The Relation Between Galaxy Evolution and Galaxy Interaction

David Patrick O'Ryan

Physics

Department of Physics

Lancaster University

January 18, 2024

A thesis submitted to Lancaster University for the degree of Doctor of Philosophy in the Faculty of Science and Technology

Supervised by Dr Brooke Simmons

Abstract

Hierarchical models and observations show that galaxy interaction and merging is of paramount importance to galaxy assembly and evolution. However, the relationship between these physical processes and the characteristics of the galaxies involved is unclear. In this thesis, we make direct constraints between the physical processes occuring in galaxy interaction - increased star formation, nuclear activation, and morphological disturbance - and the underlying parameters of galaxies - their interaction stage, stellar masses, kinematics, and orientation parameters.

To constrain this relation, we require large samples of interacting galaxies which are representive of the full underlying parameter space. First, we aim to build this sample. A clear signature of interaction or merger activity is through the morphological distortion of a galactic system. We search for such signatures through the entire *Hubble* Space Telescope (*HST*) science archive. This is only possible with ESA Datalabs. In total, we classify 92 million sources into interacting and non-interacting galaxies. We find 21,926 disturbed and interacting galaxy systems; the largest interacting galaxy sample morphologically classified to date.

We use this new sample to explore the relationship between interaction stage and numerous galactic parameters. We cross match our new sample with the Cosmic Evolutionary Survey and find ancillary data for 3,829 interacting systems. We find radical changes in the star formation rate of our sample with stage, with the complete disappearence of the red sequence at the merging stage. We find that the fraction of galaxies with an active galactic nuclei is constant with

interaction stage, except at the point of coalescence where we find it dramatically increases. By investigating the relationship between these fundamental processes and stage, we show that there is a direct relation between the dynamical timescale and these processes.

Thus, we introduce a new algorithm which will be capable of exploring this relationship. We utilise a three-body numerical simulation with a Markov-Chain Monte Carlo (MCMC) algorithm to directly map the parameter space to morphological disturbances. We constrain the underlying parameters of a sample of 51 synthetic images representing different observed interacting systems. We are able to recover the true parameters within a confidence interval of 86.4%. We apply this methodology to a subset of observational data, and explore this algorithms potential as well as its limitations.

This thesis is dedicated to those who took the time to listen to a wee boy talk about the stars.

Acknowledgements

For as long as I can remember I have aspired to be an astrophysicist and, while I have never lost that passion and fire for astonomy, it would not remain so resolute without the unwavering support of those around me. I offer my sincere gratitude and thanks to:

My supervisor Brooke Simmons. Your guidance, enthusiasm and support over these many years have made this work possible. I left every single guidance meeting inspired and encouraged, ready to work harder than before.

My many collaborators: Bruno Merin, Chris Lintott, Mike Walmsley, Becky Smethurst, Tobias Geron, Sandor Kruk, Kameswara Mantha, Karen Masters, and the whole Galaxy Zoo Team who were all so supportive, insightful and inspiring throughout my PhD.

The lecturers and academics in my department: John Stott, Julie Wardlow, Isobel Hook, Mathew Smith and Sam Oates. Their help and guidance, especially in the final months of my PhD, was essential.

The ESA trainee cohort of 2022. When my love for astronomy was wavering and committment to academia in doubt you showed me a wonderful, inquisitive and welcoming environment that brought my passion back.

The team at 1715Labs. Working with you was a brilliant experience, and gave me a peek behind the curtain into life in industry.

Those in the Lancaster PhD office, who made it such a welcoming place. To Jamie Dumayne and Izzy Garland for many nights of excellent tacos. To Heather Wade and Tom Cornish for introducing me

to Lancaster. To Rahul Rana, Pascale Desmet, Jon Carrick, Nick Amos, Harry Stevenson and Andrew Milligan for some great laughs, scientific discussion and excellent dinners.

My house mates and close friends: Amy Hewitt, Matthew Thorne, Nikita Mehta and Zach Mason. You're all scattered to the wind excelling at your own things now, but I hope that we will meet again. Until then, I wish you all the success in the world.

Most importantly, my parents. You fanned those initial flames that gave me a passion for astronomy that has never left me. Your support then, and now, has driven me forward. It is only because of you that I have made it this far.

My brother and sister, Michael and Ruth. Thank you for putting up with me, and being amazing people.

Calum Cooper and Ronan Duff, who have supported me completely throughout my life. Even when my decisions lead me to leave you in Glasgow, you stood by me and still remain my closest friends.

To the University of Glasgow, where I aspired to go throughout my childhood. Learning there gave me the tools to complete and, hopefully, excel at my PhD.

This research was funded by the Science and Technology Funding Council of the UK, without them this work would not have been possible.

Finally, to Abby. Know your unwavering support got me through the toughest parts of my PhD. Your selflessness has left an impact on me that will persist for the rest of my life.

Declaration

This thesis is my own work and no portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification at this or any other institute of learning.

The research presented in Chapter 2 has been published in the relevant scientific journals in paper format and can be found as O'Ryan et al. (2023, ApJ, 948, 1). The contents of Chapter 3 and 4 have been written in the style of journal papers as they will be submitted for publication in due course.

As my thesis is the combination of three papers, its layout reflects this. Each science chapter begins with an abstract summarising that work and the introduction will often re-introduce the concept of interacting and merging galaxies.

I shall be telling this with a sigh
Somewhere ages and ages hence:
Two roads diverged in a wood, and I–
I took the one less traveled by,
And that has made all the difference.

- Robert Frost,
The Road Not Taken

*

To an old wine or a new idea, he cannot say no.

- Berthold Brecht,
The Life of Galileo

Contents

List of Figures

Relevant First Author Publications

Chapter 2

- Harnessing the Hubble Space Telescope Archives: Creating a Catalog of 21,926 Interacting Galaxy, O'Ryan et al., ApJ, 2023, 968, 40

Relevant Co-Author Publications

- Galaxy Zoo DESI: Large-Scale Bars as a Secular Mechanism for Triggering AGN, I. Garland et al., submitted (O'Ryan: 12^{th} of 15 authors)
- Galaxy Zoo DESI: Detailed Morphology Classifications for 8.7M Galaxies in the Legacy Imaging Surveys, M. Walmsley et al. (O'Ryan: 12^{th} of 16 authors), MNRAS, 2023, 526, 3
- Zoobot: Adaptable Deep Learning Models for Galaxy Morphology, M. Walmsley et al. (O'Ryan: 13^{th} of 17 authors), JOSS, 2023, 5312, 85
- The most luminous, merger-free AGNs show only marginal correlation with bar presence, I. Garland et al. (O'Ryan: 14^{th} of 17 authors), MNRAS, 2023, 522, 1
- Galaxy and Mass Assembly: Galaxy Morphology in the Green Valley, Prominent Rings and Looser Spiral Arms, D. Smith et al. (O'Ryan: 17^{th} of 18 authors), MNRAS, 2022, 517, 3
- Preparing for low surface brightness science with the Vera C. Rubin Observatory: Characterization of tidal features from mock images, G. Martin et al. (O'Ryan: 20^{th} of 52 authors), MNRAS, 2022, 513, 1
- Gems of the Galaxy Zoos-A Wide-ranging Hubble Space Telescope Gap-filler Program, W. Keel et al. (O'Ryan: 16^{th} of 16 authors), AJ, 2022, 163, 4

- Quantifying the Poor Purity and Completeness of Mophological Samples Selected by Galaxy Colour, R. J. Smethurst et al. (O'Ryan: 9^{th} of 10 authors), MNRAS, 2022, 510, 3

- Origin of the Local Group Satellite Planes, I, Banik, D., O'Ryan & H. Zhao, MNRAS, 2018, 466, 4

Chapter 1

Introduction

In the context of the large scale structure of the universe, galaxies represent a fundamental building block of matter. Their evolution through cosmic time leads to the local universe as we see it today (Springel et al., 2005). A key part of its evolution is that of mutual interactions between galaxies. Our best cosmological model, Λ-Cold Dark Matter, dictates that galaxies assembled hierarchically (White & Rees, 1978; White & Frenk, 1991). Therefore, we must understand the effects of mutual interaction to not just understand galaxy evolution but to better understand our theories of cosmology.

We have come to understand that interaction has multiple impacts on the evolution of galaxies. The first, and most obvious, effect is that it leads to the distortion of the galaxies involved and the formation of distinct tidal features. Early numerical simulations had excellent success recreating, and thus proving, these distortions to the galaxies were from tidal interactions between galaxies (Toomre & Toomre, 1972). Thus, tidal features became the primary method by which to identify interacting galaxies. However, as our identification methods developed it was found using only visual information led to high levels of contamination by close pairs by projection (e.g Ackermann et al., 2018; Blumenthal et al., 2020; Pearson et al., 2022). Often, pairs of systems which appeared interacting or distorted were found to be at very different redshifts; their peculiar morphologies caused by other processes. A paramount requirement for the identification of interacting pairs became redshift information, to ascertain the 3D

distances from each other. However, spectroscopic information across the many million close galaxy pairs we know is limited.

From the systems we have reliably identified, we have found mutual interaction has further effects on galaxies than morphological disturbance. It is has been found that interaction also causes an increase in the measured star formation (Bushouse, 1987; Mihos et al., 1992), higher fractions of active galactic nuclei (Ellison et al., 2008; Satyapal et al., 2014) and evidence for quenching of the systems involved (Schweizer & Seitzer, 1992; Gabor et al., 2010). This is particularily true of interactions between gas-rich galaxies where the fuel for star formation is readily available and quickly used. Rejuvenation of galaxies can also occur due interactions between gas-rich and gas-poor systems (Schweizer & Seitzer, 1992; Hopkins et al., 2009b).

The scale, change and impact of these effects is dependent on a host of underlying parameters of the galaxies involved. These include the galactic masses, the orientation of the galaxies and velocity of the encounter. This link between these parameters and the characteristics of the final system have been poorly explored. In this work, we will investigate this relationship by exploring a large sample of interacting galaxies that we create and describe a new software to extract their underlying parmeters. For now, however, it would be prudent to begin this discussion by exploring the definition of a galaxy, and outline the major work already conducted into linking galaxy interaction to its underlying parameters and, furthermore, with galaxy evolution.

1.1 What is a Galaxy?

For the purposes of this work, a galaxy will be the smallest unit of mass we will consider. A galaxy is a gravitationally bound system of gas, dust, stars and dark matter. The orbits and kinematics of each of these components leads us to different classifications of galaxies. The orbits of these components is often about a supermassive blackhole (SMBH) at the galactic centre. While almost all examples of galaxies have a SMBH at their centre, there are some examples where it is speculated they do not (Gebhardt et al., 2001). If the stars, gas and

dust about this SMBH have been orbiting unperturbed for a long period of time they flatten into a disk. The disk is split into a thick and thin component. Their general sizes are dictated by the history of the galaxy, and whether they have interacted or merged with many other galaxies in their history. Such harassment by other galaxies leads to heating of this disk. This heating takes the form of increased peculiar motion in the stellar orbits. This heating contributes not only to the thick disk, but can also lead to the growth of a bulge component in the galaxy (Hopkins et al., 2010; Bell et al., 2017). The bulge can be grown by other, secular, processes but we will focus on interaction and merging.

The bulge component is composed of stars whose orbit has been heated by changes in the gravitational potential due to galaxy interaction and harassment. This leads to new orbits that form a spherical bulge component around the galactic centre. This leads the galaxy to have a bulge and disk component, with the bulge composed of a spherical component of stars while the disk is formed of stars and gas continuing in unperturbed orbits. This is the simple view of a classic disk galaxy. Many disk galaxies do not show bulges at all, or show a pseudobulge (Gadotti & Kauffmann, 2009). This remains, at least partially, dependent on the merger history of the galaxy. If a galactic interaction leads to a strong enough merger, entire galactic disk is heated and can be destroyed. This leads to a completely spheroidal galaxy called an elliptical galaxy (Toomre, 1977; De Lucia et al., 2006).

The stars, gas and dust are embedded in a much larger, spherical halo. This galactic halo has two different matter components: baryonic and non-baryonic. The baryonic component is formed of very diffuse, ancient stars that form the stellar halo, stellar streams or globular clusters which orbit around the galaxy. The non-baryonic component is formed of dark matter. What dark matter is precisely is still debated, but this spherical dark matter halo extents out to many times the luminous matter of the galaxy. In general, all examples of galaxies are embedded in a dark matter halo (although there are some debated exceptions as described in van Dokkum et al., 2018). This dark matter halo has also been imperative to our understanding of how galaxies form, evolve and interact.

For much of the history of extragalactic astrophysics, the idea that galaxies could interact and merge was thought to be very improbable except in dense

clusters. The small radii (and, therefore, cross section) of the luminous matter in galaxies led astronomers to believe that the probability of an interaction was close to 0 and that galaxies were island universes (Hubble, 1926). However, with the development of the idea of the dark matter halo, it was realised that the probability of galactic systems interacting in the field was reasonable. Over time the idea that galaxy interaction and merging would have a significant impact on galaxy evolution was developed. In fact, the merger rates of galaxies through cosmic time is a well studied concept of our cosmological models with observations and simulations.

1.2 Galaxy Assembly Across Cosmological Time

Galaxies throughout cosmic history have been assembling and accreting matter. Studies at high redshifts reveal that galaxy structure is significantly different compared to the local volume. Initially, galaxies were very small systems that formed from gravitational instability thoughout the early cosmos (Lacey & Cole, 1993). These instabilities were generated by the physical processes of the very early Universe brought about by its small scale. From these initial density perturbations arose small systems which accreted from the surrounding gas and dust. These early galactic systems were much smaller than those we see today. These galaxies had peculiar morphologies (Elmegreen & Elmegreen, 2005) with disk galaxies beginning to dominate at $z \approx 3 - 6$ (Ferreira et al., 2022). These early systems formed stars at a much higher rate than the present day. While these systems were accreting much of their matter to gain mass, the Universe was expanding. At this time the rate of galaxies interacting and merging was also greater than today (Hopkins et al., 2010; Lotz et al., 2011), reaching as high as 50% at $z \approx 6$ (Conselice & Arnold, 2009; Bluck et al., 2009). This provided a peak in the cosmic star formation of the history through $1 < z < 3$ where approximately half of all stellar mass was formed (Bundy et al., 2005).

Figure 1.1 shows the changing merger rate through cosmic time. As the merger rate increased, peaked and then declined, it had profound effects on both the star formation density of the Universe and the morphology of those galaxies within it.

