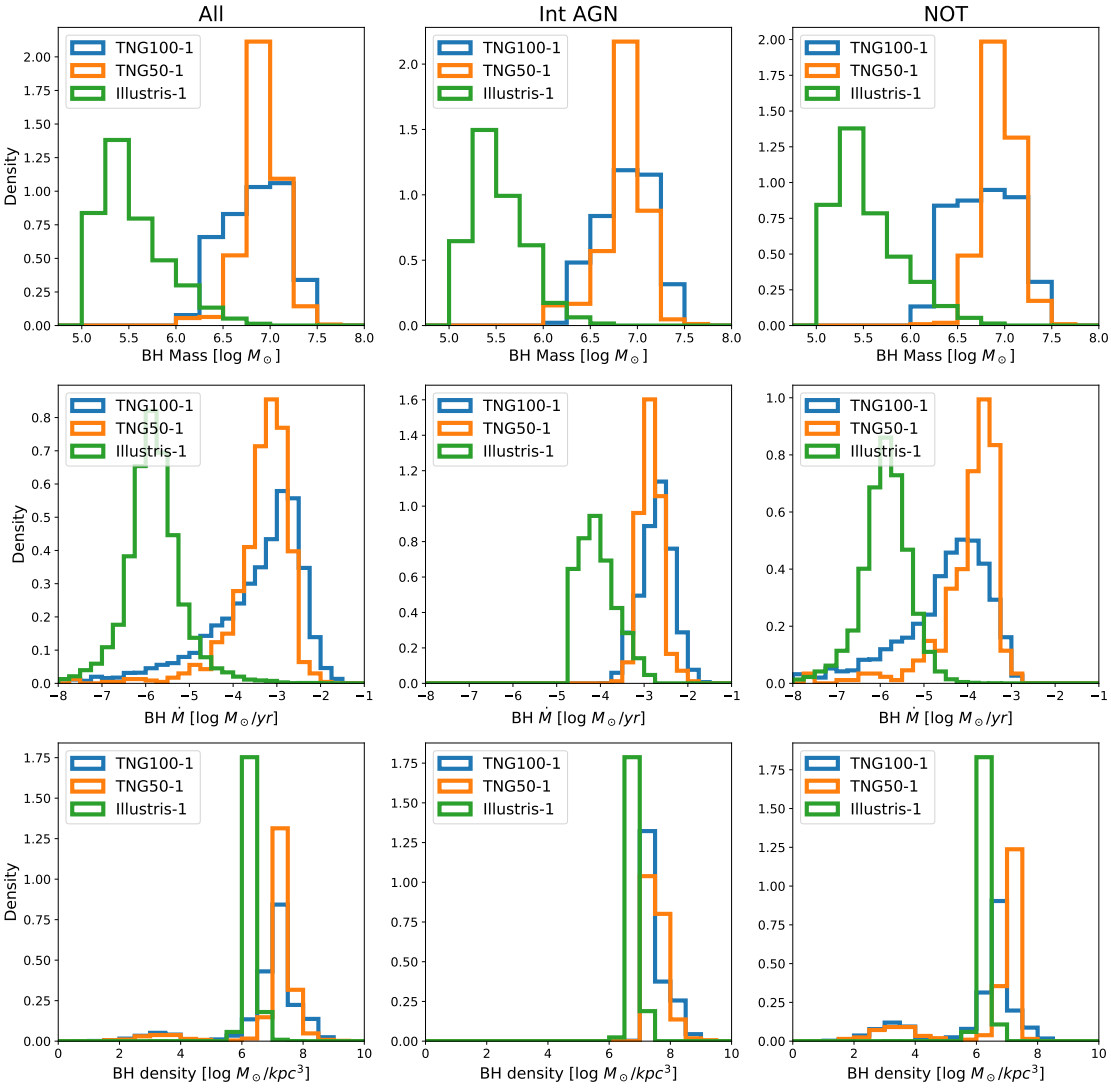

Figure 3.4: BH comparison histograms between simulations. Top row shows BH mass, middle row displays BH accretion rate, and the bottom row shows the density of local comoving gas of the BH. The columns display the full sample in the first row, Int AGN in the second column, and lastly non-AGN in the third column. TNG100-1 BH are in blue, TNG50-1 are orange, and Illustris-1 is green.

changes the properties of the accretion gas cells.

Between TNG100 and TNG50, the differences are subtle. They roughly have the same average BH mass but with TNG50 being more concentrated (TNG100: $8.70 \pm 6.10 \times 10^6$ $M_\odot$, TNG50: $8.94 \pm 4.09 \times 10^6$ $M_\odot$). Interestingly, BHs of AGN galaxies are more massive than non-AGN in TNG100 by roughly one dex, but the opposite is true for TNG50. This is also reflected in the accretion rates where BH in TNG100 on average have $\sim$ 1.6 times that of TNG50, but non-AGN in TNG50 have higher accretion rates than non-AGN in TNG100.

Regarding gas densities, there are few things to note. TNG100 and TNG50 follow similar trends with AGN having higher density gas reservoirs than non-AGNs. The similar distributions of BH densities between the two simulations suggest that resolution is unlikely to strongly impact the results. Noteworthy is a small non-AGN population of both TNG100 and TNG50 galaxies with densities around $\log \rho = 3$, i.e a population with little-to-no gas. In fact, out of 548 ($\sim$ 8.1 per cent of all dwarfs, $\sim$ 18.8 of NOT dwarfs) subhalos with $\log \rho \leq 5$, 487 of them do not have any gas cells associated with them in the group catalogue and 546 (i.e all but two) are considered red ($u - r \geq 2$, red galaxies are discussed further in Section 3.4.4). This population does not exist in Illustris-1 suggesting that differences in subhalo identification parameters and/or gas physics result in this population.

3.2.4 Time since last merger

The time since last merger (TSLM) is a measure to see if there is a time lag between current AGN activity and a past merger event. While this can already be inferred from morphological disturbances (e.g Ellison et al., 2019) or post-star burst spectra (Pawlik et al., 2019), the TSLM method provides a channel in which simulations can explore this connection, too.

This is done by following the main progenitor branch (MPB) to find the snapshots at which mergers that exceed a merger mass threshold using stellar masses. Three different minimum mass ratios are selected for analysis: 1:10, 1:4, and 1:2. The highest snapshot number is the one chosen as the TSLM.

The TSLM distributions for different samples can be seen in Figure 3.5 and are then compared against each other using the KS-testing described in Section 3.2.6.

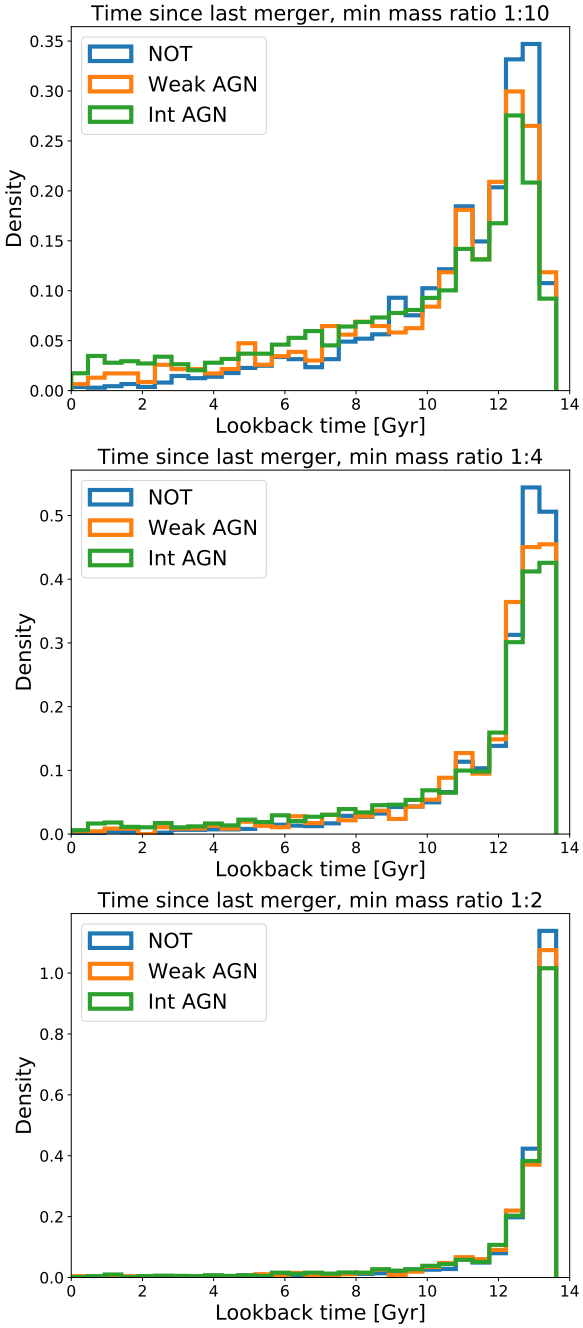

Figure 3.5: Time since last merger histogram for three selected subsamples: NOT – low mass galaxies with a black hole but no AGN activity in blue (Section 3.2.2), weak AGN galaxies in orange (Section 3.2.3), and intermediate AGN galaxies in green (also Section 3.2.3). Galaxies with no mergers have a TSLM equal to the age of the universe

Since this study is working near the resolution limit, it is vital to ensure that the chosen galaxies are not only of good quality at $z = 0$ but also at earlier times. Most (89.5 per cent) have a MPB down to snapshot 5 ($z = 9.39$, lookback time 13.286 Gyr) with a tail that includes the rest extending to snapshot 15 ($z = 5.53$, lookback time 12.767 Gyr).

This means that there is significant incompleteness at the very early times of the universe. This is further complicated by the fact that the galaxies were less massive earlier which means that e.g a 1:10 merger mass ratio would require a similarly smaller merger mass in order to be counted. This can be seen from Figure 3.6 where the majority of the merging galaxies are happening 10 Gyrs ago mostly have a stellar mass of $10^{6.5} - 10^{7.0} M_{\odot}$. This means that the merging galaxies consisted of only 10s of particles. Merging galaxies only consistantly reach reasonable resolution ($\sim$ 100 star particles, around $3 \times 10^8 M_{\odot}$) at a lookback time of 6 Gyr. As such, any comments on TSLM with lookback times larger than 6 Gyr have uncertainties due to resolution.

3.2.5 Distance to 10th nearest neighbour

Another way the environment is quantified is by the 3D distance to a galaxy's 10th nearest neighbour, D_{10}. This method is used to describe the density of the local environment and was also used in Kristensen et al. (2020). However, the line-of-sight distance in observations is calculated from redshift and is therefore not the 'true' distance like in simulations. While other types of environmental measures also exist (for a review, see Muldrew et al., 2012), this approach is used because it is a better measure of the local density rather than cluster density and it makes comparisons to previous observational work easier. While Haas et al. (2012) mention that $N^{th}-$neighbour measurements anticorrelates with halo mass, this anticorrelation is weak for $N = 10$ seeing that typical halo masses for low mass galaxies in this sample is 10^{11} $M_{\odot}$. Therefore, this environment measure probes something else than just halo mass.

For each galaxy in the sample, the distances to all other 'real' galaxies (defined as non-zero DM mass and above the minimum stellar mass of $10^9 M_{\odot}$) are calculated and then placed in ascending order. Observationally, no restrictions are placed on neighbour galaxies, which may result in more valid neighbours and thus lower $D10$. The 10th element is then chosen

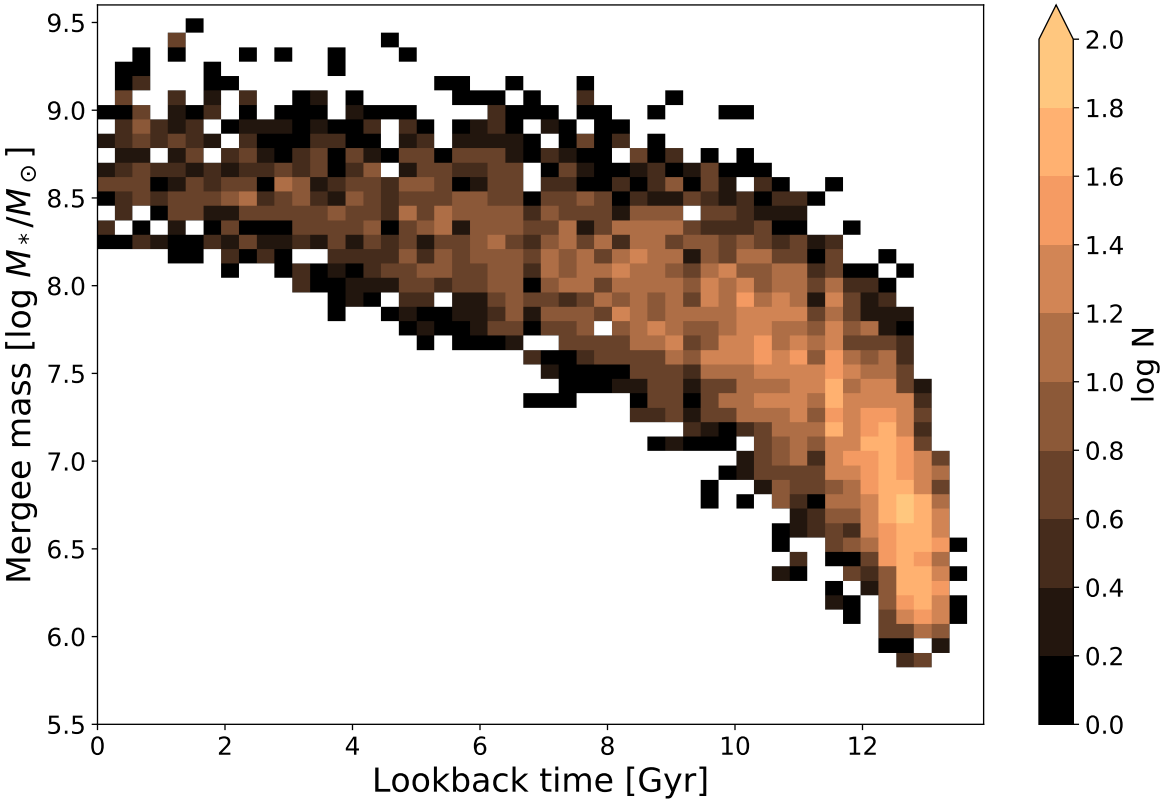

Figure 3.6: Time since last merger (1:10 mass ratio merger) versus merger stellar mass 2D histogram.

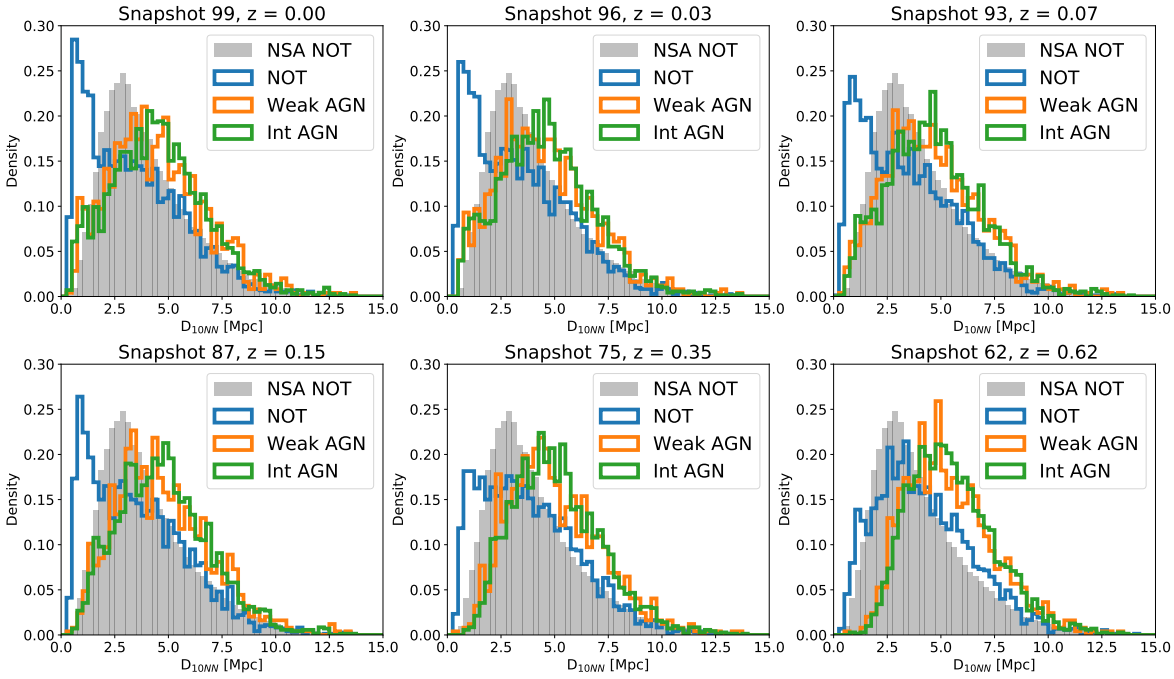

Figure 3.7: Distance to 10th nearest neighbour histogram for selected snapshots. Blue is non-AGN galaxies, orange is weak AGN while green is intermediate. Strong AGNs are not included as this sample size is small. The grey background histogram is observational data from the NASA-Sloan Atlas ($M_* \leq 3 \times 10^9$ $M_\odot$, $z \leq 0.055$). The snapshots (from high to low) roughly corresponds to lookback times of 0.00, 0.48, 1.00, 1.98, 3.97, and 6.01 Gyr.

and this distance is then chosen as the subject galaxy's distance to its 10th nearest neighbour.

This measurement is also performed for current AGN galaxies' past environments ~ 0.5, 1.0, 2.0, 4.0, and 6.0 Gyr ago (i.e at snapshot 96, 93, 87, 75, and 62 (Illustris: 132, 129, 123, 110, 97), respectively). This is done by finding their past progenitor following the main progenitor branch and then repeating the steps described in the previous paragraph.

As with TSLM, the distributions of different subsamples are then compared against each other using the KS-testing method in Section 3.2.6 and the distributions can be seen in Figure 3.7.

3.2.6 Kolmogorov-Smirnov testing

The different subsamples are compared against each other using a Monte Carlo approach to 2-sample Kolmogorov-Smirnov (KS) testing. KS testing itself examines the null hypothesis that a sample is drawn from a reference distribution. In the case of a 2-sample KS test, it tests whether the two samples differ (i.e the null hypothesis is whether the two distributions are drawn from the same reference sample).

The implementation of the test in this study follows that of Penny et al. (2018) and Kristensen et al. (2020) but slightly tweaked. First, a test measure is selected (e.g 10th nearest neighbour at snapshot 99). Then, a subject sample is selected (e.g weak AGN). A random sampling with replacement of N elements is then performed, and this new sample is then the final subject sample that is to be compared. Next, for each element in the final subject sample, a match galaxy is found in the reference sample (e.g strong AGN). The elements are matched in mass (±20%) and $u - r$ colour (±0.4). In the case of multiple matches, a random galaxy is selected and added to the final reference sample while no matches removes the subject galaxy from the final subject sample. The final subject sample and final reference sample are then compared to each other using a 2-sample KS-test.

This process is repeated 100 times and an average is calculated. This is then repeated 10 times, and the average these 10 tests are then found and subsequently plotted. The error is calculated by finding the average of the standard deviations from the 10 tests and divided by $\sqrt{10}$.

A sampling size of 500 is chosen as the primary sampling size because of its resemblance to observational data as well as to avoid under- and oversampling. However, since there only are 54 strong AGN, any comparisons involving this sample, the sampling size is set to 54. Observational data do permit a larger sample size (see a discussion of this in Section 3.3.3 and Section 3.4.5), and testing with sample sizes of N=1 000, 1 500, and 2 000 is also performed, although the size of the weak AGN is only 988 resulting in an upper sampling limit of 988 when comparisons involving weak AGN are done. However, any difference between distributions found only with a higher sampling size means that the difference is less pronounced than with a smaller sample. Therefore, tests with $p \leq 0.05$ and large sampling sizes are only used to infer trends rather than reject the null hypothesis of the two distributions belonging to the same parent distribution.

4 sampling sizes, 5 different samples, and 9 different tests (3 TSLM + 6 D_{10}) yield 900 p-values. These are fully plotted in Appendix 3.5 in Figures 3.14, 3.15, and 3.16). However, significant results are summarised in Table 3.5 which shows which sampling size the different tests reach significant levels.

3.3 Results

This section contains the results of the KS-testing for the different measures. One thing to keep in mind is that when comparing two subsamples, it is important which subsample is used as subject sample and which as a reference sample. For example for D_{10} distributions, having the NOT subsample as a subject sample and comparing it to the weak AGN subsample, the null hypothesis (i.e that the distributions of the two subsamples belong to the same parent distribution) is rejected. However, the p-value does not reach the threshold when keeping the AGN subsample as subject sample and using the NOT as the reference sample. The reason behind this can be inferred from Figure 3.8 which shows how the non-AGN and weak AGN D_{10} distributions change whether they are subject or reference samples.

While this may not be intuitive at first, it is not surprising. The NOT galaxies have a quite diverse range of masses and colours so that AGN galaxies have a large catalogue to find partners from. A larger match catalogue will smooth out the cumulative distribution of the

reference sample and thus decrease the average distance between the two distributions.

Comparisons that reach $p \leq 0.05$ with a sampling size of 500 will be called significant while comparisons that reach the threshold at larger sampling sizes are called trends. While these sampling sizes are smaller than fully allowable (i.e not oversampling neither subject nor reference sample), such large sampling sizes are not achievable with current observational data. Having several different sampling sizes allows for different observational surveys to find the closest matching sampling size. A more detailed discussion of this will be given in Section 3.3.3 and Section 3.4.5. A table overview of simulation comparisons can be found in Table 3.7

3.3.1 On time since last merger

The p-value for 1:10 mass ratio merger in tests with NOT and Int both reach the threshold of 0.05 regardless of which is subject and reference sample. Furthermore, this also holds for 1:4 mass merger ratio of NOT and Int, although it is only within error. Lastly, there is a significant difference within error between Int and weak AGN in 1:10 and a trend between them in the 1:4 case. These results also appear in TNG50-1 and Illustris-1 and will be examined further in Section 3.3.4.

Interpreting these results show that a merger of at least a 1:10 ratio has happened more recently, on average, in an intermediately active dwarf galaxy than in a non-AGN dwarf galaxy. While this does not establish a causal link (i.e that the merger events triggered the AGN activity), it is a statistically significant difference. Ellison et al. (2019) showed that nearly 60 per cent of mid-IR AGN hosts showed signs of visual disturbances and were either interacting with a close companion or in a post-merger phase with the latter contributing the most. The difference between AGN and non-AGN galaxies in this study is an excess of AGN galaxies with an at least 1:10 merger in the last 0-10 Gyr. 8.4 per cent of Int AGN has had a 1:10 mass ratio merger within the last 3 Gyr compared to only 1.6 of NOT galaxies. Roughly half (55.7 per cent) of Int AGN has not had a merger within the last 10 Gyr, while this number is 71.3 per cent for NOT.

Assuming that most tracers of past merger activity is gone after 1-2 Gyr (Eliche-Moral

et al., 2018), it is not unlikely that many of the AGN galaxies still retain some merger tracers – in line with the findings of Ellison et al. (2019). However, the majority of present day AGN have not had a recent merger within the last 6 Gyr, by which any tracers of a merger most likely are been long gone. This will be discussed further in Section 3.4.1.

3.3.2 Current and past environments

Similar across all times is that choosing non-AGN as subject sample and weak or intermediate intensity AGN as reference sample, the p-value from the results of the KS-tests dips below 0.05. The strong intensity AGN sample do not cross this threshold due to the small sample size. In the inverse situation with the weak and intermediate AGN as subject sample and non-AGN as reference, the p-values do not reach the threshold except at snapshot 62, but do show a trend (using a sampling size of 1000 and above). The similarity between a matched NOT to an Int AGN sample can be seen in Figure 3.8. These results are similar across both TNG simulations but not Illustris-1.

Worth noticing is the $u - r$ colour distribution of the two samples (see Figure 3.9) where both non-AGN and AGN galaxies have a peak around $u - r = 1.5$. However, the non-AGN sample has an additional peak near $u - r = 2.2$ suggesting that a significant sub-population of non-AGN galaxies are very red. Using an AGN sample as subject sample, very few of these red non-AGN galaxies are selected for the reference sample and the resultant D_{10} of the non-AGN reference sample does not have as large a peak of galaxies at 0.5 Mpc $\lesssim D_{10} \lesssim$ 2 Mpc. This suggests that this very red sub-population of non-AGN galaxies reside in dense environments. Conversely, very few AGN galaxies reside in these environments. This population and its effect on the results is discussed further in Section 3.4.4.

3.3.3 Sampling size

As mentioned in Section 3.3, sampling size has a direct influence on the p-value. This section will motivate that while some samples do not seem to be statistically different (i.e reach $p \leq 0.05$), a trend can still be inferred. It comes as no surprise that the p-value of a KS-test is dependant on the sampling size since the level at which the null hypothesis can be rejected

Table 3.5: Summary tables of smallest sampling sizes from KS results.

Subject	Reference	1:10	1:4	1:2
All	NOT	(△)		
	Weak AGN			
	Int AGN	(△)	(▲)	
NOT	All	(●)		
	Weak AGN	△	(△)	
	Int AGN	○	(○)	(△)
Weak AGN	All			
	NOT	(▲)		
	Int AGN	(○)	△	
Int AGN	All	(△)		
	NOT	○	(○)	(△)
	Weak AGN	(○)	△	

Subject	Reference	99	96	93	87	75	62
All	NOT						
	Weak	(○)	(○)	(○)	(○)	(○)	(○)
	Int	○	(○)	(○)	(○)	(○)	(○)
NOT	All	(●)	(●)	●	●	(△)	△
	Weak	○	○	○	○	○	○
	Int	○	○	○	○	○	○
Weak	All						
	NOT	(△)	(△)	(△)	△	△	(○)
	Int						
Int	All						
	NOT	△	△	△	△	△	(○)
	Weak						

Table 3.6: Summary tables of which smallest sampling sizes KS results are significant at for TSLM for different miniumum merger mass ratios (top table) and D_{10} at different snapshots (bottom table). Open dots, open triangles, filled dots, and filled triangles represent sampling sizes of 500, 1000, 1500, and 2000, respectively. Symbols in parenthesis indicates that the $p \leq 0.05$ is reached within error. For a full plots, see Figures 3.14, 3.15, and 3.16. Strong AGN are not included since no test reaches the threshold. Tests with weak AGN are limited to a maximum sampling size of 988.

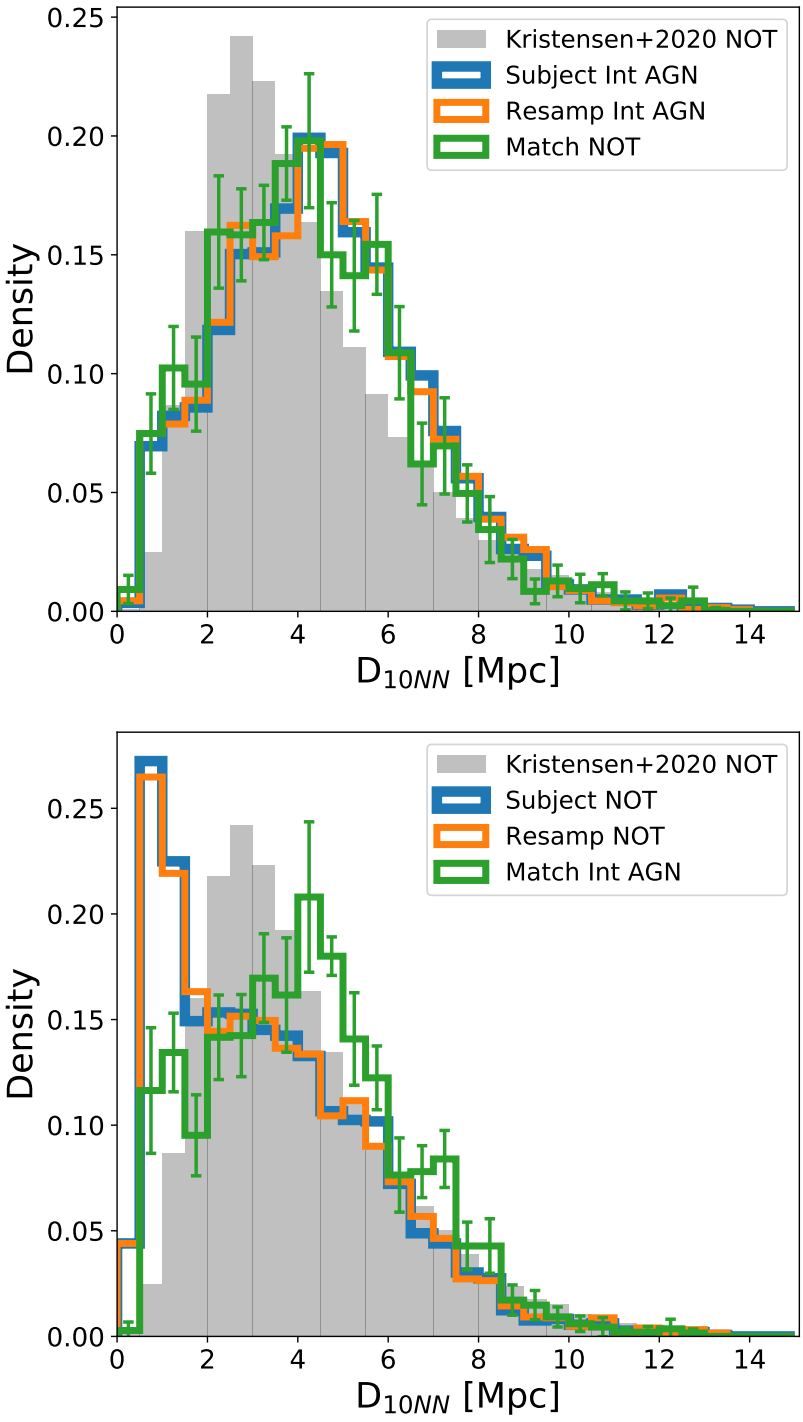

Figure 3.8: Distance to 10th nearest neighbour histogram with subject samples and their matched reference sample. Blue is the original subject sample, orange is the resampled subject sample while green is the matched reference sample. Top: Int AGN as subject sample and NOT as reference. Bottom: NOT as subject and Int AGN as reference. Errorbars are calculated as the spread of the averages in each bin of 100 resampling runs with a sampling size of 500. The peak at small distances (between 0.5 to 2 Mpc) for NOT disappears when it is used as a reference sample for AGN samples.

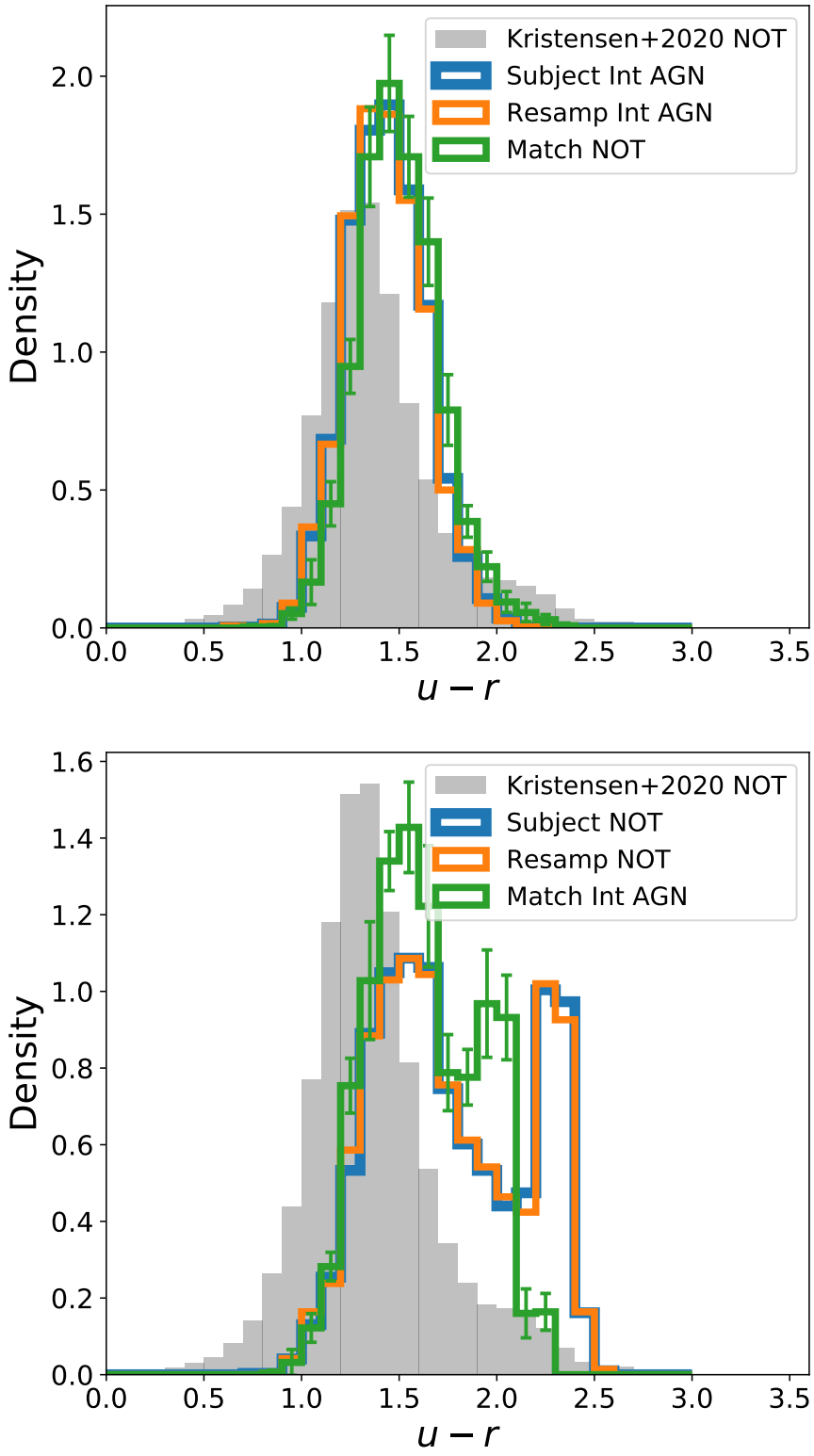

Figure 3.9: Colour histogram with subject samples and their matched reference sample - like Figure 3.8. Top plot is using Int AGN as subject sample versus non-AGN as reference sample. Bottom plot is in reverse. Errorbars are calculated as the spread of the averages in each bin of 100 resampling runs with a sampling size of 500. Few red ($u - r \geq 2.0$) NOT galaxies are selected when using AGN samples as subject samples.

scales with sample size. More specifically:

$$D_{n,m} > c(\alpha)\sqrt{(n+m)(n\cdot m)^{-1}}, \tag{3.2}$$

where $D_{n,m}$ is the maximum distance between the cumulative probability of the two distributions, α is the threshold at which to the null hypothesis is rejected, $c(\alpha) = 1.358$ for $\alpha = 0.05$, and n and m are the sample sizes. What this means for the interpretation of the above results is that some of the tests that did not yield $p < 0.05$ can reach that threshold given a larger sample size (e.g subject sample weak AGN vs NOT as reference sample with a sample size of 1000).

Table 3.5 shows a summary of the KS tests that reach $p \leq 0.05$ at what sampling size. For TSLM, few trends appear for 1:2 mass ratio mergers – only when using the largest sampling size. A similar pattern can be seen for the 1:4 mass ratio, although the trends are found for the same comparisons samples (NOT vs Int AGN, with both as both subject and reference sample) at a lower sampling size. These two comparisons ultimately reach a significant level at 1:10 mass ratio mergers within error.

For D_{10}, the only significant distribution is NOT as subject sample and AGN (both weak and intermediate) as reference samples. As subject sample, weak and intermediate AGN tend to differ at all snapshots from NOT as a reference sample at higher sampling sizes. Ultimately, the usefulness of inferring trends is to estimate the robustness of the tests and to indicate which comparisons are worth further looking into.

3.3.4 TNG50-1 and Illustris-1

The KS-testing suite with a sampling size of 500 has also been performed on TNG50-1 and Illustris-1. However, the sampling size for the KS-tests never reach a sampling size of 500 since the sizes of the different populations are all below 500. As mentioned in Section 3.2.6, the sampling size is scaled to the smallest subsample size (e.g in TNG50-1, comparing NOT (size: 417) vs intermediate AGN (size: 337), the sampling size will be 337.

First, both TNG simulations yield a similar distribution of NOT, weak, intermediate and strong AGNs, although more AGNs are considered weak AGN in TNG50 than in TNG100. The overall percentage of dwarf galaxies considered AGN is the same though (TNG100:

Table 3.7: Comparison of significant KS results between different simulations.

Reference	**NOT**	**Weak AGN**	**Int AGN**
NOT	-	○ ○ ○ / ● ● ○	● ○ ● / ● ● ○
Weak	○ ○ ○ / ○ ○ ○	-	○ ○ ○ / ○ ○ ○
Int	● ○ ● / ○ ○ ○	○ ○ ○ / ○ ○ ○	-

	TNG100	TNG50	Illustris
10:1	●	○	●
D_{10}			

Table 3.8: Columns denote the reference sample while the rows are for subject samples. Each cell is further subdivided into six cells with the columns being the different simulations (TNG100-1, TNG50-1, and Illustris-1 respectively) while the rows are the 10:1 at the top and D_{10} (z=0) is at the bottom. A labelled subtable is shown below (subject Int, reference NOT). A filled circle indicates that the test reached $p \leq 0.05$ while an open circle indicates $p \leq 0.05$ is reached within error.

$\sim$ 57 per cent, TNG50: $\sim$ 58 per cent). In Illustris-1, though, the fraction is considerably smaller ($\sim$ 5 per cent). This is not surprising considering the different BH seeding and physics between the two simulations. Despite this, results are consistent between simulations, as will be described below.

Regarding time since last merger, all three simulations yield a difference in the 10:1 mass ratio mergers between the NOT and Int samples – although TNG50 only reaches a significant level within error. However, this should be seen in the light of a low sample size where this KS-test only uses a sampling size of 337 for TNG50 (500 for TNG100). Illustris likewise

utilises a smaller sampling size (254), but still manages to reach a significant level.

Regarding NN10, only TNG simulations find significant differences AGN and NOT samples – and only with NOT as subject. As described in Section 3.3.2, a red dwarf population is largely responsible for this. This subpopulation does not exist in Illustris-1. However, while simply excluding the red dwarfs (and/or requiring a gas component) in TNG does make the NOT sample resemble the AGN distributions more, KS-tests on the NN10 parameter still yields significant differences. This will be discussed further in Section 3.4.4

3.4 Discussion

Five discussions are included below. First is whether or not recent mergers play a significant role in triggering AGN activity. This is followed up by whether there is a time lag between a past environment and current AGN activity. Two more technical discussions ensue where the black hole requirement is first and goes into the details of what effect this requirement has on the sample. Second technical discussion is about whether the dwarf galaxy selection is sufficiently restrained (i.e are the dwarf galaxies found real or are they artifacts of the simulation). Lastly is a discussion about the KS sampling size, and whether this study has used the optimal sampling size or not.

3.4.1 Mergers as a significant trigger channel

TSLM found a difference for a minimum merger mass ratio of 10:1 between AGN and non-AGN galaxies – especially for intermediate intensity AGN. The differences between the distributions are an over-abundance of AGN galaxies with a merger within the last 4 Gyrs and an under-abundance at 10+ Gyr (see Figure 3.5). However, it is only 11.2 per cent of intermediate AGN with a TSLM $\leq$ 4 Gyr and 3.1 per cent for non-AGN, so it is only a minority of all intermediate AGNs in that belong to that bin. 55.7 per cent of intermediate AGN and 71.3 per cent of non-AGN have a TSLM-value of $\geq$ 10 Gyr, which also includes no mergers. The fraction of non-mergers increases with merger mass ratio but the fraction is similar between different samples (fractions for 1:10, 1:4, and 1:2 with formatting as $\mathrm{all}_{\mathrm{Int}}^{\mathrm{NOT}}$: $2.6^{2.6}_{2.6}$, $16.7^{17.6}_{16.0}$, and $45.3^{47.7}_{43.3}$ per cent). For weak AGN, the numbers are 6.9 percent and 65.1

percent, thus showing a similar but weaker trend as intermediate intensity AGN. The fraction of galaxies that has had a merger in its past but not recently (i.e 4 Gyr $\leq$ TSLM $\leq$ 10 Gyr) is 25.6, 28.0, and 33.1 per cent for non-AGN, weak, and intermediate AGN, respectively.

As mentioned in Section 3.2.4, merger activity further than 6 Gyr ago has to be considered with a grain of salt. Therefore, we pertain ourselves to only a distinction of *recent* activity meaning TSLM $\leq$ 4 Gyr and no or unaffected by mergers TSLM $\geq$ 6 Gyr.

Interpreting on these numbers, it means that a dwarf galaxy with a recent merger is more likely to host stronger AGN activity, although it is not a requirement since the majority of active galaxies have had longer a longer time since a minor merger. In fact, the median TSLM value of non-, weak, and intermediate AGN is 11.56, 11.26, and 10.52 Gyr, suggesting that most AGN galaxies are still unaffected by merger activity. So while other factors are in play for triggering most AGN activity in dwarf galaxies, mergers seem to be associated with increased and stronger AGN activity. This is similar to findings in the EAGLE simulation by McAlpine et al. (2020) who found that mergers increase rate of luminous AGN. In our study, the sample size for strong AGN activity is too small to make convincing conclusions.

An important note to make is that this study only examined the role of mergers. Martin et al. (2020a) remark that mergers only drive 20 per cent of morphological disturbances in the NewHorizon cosmological simulations but are instead most often due to interactions that do not result in a merger. Such interactions are not picked up and looked at in this study but may be an important channel for AGN triggering.

While mergers do not appear to be the most significant trigger channel, it cannot be dismissed. Further and more complex examination of the dynamics of especially intermediate AGN dwarf subhalos may be needed to map out the cause of this increase.