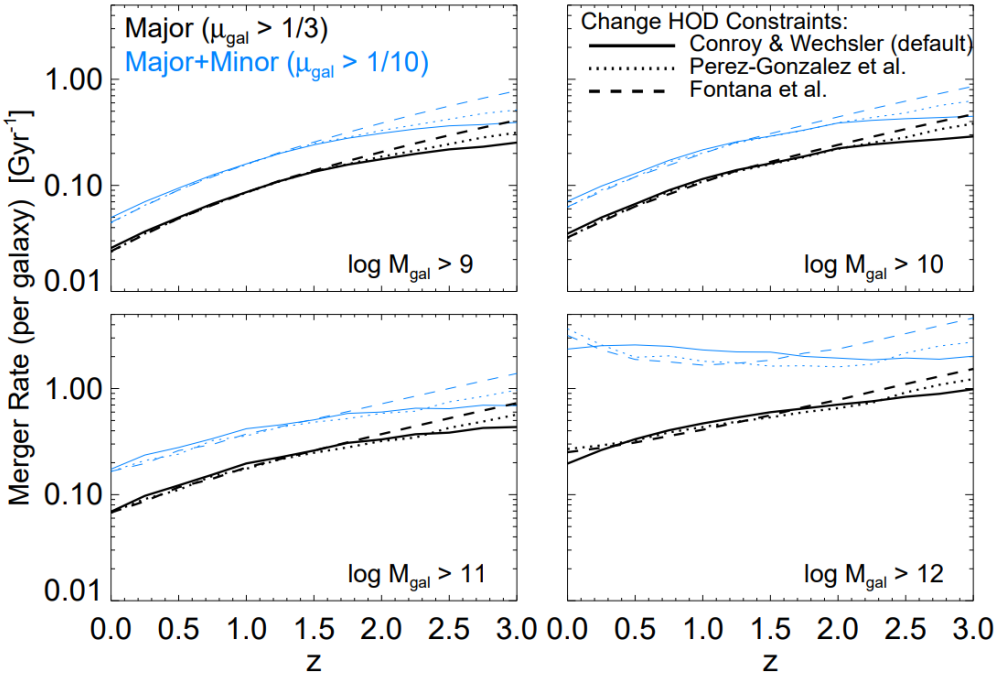

Figure 1.1: The clearly declining merger rates from $z = 3$ to $z = 0$. This is Figure 3 of Hopkins et al. (2010). This study looked at the history of a simulated set of galaxies of various stellar mass, and investigated the decrease in the merger fraction as a function of redshift and baryonic mass. As shown, for all mass bins and methods of identifying mergers (the different lines) the merger rate decreases. This is only not true for mergers between very high mass systems and low mass systems (in the bottom right of this plot). Thus, mergers between high and low mass systems may still have some driving force in the cosmic star formation rate density.

From $z = 1$, the once massive galaxies with rapid star formation have transformed into bulge-dominated galaxies containing SMBHs (Brown et al., 2007). At lower redshifts we find that the majority of the stellar mass is contained within bulge dominated systems. These galaxies are dominated by old stellar components and have little ongoing star formation (Hogg et al., 2002; Bell et al., 2004). Thus, showing that mergers played a significant role in star formation of the past.

This gives us an avenue by which to link observed galaxy morphologies to their merger histories. We split them into two distinct populations. One is of galaxies with an intense merger history, leading to the complete destruction of their disks and their domination by older stellar populations. The other is of those

with a relatively relaxed merger history. These non-destructive interactions and mergers served to feed the gas within the galactic disks and increase their mass, while maintaining their star formation rates. These systems are dominated by younger stellar populations, with star formation continuing to the present epoch. Thus, starting from the observations of morphology, we can make assumptions about the internal gas content, star formation and, therefore, colour of galaxies.

1.3 Galaxy Morphology & Galactic Properties

Galaxy morphology, and its change with time, is the study and measure of a galaxys' shape and features. A striking distinction in galaxy morphology is between disk and elliptical galaxies, the primary breakdown in the famous 'Hubble tuning fork' shown in Figure 1.2 (Hubble, 1936). This was thought to be a map of galaxy evolution with early-type elliptical galaxies on the left which would evolve to the late-type disk and spiral galaxies on the right. However, this was found not to be the case with early-type elliptical galaxies being older than late-type disk galaxies. So, while this classification scheme was not a direct evolutionary pathway it was found to be indicative of the merger history of a galaxy. Elliptical galaxies are spheroidal systems, with high internal velocity dispersions. Such systems are often created as the result of mergers and cannibalism of smaller systems to interactions with counterparts of a similar mass (Baugh et al., 1996; De Lucia et al., 2006). Disk galaxies, on the other hand, are rotationally-dominated systems with a central bulge whose size is crucially dependent on the merger history (Barnes, 1992; Hopkins et al., 2010; Bell et al., 2017). In fact, galaxies that appear to have no bulge component at all likely have had no merger event in the last few Gyrs (Martig et al., 2012).

This is also reflected in the observed colour of disk and elliptical galaxies. Observational colours in this context are the comparison of flux in two different wavelength ranges: one capturing the flux of young, star forming regions while the other captures the flux from old, low mass stars. In combination, these young and old stars, constitute a stellar population. Stellar populations have well defined spectral energy distributions (SEDs) based on their age and, indirectly,

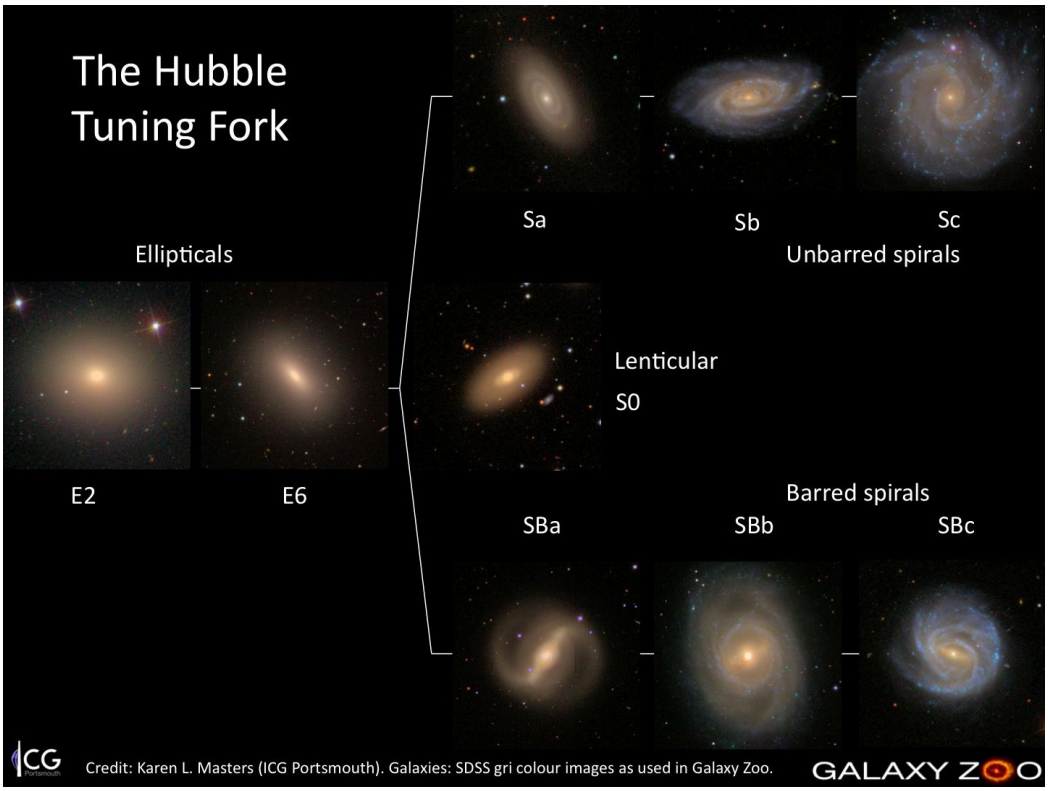

Figure 1.2: The Hubble tuning fork. Credit goes to the Galaxy Zoo collaboration (and specifically Karen Masters) for the creation of this figure. What was once thought of as an evolutionary track for galaxies, is now a widely adopted classification system for them based on morphology. On the left, we have elliptical systems - so called early-type galaxies - which are often 'red and dead' systems. They have an intense merger history, which has destroyed their disk component and caused them to use the majority of their gas in star formation. On the right, we have two different kinds of disk galaxies. Ignoring the bar or un-barred part of this, they are systems which have a less intense merger history and have only accreted smaller systems into them. This serves to enhance their gas disk, and preserve their disk component.

to a galaxys' star formation rate (SFR) and gas mass. When star formation is occurring rapidly, and the SFR is very high, lots of young, luminous OB-type stars form. The flux from these stars falls mainly in our blue filters, in the rest wavelength of ultraviolet. These stars have lifetimes of only a few million years and, therefore, die very quickly. Thus, if stars are forming slowly, and the SFR is very low, these OB-type stars will die and not be replaced. This leaves an older stellar population, mainly composed of G- to M-type stars which lie primarily in our green and red filters, with wavelengths in the optical. Thus, the SFR and age of the dominant stellar population of a galaxy influences their observed colour.

The SFR, then, is highly dependent on the gas mass present in the galaxy. This gas must be molecular gas, with little energy so it is able to form massive clouds that fragment and undergo collapse (Goldreich & Lynden-Bell, 1965; Quirk, 1972). From models of individual clouds, it was noted that the gas density was related to the density of star formation in galaxies (Schmidt, 1959). This was further refined to be a global law in disk galaxies in particular in Kennicutt (1998). In this, it was shown that the surface area of star formation is directly related to the surface area of gas by

$$\Sigma_{SFR} \propto \Sigma_{Gas}^{n}. \tag{1.1}$$

This relation has been found to be $n \approx 1.3$. Thus, two things are happening here, if a galaxy has gas, it has star formation and if it has star formation, it will be observed to be blue and vice versa.

When observing populations of red galaxies, those dominated by older stellar populations, it has been found they are often elliptical galaxies (Bower et al., 1992). These galaxies have a more violent merger history that has destroyed the ordered rotationally dominated component of a galactic disk and, in the process, either removed or used the gas within the galaxy (Faber & Gallagher, 1976). As a result of this, the current gas mass and density are very low which leads to the SFR being also low.

The opposite is true for disk galaxies. As stated previously, these systems have had a less tumultous merger history - particularily since $z \approx 1$. An indicator of the merger or interaction history in disk galaxies is the size and shape of the

bulge component (Emsellem et al., 2011). However, because of the plentiful gas, dust and ordered rotation in the galactic disk the SFR within is larger than elliptical galaxies. The surface density of gas within the disk is much larger than in elliptical galaxies providing the ideal for star formation. Spiral arms can exist in such galaxies, containing large filaments of gas and dust. They contain areas where large molecular clouds can condence, fragment and then collapse into new stars. These star forming regions will be young enough that massive OB-type stars will form, which in turn changes the underlying SED of the galaxy that we observe. The blue filter will contain more flux, and therefore, the galaxy appears blue when comparing the red and blue filters.

This gives rise to two distinct populations of galaxies: old, red elliptical galaxies and young, blue disk galaxies. When plotting colour-colour or colour-magnitude diagrams, there is clear bi-modality which separates these two populations with the 'green valley' running between them (Strateva et al., 2001). Figure 1.3 shows the colour-colour distribution if the two populations. In this Figure, we clearly can see the blue population - blue cloud - and the red population - red sequance - beneath it with the green valley in between. The green valley is a disputed area of this distribution. Some works claim that it is a rapid transition phase (Schiminovich et al., 2007; Smethurst et al., 2015), where galaxies are moving between the two populations due to different processes. Others, meanwhile, show that the picture is much more complicated, with morphology and environment playing an important role that dictates the decline in star formation (Schawinski et al., 2014). However, what is definitive is the rule elliptical galaxies are always red and disk galaxies are always blue is not ubiquitous (Smethurst et al., 2022). Examples of blue ellipticals and red spirals do exist (Schawinski et al., 2009; Masters et al., 2010; Keel et al., 2022) and they are the subject of intense debate and study in the current field. However, as a general guideline for the expected properties of a galaxy the colour and morphology are deeply interdependent on one another.

There are also many types of systems that do not fit this simple morphological definition of galaxies. Starburst galaxies are often very morphologically irregular and have measured SFRs far in excess of expectation for their mass. These then lead into post-starburst galaxies and merger remnants, who are the complete

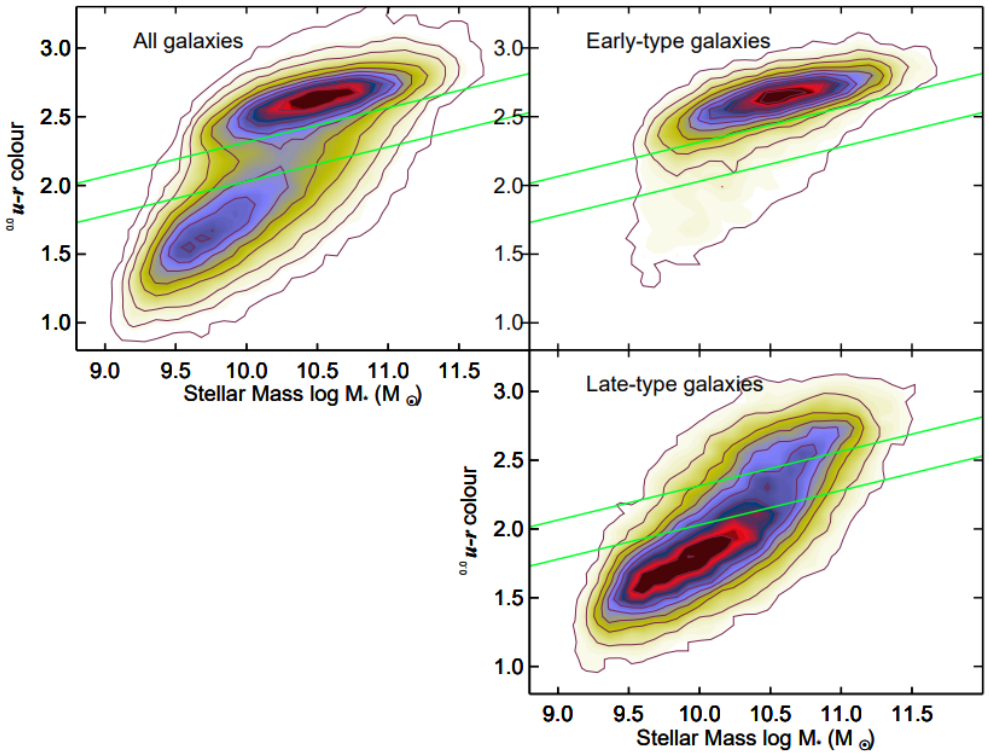

Figure 1.3: The distribution of galaxies in colour-colour space showing two distinct populations in a clear bi-modal structure. This is using the u-band and r-band filters and, therefore, the closer to 0 on the y-axis the bluer the galaxy. Left plot: the top population is the blue cloud, primarily composed of disk galaxies with young, star forming stellar populations. The bottom population is the red sequence, primarily composed of more massive elliptical galaxies which are gas poor and quiescent. Between these two populations lies the green valley, marked by the green lines. This is believed to be a transitional population moving between the blue cloud and red sequence for debatable reasons. To further show the split with morphology, the panels on the right split the colour-mass space into early-type (elliptical) galaxies and late-type (disk) galaxies. Note, this is Figure 2 of Schawinski et al. (2014)

opposite and have measured very low SFRs. These are often called quiescent galaxies. These highly irregular systems have direct interplay, however, with their merger histories and surrounding environments (Pawlik et al., 2018; Hani et al., 2020). In fact, it has been found that in cluster environments, the fraction of disk galaxies drops to $\leq$10% when compared to $\approx$60% in the field.