3.4.2 Time lag and impact from past environments

A question that is attempted to be answered in this study is whether or not the past environment can trigger or at least may lead to AGN activity further down the road. What is found is that the past environments of both $z = 0$ AGN dwarf galaxies and non-AGN galaxies are towards higher D_{10} values (i.e less dense environments, see Figure 3.10) but that they otherwise

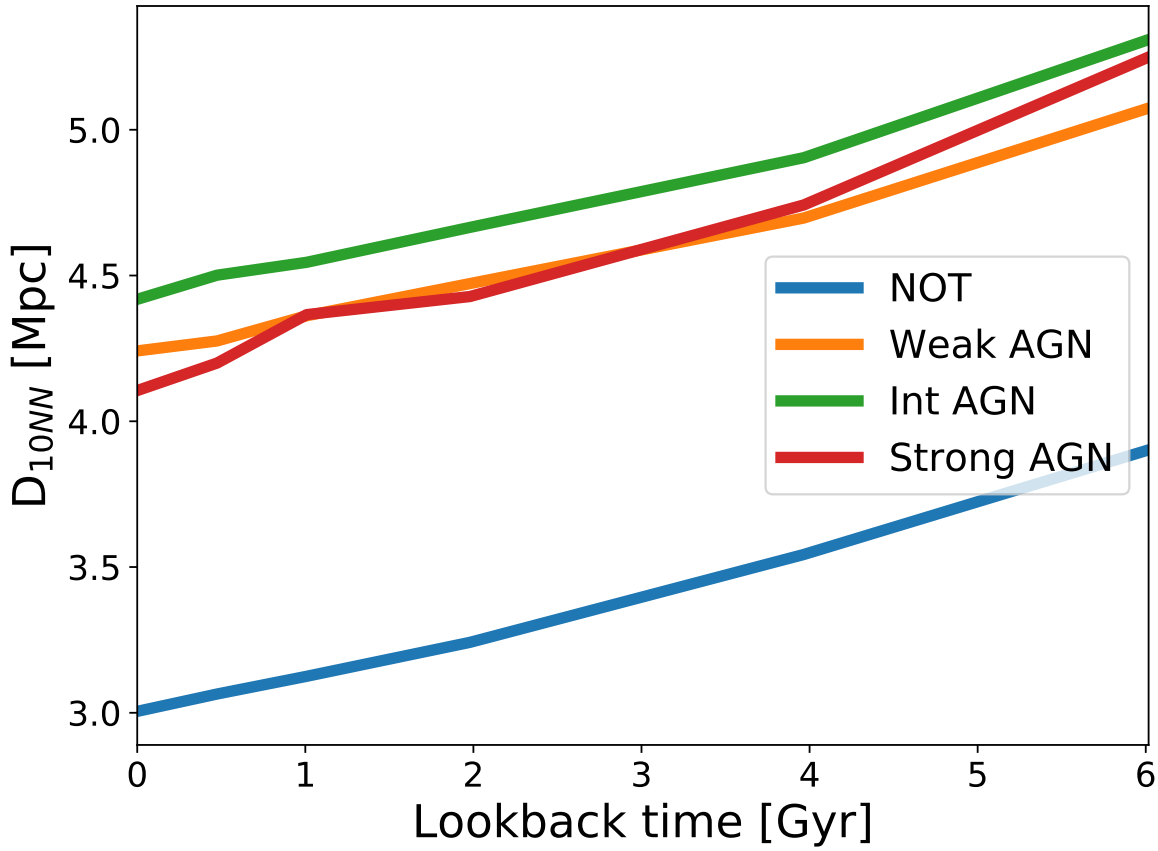

Figure 3.10: Distance to 10th nearest neighbour evolution. Each line is the median of the D_{10} distribution of different samples at different snapshots.

maintain their differences; the average D_{10} of the subsamples increase similarly, but the difference between their averages remain the same in all snapshots. Similarly, the appearance of the distributions of the subsamples all flatten going back in time, but their overall shape remains the same which means they maintain their differences.

A non-insignificant number of non-AGN galaxies that are red ($u - r \geq 2.0$) are found in dense environments ($D_{10} \leq 2$ Mpc). More specifically, around 31 per cent of non-AGN galaxies are red, and of those 31 per cent, 76 per cent of them are found in dense environments. For blue non-AGN galaxies, only 16 per cent are in dense environments, and for Int AGN galaxies, only 12 per cent are in dense environments. All of this is to say that almost all of the red non-AGN dwarf galaxies are in dense environments while only a few of blue non-AGN and AGN galaxies are in dense environments today.

Furthermore, the red peak (see Figure 3.9) is not significantly present for $z = 0$ non-AGN galaxies at $z = 0.7$ (snapshot 59) with only 1.1 per cent has $u - r \geq 2.0$. This number grows to 4.3, 11.4, 16.1, and 21.7 per cent for $z = 0.5, 0.3, 0.2, 0.1$, respectively. Of the red non-AGN dwarf galaxies at $z = 0$, 89.3 per cent of them are already in dense environments at $z = 0.7 - 0.6$. This is in contrast to both $z = 0$ AGN and blue non-AGN dwarf galaxies of which only 1.0 and 35.0 per cent are in dense environments at $z = 0.7 - 0.6$, respectively.

Given that the colour is calculated from the stellar particles in the galaxies with a dust attenuation model, a red colour suggests either an old stellar population or strong dust attenuation. For a galaxy in a dense environment that also has a shallow potential well, stripping of its gas reservoirs can quench star formation and thus be left with an aging stellar population. Sabater et al. (2015) suggest from observations that the level of nuclear activity in galaxies depends on the availability of cold gas in their nuclear regions. Applying this explanation for the results of this study, it would mean that dense environments strip dwarf galaxies of their cold gas halting star formation and AGN activity, too.

As mentioned in Section 3.2.3, all subhalos with low gas density near the BH have no gas cells associated with them and are red. However, this only accounts for slightly more than half of the red subhalos (548 out of 915 red galaxies). Generally, the red population has fewer gas cells (median red subhalo $n_{\mathrm{gas}} = 0$, red subhalos with gas cells $n_{\mathrm{gas}} = 475$, all dwarfs

$n_{\rm gas}$ = 8108). This low count of gas cells can be due to stripping, or some other physical mechanism, or due to the Subfind algorithm, although the red population does not exist in Illustris which uses a similar Subfind algorithm.

Still, this does not answer the question whether circumstances in the past has led to AGN activity now. However, a dense past environment can be a strong indicator of whether or not AGN activity is likely in the future, and if a dwarf galaxy has been in a dense environment in the past $\sim$ 6 Gyr, it is unlikely to host AGN activity.

3.4.3 Black hole requirement

As described in Section 3.2.2, only galaxies with a black hole are included in the dwarf galaxy sample. Due to the way BH seeding works in IllustrisTNG, about an eighth of dwarf galaxies are left without a black hole, which consequently means they will never show up as having AGN emission. This is despite the fact that their real life counterparts may very well host a BH.

Overall, there are two categories of no-BH dwarf galaxies, which can be inferred from Figure 3.11: 1) Those whose FoF halo are below the mass threshold for seeding (roughly 2.0 per cent), and 2) those whose FoF halo is larger than the mass threshold (roughly 98.0 per cent). These two categories will have different cosmological histories. For a minimum stellar mass cut of $10^8 M_\odot$, these percentages change to 27.4 per cent below and 72.6 per cent above the seeding threshold, which suggests that the first scenario is more common in lower mass galaxies.

The light FoF halos are assumed to have never reached the FoF halo mass threshold and thus constitute isolated galaxies that have evolved secularly and only had few to no mergers in its past. The dwarf subhalos in massive FoF halos are presumed to similarly never have been able to reach the BH seed mass threshold but whose FoF halo has merged with another halo with either a BH already (and thus are restricted from being seeded a BH) or a more massive subhalo in which the BH would then be seeded in (since if two or more subhalos exist in the same halo, only the most massive subhalo would be seeded a black hole).

The number density distribution of the dwarf subhalos with either a BH or no BH supports

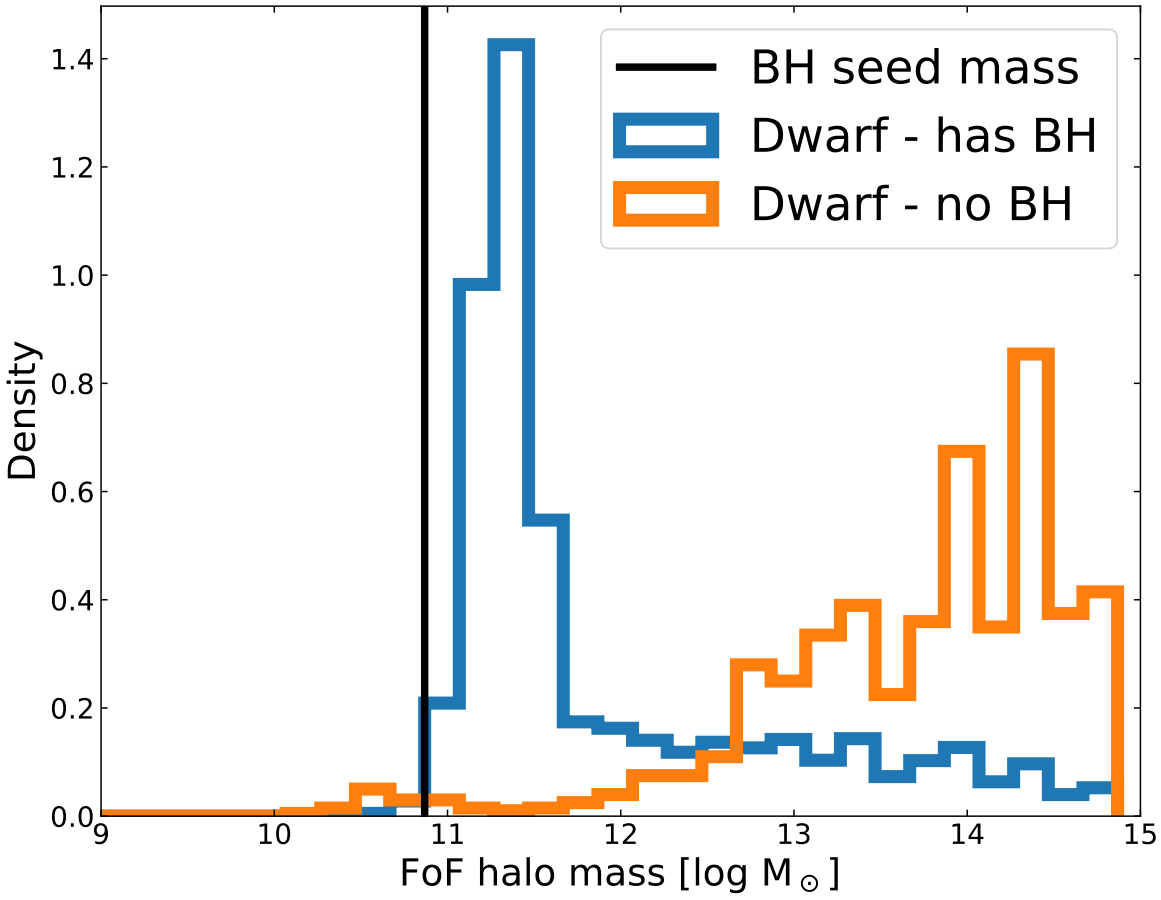

Figure 3.11: Mass distribution of low mass galaxies with (blue) and without (orange) BH. Additionally, the FoF halo mass threshold (7.38×10^{10} $M_{\odot}$) for when a BH is seeded is shown as a black line. Galaxies with (without) a BH, 0.7 per cent (2.0 per cent) have a lower mass than the seed threshold and 99.3 per cent (98.0 per cent) have a higher mass.

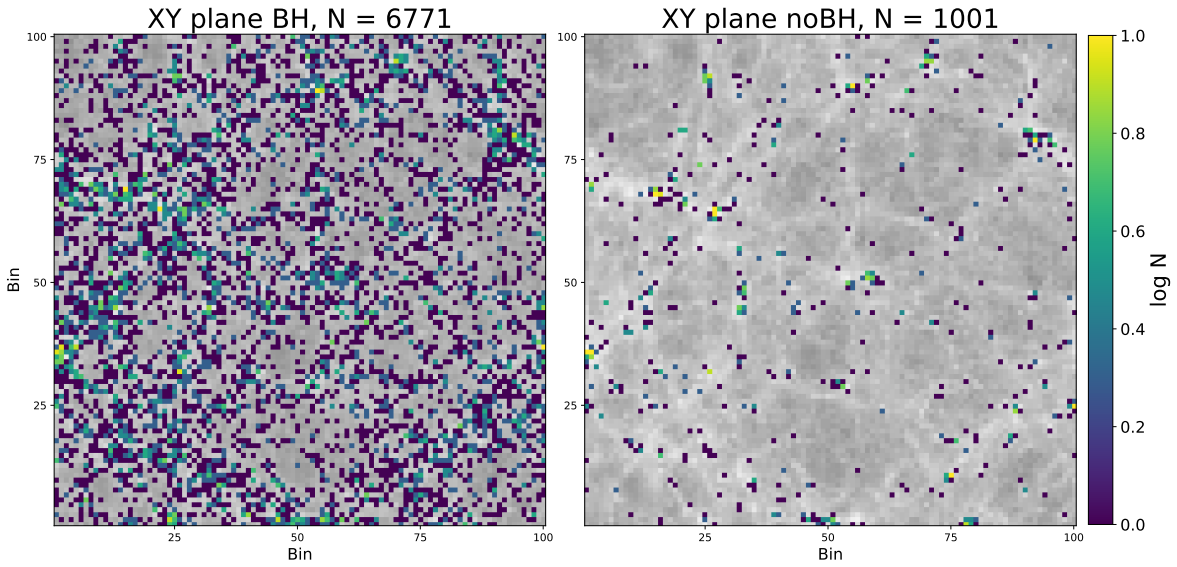

Figure 3.12: Spatial distribution on the XY plane of low mass galaxies with (left) and without (right) BH. There are 100 bins on each axis and the number of subhalos in each bin is then counted.

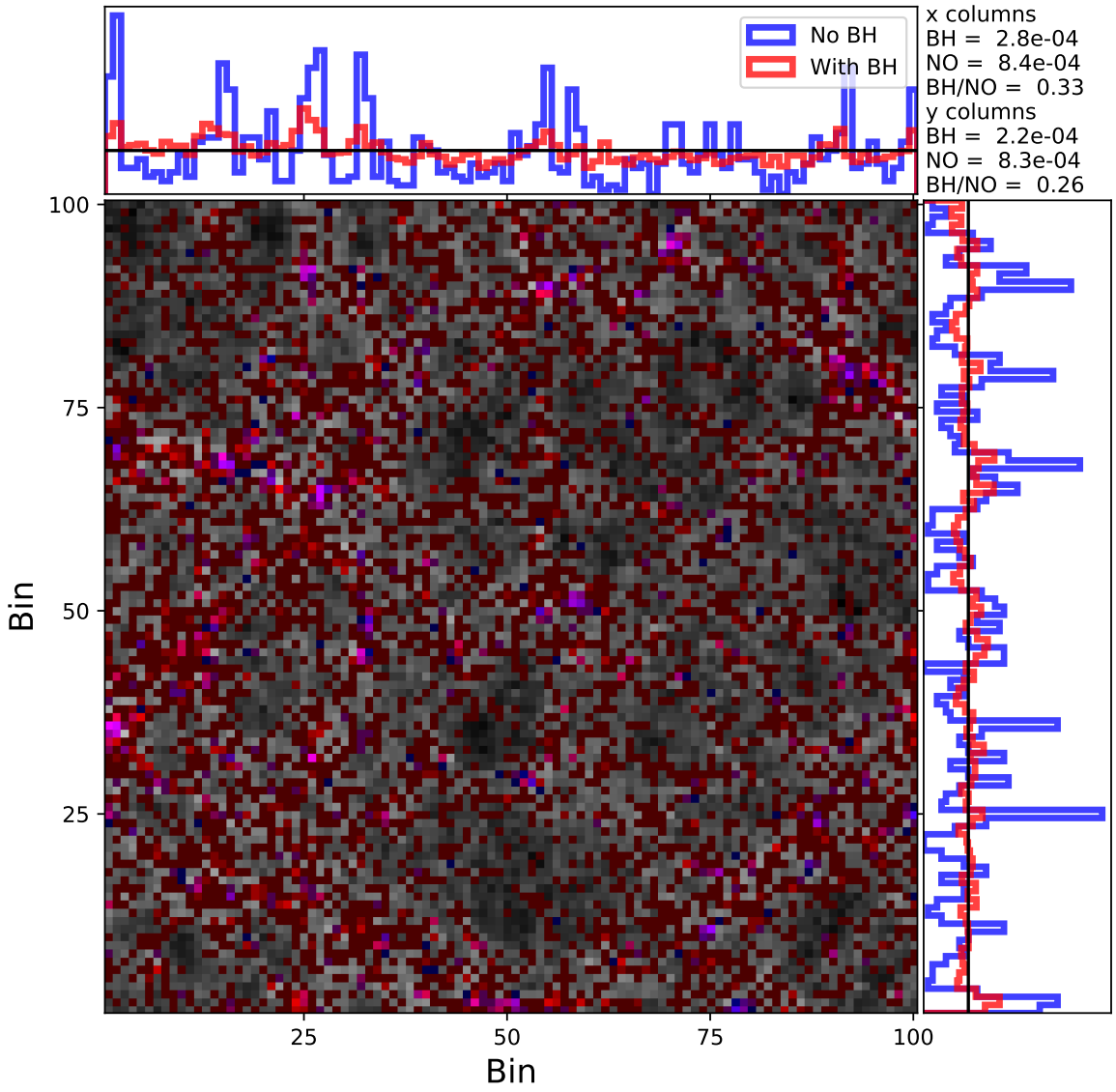

Figure 3.13: Spatial distribution on the XY plane of low mass galaxies with (red) and without (blue) BH. There are 100 bins on each axis and summed up in a normalised histogram. Each pixel bin is given a colour corresponding to the ratio between the number of BH to no-BH galaxies with more blue meaning a higher number of no-BH galaxies. The black line in the histograms show the average density, i.e the density distribution if all galaxies were spread out evenly. The departure from the this distribution of the BH and no-BH distributions is calculated and shown next to the histograms. The residual of the no-BH distribution is higher for all axes indicating that they clump together moreso than BH galaxies.

the idea of the different cosmological histories between the samples. Figure 3.12 shows the two samples side by side in a log N density plot, and despite the different sizes of the samples, the no-BH sample is more clustered with multiple spots of log N densities of around 1.0. Few or no such spots appear in the BH sample. The densities are found from binning coordinates in 100 bins and then counting the numbers of subhalos in each bin.This is further quantified in Figure 3.13 where dwarf subhalos with BH have a lower residual from an average spatial distribution compared to dwarfs with no BH.

These considerations show that certain demographics of dwarf subhalos are excluded: The ones that are very isolated and the ones in very dense environments, although the former is negligible consisting of less than 1 per cent of the total population. That means that the data set used in this study is unable to satisfyingly describe extreme scenarios of dwarf subhalos with AGNs and any conclusions are limited to the more moderate population.

Bias from exclusion of no-BH subsample

In order to quantify the bias from this exclusion, the KS tests are run with a sampling size of 500 where the non-black hole galaxies are added to either the non-AGN sample or the intermediate AGN sample. In the modified samples, the no-BH sample constitutes 25.6 per cent and 26.2 per cent in the NOT and Int samples, respectively. The no-BH population by itself is characterised by small D_{10} (and high halo masses) and a similar TSLM distribution as the NOT sample. From these considerations, it is expected that adding the no-BH sample to the NOT sample will amplify the already exisiting difference between NOT and Int regarding both mergers and D_{10} while adding the no-BH sample to Int will lessen the differences. The question is then if this is to a significant degree.

Regarding time since last merger, TSLM, the results stay the same, although adding the no-BH to Int does increase the p-value for 10:1 TSLM to just above 0.05 (with NOT as both subject as reference sample versus the modified Int sample), but the error extends below 0.05. For the 1:4 TSLM, using NOT as subject and modified Int as reference, the threshold is now linger within error. Adding the no-BH to NOT does not change the result, except making the p-value even smaller.

On the distance to the 10th nearest neighbour, D_{10}, results are also somewhat as expected but with some indication that the no-BH population is different to both the NOT and Int populations. First off, adding the no-BH sample to the Int sample, the p-value is now below 0.05 for NOT and modified Int as both subject and reference sample (compared to only NOT as subject and Int as reference in normal testing). Adding the no-BH sample to NOT maintains the original results (subject NOT and reference Int $p \leq 0.05$), but having Int as subject and modified NOT as reference now reaches the threshold within error.

The NOT and Int samples are also compared against their modified samples. For TSLM, NOT is able to produce a p-value above the threshold while Int does not (within error). This is as expected since the no-BH TSLM distribution closely resembles the NOT distribution while differing from the Int distribution. For D_{10}, neither sample is able to reach above $p = 0.05$ when their modified sample is used as subject sample. Furthermore, no tests reach the threshold when both samples are modified (e.g modified NOT vs modified Int), but this can be interpreted as subject no-BH galaxies being able to match with themselves and/or other galaxies within no-BH with whom they share characteristics with.

Summarily, in the extreme cases where the no-BH subhalos belong to either non-AGN or intermediate AGN samples, the TSLM results remain unchanged (except for 1:4 mergers) although less certain when the Int sample is modified. The picture is muddled regarding D_{10} where the no-BH sample inhabits a different parameter space compared to both NOT and Int samples. The initial results still hold but requires an added complexity to the interpretation of the results; the typical environment of the no-BH sample is very dense ($D_{10} \leq 1.0$ Mpc) and belonging to a very massive halo ($M_{halo} \geq 13\ M_{\odot}$, and this type of environment is not typically seen among the other samples. Whichever sample the no-BH is added to would introduce a unique environment and may thus yield a significant difference in D_{10} distributions. Ultimately, by not using no-BH subhalos in the main analysis and results makes us unable to gauge the impact of the very dense environments, but even when they are included, the results stay the same. Similar trends are found in TNG50-1 and even Illustris-1 despite its different BH prescription.