The effect of the galactic environment on a galaxy should also not be understated. There are many ways to define the galactic environment, but it is most commonly associated with the number density of galaxies about them (Eisenstein et al., 2003; Balogh et al., 2004). There are three broad classifications of environment: field, filament and cluster. A field galaxy is one with neighbours, and is in relative isolation. A filament galaxy has some neighbours, but their influence is minimal. A cluster galaxy is one in which a large number of other galaxies are bound in large galactic super-structures. In such an environment, SFRs are suppressed (Baldry et al., 2006) by ram pressure stripping and galaxy morphology is highly irregular due to constant interaction and merging. In these environments, the relation between galaxy morphology and the underlying processes is almost completely broken by interference from the environment.

Thus, galaxies in the field or in filaments are of particular interest. Here, we can study the direct link between morphology and the underlying processes of galaxies. And these, in turn, are highly dependent on galaxies underlying merger and interaction histories. So, what is this relation? As stated previously, the influence of interaction and merging accelerates the evolution of rotational systems to dispersion dominated systems. However, there are many more subtle affects that can be attributed to galactic merging and interaction. For instance, if two galaxies can be brought into a state of starbursting due to merging (Martig & Bournaud, 2008). This completely changes the underlying SED of the galaxies involved, and can lead to the rapid quenching both systems (Violino et al., 2018; Ellison et al., 2022). Depending on when we observe such a galaxy, we can find either a much bluer or redder galaxy than expected (Di Matteo et al., 2007). However, these effects are also debated, with some studies finding that galaxy interaction does not induce significant change in the SFR, and therefore, the underlying SED of the galaxy population (Bergvall et al., 2003). So, why is this? It transpires that the effects of a galaxy merger or galaxy interaction is dependent

on the underlying parameters of the galaxies themselves. This leads to multiple classifications of galaxy interactions which each lead to different outcomes for the galaxies involved.

1.4 Categorisation of Mergers

The effects of galaxy interaction and merging is dependent on multiple factors and underlying parameters. A key parameter to the destruction or survival of a disk is the mass ratio between the two systems (Bournaud et al., 2005; Cox et al., 2008). The impact parameter also has significant influence over the final morphology over the system, however, we will discuss this further in Section 1.5. The gas content of each galaxy also has an effect on the changes of the internal SFR of the systems. The large change in the systems we see based on these parameters gives rise to many different categorisations of galaxies. For the mass ratio, we define a major, a minor and a micro interaction. These are the ratios of the primary galaxy mass (the more massive galaxy) and the secondary galaxy mass (the less massive galaxy).

These are then further sub-divided into two seperate categories based on their gas content: a wet merger or a dry merger. This refers to the gas mass and colour of both galaxies. A wet merger involves lots of gas in the two galaxies colliding, and therefore, lots of resultant star formation. Both of the involved galaxies are blue, and often disk-dominated galaxies. A dry merger involves little gas, and is often the merger of massive, red elliptical galaxies. There are also intermediate categorisations based on the amount of gas, such as damp mergers (where there is some gas, but not enough to dramatically increase star formation) and mixed mergers (mergers between gas-poor and gas-rich systems), however, we will focus on the binary definition of wet and dry mergers.

Depending on the classification of an interaction leads to specific outcomes for the interacting galaxy system. A major interaction is when the galaxies involved have a mass ratio of approximately 1:1. Some definitions vary, however, and allow this limit to go down to 1:3. For the purposes of this work, they must have a mass ratio of at least 1:3. Major interactions are the most devastating to the

morphology of the systems involved, with the forces upon them causing severe morphological distortion and the formation of tidal features (described further in the following section). If the two systems merge, there is complete destruction of the disks in both galaxies and the post-merger remnant will be highly irregular. If this is also a wet merger one we would also expect a significant increase in the SFRs of both galaxies (Mihos & Hernquist, 1994, 1996; Woods et al., 2006). We also find that wet major mergers have an increased AGN fraction, which suggests that such a merger may play a role in starting nuclear activity (Alonso et al., 2007; Ellison et al., 2011; Koss et al., 2012). The opposite of this is true when a major interaction is dry, where only a small increase in SFR is found in the nuclear region of the galaxies (Sánchez et al., 2004; Bell et al., 2006).

A minor interaction has less catastrophic consequences for the morphology of one of the two galaxies. This is defined as an interaction where the mass ratio is less than 1:3 but greater than 1:10. Due to this large mass disparity, the morphology of the primary galaxy is relatively unaffected by such an interaction. However, the secondary galaxy will be almost completely destroyed by the encounter. If only an interaction occurs, the secondary will be higly disrupted forming a long and stretched tidal tail as it moves through the orbit. When these systems merge, we observe small increases in the SFR of the primary. There is ongoing work investigating whether such mergers were actually a primary driver of star formation in galaxies across cosmic time through rejuvenation of gas resevoirs within the primary galaxy (Bournaud et al., 2007; Kaviraj, 2014b; Jackson et al., 2022). Finally, a micro interaction is one in which the mass ratio between the primary and secondary galaxies is less than 1:10. This sees the complete destruction of the secondary galaxy, whether a flyby or a complete merger. These can form stellar streams about their primary galaxy and are often absorbed by the primary with very little change in its morphology.

Figure 1.4 displays examples of each of the different merger categories we have defined here. These, from left to right, are the Arp 240, Arp 188 and NGC 5907 systems. They each show a major, minor and micro interaction, respectively, and demonstrate the change in effect the mass ratio has on the morphology of the galaxies involved. From Arp 240, with the complete distortion of the galactic disks and formation of tidal features to NGC 5907 where the primary galaxy is

Figure 1.4: Examples of a major, a minor and a micro interaction. Each of these are wet interactions, containing lots of gas and therefore, increases the rate of formation of stars. These are the Arp 240, Arp 188 and NGC 5907 systems, respectively. Here, we show the famous double looped stellar stream of NGC 5907 but point out that the existence of the second loop is in dispute and that we only show it here for illustrative purposes (van Dokkum et al., 2019). From major to micro interactions, we see decreasing impact and change in the morphology of the primary galaxies but always the complete destruction of the secondary.

barely disturbed at all. Thus, the different categories of interactions and mergers have very different impacts on the systems involved. We will now discuss, in depth, the effects that interaction has on these galaxies and specifically explore the formation of morphological disturbances like tidal features, changes in the SFR and the increase in AGN fraction.

1.5 Effects of Galaxy Interaction

Until the work of Toomre & Toomre (1972), the idea that two galaxies would encounter each other and interact was thought to be minute. We now understand that galaxy interaction plays an important, and fundamental role in the evolution of galaxies. As stated previously, they have multiple effects upon the systems undergoing the interaction. The specific effects of interaction rests on a host of parameters. We have discussed the mass ratio and gas content and mentioned the impact parameter but there is also the orientation of the interaction, the relative sizes of the galaxies and the point in the dynamical history of the interaction we are observing. Thus, in this section, we will explore how these different underlying parameters link to the physical processes we observe in interacting and merging galaxies.

1.5.1 Morphological Distortion: Tidal Features

As previously discussed, galaxies are large gravitationally bound systems of stars, gas, dust and dark matter. Upon a close encounter between two galactic systems, these components experience strong gravitational forces which disrupt their ordered layout. As the galaxies move through their relative orbits, the gravitational potential changes at an accelerating rate. This imparts energy into the ordered components of the galaxies, and causes thermalization of their internal motions. This leads to violent relaxation, where the stellar orbits are so altered they no longer follow their prior orbits. However, the energy in the system must be conserved. Thus, as the internal stellar motions within the galaxy thermalizes, the energy in the galaxy's orbits decay via dynamical friction. If this decay is large enough, the energy in the galaxy's orbit may be sufficiently reduced to no longer escape from one another and they will eventually coalesce to leave a single merger remnant.

The changing graviational fields as the two galaxies encounter each other leads to radial distortion of each galaxy. If stellar material, including gas, dust and stars, is close to the edge of the galaxy a combination of the galactic rotation and radial elongation lead to it being sheared off and away from the galaxy. This forms two 'tidal tails' in the system, one leading and one proceding the motion of the galaxy. Dependent on the geometry of the encounter, the trailing tail of one galaxy and the preceding tail of the other can form a 'tidal bridge' - a linking of the two systems. An example of both of these is the interacting pair Arp 240, shown in left hand plot of Figure 1.4.

The geometry of the bulk motion of the galaxy is very important in forming these tidal features. As the internal galaxy rotation and bulk motion of the galaxies orbit in the encounter must match for this shearing of material to occur. Thus, the formation of tidal tails and tidal bridges is only possible in a prograde interaction whereas a retrograde interaction suppresses them. Further, depending on the relative velocities of the galaxies these tidal tails can be split off from the galaxy as the encounter continues. This can lead to the formation of 'tidal debris' about the galaxies.

The confirmation that features such as these were from a tidal origin came primarily from simulations of interactions of different systems. The first restricted numerical simulations were conducted by Toomre & Toomre (1972) who successfully modelled the features of four different systems using distributions of test particles. Numerous works since have recreated tidal features using numerical simulations (Salo & Laurikainen, 1993; Petsch & Theis, 2008; Barnes & Hibbard, 2009; Wallin et al., 2016). These have also been expanded to include a range of other properties in hydrodynamic simulations (Hopkins et al., 2013; Moreno et al., 2019, 2021; Sparre et al., 2022), and investigated in cosmological simulations (Kaviraj et al., 2015; Rodríguez Montero et al., 2019; Hani et al., 2020; Das et al., 2023b). In these more advanced simulations the key to linking the simulations to observations is the morphology of the tidal features. This is often from either direct comparison or from searching for analogues through a suite of interaction simulations in cosmological simulations.

While the most striking tidal features to form in galactic encounters are these tidal tails and bridges, we also see the formation of many other features. These include 'stellar streams' (previously discussed), shells and rings forming in the galactic disk. As stated, stellar streams are likely smaller galactic systems that have been destroyed while passing the primary galaxy. This leaves a faint stream of material about the galaxy. The right plot of Figure 1.4 shows an example of a stellar stream with the NGC 5907 system. Shells, on the other hand, are formed around galaxies and can be present in as many as 10% - 20% of elliptical and lenticular galaxies (Malin & Carter, 1983; Atkinson et al., 2013). The left panel of Figure 1.5 shows an example of shell in the elliptical galaxy NGC 1344. Investigation of the formation shells shows they are primarily composed of stars (Quinn, 1984). There are often numerous shells in a single galaxy. Shells are formed from the disruption of a small satellite around a significantly more massive galaxy - so, in a minor to micro interaction. This then forms a stellar stream which, over time, condenses to a cloud of stars orbiting the primary galaxy. This then either forms into an X-shaped structure or a annulus which, in the 2D projection of the sky, appears as a shell. Thus, observing a stellar stream or a shell is highly dependent on the time in the dynamical history of the encounter that we are observing.

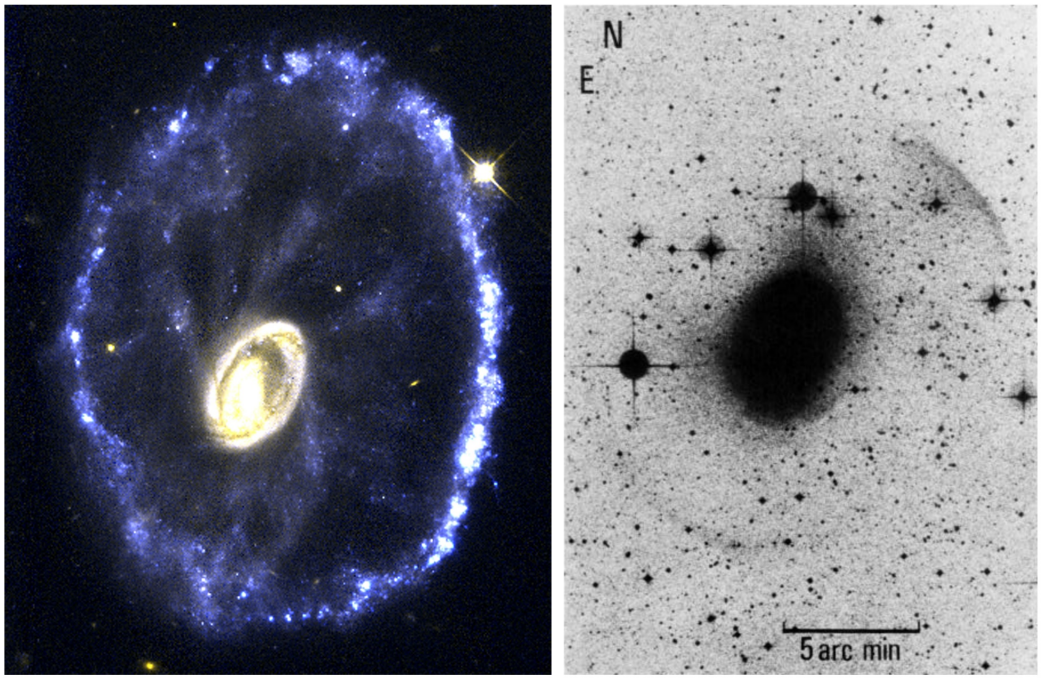

Figure 1.5: Examples of a collisional ring galaxy and a system containing shells. Left: The Cartwheel galaxy, the most famous ring galaxy to date. Ring galaxies can only be formed by a direct collision between two galaxies. The ring itself is composed of a region of the disk undergoing intense star formation due to the density wave passing through the disk from the impact of the secondary galaxy. Right: The system NGC 1344, showing two shells of stellar material surrounding it. This is formed by the condensation of stellar streams into clouds of stars. Cartwheel Galaxy Image: NASA / Hubble, NGC 1344 Image: Malin & Carter (1983)

Finally, we find rings can form from galaxy interaction. These systems are often called ring galaxies or collisional ring galaxies. Figure 1.5s right panel shows the Cartwheel Galaxy: a famous example of a ring galaxy. This ring is formed of young, hot stars and is formed only from the head-on collision with another system (Lynds & Toomre, 1976). The interaction to form this is so intense, that it causes a density wave to pass through the galactic disk which triggers intense star formation at its wake.

From the example of a ring galaxy, we see that interaction not only affects the stellar distribution of the galaxy, but also has direct impacts on the gas within it. As we have stated in previous sections, interaction and merging can induce enhancements in star formation and even lead to a starburst in a galaxy which brings about the complete quenching of the system.

1.5.2 Star Formation Enhancement

As noted earlier, it is often observed in interacting and merging galaxies that the star formation is enhanced in some way. While to what level this enhancement is often debated between observations (Barton Gillespie et al., 2003; Li et al., 2008a; Patton et al., 2011; He et al., 2022) and simulations (Di Matteo et al., 2007; Cox et al., 2008; Hopkins et al., 2013; Moreno et al., 2021), the underlying processes that lead to enhancement are well understood. First, by star formation enhancement we mean that the SFR either globally or in different regions of an interacting galaxy is higher when compared to isolated galaxies. Thus, some property of interaction is causing an increase in the star forming activity of the galaxies involved.

The rate at which stars form is known to be directly related to the surface density of gas at any given piont within the galaxy, shown in equation 1.1. Due to an interaction the torques and gravitational forces upon the gas clouds a loss of angular momentum. This causes the gas clouds to drift inwards, towards the galactic centre. As more gas clouds fall into the centre, the surface density of gas in this region increases. This in turn increases the SFR.

This is a simple explanation, and an easy idea to hold about driving the increase in star formation in interacting galaxies. As the morphology of the

galaxy is compressed into a tidal feature, we see the same increase in the surface density of gas which leads to a further increase in the SFR. This movement and compression of gas into the galactic centre also has a secondary effect on the SMBH at the galactic centre. Increasing the gas density begins the feeding of the SMBH which causes the ignition of nuclear activity.

1.5.3 Ignition of Active Galactic Nuclei

The presence of an active galactic nuclei (AGN) in a galaxy is easy to identify observationally. When viewed them face on, the AGN is so luminous that it often outshines the entire bulge and disk of the galaxy. To study the galaxies they are present in we must carefully remove all contribution of the AGN to the underlying SED flux. The bright flux we observe is from a long and complicated process of accretion of material onto the SMBH at the galactic centre (for an excellent breakdown of the structure and evolution of AGN see Beckmann & Shrader, 2012). The structure of material about the black hole is also complex, and results in many observational oddities.

An accretion disk of material that is orbiting the black hole structure exists. As this material accretes, it causes the SMBH to project powerful jets of radiation perpendicular to the accretion disk plane. These jets are highly luminous, and we observe them in the optical, infrared and X-ray. Two populations of AGN have been discovered: containing narrow and broad emission lines, named Seyfert 1 and Seyfert 2 AGN respectively. While at first thought to be two distinct types of AGN, these are believed to be one AGN population viewed from different angles (for a review of the unification, see Netzer, 2015). This is possible as surrounding the accretion disk, there is a torus of material. This material is primarily dust, and highly absorbing of the emission from the SMBH-accretion disk system. This means we only see narrow line emission (Seyfert 2) when we observe the AGN through the torus and observing but observing both narrow and broad lines (Seyfert 1) of emission when not viewed through the torus. As we move to higher luminosities, we get further populations of AGN such as blazars, quasars, radio-quiet and radio-loud systems. These classifications are all dependent upon

the inclination at which we are viewing the AGN system and the material in the line of sight.

From this description of an AGN structure we can see how galaxy interaction increases the likelihood of nuclear ignition. AGN require large volumes of gas and dust to be present at the galactic core; a surge of which is caused by interaction (Hopkins et al., 2008, provides an excellent summary of this process from the point of view of simulations). However, what is often surprising from observations of samples of interacting systems is the meagre increase in the AGN fraction within them. There have been many hypothesis as to why this might be, from a delay in the AGN ignition (Ellison et al., 2011) to AGN flickering (Schawinski et al., 2015). A delay in AGN ignition would make sense, again, in the context of the structure we have just discussed. As gas and dust are moved into the galactic core, they do not necessarily immediately move directly around the SMBH and may take time to get there. There may be other mechanisms at work within the AGN itself that causes suppression of ignition for some time. This could also produce the AGN flickering, where material may be periodically cutoff from the SMBH.

It cannot be disputed that the sudden movement of gas into the galactic centre during interaction appears like an obvious way for AGN activation to occur. However, this movement of gas around the galaxy and use in star formation and use or ejection in an AGN can start the quenching of the galaxy.

1.5.4 Quenching of Galactic Systems

During the interaction of two galaxies there is significant movement of gas, AGN activation and large increases in star formation across the galaxy. However, a galaxy only has a limited resevoir of gas available to it. Therefore, this sudden increase in gas usage in multiple different processes leads to the gas being used in a shorter timescale than expected. Once the gas is used, and the gas density drops significantly, star formation and AGN will rapidly cease and the colour of the galaxy. This process is called quenching.

This rapid quenching of systems is different to what is expected in the field. Galaxies gradually use their gas over long periods of time forming stars and slowly

quench as the average gas density across the disk reduces to the point of very low SFRs (Peng et al., 2010b) in 'mass quenching'. A system can also be quenched by the environment in 'environment quenching'. However, by concentrating gas and increasing the gas density in an interaction, even if a galaxy had previously stopped forming stars it is able to commence star formation again and use the remaining gas entirely. Thus, we often seen increases in the SFR during and immediately after an interaction but then the sudden shut off of SFR shortly after it (Ellison et al., 2022).

The sudden use of gas is not just from it being used in star formation, but also from stellar feedback around the areas where the stars are forming. As these new stellar populations form OB-type stars very quickly die and cause an increase in the rate of supernova. This, in turn, creates strong shocks and winds that drive the molecular gas out of the galaxy at very high speed (Bolatto et al., 2013; Geach et al., 2018). A similar process can occur due to the AGN in the galaxy. These objects drive strong winds across the galactic system. These winds can also act to drive out gas (Cicone et al., 2014; Cheung et al., 2016; Baron et al., 2018) and cause the cessation of star formation.

1.6 Identifying Interacting and Merging Galaxies

We have described and explored the effects of interaction and merging across the galaxy population. We have described the basic properties of galaxies and how these are differ when compared to interacting galaxies. However, we are only able to fully understand these differences by using samples of interacting galaxies and comparing them to control samples.

The first samples of interacting galaxies were identified by eye. The earliest interacting galaxy catalogue was the Arp (1966) Atlas of Peculiar Galaxies. This contained 166 interacting systems. This was quickly added to with a 2-part edition of the Vorontsov-Velyaminov (1977) catalogue, providing a further 268 systems to the original Arp (1966) catalogue. Both of these catalogues contained

major interactions, where clear tidal features easily put them into the interacting galaxy classification. These samples were also not large enough to make statistical, representative statements about the interacting galaxy population. Thus, larger catalogues had to be created.

Later catalogues often used visual classification to find interacting and merging galaxies. A very successful example is from the Galaxy Zoo collaboration which created a catalogue of 3,003 interacting galaxies identified by citizen scientists based on morphology (Darg et al., 2010a). However, a limitation was found when visually classifying interacting galaxies. With no information on the redshift, and therefore of 3D distribution of a galaxy pair, it is very difficult to distinguish 'close pairs' from truly interacting pairs. A close pair, in this context, is two galaxies which appear to be very close together in the sky in the 2D projection of the sky but are actually at different redshifts. This means they are not actually close together in 3D space and cannot be interacting. Many interacting galaxy samples created by visual classification must to throw large numbers of false positives from their sample (Blumenthal et al., 2020; Pearson et al., 2022). As machine learning algorithms began to be used in galaxy classification, this problem of contamination by close pairs continues.

Neural networks ~~have began to~~ play an increasingly central role in galaxy classification. As the size of observational samples ~~have~~ increased exponentially, so has the ~~apparence~~ of the impossibility of visually classifying each galaxy individually. Rather, a neural network can now be trained to make morphology classifications of galaxies using a much smaller subset of visually classified galaxies. A neural network is a layered structure of interconnected nodes which can pass information to one another. Each node has a weight assigned to it which allows it to alter the input and then pass on the weighted output. If the weighted output is above or below some value, the output is set zero. The input in this context can be some information about a galaxy, a section of an image or an underlying parameter estimation. These nodes all interconnect through many different layers where they all perform operations on their input and provide weighted outputs. The weighted outputs from the final layer are then passed to a classification layer which outputs either a formal classification or some value that can be mapped to a classification by a user.

A formal classification can be made in this way by training the neural network to recognise certain features of an image. Training a neural network in supervised learning involves providing it with a training set of fully annotated data so it can 'learn' classifications. To train, the neural network iterates through the provided training set and makes classifications upon each input. It then checks how many it got right and how many wrong, and then tweaks the weights of each node. It then makes the classifications again and checks whether it got more right or wrong. This is done through many epochs of the training set, it gradually tweaks its internal weights such that it should be able to make the same classifications on an unknown dataset and recognise images with similar classifications to the training set. Thus, it is important that the training set is large and is representative of the full parameters space a user wishes to classify.

In morphology classification, the workhorse neural network is the convolutional neural network (CNN). A CNN takes an image and breaks it down into smaller subsections in a convolutional layer. These smaller subsections - essentially arrays of different sections of the image - are passed through the neural network operations and activations and are altered by the weights at each node. Depending on the number of layers in the CNN, the image may be sub-divided, convolved and weighted many times before being fed into a classification layer which outputs the final classification. The real power of a CNN such as described is that it is able to recognise features in an images. For instance, tidal features, AGN or the whole morphology of the galaxy such as ellipticals or disks.

However, using a CNN for the purposes of identifying interacting galaxies still introduces the same problems as using visual identification. While the classifications it makes on the morphology of interacting galaxies, or on two galaxies being near each other in an image, it faces the same contamination from close pairs.

New methods combining CNNs with morphological parameterisations of a galaxy are being developed to overcome this issue. This has found some success in correctly identifying interacting galaxies fom morphology alone (Ren et al., 2023). Other works have turned to training CNNs on simulated data; where the training set could be controlled and large enough to span the entire parameter space of interaction. These training sets would also have the advantage of being completely free of contamination. However, when applied to observational data

it was found that the accuracy of such models completely breaks down (Bottrell et al., 2019; Ćiprijanović et al., 2020). Thus, creating large and pure interacting galaxy samples remains a task to complete, and the method being it an open question.

1.7 This Thesis

The primary focus of this thesis is investigating and constraining the relationship between galaxy interaction and evolution. The list of unresolved questions about galaxy interaction is long and complex, but we will focus on the following questions here:

- What is the relation between galaxy interaction and the enhancement in star formation?
- What is the relation between galaxy interaction and the fraction of AGN?
- At what point in an interaction do we find these relations are greater than compared to non-interacting galaxies?
- What is the conclusive relationship between the dynamical timescale of interaction, the formation of tidal features and the above effects?

To fully answer these questions three components are required: first, a large statistically significant sample of interacting galaxies; second, available ancillary data of these systems to make inferences between the underlying processes in interaction and their parameters; and thirdly, the tools constrain the dynamical timescales of each system. This thesis details methodologies which acheive these goals. First, in Chapter 2, we create the largest interacting galaxy catalogue to date through combining morphological classification with novel methods of machine learning, data extraction and analysis. We then use this catalogue, in Chapter 3, by matching to existing ancillary catalogues and conduct our own inferences about the relationship between galactic parameters and underlying processes. However, we find that we can only constrain this relationship to a

very general point. We require tighter constraints on the dynamical timescale of the interactions in our sample to further explore the relationship between galaxy evolution and interaction. In Chapter 4, we describe software we have built that will conduct such constraint in the context of Bayesian statistics and galactic parameterisation. We will describe the results we have found by applying it to a small, well constrained, interacting galaxy sample and then describe the limitations of the approach in terms of computational efficiency. Finally, in Chapter 5 we summarise our results, describe them in the context of current works and speculate about the future work to advance them.

Where necessary, we use a Flat ΛCDM cosmology with $H_0 = 70 \mathrm{kms}^{-1}\mathrm{Mpc}^{-1}$ and $\Omega_M = 0.3$. Hereafter in this thesis, when referring to an interacting galaxy we are referring to a galaxy which has undergone one or multiple flybys by a secondary galaxy and caused a tidal disturbance. A merging galaxy is the final state of these flybys, where two or more systems have coalesced to form a highly morphologically irregular system.

Chapter 2

Creating a Large Sample of Interacting Galaxies

2.1 Abstract

Mergers play a complex role in galaxy formation and evolution. Continuing to improve our understanding of these systems require ever larger samples, which can be difficult (even impossible) to select from individual surveys. We use the new platform ESA Datalabs to assemble a catalogue of interacting galaxies from the *Hubble Space Telescope* science archives; this catalogue is larger than previously published catalogues by nearly an order of magnitude. In particular, we apply the `Zoobot` convolutional neural network directly to the entire public archive of *HST* $F814W$ images and make probabilistic interaction predictions for 126 million sources from the *Hubble* Source Catalogue. We employ a combination of automated visual representation and visual analysis to identify a clean sample of 21,926 interacting galaxy systems, mostly with $z < 1$. 65% of these systems have no previous references in either the NASA Extragalactic Database or Simbad. In the process of removing contamination, we also discover many other objects of interest, such as gravitational lenses, edge-on protoplanetary disks, and 'backlit' overlapping galaxies. We briefly investigate the basic properties of this sample, and we make our catalogue publicly available for use by the community. In

addition to providing a new catalogue of scientifically interesting objects imaged by *HST*, this work also demonstrates the power of the ESA Datalabs tool to facilitate substantial archival analysis without placing a high computational or storage burden on the end user.

2.2 Introduction

Interacting and merging galaxies are important to our current theory of ΛCDM cosmology, in which structure typically assembles hierarchically (Abadi et al., 2003; Springel et al., 2005; De Lucia & Blaizot, 2007; Guo & White, 2008). Galaxy interaction leads to highly disturbed morphologies (Toomre & Toomre, 1972; Hernández-Toledo et al., 2005; Wallin et al., 2016), intense starbursts (Mihos & Hernquist, 1996; Springel, 2000; Saitoh et al., 2009; Moreno et al., 2021) and, potentially, quenching of some systems (Hopkins et al., 2013; Smethurst et al., 2018; Hani et al., 2020; Das et al., 2023a). In general, galaxies undergoing interaction are observed to have higher star formation rates than those that exist in the field (Ellison et al., 2008; Scudder et al., 2012; Pearson et al., 2019b). Interaction also has a direct impact on the gas angular momentum within each galaxy, causing it to decrease. This, potentially, leads to funnelling of gas into into their nuclear regions and igniting activity. This could be a connection with active galactic nuclei (Ellison et al., 2008; Li et al., 2008b; Ellison et al., 2011; Comerford et al., 2015). However, such a connection remains debated (Alonso et al., 2007; McKernan et al., 2010; Marian et al., 2020). Thus, understanding galaxy interaction is crucial to testing theories of galaxy evolution itself.

Interacting galaxies have long been explored with different samples of galaxies. Examples include constraining merger rates as a function of redshift (Lotz et al., 2008b), inferring the contribution of minor mergers to the cosmic star formation budget (Kaviraj, 2014b,a), and examining interactions as a function of their local environments, internal properties and AGN activity (Darg et al., 2010b). These studies (and many others; for further examples, see Barton et al., 2000; Alonso et al., 2004; Ellison et al., 2013; Holincheck et al., 2016; Silva et al., 2021) illustrate the complex parameter space involved in understanding the role of interaction in

galaxy evolution. Thus, to effectively study interacting galaxies, we need observed datasets of such a size that they can sample a wide range of various parameters of interest.