3.4.4 Are control galaxies properly constrained?

The selection of a proper control sample (in this study, it is synonymous with the 'NOT' sample) is important since comparison to this constitutes the basis of the statistical analysis. A biased control sample will give an impression that comparison samples follow different distributions and may lead to interpretations of their environment and past. Section 3.4.3 discusses one selection criteria (i.e requiring a black hole) that removes a significant amount ($\sim$ 13 per cent) of the low mass galaxy sample. Most of these are in dense environments (see e.g Figures 3.13 and 3.12) thus resulting in the overall distribution moving towards less dense environments/larger D_{10} distances.

Even after this correction, the control sample still has a second peak in the D_{10} distribution around 0.5-1.5 Mpc. While the KS-testing does reveal that NOT distributions and matched AGN distributions do not follow the same parent distribution, Figure 3.8 shows that this discrepancy can be almost nullified if only the blue NOT galaxies are considered (i.e matched NOT to weak AGN follow similar distributions). Since there are only blue AGNs, a matched reference sample to an AGN sample must also tend towards being blue. Nevertheless, there are two possible scenarios: 1) The control sample is properly chosen which means that dwarf galaxies in dense environments in TNG tend not to develop AGN activity, or 2) the control sample is not sufficiently constricted and thus that the D_{10} measure is biased.

Regarding the first point, several of the selection criteria are in place to avoid biases from technical parts of the simulation. That is not to say that further technical biases do not exist, but since the main contributors to a bias in the control sample have been identified and corrected, we conclude with this caveat in mind that dwarf galaxies in dense environments in TNG100-1 are less likely to develop AGN characteristics – possible due to a lack of gas. This similarly is the case for TNG50-1. In Illustris-1, though, this population is not present, suggesting either a systematic galaxy definition difference or dwarf galaxy population difference between TNG and Illustris – maybe due to a difference between stellar and AGN feedback models in Illustris and TNG.

Observationally, with a similar method, no such trend is found (Kristensen et al., 2020), and a double peak in the distance distribution is also not found (i.e one between 0.5-2.0 Mpc

and another near 2.5 Mpc, see NSA NOT distribution in Figure 3.7). This suggests that the control sample is not sufficiently constricted. It may be a manifestation of the missing satellites problem where there actually *is* a concentration of dwarf galaxies in dense neighbourhoods but they are not observable resulting in a missing peak at 0.5-2.0 Mpc in observational data (e.g Fattahi et al., 2020, for a Milky Way-Andromeda like system).

However, the most recent and highest resolution IllustrisTNG simulation run, the TNG50 (Pillepich et al., 2019; Nelson et al., 2019) run, has found that observations and simulations are in good agreement for Milky Way-Andromeda like systems (Engler et al., 2021), down to a stellar mass of 10^7 $M_\odot$. This stellar mass threshold is above the lower mass threshold of this study, so the missing satellites problem seems an unlikely culprit. The population also exists in the TNG50-1 data, but does not exist in Illustris-1. This indicates that this population is systematic to TNG. If this population does not exist in other simulations, it would suggest that the red dwarf galaxy population in TNG100-1 (and TNG50-1) is of a non-physical origin and should be excluded. Adding to this argument is the work of Dickey et al. (2021) that found an overestimation of the quiescent fraction of isolated dwarf in simulations compared to observations.

The significance of this red dwarf population on the results is tested by removing dwarf galaxies with $u - r \geq 2.0$ in the NOT sample, which removes 891 dwarf galaxies resulting in a modified NOT sample size of 2 017. While the peak near 1 Mpc in the D_{10} distribution shrinks, there is still a noticeable plateau between 0.5-3.5 Mpc (which is not present in AGN samples) and KS-testing (500 sampling size) does indeed still show a significant difference in distributions between modified NOT versus Int galaxies, although it changes to be only within error. TSLM results are unchanged. This lends credence to the results from this study – at least in TNG simulations

3.4.5 Optimal KS sampling size

There are several considerations when choosing the sampling size. One point is regarding the effect on the KS statistics (see Section 3.3.3) with a larger sample size yielding lower p-values. Obviously, if there is a statistical difference between two distributions, then it is

desirable that the testing shows this. However, large sample sizes (especially if *oversampling*) may exaggerate small differences that may be due to random error.

Another consideration is the resemblance to observations. In Kristensen et al. (2020), $\sim 40\,000$ low mass galaxies were used and $\sim 200 - 4\,000$ of these were classified as AGNs (depending on selection method). The data used was the NASA-Sloan Atlas, which covers around 1/3 of the sky to a very high level of completeness at $z \leq 0.055$. Assuming these numbers are near the current observational limits and that oversampling is not desired, this effectively limits our sample size to $\sim 200 - 4\,000$.

Although a fixed sample size of 152 was used in Kristensen et al. (2020), 500 is used in this study since it is well within the observational range described in the previous paragraph. Similarly, it does yield several tests with p-values below 0.05 (e.g, in D_{10} distributions, subject sample NOT vs AGN samples as reference samples, see Section 3.3). However, as mentioned in Section 3.3.3, increasing the sample size to even 1 000 provides further comparisons that drop below 0.05, which suggests at the very least that a trend exists in those comparisons.

McAlpine et al. (2020) remark that both minor ($1 : 10 \leq M_1/M_2 \leq 1 : 4$ and major mergers ($M_1/M_2 > 1 : 4$) play a role in black hole activity (but not significantly in black hole growth) in the EAGLE simulation – a relation that is also can be inferred in this study with a sample size of more than 1 000, although not significantly with a sample size of 500. Thus a sample size in this range reproduce results from similar studies. Observationally, Ellison et al. (2019) similarly find that AGN activity is enhanced in mergers and disturbed systems lending further credence to a sample size in this range.

However, Shah et al. (2020) find no enhancement of AGN activity for close pairs interacting, except for visually identified systems or those that has already coalesced, although they remark that their results are also consistant with low-level AGN enhancement. As discussed in Section 3.4.1, mergers do not seem to be a major trigger channel in IllustrisTNG data, but minor enhancement can be interpreted from the results of this paper.

3.5 Summary

The environments of non-AGN and matched AGN (of both weak and intermediate intensity) galaxies are different from each other with non-AGN prefering denser environments ($D_{10} \leq 2$ Mpc). Environments of non-AGN matched in stellar mass and colour to AGN galaxies are not significantly different to each other.

Around 31 per cent of dwarf galaxies that do not develop AGN characteristics are red ($u - r \geq 2$) and 76 percent of those are located in dense environments at $z = 0$. Around 6.2 Gyr ago, most were blue (99 per cent) but around half (47 per cent) were already in dense environments (and 89 per cent within $D_{10} \leq 4$ Mpc). However, even ignoring red dwarf galaxies yields a significant difference between environments of AGN and non-AGN galaxies. This suggests that prolonged exposure to dense environments is not only detrimental to star formation (from the increasing fraction of red galaxies) but also AGN activity – at least in TNG.

1:10 mass ratio mergers are to a significant degree different between intermediate intensity AGN and non-AGN galaxies. The difference is primarily recent mergers (TSLM ≤ 3 Gyr) and distant/no mergers (TSLM ≥ 10 Gyr or no merger) with intermediate intensity AGN having had more recent merger activity than non-AGN. No such difference is seen in other mass ratios, although a 1:4 ratio is following the same trend but is not significant.

This suggests that for a minority (around 8.4 per cent) of $z = 0$ intermediate intensity AGN, a small merger can lead to increased AGN activity in dwarf galaxies, although it is not always the case since 1.6 per cent of non-AGN galaxies have also had a recent merger. Observations point in both ways depending on the method employed to determine recent merger history and statistical significance level.

Lastly, there are caveats working in this mass regime in cosmological simulations. One seventh of the dwarf galaxies are not included as the TNG seeding criterion does not assign a BH to these dwarfs and they mostly belong to very dense environments leaving this environments unexplored. However, the bias from excluding this population is negligible. Also, a population of non-AGN galaxies in dense envrionments is present in TNG runs but not Illustris, suggesting either a systematic galaxy definition difference or dwarf galaxy population

difference between TNG and Illustris.

We acknowledge the support of STFC through the University of Hull Consolidated Grant ST/R000840/1, and access to VIPER, the University of Hull High Performance Computing Facility.

Full KS results visualisations

This section contains the full page visualisations of all the KS-results for TNG100-1.

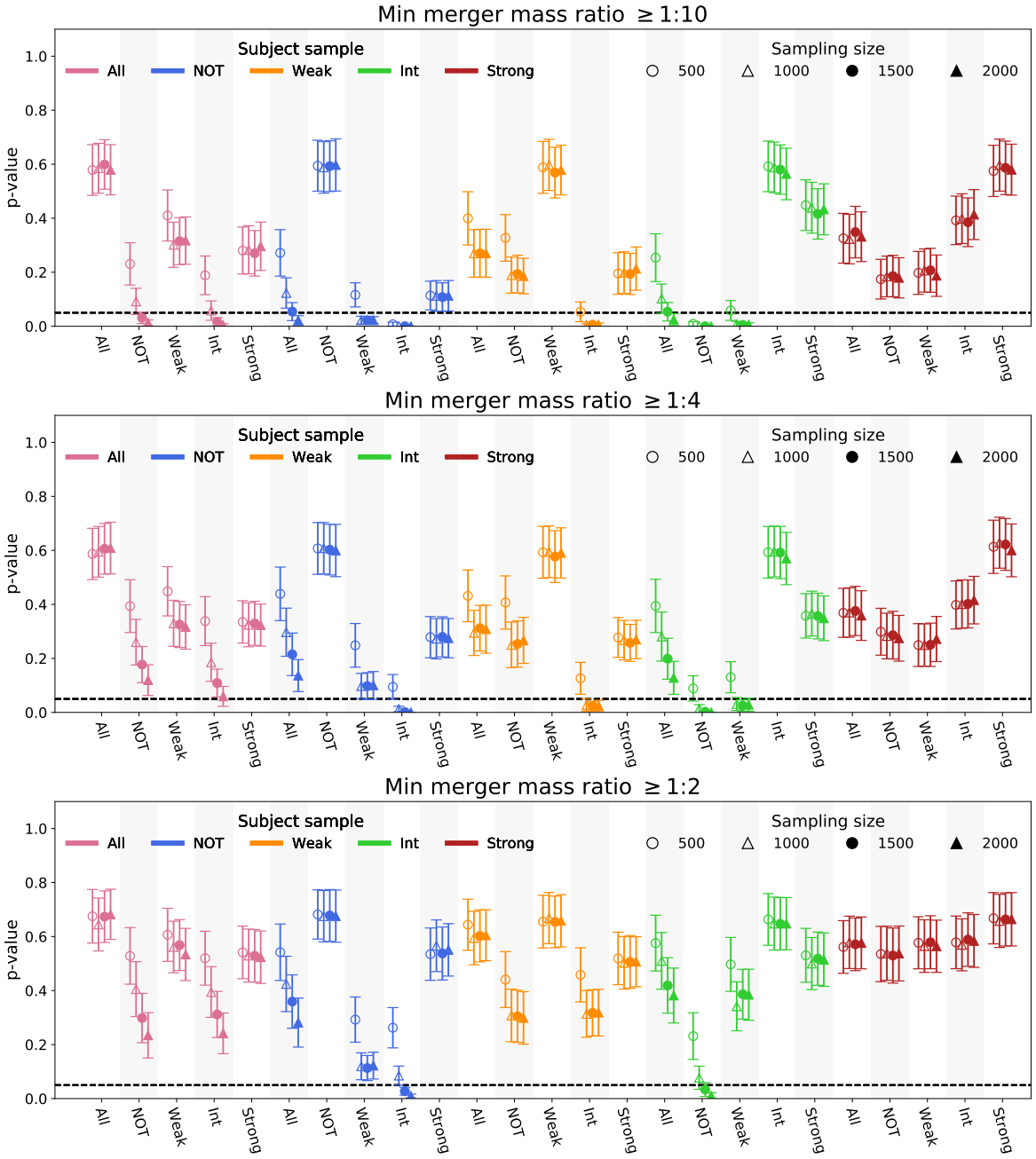

Figure 3.14: KS-testing results of merger mass ratio. Details on how the p-value and its error is calculated can be found in Section 3.2.6. Colour indicates what the subject sample is with violet being all dwarf galaxies (i.e NOT+Weak+Int+Strong), blue being non-AGN, orange is weak AGN, green is intermediate AGN, and red is strong AGN. On the x-axis is the reference samples with the marker style indicating sample size. Background shading indicates a group of data points with the same subject and reference sample. For example, if you are to look up what the p-value is for non-AGN as subject and weak AGN as reference using a sampling size of 500, it is found as the orange open circle at the 8th tick mark on the x-axis.

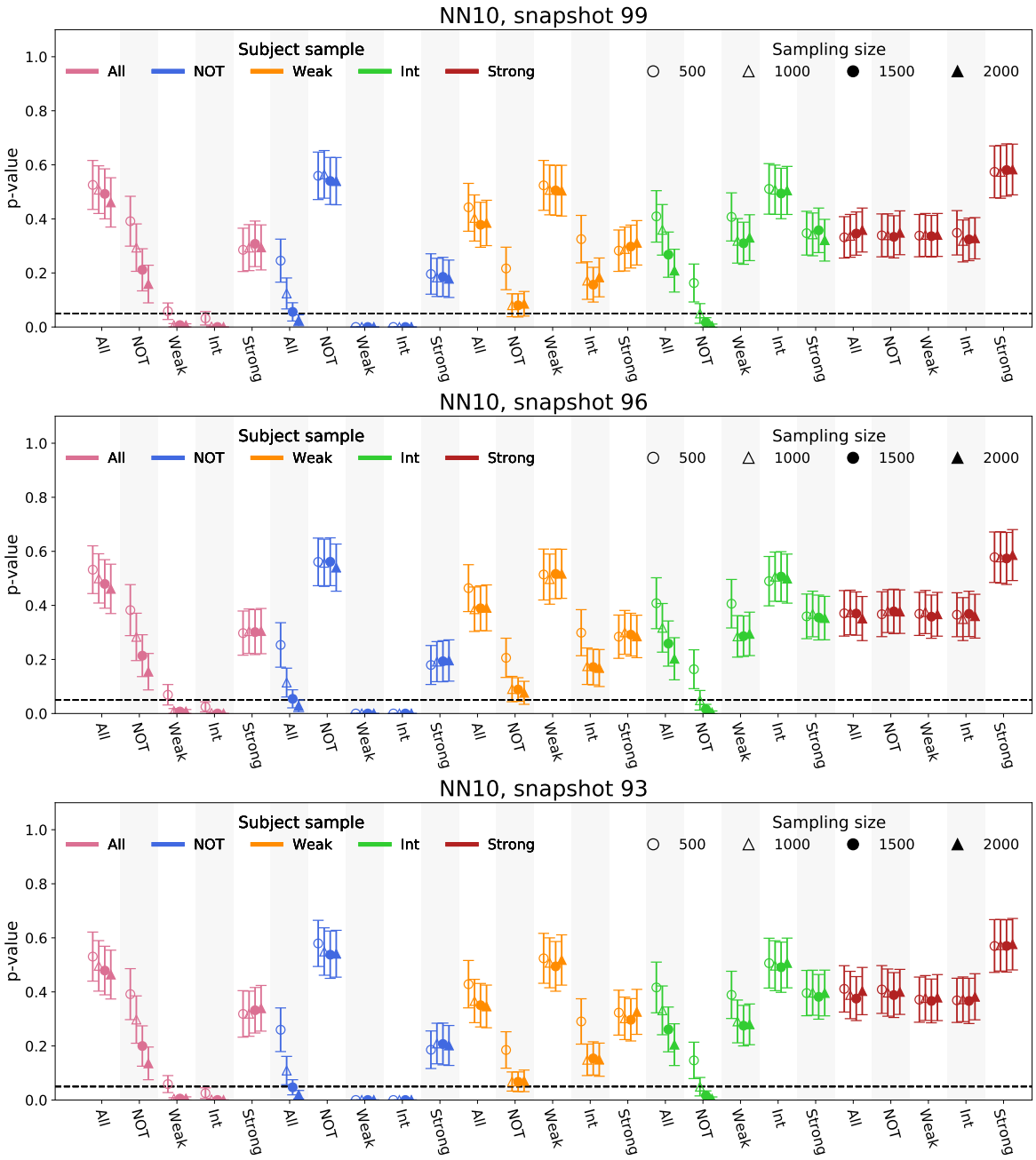

Figure 3.15: Same as Figure 3.14 but for D_{10} for snapshot 99, 96, and 93

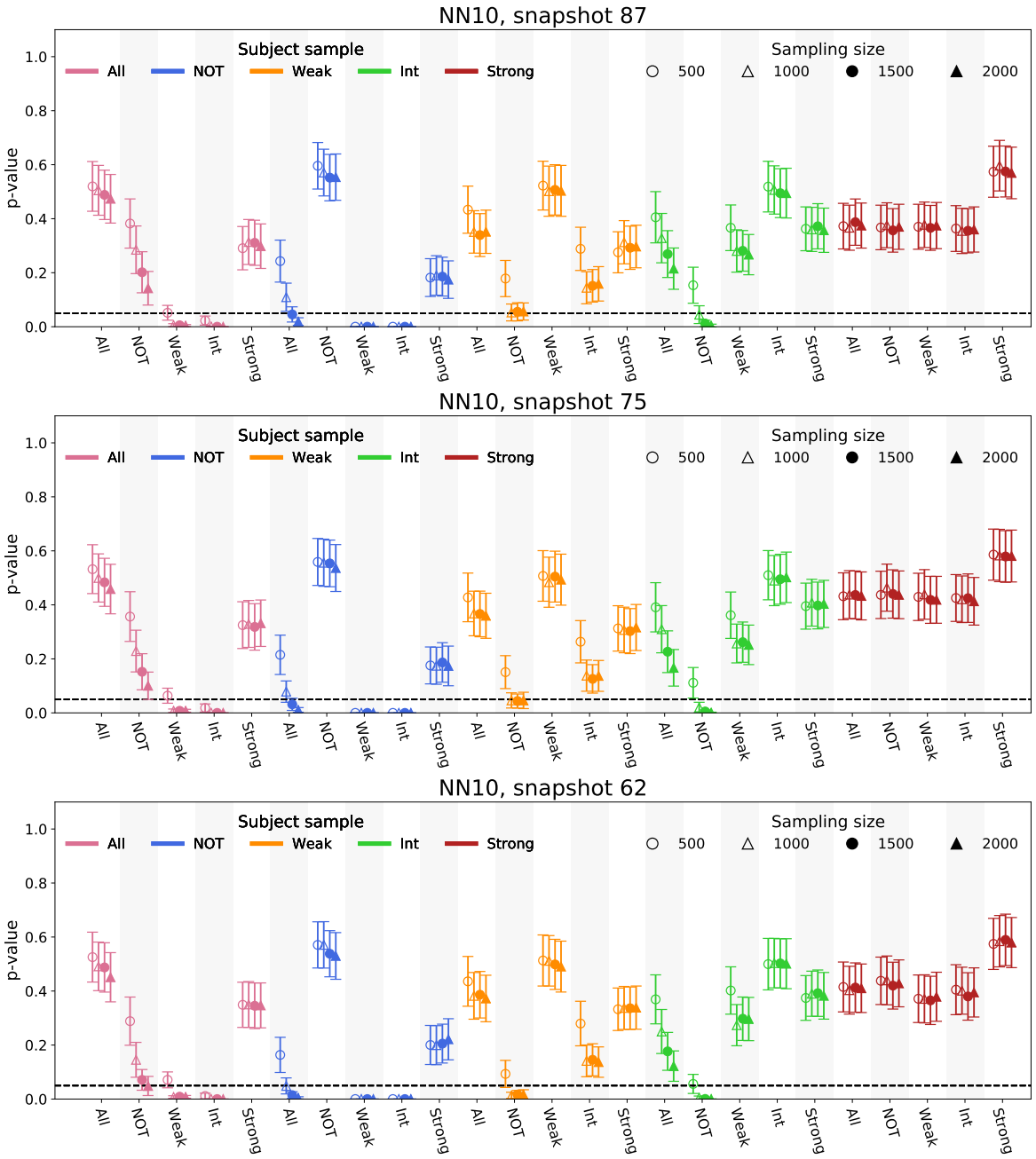

Figure 3.16: Same as Figure 3.14 but for D_{10} for snapshot 87, 75, and 62

4. Dwarf AGN and their environments – a machine learning approach

This chapter details the preliminary work in an effort to use environmental and spatially resolved parameters to classify AGN in dwarf galaxies using a machine learning approach. It has not yet been submitted for peer review. It is done in collaboration with Kevin A. Pimbblet (University of Hull) and Samantha Penny (University of Portsmouth).

Abstract

While identification, characterisation, and triggering mechanisms of active galactic nuclei (AGN) have been since the 80's, the discussion has only been extended to include dwarf galaxies within the last decade. This study aims to explore a novel AGN identification technique using a random forest (RF) classification technique, compare it to established identification methods, and investigate which set of properties/features constitute the best RF model. Data is sourced from multiple catalogues: MaNGA (and its value added catalogue, Firefly) provides spatially resolved spectra of 10 104 galaxies of which 1 149 are dwarf galaxies. These galaxies constitute the base data set, and infrared (WISE) and X-ray (XMM) observations are matched to these. The NASA-Sloan Atlas is used for estimating environmental parameters. The best model (from F1 score alone) is using internal features only of more massive galaxies. This model tends toward weighing fewer features higher and ignoring parameters that are less directly related to AGN ionisation. Conversely, this model disagrees the most with observations when it comes to dwarf galaxies, but provide twice as many dwarf AGN candidates as observations, and up to thrice as many compared to using intermediate mass galaxies as training set. This approach provides a novel and interesting venue for identification of AGN in dwarf galaxies, but the method still requires fine tuning such as feature selection optimisation and validity assessment – are the predicted AGN actually AGN? If so, RF can be used to increase

the sample size of known dwarf AGN and to adjust observational diagnostic diagrams in the low mass regime.