The first large-scale catalogues of interacting galaxies are from the mid 20th century (Arp, 1966; Vorontsov-Velyaminov, 1959, 1977, hereafter VV). These catalogues primarily used visual inspection to identify mergers (e.g., de Mello et al., 1997; Nair & Abraham, 2010) and generally found from hundreds to thousands of systems. The largest set of interacting galaxies identified by a single expert classifier contains 2,565 relatively nearby systems (Arp & Madore, 1987). Citizen science techniques can extend this number, as was presented by Darg et al. (2010b) who used them to find a catalogue of 3,003 interacting galaxies.

The inclusion of automated classification shows promise to continue this expansion. The use of machine learning in classifying galaxy morphology is well established (Ardizzone et al., 1996; Abd El Aziz et al., 2017; Barchi et al., 2020; Ghosh et al., 2020; Cheng et al., 2021). The workhorse algorithm is the convolutional neural network (CNN; for an introduction, see O'Shea & Nash, 2015), most often used in image recognition and feature extraction. CNNs can be used for general classification (e.g. early- versus late-type galaxies) or to extract specific morphological features of galaxies, such as bars, spiral arms, etc; many works have demonstrated their effectiveness at this (e.g. Ackermann et al., 2018; Jacobs et al., 2019; Bickley et al., 2021; Buck & Wolf, 2021; Walmsley et al., 2022a). Pearson et al. (2022) demonstrated the power of CNNs for finding interacting and merging galaxies specifically, finding 2,109 in 5.4 $\deg^2$ of Hyper Suprime-Cam imagery - a large sample for the small area covered.

However, issues with using CNNs in classifying interacting galaxies have been found on numerous occasions. The primary concern, is that - without due care - classifying interacting galaxies by morphology alone can be highly contaminated. For example, CNNs often confuse chance alignments of galaxy pairs on the sky for interacting systems. This leads to many predicted interacting systems being thrown away after visual inspection (in some cases up to 60%; Bottrell et al. (2019); Pearson et al. (2022)).

In this work, we aim to use machine learning to create a large, high-confidence catalogue of interacting systems, drawn entirely from existing astronomical im-

agery. We search through the European Space Agency's *Hubble Space Telescope* Science Archive[1] using a CNN to predict whether an image contains an interacting system, from among the 126 million extended objects in the *Hubble* Source Catalogue (HSC; Whitmore et al., 2016). The feature extraction we implement is focused on finding tidal features or morphological disturbance caused by the interaction. The tidal features prioritised include tidal tails, tidal bridges or tidal debris. As stated previously, this runs the risk of introducing high levels of contamination by close pairs. We thus implement further automated and manual methods, which significantly reduce this. The systems we find are often in the background of previous deep surveys (such as the Cosmic Evolution Survey, COSMOS, Scoville et al. 2007; the Great Observatories Origins Deep Survey, GOODS, Giavalisco et al. 2004; and the Pancromatic Hubble Andromeda Treasury Survey, PHAT, Dalcanton et al. 2012), where spectroscopic coverage varies. Therefore, while our final catalogue reduces contamination to $\sim 3\%$, definitively removing all contamination by close pairs remains a challenge following this work.

This paper is laid out as follows: Section 3.2 describes the HSC and all the criteria we applied to create the images we predict over. This Section also introduces ESA Datalabs[2]; a new platform which allows the user to directly access the *Hubble* Science Archive. Section 2.4 gives an in depth description of the `Zoobot` CNN we utilise for our predictions, and how it differs from a commonly used CNN. Section 2.5 explains the process of creating the training set for our CNN to find interacting galaxies, with Section 2.6 showing how well it performed and providing the diagnostics of the CNN. We also use this Section to investigate the contamination in our catalogue. Section 2.7 describes our results and discusses the final catalogue as well as define interesting systems or objects that we have found. We also explore some basic properties of the catalogue here. Finally, Section 3.7 summarises our results and conclusions.

[1] See `http://hst.esac.esa.int/ehst/`
[2] `https://datalabs.esa.int/`

2.3 Data

2.3.1 The *Hubble* Archives & ESA Datalabs

The observational data is directly from the *Hubble* Science Archive and is accessed from the new ESA Datalabs platform. The repository contains approximately 100TB of data from the *Hubble Space Telescope* (*HST*). This repository spans all *HST* instruments and filters. ESA Datalabs provides a direct interface between users and the data. On this platform, every observations' FITS file can be accessed. To streamline our pipeline, we applied criteria to the observations as not all filters have the same number of observations, some instruments are not as sensitive to the low surface brightness regime as others or the field of view of certain instruments would not be ideal for measuring galaxy morphology. Finally, we do not conduct source extraction from each FITS file ourselves but use the *Hubble* Source Catalogue (Whitmore et al., 2016, hereafter HSC) to define the centre of each source cutout.

The criteria we apply are: the observational data must be from the Advanced Camera for Surveys (ACS), it must be final product data of *HST* (i.e. within a .drc file, where the data has been drizzle (Avila et al., 2015) combined and had charge-transfer-efficiency corrections applied), observed within the $F814W$ filter and must be flagged as an extended source in the HSC. This offloads sky subtraction, cosmic ray rejection and charge efficiency calculations to the original *HST* pipeline and removes costly steps from our cutout creation process. We utilise all final product data of the $F814W$ filter from *HST* as this was the filter which contained the most FITS files, and therefore observations. The $F814W$ filter contained 9,527 final product FITS files which could be used for source extraction, whereas the closest second (the $F606W$ filter) contained ≈6000. By using the filter with the most files, we are confident that we cover a majority of the HSC. Applying this criteria gives 126 million sources to predict over.

We must create 126 million source cutouts from 9,507 different FITS files. Creating a dataset of cutouts at this magnitude in conventional methods (such as `AstroQuery` or Table Access Protocol (TAP) services) would be impractical due to making many network calls and long FITS file download times. Instead, we

use the ESA Datalabs platform, which is due to be released in Q3 of 2023. This platform has been developed to allow us to 'mount' the *Hubble* Science Archive onto it. In practice, providing access to the entire *Hubble* Science Archives as local files for the user to manipulate while on the platform. This bypasses network calls to servers to download our required FITS files, a process which could have taken minutes per download. Having direct access to the files, and quickly matching source coordinates to FITS files (described in Section 2.3.2) allows us to open a FITS file and create all source cutouts from it without having to close or reopen it. Therefore, we were able to create on the order of 10k cutouts in the same order of time taken to download a single file.

The source cutouts were created as $F814W$ gray scaled 150x150 (7.5″x7.5″) pixel images using the HSC source coordinates as the centre. The image size was set and standardized to streamline the pipeline. The majority of cutouts are centered on the source but, in a minority, misalignment between source and image centre occurs. This is a result of the drizzling process, with incorrect alignment sometimes being significant. However, the target source was always present in the cutout and we, therefore, did not attempt to rectify this. A ZScaleInterval with a hard set contrast of 0.05 and a LinearStretch following the default parameters in the `Astropy` (Astropy Collaboration et al., 2013a, 2018a) package. These were binned to 300x300 pixels (pixel resolution is 3.25″x3.25″) with a linear interpolation from the `CV2` python package. The images were created at 150×150 to minimise storage required on the early version of ESA Datalabs being used. Creating the images at half the size allowed us to scale up to 300×300 pixels without any effects of the interpolation.

2.3.2 The Shapely Python Package

A large computational expense in our pipeline was matching FITS files to sources. Conventionally, the `Astropy` CONTAINS function would be used to match source coordinates to the FITS file WCS. We instead use the `Shapely`[1] Python package. `Shapely` is a geometry orientated package primarily focused on geospatial data. We found converting the FITS image footprints into `Shapely` Polygons and the

[1] `Shapely` docs: `https://shapely.readthedocs.io/en/stable/manual.html`

source coordinates to `Shapely` Points and then checking if they overlapped had significant speed up. Per iteration, Astropy's CONTAINED_BY function matches a source to a FITS file on the order of 500ms. Using `Shapely's` CONTAINS function, the same process is on the order of 6μs.

2.4 Utilising a Convolutional Neural Network

We must choose a CNN which would best suit our needs to classify them into interacting galaxies or not. We select the newly developed CNN `Zoobot` (Walmsley et al., 2022a, 2023). `Zoobot` is a CNN specifically trained to classify galaxies based on morphology into many different types (spiral, disk, elliptical, barred, non-barred, etc). We retrain it to only classify galaxies into interacting or non-interacting. Instead of training `Zoobot` from scratch and creating a new model, we use transfer learning to finetune existing `Zoobot` models to classify our data for our particular question. This allows us to retain information from `Zoobot's` previous training. More importantly, it requires a significantly smaller training set to achieve high accuracy.

2.4.1 Zoobot

The version of `Zoobot` we use is a deep CNN which was trained on Galaxy Zoo volunteer classifications over three different Galaxy Zoo: DECaLS (GZD)(Dark Energy Camera Legacy Survey, described in Dey et al., 2019) campaigns. These were GZD-1, GZD-2 and GZD-5 - each number corresponding to the DECaLS data release. For training `Zoobot`, DECaLS imaging was selected using the NASA-Sloan Atlas (NSA), which was itself constructed with SDSS Data Release 8 (DR8) images. This also introduced implicit cuts to the training data, as SDSS can not get to the depths of DECaLS. This introduces implicit magnitude and redshift cuts on the training data. Specifically, SDSS DR8 and the NSA cover galaxies brighter than $m_r > 17.77$ and closer than $z < 0.15$. In Section 2.4.2 we describe using transfer learning to use Zoobot effectively outside of this magnitude and redshift range.

Walmsley et al. (2022a) use the 249,581 volunteer classifications from GZD-5 campaign to train `Zoobot` to answer all 34 questions (example shown in Figure 4 of Walmsley et al., 2022a) in the remaining campaigns. GZD-5 was used as it had a slightly different volunteer decision tree, having an expanded question on potential different galaxy merger stages. Each galaxy image had been shown to volunteers as a 3-colour (g,r,z) of 424×424 cutout. Each images pixel scale was an interpolation between the measured Petrosian 50%- and 90%-light radius. The measured full Petrosian radius had to be at least 3″ to be shown to the volunteers. When inputting into `Zoobot`, these cutouts were scaled and grayscaled to 300×300×1 images, averaging over the 3-colour channels to remove colour information and avoid biasing the morphology predictions. `Zoobot` utilised the Adam (Kingma & Ba, 2014) optimizer to train.

By training `Zoobot` in this way, combining the approach of answering many questions at once with Bayesian representation learning, it learns a generalisable summary of many types of galaxies. These generalised summarys are lower-dimensional descriptions of galaxy types and are referred to as representations. These representations change depending on the galaxy type, morphology or environment in an image and lead to similar images being closer together in a representation space than dissimilar ones. This representation approach on a very broad classification problem is found to increase accuracy and generality of `Zoobot`, giving it an edge over conventional CNNs. A more detailed breakdown of this approach, as well as further details about `Zoobot`s' architecture, can also be found in Walmsley et al. (2022a).

`Zoobot` was trained to give a prediction score to an image of a galaxy based on the question it is answering. The type of prediction score is set by the users choice of the model final layer in `Zoobot`. We elect to use a SoftMax output, which returns an output score as a float between 0 and 1. This prediction score is not a probability score, although it may seem analogous. A well behaved prediction score will map to probability, though not necessarily linearly. The mapping between prediction score and probability is not considered in this work, and we use the prediction score as an indicator of `Zoobot`'s confidence a source is an interacting galaxy.

We are only interested in the 'Is the galaxy merging or disturbed?' question from the Galaxy Zoo: DECaLS workflow, where the answer can be 'merging', 'major disturbance', 'minor disturbance' or 'None', and only want our version of `Zoobot` to return the answer to this. Our version of `Zoobot` is also not trained to predict over *HST* data which differs from DECaLS data (different resolutions, filter bandwidths, etc). If we were to use our version of `Zoobot` as downloaded we would likely lose accuracy. We utilise transfer learning to optimise accuracy of just our question as well as to classify *HST* data. Since this work, `Zoobot` has been trained on *HST* data so the transfer learning step would not be needed in future with the new models. How we apply transfer learning is discussed in the following Section, but an excellent review and discussion of applying transfer learning for detecting galaxy mergers can be found in Ackermann et al. (2018).

2.4.2 Transfer Learning

Transfer learning (or finetuning) is a method of applying the same machine learning model to a similar problem that it was originally trained on. Rather than having to completely retrain all parameters in a model and essentially create a new one, we can use the original model architecture and the parameters it has learned from its previous training. In the case of `Zoobot`, we keep the parameters it has learned from training on the DECaLS dataset and freeze all sections of the model responsible for feature extraction and recognition.

We construct a classification section that maximises accuracy and only allow the weights of this section to change. As the classification section has fewer parameters than the feature extraction section (the classification section contains 86,209 parameters compared to the feature extraction sections' 4,048,989 parameters) we need significantly less data to completely retrain it (in our case, a factor of 15 less). Once this retraining is complete, the weights of the feature extraction sections of the model can be unfrozen and tweaked using our smaller dataset with a very low learning rate to further boost overall model accuracy.

An example of taking an existing model and applying it to a new problem with transfer learning is shown in Walmsley et al. (2022b). Here, they take the trained model and finetune it to finding ring galaxies. They retain an accuracy of

89% while only needing to train the model on 10^3 ring galaxies. This significantly reduces computational expense and training time of the model, while keeping the required training set very small. Interacting galaxies are rare, and interacting galaxy catalogues not expansive. So retraining the full network on hundreds of thousands of interacting galaxies is not feasible. Using transfer learning, and following the example from Walmsley et al. (2022b), we only need to create a training set of 10^3 - 10^4 interacting galaxies to achieve an accuracy of ≈90%.

2.5 Creating the Training Set

We create a large training set of interacting galaxies following the criteria described in Section 3.2 to train our model. Therefore, we need a large, labelled set of interacting and non-interacting galaxies. We elect to follow the methodology of finetuning as described in Walmsley et al. (2022b), and aim to create a balanced training set. This has the advantage that it significantly improves the performance and accuracy of machine learning classifiers, but the disadvantage that it can bias our final model if few interacting galaxies exist compared to the general population. However, such a bias will be mitigated by using a high prediction cutoff to define an interacting galaxy. This is discussed in in Section 2.6.1. To create this large training set we use the Galaxy Zoo collaboration (initial data release described in Lintott et al., 2008).

2.5.1 Interacting Galaxies and Galaxy Zoo

The data in Galaxy Zoo is volunteer classifications on galaxy images spanning multiple projects. We incorporate classifications from all major Galaxy Zoo projects; Galaxy Zoo 1 (Lintott et al., 2008), Galaxy Zoo 2 (Willett et al., 2013), Galaxy Zoo: *Hubble* (Willett et al., 2017), Galaxy Zoo: CANDELS (Simmons et al., 2017) and Galaxy Zoo: DECaLS (Walmsley et al., 2022a). These projects contain a total of 1,367,760 labelled galaxy images that we must extract the interacting galaxies from. We only use labels that are from citizen scientists, and no labels generated by previous versions of `Zoobot`. We apply three criteria to each

interacting or non-interacting label. Firstly, it must have greater than 20 volunteer votes on it. Applying this allows us to use a statistically robust weighted vote from a crowd answer rather than trusting any volunteers individually. Secondly, the calculated weighted vote (i.e. the combination of the 20 or greater votes) must then be greater than 75% in favour of being an interacting galaxy or less than or equal to 25% for it not to be; this ensured purity in our training set. If the question given to volunteers was more specific (such as 'Is this a minor disturbance?' and 'Is this a major disturbance?') then if either answer was the majority vote we classified it as an interacting galaxy. Thirdly, the object must exist in the *Hubble* footprint so that we could make a cutout of it.

Checking if each training source existed in the *Hubble* footprint was only possible in an efficient way because of ESA Datalabs. Rather than having querying every coordinate and make network calls to TAP services, we extract every final product $F814W$ observation footprint and check if each labelled galaxy exists in at least one file. We make this check by creating a `Shapely` Polygon for each observational footprint and a `Shapely` Point for each labelled galaxy central coordinate. Using the `Shapely` Polygon CONTAINS function, we check if a labelled galaxy's Point overlaps with an observations' footprint Polygon. This returns a list of files which contain the training source. If a training source was not found in any observational footprint we discard it. We make no attempt here to check if our sources have other photometry available to them, and only create 1-colour images with the $F814W$ data. We provide the images to `Zoobot` as 1-colour grayscaled cutouts.

Upon applying these criteria we find 3,167 labelled interacting galaxies in Galaxy Zoo: *Hubble* project, the largest contribution to our training set. These were paired with 3,167 labelled non-interacting systems (following the previous criteria) to balance the training set. From all other projects, we find 869 labelled interacting systems which fitting the creation criteria. The primary limiting factor for Galaxy Zoo's 1 and 2 was that many found interacting galaxies did not exist in the Hubble footprint. For Galaxy Zoo: CANDELS and Galaxy Zoo: DECaLS the limiting factor was the required calculated weighted vote. These labelled interacting systems were then paired with 869 labelled non-interacting systems,

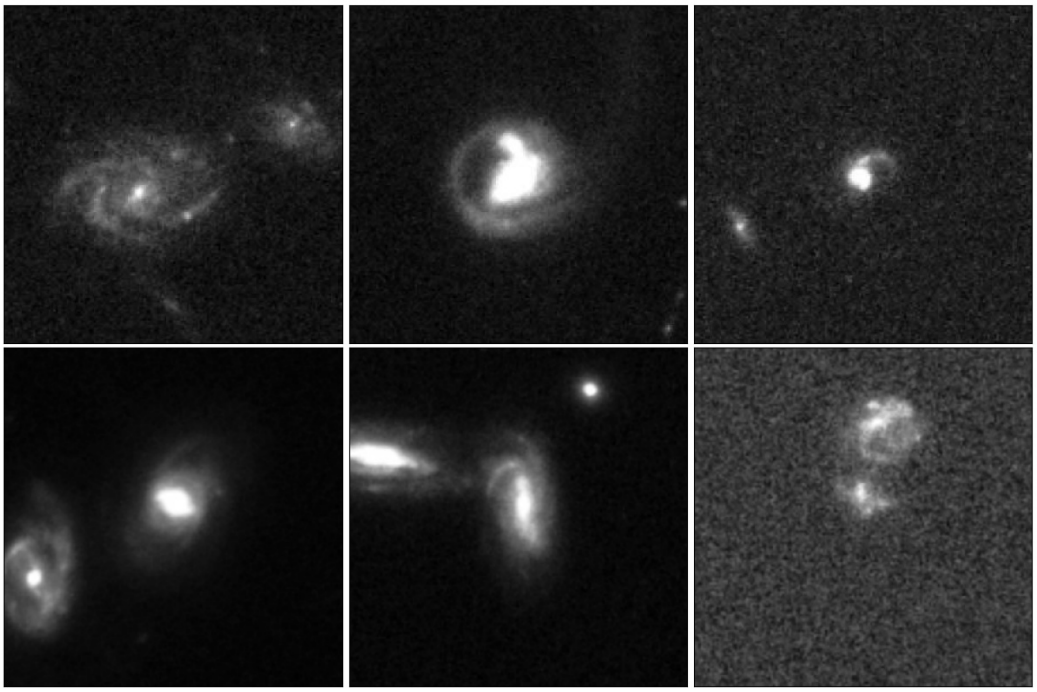

Figure 2.1: Example images of the labelled interacting galaxy systems used to train `Zoobot`. Each galaxy had a weighted vote fraction $\geq$0.75 in Galaxy Zoo. *Top Row*: Three examples from the Galaxy Zoo: *Hubble* project of the training set. *Bottom Row*: Three examples from the other Galaxy Zoo projects. These are, from right to left, Galaxy Zoo 2, Galaxy Zoo CANDELS and Galaxy Zoo DECaLS. The priority with this training set was that the interactors had clear tidal features and disruption so `Zoobot` would learn to highly weight them and not misclassify close pairs.

ensuring that each labelled non-interacting system came from the same project as its labelled interacting system counterpart.

Each of these projects has a varied redshift range: Galaxy Zoo: *Hubble* is $z < 1$, Galaxy Zoo: CANDELS $1 < z < 3$ and Galaxy Zoo's 1, 2 and DECaLs are $z < 0.15$. This introduces a redshift bias into our model, where the morphology and brightness of interacting sources changes with a $z > 1$. This is only partially rectified by including Galaxy Zoo: CANDELS, which provided 322 labelled interacting systems.

From all Galaxy Zoo projects, we find a training set of 4,036 labelled interacting galaxies and combine them with their matched 4,036 labelled non-interacting galaxies giving a total training set size of 8,072. Figures 2.1 and 2.2 show six

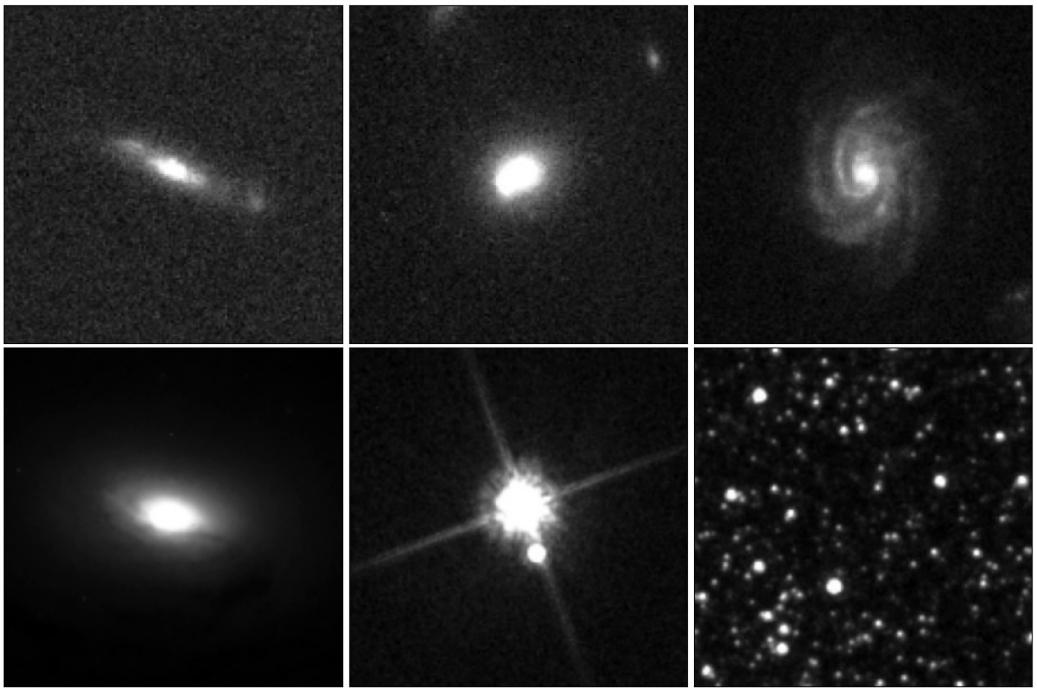

Figure 2.2: Example images of the labelled non-interacting galaxy systems used to train `Zoobot`. *Top Row*: Three examples from the Galaxy Zoo: *Hubble* project of the training set. *Bottom Row*: Three examples from the other Galaxy Zoo projects. These are, from right to left, Galaxy Zoo 2, Galaxy Zoo CANDELS and a starfield from the active learning cycle. Starfields/globular clusters/open clusters existed throughout the HSC flagged as extended sources. 1,000 images of starfields were added to the training set so `Zoobot` would give them a very low score.

examples of our labelled interacting and non-interacting galaxy training set. As we require `Zoobot` to learn to weight tidal features or disturbances highly, it is important that such structures dominate the training set. Previous works, such as Pearson et al. (2019b), have found that final catalogues produced by CNNs are often heavily contaminated by sources which are simply close pairs by projection effects and chance alignment in the sky. By focusing our CNN on tidal features, we aim to minimise this contamination. We ran an initial test of the prediction pipeline on the first 500,000 sources that had been created from the HSC to initially test our `Zoobot` model. We investigate any source which was given a prediction score ≥ 0.75 and, to further increase the size of our training set, conduct one step of active learning.

2.5.2 One Active Learning Cycle

To enlarge our training set further, we conduct one step of active learning to find interacting galaxies. An active learning cycle involves an 'expert' checking the predictions made by the model, correcting any incorrect predictions and then feeding it back into the model as additional labelled images to a training set. We complete finetuning of `Zoobot` on our initial training set of 8,072 galaxies and make predictions on the first 500,000 sources from the HSC (created under the criteria previously discussed). We visually inspect the sources `Zoobot` gives a prediction score ≥ 0.75 and correct any wrong predictions. These corrected labelled sources and those `Zoobot` correctly labelled are then added to the training set. Not only does this step allow us to add more labelled interacting galaxies to the training set, but it also allows us to evaluate `Zoobot`'s behaviour and check if it consistently predicts a type of source or galactic morphology incorrectly.

From the first 500,000 sources, a total of 6,198 sources were given a prediction score of $\geq$0.75. We correct the predictions `Zoobot` made and balance this set to 5,698. During this cycle, a large number of globular clusters/starfields/open clusters were given a very high prediction score. Figure 2.2 shows an example of these contaminating star fields. We created sources of 1,250 star fields and added these into the training set, labelling them as non-interacting. Adding the balanced 5,698 sources plus the 1,250 starfields to our training set gave us an

unbalanced training set of 15,020 sources. To then balance the training set, we took 1,250 labelled interacting galaxies from the Galaxy Zoo: *Hubble* project and made random image augmentations with the `TensorFlow` Python package. These augmentations were simple rotations, cropping and resizing. With these extra sources, our training set contains 16,270 sources. Of these, 50% (8,135) were labelled images of interacting galaxy systems.

2.6 Diagnostics

2.6.1 Model Performance

Upon finetuning `Zoobot` we validate its performance. We reuse the validation set that `Zoobot` automatically creates when training. This set is created by putting aside a random set of 20% of the training set. `Zoobots` then uses it to validate its performance in training. We record which images `Zoobot` selected, and extract these from the training set for further diagnostics. This provides us with a validation set of 3,270 images, containing 1,648 non-interacting galaxies and 1,622 interacting galaxies.

`Zoobot` gave a prediction score between 0 and 1 to each of the validation images, Figure 2.3 shows the resulting distribution. This distribution shows that our model has high confidence in what is or isn't an interacting system due to the high counts at very low and very high probability scores. It is likely the use of a balanced training set, and the very low volunteer score needed to define a source as non-interacting that leads to a strongly bi-model prediction score distribution. Using a balanced training set is an intrinsic trade off between ease of training, and potential biases introduced. Having a balanced dataset does not reflect reality, and leads `Zoobot` to over-predict interacting galaxies. Using very stringent volunteer classification cutoffs also leaves few ambiguous systems in the validation set, further enhancing this bi-modality.

The prediction score must be reduced to a binary classification for our problem. We use Figure 2.3 to define a prediction score above which a source is

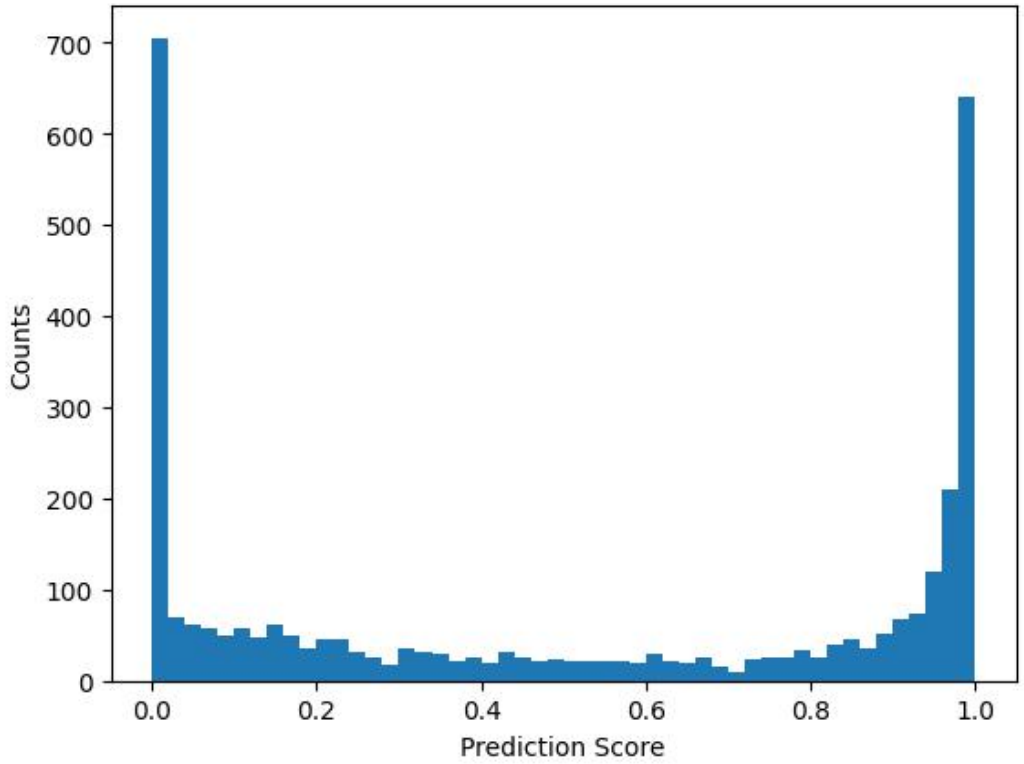

Figure 2.3: The distribution of prediction scores given to our validation set of 3,270 labelled sources set aside by `Zoobot` in training. These were split into 1,648 non-interacting sources and 1,622 interacting sources. As can be seen from the distribution, our model is often confident when a source does or does not contain an interacting galaxy by the strong bi-modality. This is likely due to the very stringent vote weightings used when selecting the training set. Using this distribution, we decide the prediction score to use as a cutoff to give us our final binary classification: interacting galaxy or not.

classified as an interacting galaxy. We measure the accuracy of `Zoobot` for different cutoffs, where the accuracy is the fraction of labels correctly predicted over the total number of labels predicted on. Figure 2.4 shows this change in accuracy. We find that our model is most accurate with a prediction score cutoff of 0.55 with an accuracy of 88.2%. Figure 2.4 also shows the change in the purity of our catalogue with changing prediction cutoff. Here, purity is the ratio of number of true interacting galaxies to total sources in the final catalogue. These scores can be combined into the F1 score of our model, shown in Figure A.2 in the Appendix.