4.1 Introduction

Dwarf galaxies constitute the first link in the chain of hierarchical structure formation. Furthermore, they are the among the smallest observable structures forming shortly after Big Bang, and are thus good probes of the primordial conditions and early stages of galaxy evolution. They grow and evolve in size and mass content through mergers and accretion of intergalactic medium (IGM) to become massive galaxies such as the Milky Way and M31 of today. Thus, studying dwarf galaxies are vital for both understanding the first steps in galaxy evolution and to help explain the massive galaxies of today.

A central, literally and figuratively, component of galaxy evolution is the (super)massive black hole ((S)MBH) found in the center of most, if not all, galaxies. These black holes grow alongside their host galaxy (Kormendy & Ho, 2013) and can regulate and stunt the growth of their host galaxies in the case of massive galaxies, while their effect on lower mass galaxies is less strong. Conversely, bulgeless host galaxies and isolated galaxies host active galactic nuclei (AGN) less often, which is a phase where the SMBH is actively accreting, while merging galaxies host AGN more often. This suggests that feedback and interactions between SMBHs and host galaxies go both ways.

The physical processes that affect galaxies can roughly be split into two categories: internal (secular) and external (environmental) processes. Internal processes include supernova (SN) feedback (Larson, 1974; Dekel & Silk, 1986; Kormendy et al., 2009) and AGN feedback (Fabian, 2012), and the net effect appears to be negative on the star formation rates of their host galaxies. External processes include ram-pressure stripping (Gunn & Gott, 1972), which is caused by the forces exerted on the gas contents of the galaxy by the intergalactic medium, and tidal interactions with other galaxies.

One method to study spatially resolved internal details is by using integrated field unit (IFU) spectroscopy such as Mapping Nearby Galaxies at APO (MaNGA; Bundy et al., 2015). Instruments like this provide two-dimensional maps of stellar velocities, mean stellar ages,

element abundance ratio, and more. Such a wealth of information can be used to characterise different populations of galaxies in novel ways that may reveal, aid, or fine tune new or existing relations.

The increasing number of parameters presents new challenges. One challenge is how to include these new parameters and how to evaluate their importance. A solution to this challenge is using machine learning (ML), which is well-utilised in general data science already. Although using ML techniques is still in its infancy in astronomy and astrophysics, several methods have already been employed successfully for e.g detecting neutral hydrogen (Fumagalli et al., 2020), damped Lyα systems (Parks et al., 2018; Garnett et al., 2017), and broad absorption line quasars (Guo & Martini, 2019) – and even unsupervised galaxy morphology classification (Martin et al., 2020b)

Further challenges exist when working with dwarf galaxies. They occupy the faint end of the galaxy distribution function, which makes it a difficult task to find them. Further to this, their low surface brightness requires long integration times to obtain a robust signal to noise ratio on both continuum emission and spectral line fluxes. This means that the spectral fingerprints used for e.g classification of AGN are vague or drowned in noise from star formation processes (Lupi et al., 2020).

AGN classification is further complicated by the fact that the usual tools are fine-tuned to regular galaxies (Sartori et al., 2015; Mackay Dickey et al., 2019; Cann et al., 2019; Lupi et al., 2020), which means that AGN in dwarf galaxies are more likely to be missed. Approaches to overcome this problem are to use multi-wavelength observations (e.g Cann et al., 2019; Cann et al., 2020) or outflows (Manzano-King et al., 2019). As such, there may be several other proxy parameters related to AGN activity that could be exploited to better identify AGN in dwarf galaxies.

A naive approach to improve AGN classification of dwarf galaxies is to train a random forest (RF) ML classifier on regular galaxies with robust AGN classification using a wide range of features. The trained RF classifier can then used to predict the labels (i.e whether galaxies have AGN or not) of dwarf galaxies.

This science project explores the use of an RF classifier trained on regular AGN galaxies

and quantifies what features the models consider important. Furthermore, the models are then used to predict the labels of dwarf galaxies and compared to their classification from traditional diagnostic tools. Features are derived from four catalogues or sources: MaNGA, a derivative catalogue of MaNGA, Firefly, NASA-Sloan Atlas, and WISE. The features and properties are a mix of inner properties and environmental ones. This study assumes a cosmology of Λ-CDM Universe with $\Omega_{\Lambda,0} = 0.6911$, $\Omega_{m,0} = 0.3089$, $\Omega_{b,0} = 0.0486$, and $h = 0.73$.

4.2 Data

Data is sourced from multiple surveys, but the basis catalogue containing all potential sources is the NASA-Sloan Atlas (NSA, `v1_0_1`), which contains 641 409 sources. Spatial properties and kinematics as well as optical emission lines are obtained using the MaNGA dataset (10 104 matches). A value-added catalogue (VAC) derived from MaNGA data is also used, namely Firefly. Infrared data is from the AllWISE survey and matched to sources in the NSA while X-ray data is from 4XMM-DR11. Environmental properties are obtained from NSA while results and RF training and predictions are only carried out on MaNGA sources.

4.2.1 NSA and MaNGA

The NASA-Sloan Atlas (NSA) is derived from the Sloan Digital Sky Survey (SDSS) using a different data processing pipeline (Blanton et al., 2011) than the standard SDSS one, including using elliptical Petrosian aperture (instead of circular) and improved photometric estimates of nearby galaxies. In particular, the `nsa_v1_0_1.fits` catalogue is used since it also constitutes the target catalogue for Mapping Nearby Galaxies at APO (MaNGA; Bundy et al., 2015) survey. The redshift limit is $z = 0.15$ and contains 641 409 sources.

The NSA catalogue contains absolute magnitude measurements in FNugriz filters, redshifts sourced from various catalogues (alfalfa, ned, sdss, sixdf, twodf, or zcat), and sky positions. A number of other properties are also available, but they are not relevant for this science project. Environmental descriptors are based on the sources in this catalogue and their positions.

MaNGA forms the backbone of the optical emission line analysis as well as spatial stellar ages. It uses integral field unit (IFU) spectroscopy to obtain and measure spectra at hundreds of points in galaxies. The first MaNGA survey was released alongside the SDSS DR13 and contained 1 390 galaxies, but the newest catalogue released alongside DR17 contains 10 104 galaxies, which is the version used in this study. The observation targets are selected from the v1 NSA catalogue and observed in a dithering pattern (Law et al., 2015) with the IFUs having 19 to 127 fibers with each fiber having a 120 micron diameter equating to $\sim$ 2" sky cover. The dithering pattern allows for coverage of the gaps between the fiber bundles, and the spatial elements of the data product are called *spaxels*. The wavelength coverage is 360-1000 nm with a resolution of $R \sim 2\,000$.

Galaxies in the MaNGA data are equipped with a NSA-ID, which corresponds to the ID the galaxies have been given in the NSA catalogue. Matching these catalogues are thus done by matching these columns in the two data sets.

Emission line measurements

MaNGA data provides emission line fluxes, and there are several ways of obtaining these. First is to use the the MaNGA Data Analysis Pipeline (DAPall, Westfall et al., 2019; Belfiore et al., 2019) catalogue, which is a summary catalogue of all observations in the survey. The catalogue contains fluxes for 35 emission lines measured either by summing or gaussian fitting at either the inner 2.5", within 1 effective radius, or within the full IFU coverage. Emission line measurements are carried out on the spaxels.

However, spatial information is relevant for this study, and we therefore obtain emission lines differently. We differentiate between the inner and outer parts of the galaxies by grouping the innermost 20 per cent of the spaxels as the inner region and the rest as the outer region and summing the fluxes in these regions. Spaxels are considered inner spaxels if their centers are inside $0.2 \times$ n-spaxels. As such, it is similar to the measurements from the DAPall catalogue but with slightly different cut-offs. The advantage of this method is that noise readings are readily available, which they are not for the DAPall catalogue. Noise readings are important for robust AGN classification.

While this method is not perfect (e.g it does not take inclination into consideration), it utilises the strength of IFU spectroscopy by only considering the innermost spaxels for AGN classification thus leaving out emission from the outer parts of the galaxy that might otherwise drown out the AGN signal.

4.2.2 Firefly

Fitting IteRativEly For Likelihood analYsis, or Firefly (Wilkinson et al., 2017), is a spectral fitting code that derives stellar properties of stellar systems such as galaxies based on a chi-squared minimisation approach. The code fits combinations of simple stellar population (SSP) models and outputs information such as age, metallicity, and stellar mass.

The MaNGA FIREFLY value-added catalogue (VAC; Goddard et al., 2017; Maraston et al., 2020) contains both spatial and global information about the stellar properties. Only light-weighted parameters are used (and not mass-weighted), which means that the data and model fluxes are normalised before fitting and adjusted to match the actual flux values afterwards. Wilkinson et al. (2017) remark that differences between light-weighted and mass-weighted results can be interpreted as more or less extended episodes of star formation.

More specifically, the six properties used in this study are light-weighted age and metallicity within 1 effective radius, and the gradient and zero point (from a linear fit) to these within 1.5 effective radius. 6 objects do not have associated Firefly data, and since this data is used for the RF training and prediction, their Firefly data values are set to 0. Since this concerns only a small number of galaxies, it is not expected to affect training the classifiers or the predicted classes.

4.2.3 AllWISE

Mid-infrared measurements are often used for AGN identification, especially obscured AGN, and showcase unique aspects of AGN signatures. The infrared data used is from the Wide-field Infrared Survey Explorer (WISE; Wright et al., 2010), and more specifically, the AllWISE data release. Galaxies are matched using sky coordinates and matched to within 5 arcseconds using TOPCAT – a similar procedure as done by Kaviraj et al. (2019). Some concerns exist since the spatial resolution of the data is such that there is a high incidence rate of

mis-attributing observations to sources, especially for dwarf galaxies (Lupi et al., 2020). 57 MaNGA galaxies do not have WISE coverage. These galaxies are treated as though they have no infrared emission, i.e emission set to a value of 0. Two channels are used in this study, W1 and W2, which are centered on 3.4 μm and 4.6 μm, respectively, and a SNR of 5 is required for robust emission measurements

4.2.4 4XMM-DR11

X-ray observations are almost unequivocally related to AGN phenomena, and the 4XMM-DR11 (Traulsen et al., 2020; Webb et al., 2020) is a serendipitous X-ray source catalogue from the XMM-Newton Observatory and created by the XMM-Newton Survey Science Centre (SSC) on behalf of the European Space Agency (ESA), who is the owner of the observatory. It contains 602 543 unique X-ray sources, but only 4 585 of them are matched to an NSA source within a 10 arcsec radius (similar matching procedure as Birchall et al., 2020), and 314 of those are also found in MaNGA data.

4.3 Methods

AGN are chosen in multiple wavelength regimes. In optical wavelengths, two Baldwin-Phillips and Terlevich (BPT; Baldwin et al., 1981) diagrams are used ([N II] and [S II] versions) with the Kewley et al. (2001); Kauffmann et al. (2003) selection criteria. WISE Infrared selection is based on Jarrett et al. (2011); Stern et al. (2012). X-ray selection is following the methodology of Birchall et al. (2020). The RF approach used is from the scikit library in Python.

4.3.1 Dwarf galaxy selection and mass splits

NSA provides two mass estimates (given in $h^{-2}M_{\odot}$) estimated from K-correction fits for different apertures: Sersic and elliptical Petrosian. The same selection criteria as Kristensen et al. (2020, 2021) is used with a stellar mass cut of $M_* \leq 3 \times 10^9 M_{\odot}$, but this version of the NSA catalogue does not contain velocity dispersions since spectroscopic measurements are not available. Furthermore, the aperture mass used is the elliptical Petrosian since it may give a better coverage of dwarf galaxies compared to Sersic aperture. Using $h = 0.73$, this yields

Table 4.1: Overview of AGN selection numbers of MaNGA data

AGN type	Dwarf		Intermediate1		Intermediate2		Massive	
Total	1 149	(100 %)	3 258	(100 %)	3 258	(100 %)	2 439	(100 %)
Non-AGN[a]	935	(81.4 %)	2 074	(63.7 %)	1 202	(36.9 %)	470	(19.3 %)
[N II] BPT[b]	97	(8.4 %)	603	(18.5 %)	1 218	(37.4 %)	1 457	(59.7 %)
[S II] BPT[b]	178	(15.5 %)	1 011	(31.0 %)	1 827	(56.1 %)	1 590	(65.2 %)
WISE	4	(0.3 %)	15	(0.5 %)	24	(0.7 %)	12	(0.5 %)
XMM	12	(1.0 %)	48	(1.5 %)	126	(3.8 %)	89	(3.6 %)

[a]: Excludes also galaxies with low SNR AGN. [b]: Only innermost 20 per cent of spaxels used.

63 656 dwarf galaxies in NSA (~ 9.9 per cent of NSA sources) and 1 149 dwarfs in MaNGA data (~ 11.4 per cent of MaNGA sources).

Further mass splits are also used for training purposes of the RF classifier. Intermediate mass galaxies are defined as having a stellar mass of $\leq 3\times10^9 M_\odot < M_* \leq 8\times10^{10} M_\odot$, which is roughly halfway between the Milky Way and the Andromeda galaxy and yields 493 479 NSA galaxies (~ 76.9 per cent) and 6,516 MaNGA galaxies (~ 64.5 per cent). The MaNGA galaxies are further split in two (at $M_* \simeq 1.75 \times 10^{10} M_\odot$) in order to construct samples of similar size to the sample of massive galaxies. High mass galaxies are the ones above the $M_* > 8 \times 10^{10} M_\odot$ threshold and yields 84 250 NSA galaxies (~ 13.1 per cent) and 2 439 MaNGA galaxies (~ 24.1 per cent).

4.3.2 AGN selection

Several AGN selection techniques are employed. An overview of the numbers from each mass bin and selection is shown in Table 4.1.

Optical selection is done by the [N II] and [S II] BPT diagrams with emission line data described in Section 4.2.1. The selection critera for AGN follows the definitions laid out in Kauffmann et al. (2003); Kewley et al. (2006):

$$\log([\mathrm{O\,III}]/\mathrm{H}\beta) > \frac{0.61}{\log([\mathrm{N\,II}]/\mathrm{H}\alpha) - 0.47} + 1.19, \tag{4.1}$$

$$\log([\mathrm{O\,III}]/\mathrm{H}\beta) > \frac{0.72}{\log([\mathrm{S\,II}]/\mathrm{H}\alpha) - 0.32} + 1.30, \tag{4.2}$$

where the emission line fluxes are measured for the 20 per cent innermost spaxels. Furthermore, a signal-to-noise ratio (SNR) of at least 3 on all emission lines is required. Using the innermost 20 per cent of spaxels, this yields 97 [N II] AGN dwarf galaxies ($\sim$ 8.4 per cent of dwarf population) and 178 [S II] AGN dwarf galaxies ($\sim$ 15.5 per cent). These occupation fractions are higher than other studies Kristensen et al. (2020) that rely on the integrated flux from the whole galaxy (or, what is covered by the SDSS fiber).

WISE AGN selection follows that of Stern et al. (2012) using:

$$W1 - W2 \geq 0.8, \tag{4.3}$$

where W1 and W2 are the WISE channels centered on 3.4 μm and 4.6 μm, respectively, and a SNR of 5 is required on both channels to ensure a robust classification.

XMM selection follows that of Birchall et al. (2020) where XMM observations are first matched to optical sources (NSA in this study), and a position-error-normalised separation is calculated for each and a value of less than 3.5 is required in order for the match to be robust. X-ray fluxes are then calculated for X-ray binary and gas contamination, and the XMM flux is then required to be at least three times of the combined contamination in order to receive an X-ray AGN classification.

4.3.3 Environment estimations

Environment is estimated from the NSA by calculating the distances to the 10 nearest neighbours with two mass cuts in the neighbouring galaxies: Either all galaxies are included or only high mass galaxies ($M_* \geq 5 \times 10^{10}$ M$_\odot$). Using all galaxies for environment estimation provides more detail about the large scale environment and local galaxy density while using only massive galaxies is more often used as an estimator of (strong) tidal forces and satellite status. For a review of different environmental estimates, see Muldrew et al. (2012).

Additionally, the gradient of the distances to nearby neighbours is also obtained. While this can be done by simply taking the distance to the N'th nearest neighbour and divide it by N, a linear fit is made to all 10 distances of the neighbouring galaxies instead. This is more robust to extreme values of distances, e.g if the 10th nearest neighbour is in another group. This linear fit is performed on both all galaxies and massive ones, and each fit provides two parameters: a gradient and a zero point, both of which are used in the machine learning classification.

4.3.4 Machine learning classification

A random forest (RF, Ho, 1995) approach is utilised and ultimately used to classify dwarf galaxies based on different training sets with different sets of properties and labels used. For training purposes, labels are the true classification of the sources in the dataset (e.g an element with a label equal to 0 means it is a non-AGN and 1 if it has AGN characteristics). More specifically, training sets are first divided into different mass bins: Light intermediate mass (intermediate1), heavy intermediate mass (intermediate2), and massive galaxies (massive). The labels are the different AGN classifications: BPT [N II], BPT [S II], WISE, and XMM. Lastly, three different sets of properties (or also called features) are used: Inner, outer, and all. Inner features are internal properties such as emission line ratios and MaNGA data while outer features are environmental estimators. All features are the combination of inner and outer features.

The features used in the different scenarios are listed in Table 4.2 and Table 4.3. The RF algorithm requires that all elements have all the features used in training and classification, but not all galaxies have high SNR on some emission lines, especially Hβ. In this naive model, SNR is not taken into account when training or classifying galaxies, but SNR is used in AGN selection (i.e label specification in training sets).

The combination of mass bins ($n = 3$), labels ($n = 4$), and property selection ($n = 3$) yields 36 different models and training sets. Each training set is evaluated using K-folding cross validation (CV) with $K = 10$. This also produces 10 estimators, or classifiers, that are used on both the training sets and dwarf galaxies to estimate the uncertainty in the model.

Table 4.2: Overview of outer features used for RF

Feature	Scenario	Source	Description
`linallenv_a`	Outer	NSA	Gradient from linear fit using all galaxies
`linallenv_b`	Outer	NSA	Zero point from linear fit using all galaxies
`linmasenv_a`	Outer	NSA	Gradient from linear fit using massive galaxies
`linmasenv_b`	Outer	NSA	Gradient from linear fit using massive galaxies
`1nnall`	Outer	NSA	Distance to 1st nearest neighbour, all galaxies [Mpc]
`3nnall`	Outer	NSA	Distance to 3rd nearest neighbour, all galaxies [Mpc]
`5nnall`	Outer	NSA	Distance to 5th nearest neighbour, all galaxies [Mpc]
`10nnall`	Outer	NSA	Distance to 10th nearest neighbour, all galaxies [Mpc]
`1nnmassive`	Outer	NSA	Distance to 1st nearest neighbour, massive galaxies [Mpc]
`3nnmassive`	Outer	NSA	Distance to 3rd nearest neighbour, massive galaxies [Mpc]
`5nnmassive`	Outer	NSA	Distance to 5th nearest neighbour, massive galaxies [Mpc]
`10nnmassive`	Outer	NSA	Distance to 10th nearest neighbour, massive galaxies [Mpc]

Table 4.3: Overview of inner features used for RF

Feature	Scenario	Source	Description
`inoutn2ha`	Inner	MaNGA	Ratio of log([N II]/Hα) of the inner and outer region
`inouts2ha`	Inner	MaNGA	Ratio of log([S II]/Hα) of the inner and outer region
`inouto3hb`	Inner	MaNGA	Ratio of log([O III]/Hβ) of the inner and outer region
`lwage1re`	Inner	Firefly	Light-weighted age within a shell located at 1 effective radius
`lwagegrad`	Inner	Firefly	Light-weighted age gradient of linear fit obtained within 1.5 effective radii
`lwagezp`	Inner	Firefly	Light-weighted age zeropoint of linear fit obtained within 1.5 effective radii
`lwz1re`	Inner	Firefly	Light-weighted metallicity [Z/H] within a shell located at 1 effective radius
`lwzgrad`	Inner	Firefly	Light-weighted metallicity [Z/H] gradient of linear fit obtained within 1.5 effective radii
`lwzzp`	Inner	Firefly	Light-weighted metallicity [Z/H] zeropoint of linear fit obtained within 1.5 effective radii
`w12colour`	Inner	WISE	WISE colour from channels W1-W2
`w34colour`	Inner	WISE	WISE colour from channels W3-W4

K-folding CV works by splitting the data set up in K subsets and using $K-1$ to train the classifier to evaluate the remaining subset. Then, a different subset is chosen for evaluation and the remaining $K-1$ subsets are used for training, and so on, until K scores/evaluations have been carried out. From the CV, an average F1 score and its standard deviation. The F1 score is the harmonic mean of precision and recall rate:

$$F1 = \frac{2 \cdot P \cdot R}{P + R}, \tag{4.4}$$

where P is the precision, i.e how many of the predicted AGN actually are AGN, and R is the recall rate, i.e how many of the actual AGN that are correctly identified:

$$R = \frac{true\ positive}{true\ positive + false\ negative}, \tag{4.5}$$

$$P = \frac{true\ positive}{true\ positive + false\ positive}, \tag{4.6}$$

The average and standard deviation of the importance of the features used are also calculated.

Each model is then used to predict the label for the sample of dwarf galaxies. Since this is a novel AGN identification method, a predicted dwarf AGN is not necessarily a true AGN and similarly, some predicted non-AGN are potentially true AGN – the RF approach assumes no prior knowledge of the AGN nature of the dwarfs.