Figure 2.5 also shows a measure of accuracy for our model at different cutoffs using confusion matrices. Importantly, it also shows how our model is getting labels wrong: either giving false positives (where a labelled non-interacting galaxy is predicted to be interacting) or false negatives (where a labelled interacting galaxy is predicted to be a non-interacting). The number of incorrect positive and negative predictions change based on the prediction cutoff, with a very low cutoff giving many false positives and a very high cutoff giving many false negatives. Figure 2.5 shows that with a cutoff of 0.50, we would return a high level contamination in our final catalogue. Of the 1,622 galaxies predicted to be interacting, 218 would be non-interacting systems - approximately 13%. Our main aim in this work is to present a highly pure, large interacting galaxy catalogue that can be used for statistical exploration of interacting galaxy parameter space. Therefore, we use a very stringent cutoff of 0.95.

Using a cutoff of 0.95 reduces contamination significantly. Figure 2.5 shows the final contamination in our validation catalogue would be $\approx$2%, where Figure 2.4 shows that we are maximising the purity in our sample at the expense of accuracy. The aim of this work is not to create a general tool to be used by the community, but to find a large catalogue of interacting galaxies. As we are investigating 126 million sources, despite removing $\approx$50% of interacting galaxies from the final catalogue, we are certain that we can find a catalogue larger than previous works.

Using such a high cutoff also reduces any risk of any biases introduced by using a balanced training set. While using such a training set often increases the accuracy and speeds up training, it can bias the model towards one conclusion.

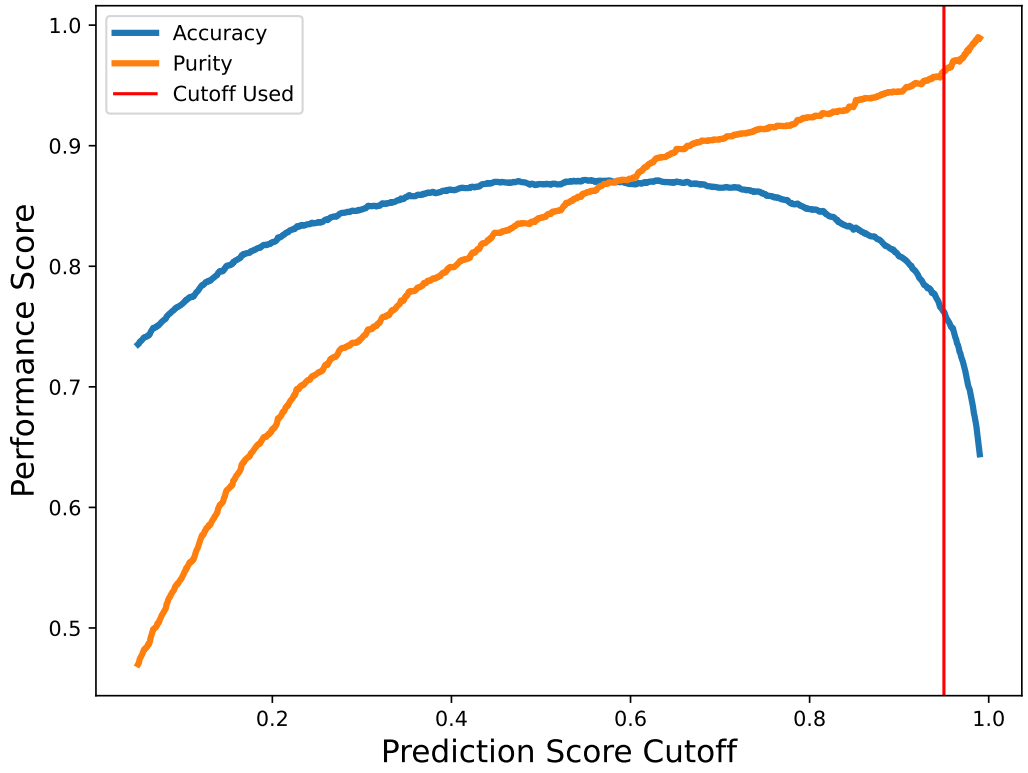

Figure 2.4: A measure of accuracy and purity against prediction score. The accuracy (in blue) is a direct measure of the number of sources `Zoobot` correctly predicted vs the total number of predictions made. The measure of purity (in orange) is the the number of predictions `Zoobot` correctly made vs the total number of predictions for an interacting galaxy. The cutoff score (in red) shows the point above which we would define an interacting galaxy and below which we would not. At this point, the accuracy appears lower due to `Zoobot` making many false negative predictions while successfully making true negative predictions. This is confirmed by the maximisation of purity. Due to the number of sources `Zoobot` is predicting over, the size of the catalogue will exceed any previous catalogues. Therefore, we use this very conservative cutoff to maximise purity over the completeness of our catalogue. These measures can also be shown with the F1 score. Figure A.2 shows this change with prediction cutoff in the Appendix.

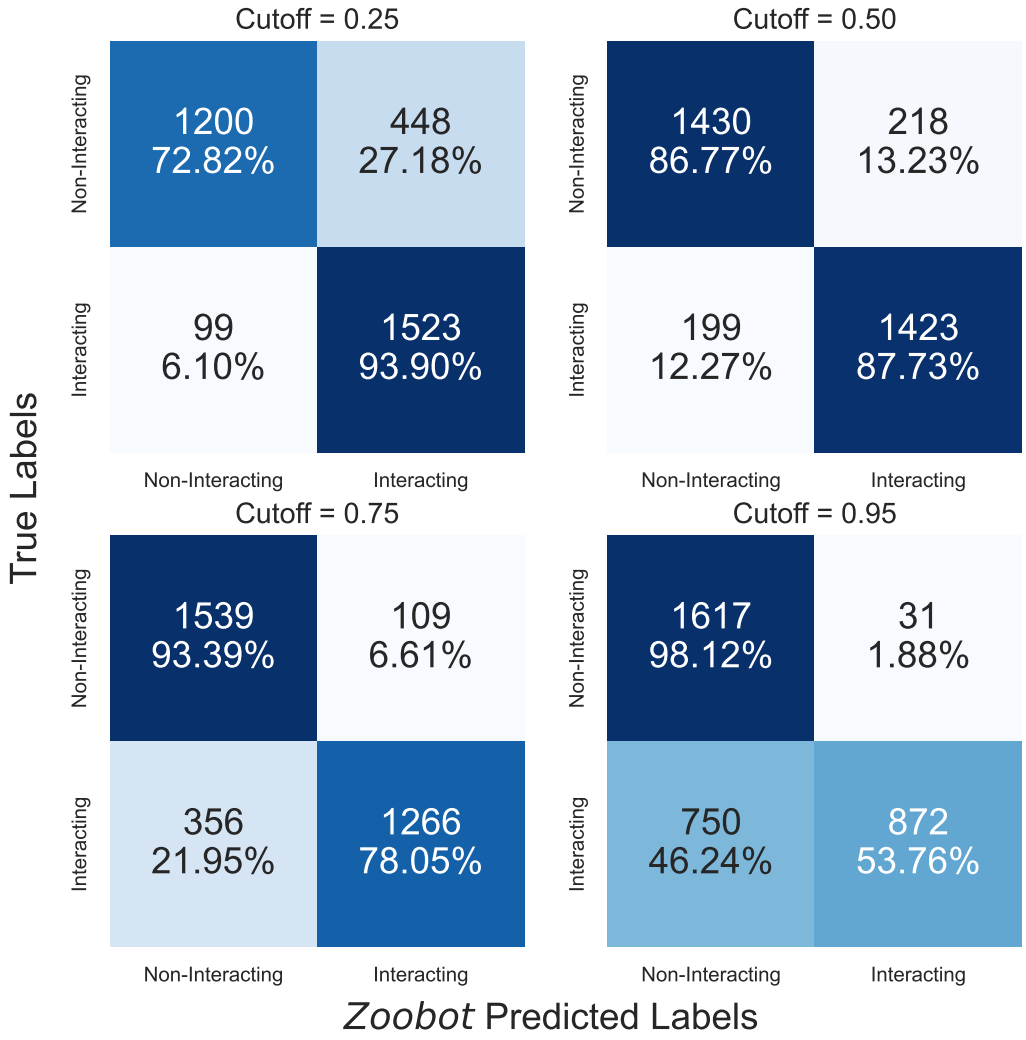

Figure 2.5: Confusion matrices of four different cutoffs of prediction score defining a binary classification of interacting galaxy or not. Confusion matrices break down our accuracy measurement into how `Zoobot` is misclassifying sources. At a cutoff of 0.50, the accuracy is highest at 88.2%. However, at this cutoff, ≈10% of our final catalogue would contain contamination. We elect to use the very stringent prediction cutoff of 0.95 for the rest of this work as it will return the lowest contamination.

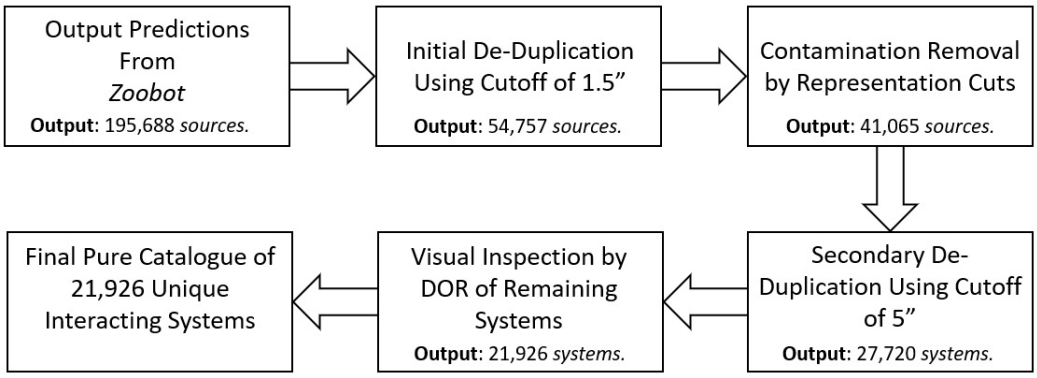

Figure 2.6: Flow diagram of our contamination and duplication removal process. De-duplication used agglomerative clustering based on sky separation. The first step of de-duplication uses a cutoff of 1.5″. This significantly reduced duplication in the catalogue, as well as the size of the catalogue to 54,757 interacting galaxies. We then applied contamination removal to this de-duplicated catalogue. Upon visual inspection, a small number of duplicated systems still existed in the catalogue. To ensure a pure catalogue of unique systems, we applied a agglomerative clustering again with a cutoff of 5″. This gave us a catalogue of 27,720 unique interacting systems. The final step to ensure purity was visual inspection by DOR, removing any remaining contamination. This gave the final pure catalogue of 21,926 unique interacting systems.

In our case, the true rate of interacting galaxies will be much smaller than 50%. Therefore, our model will be biased to labelling a source as an interacting galaxy. This will be particularly true for edge cases, which could be ambiguous to even an expert classifier. By using such a high cutoff score, this bias will be mitigated by only labelling the most clearly interacting objects as interacting.

2.6.2 Duplication Removal

The fully trained `Zoobot` made predictions on ≈126 million extended sources from the HSC that had passed our creation criteria. Of these, 195,688 sources were given a score of 0.95 or greater, ≈0.2% of the total number of sources. Upon visually inspecting a subset of sources, it is clear that our `Zoobot` model had predicted for an interacting galaxy even if it was not the central (and, therefore, target) source in the image. This is due to the misalignment of sources from the centre in the training set as described in Section 2.5. `Zoobot` learned to classify

an image as an interacting galaxy if it contained one, and not just if it was the central source. Therefore, many interacting systems were duplicated in our final catalogue, appearing in cutouts were the central source was not interacting.

Another source of further duplication was the HSC itself. In the HSC, many extended objects have multiple source IDs applied to them. This is due to bright clumps in extended sources being assigned a new ID, sources which had been found but did not exist in reality or background sources which existed in extended systems. We find that of the 195,688 Source IDs given a prediction score of 0.95 or greater, approximately 3.6 Source IDs were matched to a single real object. To refine the catalogue and remove the duplication we use spatial clustering of each source with agglomerative clustering (an introduction and description of hierarchical clustering, including agglomerative clustering, can be found in Nielsen, 2016).

Agglomerative clustering is a method of hierarchical clustering based on a distance metric between the sources. We set the maximum distance between points to define a cluster. i.e. any sources within a defined distance on the sky from each other will be merged under one source ID. This approach means we do not need any knowledge of how many cluster of sources exist in the dataset or the level of duplication within it, as would be the case in many other clustering approaches. We create distance matrices of the angular separation of every source using the `Astropy` Python package. These projected sky separations are then used as a euclidean distance in the clustering algorithm with an EUCLIDEAN_LINKAGE. The new ID of a cluster is the first source ID in the cluster.

Initially, we utilise a limiting sky separation of $1.5''$ to remove the duplication. This reduced the size of our potential catalogue to 54,757 interacting galaxy candidates. We then applied contamination removal as described in Section 2.6.3. Once contamination removal was completed, the catalogue size was 41,065 interacting galaxies. Visual inspection found further duplication, so our initial de-duplication had not been aggressive enough. To ensure the catalogue was of unique systems, we opted to use a final aggressive limiting sky separation of $5''$ completely removing the duplication in our catalogue. This aggressive de-duplication further reduced the size of our catalogue to 27,720 candidate interacting systems. However, we could be certain that each of these candidate

systems was unique. Figure 2.6 shows a full breakdown of the steps in our de-duplication and contamination removal process.

2.6.3 Bad Predictions & Removal

After the initial step of de-duplication we begin removal of contamination from the catalogue. A major, and expected, source of contamination is by close pairs of galaxies. These are systems where chance alignment in the sky appears that galaxies are close together but are actually at different redshifts. Other sources of contamination include large central galaxies with satellite galaxies about them, star fields with extended sources in them and objects with strange morphologies that `Zoobot` predicted were tidal features.