Nevertheless, using the predicted labels in combination with the pre-existing labels (from Section 4.3.2) creates four sub-populations of dwarf galaxies: True positive, true negative, false positive, and false negative, where the true/false flag refer to whether or not the predicted label is the same as the pre-existing one and the positive/negative flag refers to the predicted RF AGN status (with positive meaning it has an AGN). The false/true positive/negative designations are not taken as actual true predictions and as a test of the validity of the model, but the designations that can highlight overlaps and differences between the usual diagnostic tools and RF predictions.

4.4 Results

Each of the 36 models provide unique results, but there are several similarities between multiple models that do not warrant a separate discussions. For example, using only outer

features yields low model scores across the board and proves to be a poor choice for accurately predicting AGN, and the small differences in feature importance between the outer features only models are so small that it may just be random.

This section presents the results from the RF pipeline such as CV scores and dwarf AGN predictions. The different predicted classes of dwarf AGN are then characterised briefly at the end.

4.4.1 Cross-validation

The CV scores are given in Table 4.4. The first table contains the scores from using all features for the training model while the second and third table contains inner and out features, respectively.

Due to the low number of WISE and XMM AGN in all mass bins, the scores are very extreme: Either they have close to 0 or close to 1 with little error, or they have errors close to 50 per cent of the score. The high scores of WISE is due to the fact that WISE colours are used as features for training, and they are weighed heavily resulting in very accurate predictions. No or few XMM AGN are recalled ultimately leading to poor scores.

Using only outer features yields poor model scores except for massive BPT galaxies and [S II] BPT intermediate 2 galaxies. However, these classes have high AGN occupation fractions (56-65 per cent) and a higher score is therefore expected since predicting completely randomly for an almost even class balance will yield a score of around 0.50. The scores using only inner features or all features similarly show high scores for massive galaxies, although they are slightly higher (0.05-0.10) than only outer feature models suggesting that inner features improve the model predictions. The score differences between using only inner or all features are within error, which points towards environmental features playing little to no role in order to identify AGN correctly.

The model with the highest score (excluding WISE and XMM ones) is intermediate 2 [S II] BPT using only inner features. Intermediate 2 models have the highest score difference between [N II] and [S II] BPT models, but this mass bin also has the highest difference in AGN occupation fraction between BPT models, which may explain the difference.

Table 4.4: Overview of F1 scores from cross validation for models using different AGN labels and feature sets using three different mass selections for training set.

AGN label	Mass selection, all features		
	Intermediate 1	Intermediate 2	Massive
[N II] BPT[a]	0.66 ± 0.04	0.73 ± 0.03	0.78 ± 0.02
[S II] BPT[a]	0.69 ± 0.03	0.82 ± 0.02	0.79 ± 0.02
WISE	0.75 ± 0.39	0.98 ± 0.06	0.60 ± 0.49
XMM	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
AGN label	Mass selection, inner features		
	Intermediate 1	Intermediate 2	Massive
[N II] BPT[a]	0.69 ± 0.04	0.74 ± 0.03	0.77 ± 0.01
[S II] BPT[a]	0.70 ± 0.03	0.82 ± 0.01	0.80 ± 0.02
WISE	0.91 ± 0.14	0.98 ± 0.06	1.00 ± 0.00
XMM	0.00 ± 0.00	0.03 ± 0.06	0.02 ± 0.06
AGN label	Mass selection, outer features		
	Intermediate 1	Intermediate 2	Massive
[N II] BPT[a]	0.08 ± 0.07	0.24 ± 0.04	0.67 ± 0.02
[S II] BPT[a]	0.18 ± 0.06	0.62 ± 0.02	0.75 ± 0.01
WISE	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
XMM	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00

[a]: Only innermost 20 per cent of spaxels used.

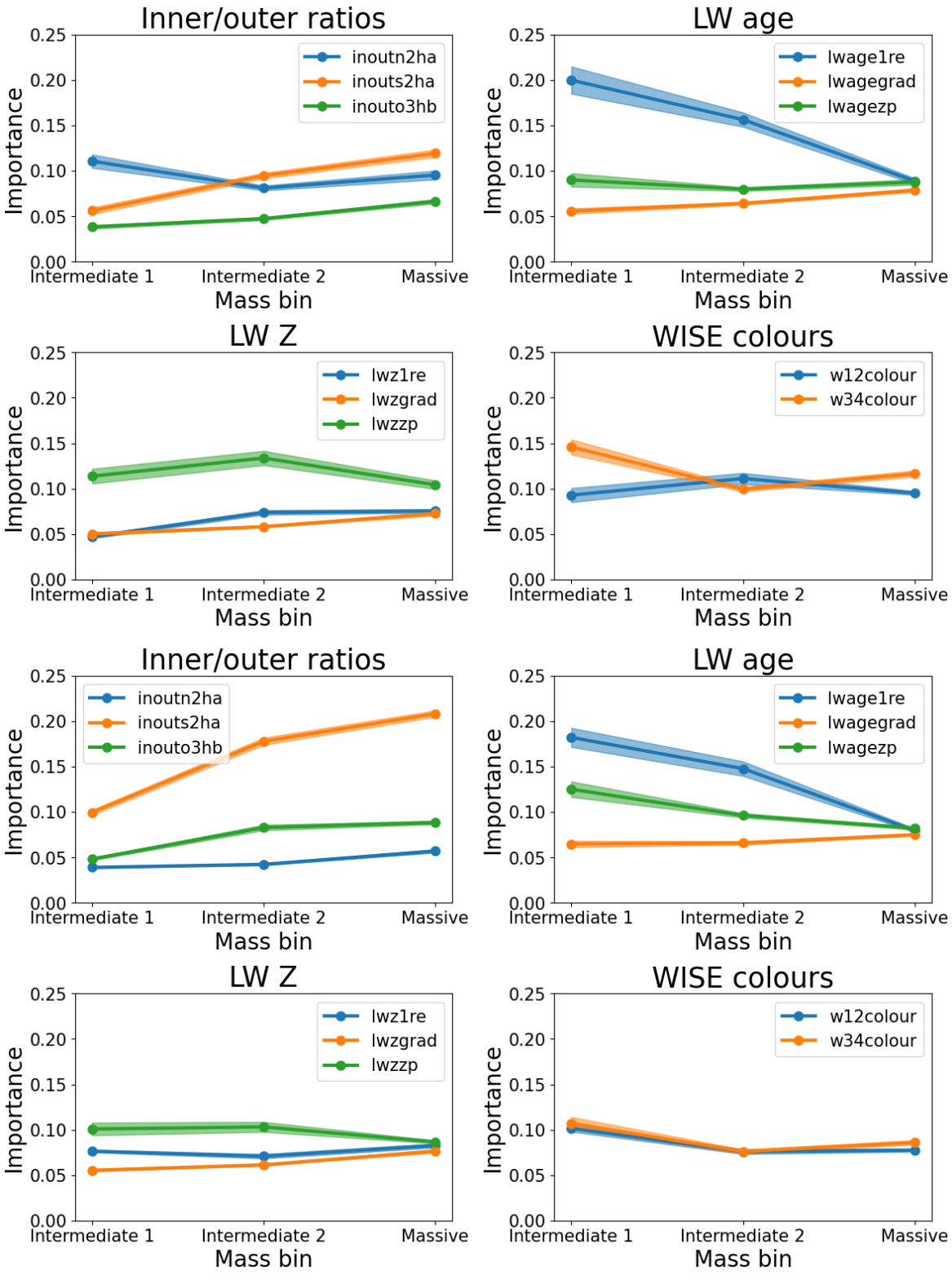

Figure 4.1: Feature importance evolution from CV for [N II] BPT (top 4 figures) and [S II] BPT (bottom 4 figures) galaxies as a function of training mass, inner features only.

4.4.2 Feature importance

As mentioned briefly in last section, outer features are poor choices for identifying AGN, which is supported by the feature importance when using only outer features – all outer features have the same weight within error. This trend holds for all AGN types and mass bins.

Inner features are more diverse in their importance. Figure 4.1 shows the feature importance of inner features and how they evolve with model mass. [N II] BPT features are shown on top and [S II] BPT are shown on the bottom. Common for both AGN selections is the importance of the light-weighted average stellar age within 1 effective radius. The light-weighted zero point of the linear stellar age fit is also of some importance, but it is a similar measure as the age within 1 effective radius and therefore also important. However, the importance decreases towards higher training masses and is effectively weighted equally with the rest of inner features for massive galaxies.

For [N II] BPT, the light-weighted zero point of the linear metallicity fit and the WISE colours are weighed higher than other features for all mass bins and are fairly constant in their importance. This trend is only weak, if not non-existent for [S II] BPT. Instead for [S II] BPT, the [S II]/Hα ratio between the inner and outer regions increase in importance and dominate at massive galaxies. All other features equalise towards more massive galaxies, which happens to [N II] BPT but to a smaller degree. Interestingly, the [S II]/Hα ratio ends up being weighed the highest for the massive [N II] BPT model.

For WISE models, the W1-W2 colour dominates, which is not surprising since the selection criteria is based on this colour. If anything, it shows that the RF algorithm correctly identifies obvious and important features and correlations. XMM models have a flat feature importance evolution with the exception of W1-W2 colour, but the low scores for XMM models suggest that the features, no matter how they are weighed, are poor predictors of XMM AGN.

4.4.3 Dwarf predictions - recall rate and precision

Although the RF predictions assume that the labels of the dwarf galaxies are unknown, preliminary classifications exist from the AGN identification used for the rest of the galaxies. These classifications can then be used to calculate the recall rate, i.e how many of the dwarf

Table 4.5: Number of dwarf galaxies in different categories according to pre-existing labels and predicted label. The numbers are the average and standard deviation of the predicted numbers obtained from the 10 CV estimators.

Training mass	Label	True +	True -	False +	False -
Intermediate 1	[N II] BPT	42.9 ± 3.7	1 037.4 ± 2.7	14.6 ± 2.7	54.1 ± 3.7
	[S II] BPT	105.2 ± 4.2	936.1 ± 1.3	34.9 ± 1.3	72.8 ± 4.2
	WISE	3.3 ± 0.5	1 145.0 ± 0.0	0.0 ± 0.0	0.7 ± 0.5
	XMM	0.0 ± 0.0	1 137.0 ± 0.0	0.0 ± 0.0	12.0 ± 0.0
Intermediate 2	[N II] BPT	43.4 ± 4.3	1 028.2 ± 4.9	23.8 ± 4.9	53.6 ± 4.3
	[S II] BPT	95.4 ± 7.2	894 ± 13.4	76.1 ± 13.4	82.6 ± 7.2
	WISE	4.0 ± 0.0	1 145.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
	XMM	0.1 ± 0.3	1 136.3 ± 0.5	0.7 ± 0.5	11.9 ± 0.3
Massive	[N II] BPT	58.2 ± 6.4	922.0 ± 29.1	130.0 ± 29.1	38.8 ± 6.4
	[S II] BPT	114.4 ± 7.9	703.8 ± 23.8	267.2 ± 23.8	63.6 ± 7.9
	WISE	2.9 ± 0.3	1 145.0 ± 0.0	0.0 ± 0.0	1.1 ± 0.3
	XMM	0.0 ± 0.0	1 135.8 ± 0.8	1.2 ± 0.8	12.0 ± 0.0

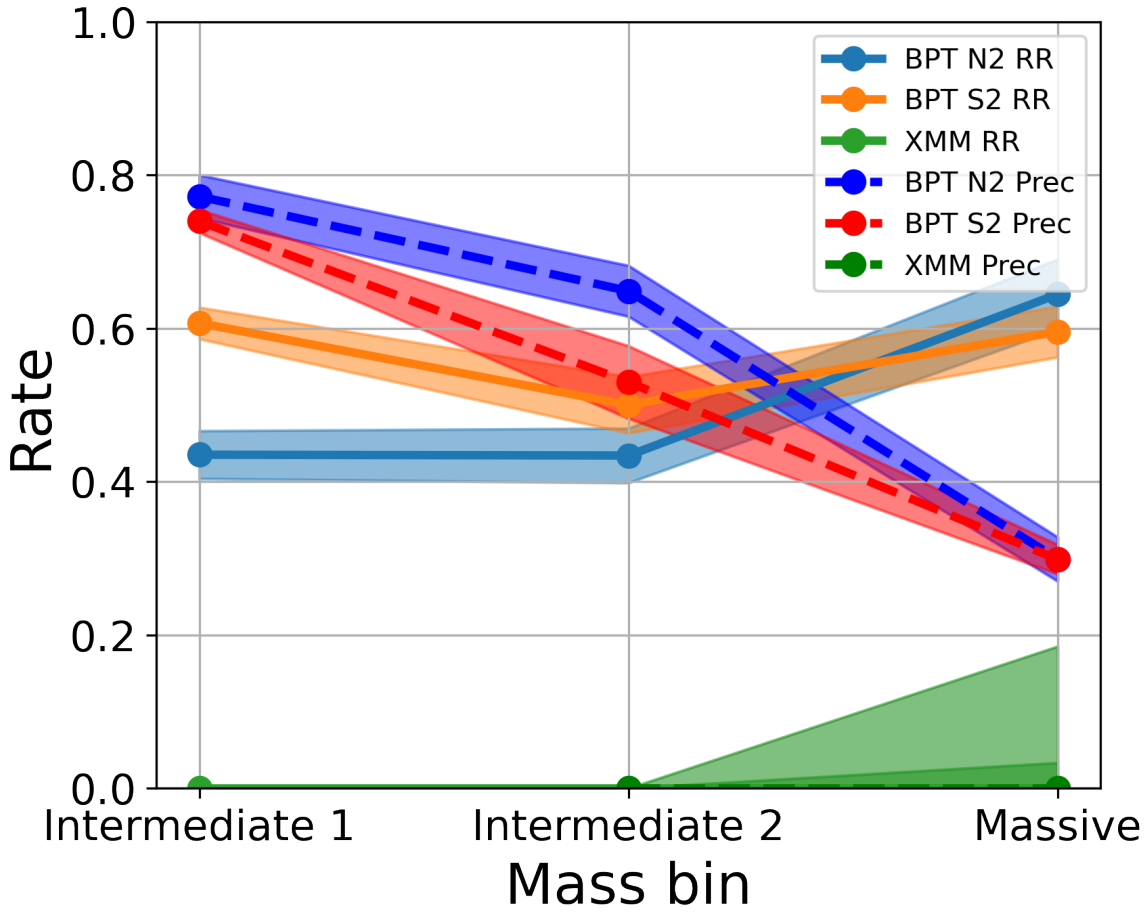

Figure 4.2: Recall rate and precision for dwarf galaxies for different mass bins. Errors are the standard deviation from the sub-classifiers of the CV testing.

galaxies with a preliminary AGN classification also receive an AGN label by the RF classifier, and the precision, i.e how many of the predicted AGN have preliminary AGN classification from MaNGA.

From the CV testing with a K-folding of 10, 10 estimators are constructed, and these are then used to predict the label of the dwarf galaxies. An average and standard deviation for are then found for each model with Table 4.5 showing the absolute numbers for all models and Figure 4.2 showing how the recall rate and precision evolves for different selection methods with mass bin.

XMM and WISE selection criteria yields very few dwarf AGN (XMM: 12, WISE: 4), so the RF models based on these as training sets are bound to fail or produce poor results. As such, these are not treated in further detail, but the recall rate and precision of XMM models are discussed briefly. The following discussions mostly focuses on [N II] BPT and [S II] BPT models.

The [N II] BPT recall rate is lower than [S II] BPT but increases towards higher mass bins. For [S II] BPT, the recall rate is similar for the lightest and most massive bin with a small dip in the intermediate 2 mass bin. In the lightest bin, the recall rate is around 15 percentage points higher than [N II] BPT , but they both evolve towards a recall rate of 60 per cent using massive galaxies as a training set.

The precision of both selection techniques follow the same with a high precision around 75 per cent but decrease towards higher training masses and reach 30 percent for massive galaxies. The [N II] BPT model is more precise than [S II] BPT for the intermediate 2 model, but decrease to the same value for the massive model.

The increase in recall rate towards higher mass models is correlated with the decrease in precision. Massive models predict higher numbers of AGN where intermediate 1 predict around 57 ± 6 [N II] BPT and 140 ± 6 [S II] BPT, massive models 188 ± 35 [N II] BPT and 381 ± 32 [S II] BPT. As such, the recall rate and precision indicate that the features are less fine tuned using massive galaxies and more lenient on which galaxies the model considers to be AGN. Conversely, it could be interpreted to mean that usual diagnostics are too restrictive for dwarf galaxies and that less restrictive selection criteria uncovers a large population of

hidden dwarf AGN.

4.4.4 Characterisation of dwarf galaxy distributions

This section provides a overview of average properties of the four different dwarf galaxy classifications from the RF predictions. It focuses on BPT galaxies since the number of WISE and XMM AGN are too low to properly characterise. Figure 4.3 shows the characterisation matrix for [N II] BPT galaxies using intermediate 2 as training set.

A recurring problem with BPT classification is the low SNR on [O III] and Hβ. While the selection criteria imposes a SNR of at least 3 on the inner region, no such requirement is in place for the outer regions. As such, the ratio of [O III]/Hβ between inner and outer regions may include using emission from the outer regions with a low SNR. This does indeed happen: populations of negative BPT predictions (i.e non-AGN according to RF), both true and false ones (i.e are in agreement and disagreement with MaNGA, respectively), have between 25-49 per cent of their population with low SNR on this ratio. This is the case for all models regardless of which mass is used for training. In half of the models, one of either the true positive or true negative populations contain only galaxies with a high SNR in both regions, which suggests that strong [O III] and Hβ signals from the whole galaxy is important in order for RF to agree with observations.

As described in Section 4.4.2, light-weighted stellar age within 1 effective radius ranks highly in the deciding the class for a galaxy. Indeed, the average age for positive (both true and false predictions) are higher than negative predictions, and positive populations are more confined than negative ones. This means that young stellar populations are anti-correlated with AGN activity, which is in line with the M_{BH}-M_{Bulge} relation – bulges are usually redder and have older stellar populations, and they are more often found with AGN Martin et al. (e.g 2018). A similar trend exists for metallicity, although weaker: positive predictions have on average a higher metallicity than non-AGN galaxies.

Despite environmental features ranking very low on feature importance, there are differences between positive and negative populations for both [N II] and [S II] BPT when considering only massive galaxy neighbours. Positive dwarf galaxies have on average smaller

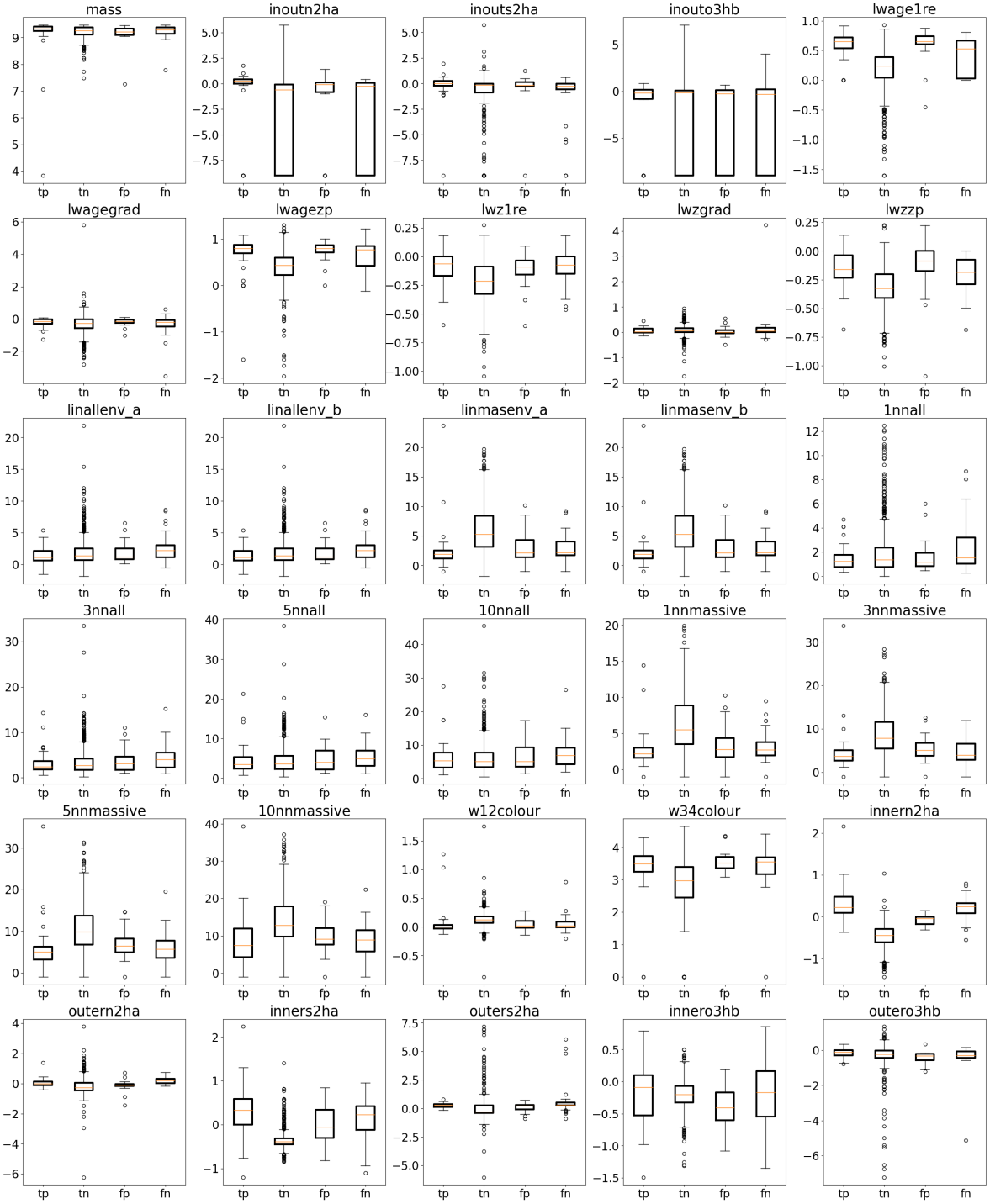

Figure 4.3: Characterisation matrix of dwarf galaxies using intermediate 2 galaxies as training set. Each plot has four box plots, one for each of the different true/false positive/negative populations.

distances to nearby neighbours (both 1st, 3rd, 5th and 10th nearest neighbour), and a smaller gradient and zero point to the linear fits, although this trend is weaker when going towards higher *N*'th neighbour. This suggests that dwarf galaxies with AGN are generally in denser environments and/or are more often satellite galaxies. Using all galaxies for environmental estimations shows no difference between populations.

4.5 Discussion

RF identifies fewer candidates using the lowest mass training set than diagnostic diagrams (around 40 per cent fewer [N II] BPT and around 25 per cent fewer [S II] BPT). Using intermediate 2 as the training mass, around the same number of [S II] BPT AGN are found, but around 45 per cent of them are not classified as AGN in MaNGA, while the [N II] BPT numbers are the same as using the intermediate 1 mass for RF training. Using massive galaxies, the number of AGN for both [N II] and [S II] selection is higher by almost 100 per cent with two thirds of them not being classified as AGN in MaNGA.

As such, lower mass training sets can be used to find new AGN candidates that usual BPT diagnostics does not find, but these models alone yield fewer candidates and with less robust classification since the used parameters are not directly related to the AGN emission. High mass training sets provide a large number of new candidates, which may turn out to be a good venue to improve AGN selection in dwarf galaxies. However, this requires follow-up analysis on the new candidates in order to verify their actual status as AGN, especially since this is a novel classification method that requires more fine-tuning.

One way to fine tune the RF is to optimise the features used. For example, the distance to the 3rd and 5th nearest neighbour both represent the same thing albeit slightly different, but not different enough to justify using both parameters. Ideally, the features used should be as orthogonal to each other as possible. MaNGA provides more data than just emission lines – e.g stellar velocities and dynamics (e.g Penny et al., 2018), outflows (Wylezalek et al., 2020; Avery et al., 2021), and star formation histories (Zhou et al., 2020). This study only explored a subset of these parameters, and three combinations of features (inner, outer, and all) were used. As such, MaNGA and IFUs are great options for RF and fine tuning of dwarf AGN

selection.

Another problem with using RF to identify AGN is apparent judging from the low scores from the CV. Higher scores are found in models with a more equal class balance between AGN and non-AGN, which can be caused by overfitting of parameters. However, massive galaxies are the ones with the highest test scores and in those models, the importance of internal features that are more directly related with the classification criteria (i.e ratio of the log [N II]/Hα and [S II]/Hα between inner and outer regions) increase and dominate. The dependence and effect of class balance can be tested by constructing training sets of each mass bin with the same class balance.

4.6 Future work

The results and discussion in this study represents a preliminary methodology into using RF to improve and classify AGN in dwarf galaxies, and several improvements and changes for future modelling are already identified. This section goes through several of these suggestions, and some of them have already been mentioned in Section 4.5.

The parameter/feature space used for training and identifying AGN can be fine tuned. Already established in this study, environmental features are poor estimators for galaxy type and when feature optimising, they can reasonably be reduced or even removed. Similarly, the zero point of the linear fits of metallicity and stellar age from Firefly are to some degree degenerate with those obtained from within 1.5 and 1 effective radii. Kinematically offset cores have also been found in MaNGA and are associated with AGN activity Penny et al. (2018), which supports the idea of including kinematic parameters as features in AGN models.

Besides including additional properties and features, multi-wavelength data can also used used differently. For example, WISE colours have an above average weight in both BPT selection schemes proving their relevance in a RF approach. The selection of colours can be expanded to include UV and optical colours as well. Additionally, X-ray is a strong indicator of AGN activity, even in dwarf galaxies (Baldassare et al., 2017), but X-ray features are not utilised in the RF feature scheme. In part, this is due to incompleteness in the overlap with the MaNGA data set. Nevertheless, improved inclusion of multi-wavelength data sets is expected

to boost the quality of the RF models.

Proper assessment of the different models can also be boosted by more detailed characterisation of the subdivisions of the predicted labels of the dwarf galaxies. For example, what sets the false positives (i.e the AGN that are predicted to be AGN by RF but not emission lines) apart from true positives? Qualitative assessments of this sub-population will provide important clues as to where they lie on regular diagnostic diagrams and whether they hide an AGN – and if so, adjust the diagnostic criteria accordingly for the low mass regime.

Lastly, a single approach has been used for determining emission line strengths: using the inner 20 per cent of spaxels of MaNGA data. However, this does not take into account orientation or angular size of each galaxy, and the effect of changing the way inner and outer regions are defined is worth examining. Further to this, other studies point towards off-nuclear AGN emission in dwarf galaxies using MaNGA (Mezcua & Domínguez Sánchez, 2020), so there are other spatial effects to be aware of.

4.7 Conclusions

This preliminary study into the use of RF to identify dwarf galaxies has shown promise in the approach. While further fine tuning of the model and assessment of the predictions will boost the validity of the method, several key findings can already be presented now. These findings can be summarised as follows:

- Environmental features are poor predictors of AGN activity, even when used in combination with internal features. However, dwarf [N II] and [S II] BPT AGN that are identified as AGN in both RF and MaNGA data are on average closer to a massive galaxy than non-AGN.

- The more massive galaxies used for training, the better model, if going by F1 score alone. Going by score alone, though, washes out nuances such as class balance and feature importance. More massive galaxy models rely on fewer parameters that are more directly related to AGN activity. Specifically, [S II] BPT ranks the [S II]/Hα ratio between inner and outer regions the highest – a feature that [N II] BPT also rank

highly but with a more flat feature importance distribution in general (i.e close to equal weighing of all features)

- Average stellar age within 1 effective radius is weighed the highest using lower mass galaxies as training sets but decreases in importance towards higher masses. For [N II] BPT, metallicity, infrared colours, and [N II]/Hα inner/outer ratio are also ranked highly, while for [S II] BPT, metallicity is less important and feature importance is ultimately dominated by [S II]/Hα inner/outer ratio.

- Predictions of dwarf galaxy classification from lower mass training sets yields results that most accurately resemble the observations, but the AGN occupation fraction is lower. As such, this training set alone is a poor choice for classifying AGN but can be used to find otherwise hidden AGN

- Conversely, higher mass training sets yield a high number of AGN that observations do not predict and generally disagrees the most with observations out of all models – despite their good CV scores. The question is then whether the predicted dwarf AGN actually are AGN.

- Since this study has been a preliminary approach, there are several improvements to the method that can be implemented for future work. Highest on the list is improving feature optimisation and inclusion. This includes reducing the number of features that are degenerate (e.g certain environmental parameters) while also including more relevant features (such as multi-wavelength data and colours).

- Further exploration of the different true/false postive/negative subcategories of dwarf AGN are similarly a venue for improving the validity of the method, but it can also be used to adjust regular AGN selection methods such as the BPT diagram to better fit the dwarf mass regime.

5. Conclusions

5.1 Discussion

The seeming disagreement between the impact of environment between observations (Kristensen et al., 2020) and simulations (Kristensen et al., 2021) on AGN in dwarf galaxies warrants further investigation. Part of the disagreement stems from numerical reasons rather than physical ones, and further work in properly implementing dwarf galaxies and BH growth in simulations – work which is already being carried out in the form of projects and simulation suites such as FABLE and Horizon-AGN (Koudmani et al., 2021). Mergers, however, seem to consistently be coupled to increased AGN activity in both simulations and observations. Using cosmological simulations, the full cosmological histories of present day AGN galaxies can be obtained, and indeed, Kristensen et al. (2021) found that recent mergers ($\leq$ 4 Gyr) are associated with increased AGN activity. This is in line with some observations (Ellison et al., 2019), but other studies find that dwarf galaxies with AGN characteristics do not show an excess in merger fraction (Kaviraj et al., 2019). While these studies seem to be at odds with each other, it is possible to reconcile them by acknowledging that they use different merger measures, samples, and statistical treatments. Furthermore, the triggering process may not lead to immediate AGN activity (Hopkins, 2012) which makes an environment-AGN connection even more elusive.

This illuminates another obstacle in the question of whether environment can trigger AGN activity. Environment can be quantified in many ways (Muldrew et al., 2012), and the lack of a standard measure means that both results finding a connection and results finding none are published. The threshold for when a finding is significant also varies from study to study. Some simulation studies require very few minimum particles to constitute a structure($\sim$ 10, Fattahi et al. (2020), $\sim$ 50 (Martin et al., 2020a)), while e.g Kristensen et al. (2021) requires an order of magnitude more. Further to this translation problem between studies is the treatment of AGN and black holes. Two prominent simulations that focus on and include up-to-date BH

modelling, IllustrisTNG and Horizon-AGN, treat accretion and black hole seeding differently, making results even harder to compare

Furthermore, BH models in simulations are not calibrated to dwarf galaxies, and findings from these are not necessarily transferable to observations. For example in IllustrisTNG, the black hole seed masses are overmassive compared to the M_{BH}-σ_{bulge} relation (Xiao et al., 2011; Baldassare et al., 2020) by around an order of magnitude, which results in dwarf galaxies having more massive BH than what is inferred from observations. This leads to higher accretion rates and thus stronger AGN, but the BH ultimately grow over time so that they follow the relation for massive galaxies. Alternatively, while using a lower seed, the accretion rate can be artificially boosted by a factor of 10 (Illustris, and others), which will similarly yield the BHs following established M_{BH} relations. However, using non-physical parameters to make the model fit is not a satisfying solution either. Even ignoring the problems of the BH model in relation to dwarf galaxies, AGN identification in simulations rely on different parameters than observations – some rely on the Eddington ratio, while others employ semi-analytical models using the local gas parameters and energy output from the BH.

Proper identification of dwarf AGN in observations is not without difficulties either. Some effort has been put towards identifying AGN in dwarf galaxies that are missed by usual measures. For example, Birchall et al. (2020) used X-ray data and found 61 dwarf AGN of which 85 per cent were not identified by the more commonly used BPT diagram. Mid-IR have also proven a fruitful venue for detecting AGN missed by other methods (Hainline et al., 2016; Lupi et al., 2020), so a multi-wavelength approach seems to be the best to recover as many dwarf AGN as possible. However, as noted by e.g Lupi et al. (2020), AGN identified by different wavelengths vary in properties and do not necessarily represent the same class. As such, grouping together all AGN identified by different selection methods may yield a diverse set of galaxies making calibration adjustments to e.g the [N II]/Hα and [O III]/Hβ BPT diagram questionable (as proposed by e.g Cann et al., 2019).

A promising and currently underutilised channel for improving dwarf AGN identification is machine learning. Chapter 4 shows that even a simple implementation of a random

forest classifier is worth pursuing – both in regards to uncovering hidden AGN but also to adjust existing selection methods to the low mass regime. However, the early stages of this endeavour still require fine tuning. Feature selection and model optimisation is needed in order to prevent overfitting and increase accuracy. One finding already is that environmental parameters constitute poor features for AGN classification, although galaxies with both affirmative observational and RF AGN are in denser environments than their non-AGN counterparts. This seeming contradiction could be due to using degenerate features of environment instead of a select few, supporting the notion of improved feature selection. A big improvement would be multi-wavelength features, but the coverage of e.g infrared and X-ray is limited compared to optical.

5.2 Summary

The effect of environment on AGN activity in dwarf galaxies remains an open question, but it depends on what exactly is being asked and which AGN diagnostic is being used. Using the NASA-Sloan Atlas, a sample of low redshift ($z \leq 0.055$) dwarf galaxies are found, [N II] BPT and WHAN are used as AGN criteria, a three dimensional distance measure to the 10th nearest neighbour constitute the environmental estimator, and applying a strict Kolmogorov-Smirnov statistical testing method, a binary result is found showing that there is no connection between environment (both immediate and large scale) and AGN activity. However, a redshift-limited sample to correct for SDSS fiber coverage bias does reach a statistical significance within error, making it difficult to conclusively rule out an environmental connection.

The limitations of this approach is that it naively uses commonly used AGN diagnostics, which will provide an incomplete set of dwarf AGN, it uses a simple environmental measure providing only a snapshot in time of their evolution, which does not take their histories and earlier interactions into account, and integrated flux of the whole galaxy, which enables star formation signatures to drown out AGN signatures.

One way to overcome several of these biases is to use a data set in which the full extent of properties of the galaxies are known. This can be achieved by turning towards simulations. Using the IllustrisTNG simulation suite (TNG100 and TNG50 more specifically)

and constructing two environmental measures: distance to the 10th nearest neighbour and time since last merger with three different merger mass ratios, environment is found to have a significant effect under certain circumstances using a KS-testing suite. AGN galaxies with similar properties as non-AGN in terms of stellar mass and colour are found to prefer denser environments, but a sample of non-AGN constructed to match AGN galaxies show no difference in environment to the subject AGN sample. Recent minor mergers are found to unequivocally be associated with an increase in intermediate strength AGN activity.

However, the testing suite is different compared to the one ran on the NSA sample (constructing subject and reference samples, and then reversing their roles is only done for the simulation study) as is the AGN selection criteria (emission lines for NSA, Eddington ratio for IllustrisTNG), and as such, a one-to-one comparison between the results is misleading. Furthermore, working with dwarf galaxies in the IllustrisTNG simulation suite is working on the edge of its resolution limits. This has the potential to introduce numerical biases that are not physical in nature and will skew the findings. For example, a subpopulation of dwarf galaxies are found to be very red and reside in dense environments, but this population exists only in TNG simulations and not Illustris-1, the precurser to TNG.

A venue to improve dwarf AGN selection is to consider a wide range of properties (instead of just two emission line ratios) of regular AGN and label dwarf galaxies that exhibit similar behaviour. These properties can be both local (e.g only core region), global (i.e across the whole galaxy), or external (such as neighbourhood density). The spatially resolved spectrospic MaNGA data in combination with matched mid-IR and X-ray data from AllWISE and XMM provide a data set of internal features for 10 104 galaxies with environmental estimations from an updated NSA catalogue, and using a simple random forest model trained on more massive galaxies has shown promise in this approach.

The preliminary findings point towards environmental features as poor predictors of AGN activity, but dwarf AGN identified by both RF and observations generally prefer denser environments. More massive galaxies constitute the best training sets as far as scores go, but simultaneously disagree the most with observations when it comes to dwarf galaxies, although they provide the highest number of dwarf AGN candidates.

Ultimately, there are a number of improvements that should be considered for further studies. A wider range of internal parameters is desirable, such as multi-wavelength data, but availability and coverage is limiting this. Degenerate features need to be considered more carefully and avoided, and iterative models with feature elimination is one approach to limit this. Even using different ML models can be an improvement such as using ensemble methods rather than a simple random forest. Lastly, the predictions from ML need to be double checked and verified, which can then be used to fine tune exisiting AGN diagnostics.

5.3 Future work

The two main goals of this research has been to quantify the effect of environment and to highlight and improve AGN selection in dwarf galaxies, but the research has certain limitations. This section will outline future work that can be adopted to improve the findings and conclusions presented in this study – both short term and long term.

One problem with dwarf galaxies with AGN is the fact that different selection methods yield samples with little overlap (Hainline et al., 2016; Baldassare et al., 2017; Mackay Dickey et al., 2019; Birchall et al., 2020), which suggests that each of these populations are heavily biased. For example, mid-IR selection are often associated with obscured AGN while X-ray and optically selected AGN are unobscured, and limiting environmental analyses to only one population will bias the findings. While this problem also exists for massive galaxies, it seems more pronounced for dwarf galaxies and may be due to the fact that selection of AGN in dwarf galaxies is not fine tuned to this mass regime.

This bias can be overcome by constructing a more complete set of AGN in dwarf galaxies, for example by using more sensitive instruments and surveys such as the X-ray eRosita (Latimer et al., 2021), the mid-IR James Webb Space Telescope (Richardson et al., 2022), and optical MaNGA (Comerford et al., 2022). However, even employing multi-wavelength data in identifying dwarf AGN is troublesome. Cann et al. (2020) remark that the intrinsic low X-ray luminosity and emission line ratios of low-metallicity and low-mass galaxies like J1056+3138 are challenging for commonly employed diagnostics.

The question then becomes how to adjusting the commonly employed diagnostics to dwarf

galaxies. Chapter 4 proposes using a RF approach by using several internal features (such as emission, kinematics, and spatially resolved properties), training a classifier on massive galaxies on whom the common diagnostics are more fine tuned, and ultimately predicting the labels of dwarf galaxies. While this has shown that it is feasible, optimisation of this model will greatly boost the confidence in the findings. Optimisation covers subjects such as feature selection (e.g removing degenerate and irrelevant features), finding the best ML algorithm (i.e using ensemble methods rather than a simple random forest method), and test the validity of the predictions.

Another complimentary approach to find dwarf AGN is to look for signatures that are not related to the ionisation fingerprints of the nuclear emission. The upcoming 10 year Legacy Survey of Space and Time provides an interesting venue of AGN identification, namely variability, which has good prospects (Baldassare et al., 2018, 2020; Ward et al., 2021) since it finds a unique set of AGNs, but Baldassare et al. (2018) comment that perhaps low mass AGN are less variable than their more massive counterpart, thus limiting the usefulness of this method. Another interesting research venue, mostly related to dwarf galaxies, is the findings of off-nuclear AGN emission (Reines et al., 2020; Mezcua & Domínguez Sánchez, 2020; Ricarte et al., 2021), which is being explored in both observations and simulations.

In more and more simulations, dwarf galaxies have become a central aspect (Koudmani et al., 2019; Koudmani et al., 2021; Jahn et al., 2022; Fattahi et al., 2020; Sharma et al., 2020). This effort makes for a better understanding of the observable signatures in dwarf galaxies and how they might be different from massive galaxies. Not only is the resolution and dynamics being improved upon, the BH modelling is also under scrutiny. Most accretion models use a Bondi-Hoyle accretion rate, which is assuming spherical accretion only dependant on gas density, black hole mass, and local sound speed, but this may not be the most suitable model for dwarf galaxies since they either require a overmassive black hole seed (e.g IllustrisTNG) or artificially boosted accretion rate (e.g FABLE, Koudmani et al., 2021) to match observations.

Bibliography

Aihara H., et al., 2011, ApJS, 193, 29

Aird J., et al., 2012, ApJ, 746, 90

Aird J., Coil A. L., Georgakakis A., 2017, MNRAS, 465, 3390

Alexander D. M., Hickox R. C., 2012, New Astronomy Reviews, 56, 93

Alonso S., Coldwell G., Duplancic F., Mesa V., Lambas D. G., 2018, A&A, 618, A149

Amiri A., Tavasoli S., De Zotti G., 2019, ApJ, 874, 140

Anglés-Alcázar D., Faucher-Giguère C.-A., Quataert E., Hopkins P. F., Feldmann R., Torrey P., Wetzel A., Kereš D., 2017, MNRAS, 472, L109

Arsenault R., 1989, A&A, 217, 66

Avery C. R., et al., 2021, MNRAS, 503, 5134

Azadi M., et al., 2017, ApJ, 835, 27

Baade W., Hubble E., 1939, PASP, 51, 40

Baldassare V. F., Reines A. E., Gallo E., Greene J. E., 2015, ApJL, 809, L14

Baldassare V. F., Reines A. E., Gallo E., Greene J. E., 2017, ApJ, 836, 20

Baldassare V. F., Geha M., Greene J., 2018, ApJ, 868, 152

Baldassare V. F., Dickey C., Geha M., Reines A. E., 2020, ApJL, 898, L3

Baldry I. K., Glazebrook K., Brinkmann J., Ivezić Ž., Lupton R. H., Nichol R. C., Szalay A. S., 2004a, ApJ, 600, 681

Baldry I. K., Balogh M. L., Bower R., Glazebrook K., Nichol R. C., 2004b, in Allen R. E., Nanopoulos D. V., Pope C. N., eds, American Institute of Physics Conference Series Vol. 743, The New Cosmology: Conference on Strings and Cosmology. pp 106–119 (arXiv:astro-ph/0410603), doi:10.1063/1.1848322

Baldry I. K., Balogh M. L., Bower R. G., Glazebrook K., Nichol R. C., Bamford S. P., Budavari T., 2006, MNRAS, 373, 469

Baldwin J. A., Phillips M. M., Terlevich R., 1981, PASP, 93, 5

Balogh M. L., Schade D., Morris S. L., Yee H. K. C., Carlberg R. G., Ellingson E., 1998, ApJL, 504, L75

Barai P., de Gouveia Dal Pino E. M., 2019, MNRAS, 487, 5549

Barth A. J., Ho L. C., Rutledge R. E., Sargent W. L. W., 2004, ApJ, 607, 90

Baum W. A., Hiltner W. A., Johnson H. L., Sandage A. R., 1959, ApJ, 130, 749

Belfiore F., et al., 2019, AJ, 158, 160

Bellovary J. M., Cleary C. E., Munshi F., Tremmel M., Christensen C. R., Brooks A., Quinn T. R., 2019, MNRAS, 482, 2913

Bhowmick A. K., Blecha L., Thomas J., 2020, ApJ, 904, 150

Bilton L. E., Pimbblet K. A., 2018, MNRAS, 481, 1507

Birchall K. L., Watson M. G., Aird J., 2020, MNRAS, 492, 2268

Blanton M. R., Kazin E., Muna D., Weaver B. A., Price-Whelan A., 2011, AJ, 142, 31

Blanton M. R., et al., 2017, AJ, 154, 28

Blecha L., Snyder G. F., Satyapal S., Ellison S. L., 2018, MNRAS, 478, 3056

Bower R. G., Benson A. J., Malbon R., Helly J. C., Frenk C. S., Baugh C. M., Cole S., Lacey C. G., 2006, MNRAS, 370, 645

Bradford J. D., Geha M. C., Greene J. E., Reines A. E., Dickey C. M., 2018, ApJ, 861, 50

Bundy K., et al., 2015, ApJ, 798, 7

Cann J. M., Satyapal S., Abel N. P., Blecha L., Mushotzky R. F., Reynolds C. S., Secrest N. J., 2019, ApJL, 870, L2

Cann J. M., et al., 2020, ApJ, 895, 147

Cerny W., et al., 2021, ApJ, 910, 18

Cheung E., et al., 2015, MNRAS, 447, 506

Chiappini C., Matteucci F., Gratton R., 1997, ApJ, 477, 765

Chown R., et al., 2019, MNRAS, 484, 5192

Cid Fernandes R., Stasinska G., Schlickmann M. S., Mateus A., Asari N. V., Schoenell W., Sodre L. J., (the SEAGal collaboration) 2010, MNRAS, 403, 1036

Cid Fernandes R., Stasińska G., Mateus A., Vale Asari N., 2011, MNRAS, 413, 1687

Comerford J. M., Negus J., Barrows R. S., Wylezalek D., Greene J. E., Müller-Sánchez F., Nevin R., 2022, ApJ, 927, 23

Croton D. J., et al., 2006, MNRAS, 365, 11

Curtis H. D., 1915, PASP, 27, 214

Dashyan G., Silk J., Mamon G. A., Dubois Y., Hartwig T., 2018, MNRAS, 473, 5698

Davidzon, I. et al., 2016, A&A, 586, A23

Dekel A., Silk J., 1986, ApJ, 303, 39

Deng X.-F., Wu P., Qian X.-X., Luo C.-H., 2012, PASJ, 64, 93

Di Matteo T., Springel V., Hernquist L., 2005, Nature, 433, 604

Di Teodoro E. M., et al., 2019, MNRAS, 483, 392

Dickey C. M., et al., 2021, ApJ, 915, 53

Donley J. L., et al., 2018, ApJ, 853, 63

Dubois Y., et al., 2021, A&A, 651, A109

Dunlop J. S., McLure R. J., Kukula M. J., Baum S. A., O'Dea C. P., Hughes D. H., 2003, MNRAS, 340, 1095

Eliche-Moral M. C., Rodríguez-Pérez C., Borlaff A., Querejeta M., Tapia T., 2018, A&A, 617, A113

Ellison S. L., Patton D. R., Simard L., McConnachie A. W., 2008, AJ, 135, 1877

Ellison S. L., Viswanathan A., Patton D. R., Bottrell C., McConnachie A. W., Gwyn S., Cuillandre J.-C., 2019, MNRAS, 487, 2491

Engler C., et al., 2021, MNRAS, 507, 4211

Fabian A. C., 2012, ARA&A, 50, 455

Fath E. A., 1909, Popular Astronomy, 17, 504

Fattahi A., Navarro J. F., Frenk C. S., 2020, MNRAS, 493, 2596

Ferguson H. C., Binggeli B., 1994, The Astronomy and Astrophysics Review, 6, 67

Ferrarese L., Merritt D., 2000, ApJL, 539, L9

Feruglio C., Maiolino R., Piconcelli E., Menci N., Aussel H., Lamastra A., Fiore F., 2010, A&A, 518, L155

Filippenko A. V., Ho L. C., 2003, ApJ, 588, L13

Filippenko A. V., Sargent W. L. W., 1985, ApJS, 57, 503

Filippenko A. V., Sargent W. L. W., 1989, ApJL, 342, L11

Fillingham S. P., Cooper M. C., Pace A. B., Boylan-Kolchin M., Bullock J. S., Garrison-Kimmel S., Wheeler C., 2016, MNRAS, 463, 1916

Fujita A., Martin C. L., Mac Low M.-M., Abel T., 2003, ApJ, 599, 50

Fumagalli M., Fotopoulou S., Thomson L., 2020, MNRAS, 498, 1951

Gallagher John S. I., Wyse R. F. G., 1994, PASP, 106, 1225

Galloway M. A., et al., 2015, MNRAS, 448, 3442

Garnett R., Ho S., Bird S., Schneider J., 2017, MNRAS, 472, 1850

Garrison-Kimmel S., Rocha M., Boylan-Kolchin M., Bullock J. S., Lally J., 2013, MNRAS, 433, 3539

Gebhardt K., et al., 2000, ApJL, 539, L13

Geha M., Blanton M. R., Yan R., Tinker J. L., 2012, ApJ, 757, 85

Genel S., et al., 2014, MNRAS, 445, 175–200

Goddard D., et al., 2017, MNRAS, 466, 4731

Gordon Y. A., et al., 2018, MNRAS, 475, 4223

Goulding A. D., et al., 2017, Publications of the Astronomical Society of Japan, 70

Grebel E. K., 1999, Symposium - International Astronomical Union, 192, 17–38

Greene J. E., Ho L. C., 2004, ApJ, 610, 722

Greene J. E., Ho L. C., 2007, ApJ, 670, 92

Greene J. E., Barth A. J., Ho L. C., 2006, New Astronomy Reviews, 50, 739

Gültekin K., et al., 2009, ApJ, 698, 198

Gunn J. E., Gott J. Richard I., 1972, ApJ, 176, 1

Guo Z., Martini P., 2019, ApJ, 879, 72

Haas M. R., Schaye J., Jeeson-Daniel A., 2012, MNRAS, 419, 2133

Habouzit M., Volonteri M., Dubois Y., 2017, MNRAS, 468, 3935

Habouzit M., et al., 2021, MNRAS, 503, 1940

Haines C. P., Gargiulo A., La Barbera F., Mercurio A., Merluzzi P., Busarello G., 2007, MNRAS, 381, 7

Hainline K. N., Reines A. E., Greene J. E., Stern D., 2016, ApJ, 832, 119

Hernquist L., 1989, Nature, 340, 687

Ho T. K., 1995, in Proceedings of 3rd International Conference on Document Analysis and Recognition. pp 278–282 vol.1, doi:10.1109/ICDAR.1995.598994

Ho L. C., Filippenko A. V., Sargent W. L. W., Peng C. Y., 1997, ApJS, 112, 391

Hodge P. W., 1971, ARA&A, 9, 35

Holmberg E., 1941, ApJ, 94, 385

Holmberg E., 1958, Meddelanden fran Lunds Astronomiska Observatorium Serie II, 136, 1

Hopkins P. F., 2012, MNRAS, 420, L8

Hopkins P. F., Hernquist L., Cox T. J., Di Matteo T., Robertson B., Springel V., 2006, ApJS, 163, 1

Hubble E. P., 1926, ApJ, 64, 321

Jackson F. E., Roberts T. P., Alexander D. M., Gelbord J. M., Goulding A. D., Ward M. J., Wardlow J. L., Watson M. G., 2012, MNRAS, 422, 2

Jahn E. D., Sales L. V., Wetzel A., Samuel J., El-Badry K., Boylan-Kolchin M., Bullock J. S., 2022, MNRAS,

Janesick J. R., 2001, Scientific charge-coupled devices

Jarrett T. H., et al., 2011, ApJ, 735, 112

Ji Z., Giavalisco M., Kirkpatrick A., Kocevski D., Daddi E., Delvecchio I., Hatcher C., 2022, ApJ, 925, 74

Juneau S., et al., 2014, ApJ, 788, 88

Kauffmann G., et al., 2003, MNRAS, 341, 54

Kauffmann G., White S. D. M., Heckman T. M., Ménard B., Brinchmann J., Charlot S., Tremonti C., Brinkmann J., 2004, MNRAS, 353, 713

Kaviraj S., Martin G., Silk J., 2019, MNRAS, 489, L12

Kewley L. J., Heisler C. A., Dopita M. A., Lumsden S., 2001, The Astrophysical Journal Supplement Series, 132, 37

Kewley L. J., Groves B., Kauffmann G., Heckman T., 2006, MNRAS, 372, 961

Klypin A., Kravtsov A. V., Valenzuela O., Prada F., 1999, ApJ, 522, 82

Kocevski D. D., et al., 2012, ApJ, 744, 148

Kormendy J., Ho L. C., 2013, Annual Review of Astronomy and Astrophysics, 51, 511

Kormendy J., Fisher D. B., Cornell M. E., Bender R., 2009, ApJS, 182, 216

Koudmani S., Sijacki D., Bourne M. A., Smith M. C., 2019, MNRAS, 484, 2047

Koudmani S., Henden N. A., Sijacki D., 2021, MNRAS, 503, 3568

Kristensen M. T., Pimbblet K., Penny S., 2020, MNRAS, 496, 2577

Kristensen M. T., Pimbblet K. A., Gibson B. K., Penny S. J., Koudmani S., 2021, ApJ, 922, 127

Kruk S. J., et al., 2017, MNRAS, 469, 3363

Larson R. B., 1974, MNRAS, 169, 229

Larson R. B., Tinsley B. M., 1978, ApJ, 219, 46

Latimer C. J., Reines A. E., Bogdan A., Kraft R., 2021, ApJL, 922, L40

Law D. R., et al., 2015, AJ, 150, 19

Lehmer B. D., et al., 2016, ApJ, 825, 7

Lewis I., et al., 2002, MNRAS, 334, 673

Li Y., et al., 2020, ApJ, 895, 102

Lindblad P. O., 1960, Stockholms Observatoriums Annaler, 4, 4

Liu W., Veilleux S., Canalizo G., Rupke D. S. N., Manzano-King C. M., Bohn T., U V., 2020, ApJ, 905, 166

Lupi A., Sbarrato T., Carniani S., 2020, MNRAS, 492, 3255

Lupton R., Gunn J. E., Ivezić Z., Knapp G. R., Kent S., 2001, The SDSS Imaging Pipelines. p. 269

Mackay Dickey C., Geha M., Wetzel A., El-Badry K., 2019, ApJ, 884, 180

Man Z.-y., Peng Y.-j., Kong X., Guo K.-x., Zhang C.-p., Dou J., 2019, MNRAS, 488, 89

Manzano-King C. M., Canalizo G., 2020, MNRAS, 498, 4562

Manzano-King C. M., Canalizo G., Sales L. V., 2019, ApJ, 884, 54

Maraston C., et al., 2020, MNRAS, 496, 2962

Marconi A., Hunt L. K., 2003, ApJL, 589, L21

Marian V., et al., 2020, ApJ, 904, 79

Martin G., et al., 2018, MNRAS, 476, 2801

Martin G., et al., 2020a, MNRAS,

Martin G., Kaviraj S., Hocking A., Read S. C., Geach J. E., 2020b, MNRAS, 491, 1408

Martínez-Delgado D., et al., 2021, A&A, 652, A48

Mateo M. L., 1998, ARA&A, 36, 435

McAlpine S., Harrison C. M., Rosario D. J., Alexander D. M., Ellison S. L., Johansson P. H., Patton D. R., 2020, MNRAS, 494, 5713

McConnachie A. W., 2012, AJ, 144, 4

McCracken H. J., et al., 2012, A&A, 544, A156

Mendez A. J., et al., 2013, ApJ, 770, 40

Mezcua M., Domínguez Sánchez H., 2020, ApJL, 898, L30

Mezcua M., Civano F., Fabbiano G., Miyaji T., Marchesi S., 2016, ApJ, 817, 20

Micic M., Holley-Bockelmann K., Sigurdsson S., Abel T., 2007, MNRAS, 380, 1533

Miller C. J., Nichol R. C., Gomez P. L., Hopkins A. M., Bernardi M., 2003, ApJ, 597, 142

Monaco P., Giuricin G., Mardirossian F., Mezzetti M., 1994, ApJ, 436, 576

Moore B., Katz N., Lake G., Dressler A., Oemler A., 1996, Nature, 379, 613

Moore B., Lake G., Katz N., 1998, ApJ, 495, 139

Moore B., Quinn T., Governato F., Stadel J., Lake G., 1999, MNRAS, 310, 1147

Moran E. C., Filippenko A. V., Chornock R., 2002, ApJL, 579, L71

Moran E. C., Shahinyan K., Sugarman H. R., Vélez D. O., Eracleous M., 2014, AJ, 148, 136

Mulchaey J. S., Regan M. W., 1997, ApJL, 482, L135

Muldrew S. I., et al., 2012, MNRAS, 419, 2670

Mutlu-Pakdil B., et al., 2022, ApJ, 926, 77

Nelson D., et al., 2018, MNRAS, 475, 624

Nelson D., et al., 2019, Computational Astrophysics and Cosmology, 6, 2

Oh S., Oh K., Yi S. K., 2012, ApJS, 198, 4

Onions J., et al., 2012, MNRAS, 423, 1200

Padilla N., Lambas D. G., González R., 2010, MNRAS, 409, 936

Padovani P., 2017, Nature Astronomy, 1, 0194

Parks D., Prochaska J. X., Dong S., Cai Z., 2018, MNRAS, 476, 1151

Pawlik M. M., McAlpine S., Trayford J. W., Wild V., Bower R., Crain R. A., Schaller M., Schaye J., 2019, Nature Astronomy, 3, 440

Peng Y., et al., 2010, ApJ, 721, 193

Peng Y., Lilly S. J., Renzini A., Carollo M., 2012, ApJ, 757, 4

Penny S. J., et al., 2016, MNRAS, 462, 3955

Penny S. J., et al., 2018, MNRAS, 476, 979

Pillepich A., et al., 2018, MNRAS, 473, 4077

Pillepich A., et al., 2019, MNRAS, 490, 3196

Pilyugin L. S., Vílchez J. M., Mattsson L., Thuan T. X., 2012, MNRAS, 421, 1624

Pimbblet K. A., Shabala S. S., Haines C. P., Fraser-McKelvie A., Floyd D. J. E., 2013, MNRAS, 429, 1827

Planck Collaboration et al., 2020, A&A, 641, A1

Radcliffe J. F., Barthel P. D., Garrett M. A., Beswick R. J., Thomson A. P., Muxlow T. W. B., 2021, A&A, 649, L9

Reddish J., et al., 2022, MNRAS, 512, 160

Reines A. E., Volonteri M., 2015, ApJ, 813, 82

Reines A. E., Greene J. E., Geha M., 2013, ApJ, 775, 116

Reines A. E., Condon J. J., Darling J., Greene J. E., 2020, ApJ, 888, 36

Ricarte A., Tremmel M., Natarajan P., Quinn T., 2019, MNRAS, 489, 802

Ricarte A., Tremmel M., Natarajan P., Zimmer C., Quinn T., 2021, MNRAS, 503, 6098

Richardson C. T., Simpson C., Polimera M. S., Kannappan S. J., Bellovary J. M., Greene C., Jenkins S., 2022, ApJ, 927, 165

Riess A. G., et al., 1998, AJ, 116, 1009

Roberts M. S., Haynes M. P., 1994, ARA&A, 32, 115

Rodriguez-Gomez V., et al., 2015, MNRAS, 449, 49

Sabater J., Best P. N., Argudo-Fernández M., 2013, MNRAS, 430, 638

Sabater J., Best P. N., Heckman T. M., 2015, MNRAS, 447, 110

Sartori L. F., Schawinski K., Treister E., Trakhtenbrot B., Koss M., Shirazi M., Oh K., 2015, MNRAS, 454, 3722

Satyapal S., Ellison S. L., McAlpine W., Hickox R. C., Patton D. R., Mendel J. T., 2014, MNRAS, 441, 1297

Satyapal S., Abel N. P., Secrest N. J., 2018, ApJ, 858, 38

Schawinski K., Thomas D., Sarzi M., Maraston C., Kaviraj S., Joo S.-J., Yi S. K., Silk J., 2007, MNRAS, 382, 1415

Schmidt M., 1963, Nature, 197, 1040

Schmidt M., et al., 1998, A&A, 329, 495

Schutte Z., Reines A. E., 2022, Nature, 601, 329

Seyfert C. K., 1943, ApJ, 97, 28

Shabala S. S., et al., 2012, MNRAS, 423, 59

Shabala S. S., Deller A., Kaviraj S., Middelberg E., Turner R. J., Ting Y. S., Allison J. R., Davis T. A., 2017, MNRAS, 464, 4706

Shah E. A., et al., 2020, ApJ, 904, 107

Shapley H., 1919, Publications of the Astronomical Society of the Pacific, 13, 438

Shapley H., 1938, Nature, 142, 715

Sharma R. S., Brooks A. M., Somerville R. S., Tremmel M., Bellovary J., Wright A. C., Quinn T. R., 2020, ApJ, 897, 103

Sijacki D., Springel V., Haehnelt M. G., 2009, MNRAS, 400, 100

Sijacki D., Vogelsberger M., Genel S., Springel V., Torrey P., Snyder G. F., Nelson D., Hernquist L., 2015, MNRAS, 452, 575–596

Silk J., 2017, ApJL, 839, L13

Silverman J. D., et al., 2009, ApJ, 696, 396

Simmons B. D., Smethurst R. J., Lintott C., 2017, MNRAS, 470, 1559

Slipher V. M., 1917, Proceedings of the American Philosophical Society, 56, 403

Smethurst R. J., Simmons B. D., Lintott C. J., Shanahan J., 2019, MNRAS, 489, 4016

Spindler A., et al., 2018, MNRAS, 476, 580

Springel V., 2010, MNRAS, 401, 791

Springel V., White S. D. M., Tormen G., Kauffmann G., 2001, MNRAS, 328, 726

Springel V., Di Matteo T., Hernquist L., 2005a, MNRAS, 361, 776

Springel V., et al., 2005b, Nature, 435, 629

Stasińska G., et al., 2008, MNRAS, 391, L29

Stasińska G., Costa-Duarte M. V., Vale Asari N., Cid Fernandes R., Sodré L. J., 2015, MNRAS, 449, 559

Steinborn L. K., Hirschmann M., Dolag K., Shankar F., Juneau S., Krumpe M., Remus R.-S., Teklu A. F., 2018, MNRAS, 481, 341

Stern D., et al., 2005, ApJ, 631, 163

Stern D., et al., 2012, ApJ, 753, 30

Terrazas B. A., et al., 2020, MNRAS, 493, 1888

Tolstoy E., Hill V., Tosi M., 2009, ARA&A, 47, 371

Toomre A., Toomre J., 1972, ApJ, 178, 623

Traulsen I., et al., 2020, A&A, 641, A137

Trebitsch M., Volonteri M., Dubois Y., Madau P., 2018, MNRAS, 478, 5607

Treister E., Schawinski K., Urry C. M., Simmons B. D., 2012, ApJL, 758, L39

Trump J. R., et al., 2015, ApJ, 811, 26

Villforth C., et al., 2016, MNRAS, 466, 812

Vogelsberger M., et al., 2014, Nature, 509, 177–182

Volonteri M., Bellovary J., 2012, Reports on Progress in Physics, 75, 124901

Vulcani B., et al., 2018, ApJL, 866, L25

Wadepuhl M., Springel V., 2011, MNRAS, 410, 1975

Ward C., et al., 2021, arXiv e-prints, p. arXiv:2110.13098

Webb N. A., et al., 2020, A&A, 641, A136

Weinberger R., et al., 2018, MNRAS, 479, 4056

Weisz D. R., et al., 2011, ApJ, 743, 8

Westfall K. B., et al., 2019, AJ, 158, 231

Wethers C. F., et al., 2022, ApJ, 928, 192

Wetzel A. R., Tinker J. L., Conroy C., van den Bosch F. C., 2013, MNRAS, 432, 336

Wilkinson D. M., Maraston C., Goddard D., Thomas D., Parikh T., 2017, MNRAS, 472, 4297

Wirtz C., 1916, Astronomische Nachrichten, 203, 293

Wirtz C., 1917, Astronomische Nachrichten, 204, 23

Woo J., et al., 2013, MNRAS, 428, 3306

Wright E. L., et al., 2010, AJ, 140, 1868

Wylezalek D., Flores A. M., Zakamska N. L., Greene J. E., Riffel R. A., 2020, MNRAS, 492, 4680

Xiao T., Barth A. J., Greene J. E., Ho L. C., Bentz M. C., Ludwig R. R., Jiang Y., 2011, ApJ, 739, 28

Xin Y., Deng X.-F., 2021, Astrophysics,

Yan R., 2011, AJ, 142, 153

Yan L., et al., 2013, AJ, 145, 55

Yang X., Mo H. J., van den Bosch F. C., Pasquali A., Li C., Barden M., 2007, ApJ, 671, 153

Yang G., Brandt W. N., Darvish B., Chen C. T. J., Vito F., Alexander D. M., Bauer F. E., Trump J. R., 2018, MNRAS, 480, 1022

York D. G., et al., 2000, AJ, 120, 1579

Zhang Z., et al., 2021, A&A, 650, A155

Zhou S., Mo H. J., Li C., Boquien M., Rossi G., 2020, MNRAS, 497, 4753

Zwicky F., 1957, PASP, 69, 518

van Maanen A., 1916, ApJ, 44, 210