Upon applying the clustering by sky projection of $1.5''$, the catalogue contained 54,757 candidate interacting galaxies. Our primary concern is contamination by close pairs. Creating catalogues of interacting galaxies with CNNs are notorious for suffering from this problem, where a significant number of candidates must be removed from otherwise large final catalogues (Bottrell et al., 2019; Pearson et al., 2022). The decisive way to remove this contamination is to compare redshift measurements of each galaxy in the candidate interacting system. However, this is impractical for our catalogue where the majority of candidates have no redshift measurements. To find close pairs, and remove them effectively, we take advantage of the representations `Zoobot` learns of each image. As described previously, `Zoobot` was trained to answer every question in Galaxy Zoo: DECaLS simultaneously for every galaxy. It therefore learns a generaliseable representation of many kinds of galaxies. In this representation space, morphologically similar galaxies will exist close together in clusters while those that are dissimilar will be further apart. We extract the features `Zoobot` has learned of each candidate, and plot its representation.

We remove the classification head of `Zoobot` and directly output the final layer of the feature learning section of the model. This gives 1,280 features (the representations) for each of our 27,720 candidate systems. However, there will be much redundant information in this very high dimensional feature space. We compress this using incremental principal component analysis (PCA) (Ross et al.,

2008). An excellent demonstration of using this approach can be found in Walmsley et al. (2022b). We reduce the dimensionality from 1,280 to 40(as in Walmsley et al. (2022b)), and input the resultant components into the Auto-Encoder `UMAP` (McInnes et al., 2018). `UMAP` projects the 40 dimensional components of each candidate system onto a 2 dimensional manifold. The position of each galaxy on this manifold is directly linked to its visual morphology. Close pairs have similar visual features which will then appear as a cluster in our representation space.

Figure 2.7 shows the representation distribution of our 54,757 candidates after compression with `UMAP`. A random image in each bin has been selected to show the morphology of the objects within the bin. There are three clear gradients that exist in the representation distribution: one of source size, one of the source inclination and one of image contrast between the source and the background. The gradient of source size is clear from left to right. This is also true of contrast between the source and background. The gradient of source inclination is from top to bottom. The top shows very inclined sources, and even the diffraction spikes of stars, while along the bottom we find face on sources which take up a larger part of the cutout centre. At the very bottom of the figure (away from the main body) a cluster of very poorly contrasted sources with the background that are face on are found. The gradients of inclination and source size are expected while that of contrast is less so. This gradient is likely a result of how we created our images using a Linear Stretch with fixed contrast. The effect of this is that dimmer sources have brighter backgrounds, a particular issue at high redshift.

Figure 2.7 has many areas of similar morphology. On the left, we have isolated objects: disturbed spirals or large galaxies with tidal disturbance to them. Along the bottom, we see isolated bright objects with satellites about them. On the bottom right, we see our area of representation space dominated by close pairs. In the centre, we see the population of interacting galaxies that `Zoobot` was trained to find. The areas of representation space which are dominated by clear sources of contamination are cut. Figure 2.8 shows a scatter plot of the representation distribution and the cuts we make. They are made such that any source with a Y Mapping of $-2 \leq Y \leq 4.75$ will be kept in the catalogue. The choice of these cuts has been made by eye, and then bootstrapping the remaining images to check

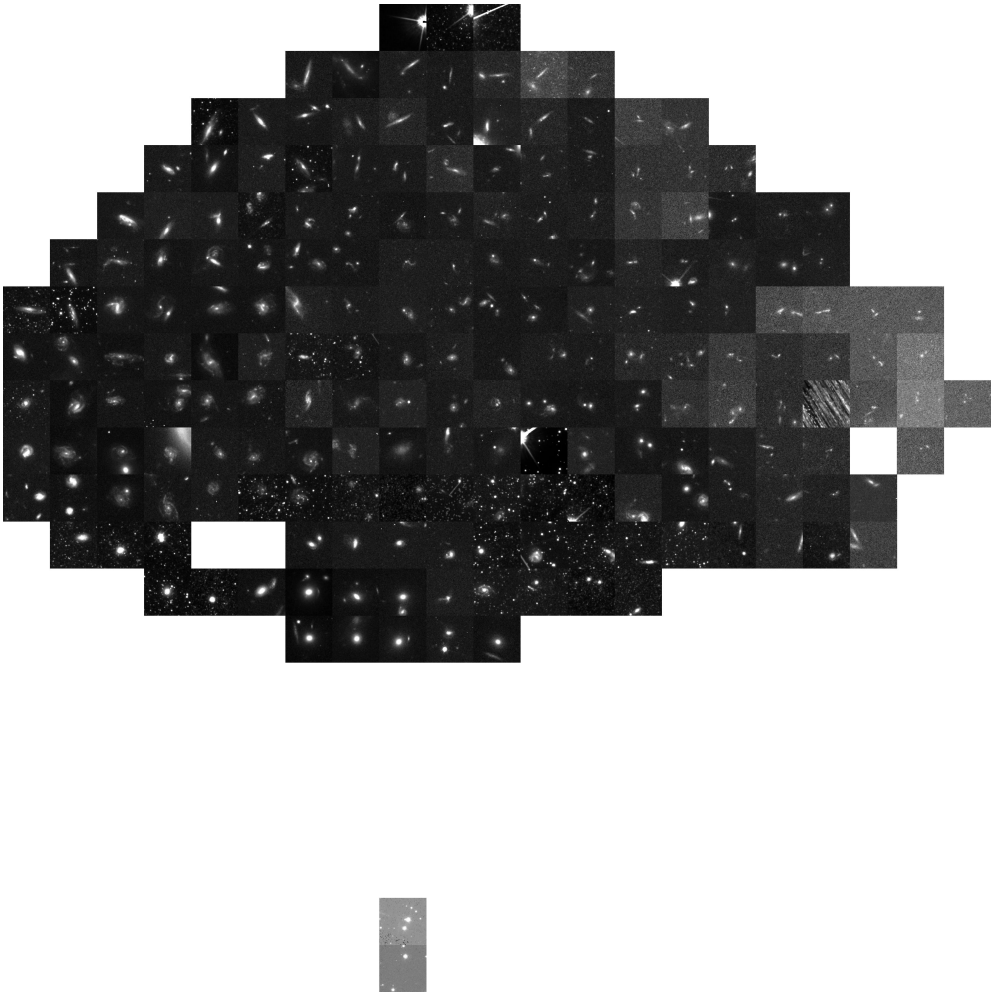

Figure 2.7: The representation distribution of 54,757 candidate interacting galaxies. This distribution is the compressed 2D representation of the 1,280 dimensional representation that `Zoobot` has learned of each image. Each image is a randomly selected one from sources within each bin in the distribution. The X and Y axis on this plot are the 2D mapping on the manifold given by UMAP for the 40 dimensional principal components of each source, and not physical parameters. Three gradients are clear in this distribution: first; from the left to right there is a distinct gradient in the contrast of the images. The images to the left are local galaxies with low redshift, while those on the right are dimmer sources at much higher redshift. This is an effect of how the images are created using a linear scaling function and a fixed contrast. The second feature, also from left to right, is a gradient of larger source size to smaller source size. This is a feature `Zoobot` has learned based on the redshift of the source as well. The third, from top to bottom, is a gradient of the inclination of the source. With the most inclined (and even diffraction spikes) of the sources appearing at the top, while at the bottom the sources are face on. Along the bottom of the representation plot, there are close paired sources as well as many star fields. Along the very top, there is contamination in the form of isolated stars in star fields. Thus, we make aggressive cuts along the top and bottom of our representation space to remove as much contamination in a general way. The full representation plot, with all sources and the cuts, is shown in Figure 2.8.

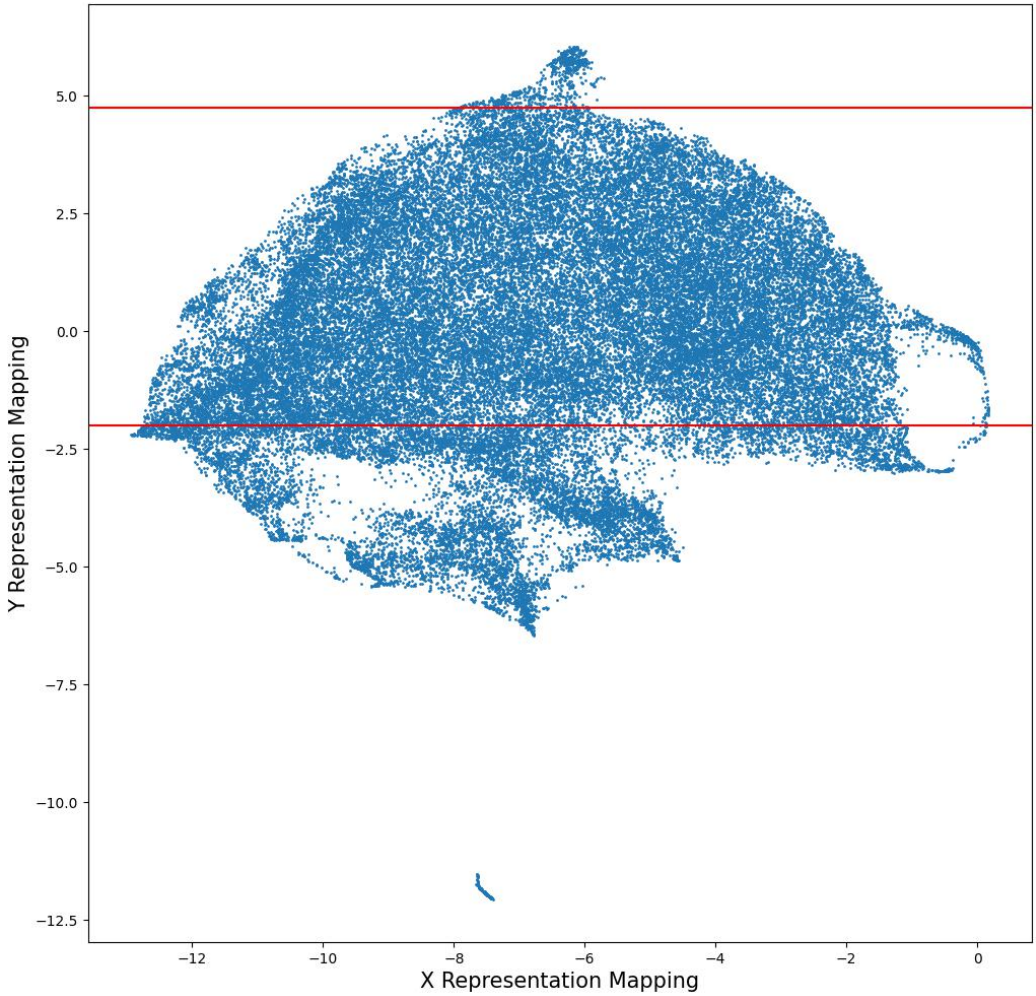

Figure 2.8: Scatter plot showing the precise distribution of each representation of sources in the remaining 54,757 sources. This is the unbinned version of Figure 2.7. The two red lines show the cutoffs utilised to remove the majority of close pairs by projection as well as the very obvious contamination of stars and stellar fields at the top of the representation distribution. The number of candidate interacting systems in the catalogue was reduced to 41,065 systems.

contamination removed. After applying these cuts, we retain 41,065 systems in our catalogue.

We estimate $\approx 25\%$ of sources in the greater than 0.95 prediction bin are close pairs. This may seem lower than previous works, but is due to our very conservative prediction cutoff. The general cuts to our population based on their position in representation space makes it very likely that we retain some close pairs in the catalogue, while also removing interacting galaxy systems.

As described in Section 2.6.2, we then apply a $5''$ to the 41,065 remaining candidates, further reducing our catalogue to 27,720 systems. With such an aggressive sky projection cut, many individual interacting galaxies are now identified under the same ID as the secondary galaxy in the system. To remove remaining contamination in the catalogue, a final visual classification step was conducted. This visual inspection was conducted by DOR. Any systems removed at this stage were classified into three categories: interacting system, contamination and gems. The gems sub-category became necessary as many sources of contamination that were being removed were objects of other astrophysical interest, and is described in Section 2.7.2.

2.7 Results & Discussion

2.7.1 An Interacting Galaxy Catalogue

Upon de-duplication and contamination removal described in Sections 2.6.2 and 2.6.3, our final catalogue contains 21,926 interacting systems. Figure 2.9 shows a random sample of 50 of the systems from our catalogue. In these examples we can see highly distorted or currently interacting systems, precisely what we trained `Zoobot` to highly predict. Some cutouts are of the full interacting system, containing both the primary and secondary galaxies in the interaction. Some source cutouts only show one of the interacting galaxies, though these systems remain highly disturbed. Due to the constraints in our training set, so highly weighting disturbance or tidal features in our predictions, we are sampling interaction from all epochs except the approach to the initial pass. At this initial stage, there

will be no tidal features formed or disturbance in the disks as the two galaxies approach each other. Separating them from close pairs would be difficult without kinematic or redshift information, not available for the majority of these sources.

We investigate which of the systems in our catalogue have previous references in the astrophysical literature. To search the literature, we use the `AstroQuery` Python package with a coordinates based search of cutoff radius $5''$. We search the astronomical databases Simbad (Wenger et al., 2000), the NASA Extragalactic Database (NED; Helou et al., 1991) and ViZieR (Ochsenbein et al., 2000) for references to our interacting systems. These return either a list of references, or an empty list showing no references associated with the system. We find that 7,522 of our systems have at least 1 reference associated with them, while and 14,404 do not. A flag exists in the catalogue data release which shows whether a system has references associated with it or it could be considered a 'new' system. We, however, do not claim that these systems are discovered by ourselves. These systems have always existed in the backgrounds of large surveys or observations and been discovered by others, it is only with ESA Datalabs that we can apply a methodology such as in this work to extract those systems from these observations. We also do not claim that these unreferenced systems are particularly interesting or phenomenal. It is most likely that these systems are the very faint background galaxies in surveys or observations whose main objective was something other than finding interacting galaxies. This will be further discussed in Section 2.7.3.

Figure 2.10 shows the distribution of our catalogue in the sky. The *HST* is able to observe the majority so the catalogue sources are scattered throughout it. We find that the sources cluster in different parts of the sky which correspond to major surveys conducted using the *HST* involving ACS/WFC and the $F814W$ filter. We also mark the centres of the seven surveys which correspond to the major clustering of interacting systems in the sky. These were the COSMOS, the GOODS North, GOODS South, PHAT, CANDELS, AEGIS and Spitzer Space Telescope FLSv Region (Morganti et al., 2004) surveys.

The full catalogue and data product are found on Zenodo at the following DOI where it is freely accessible to the community: doi:10.5281/zenodo.7684876. Table 2.1 shows an example of the data and format of the 50 sources shown in

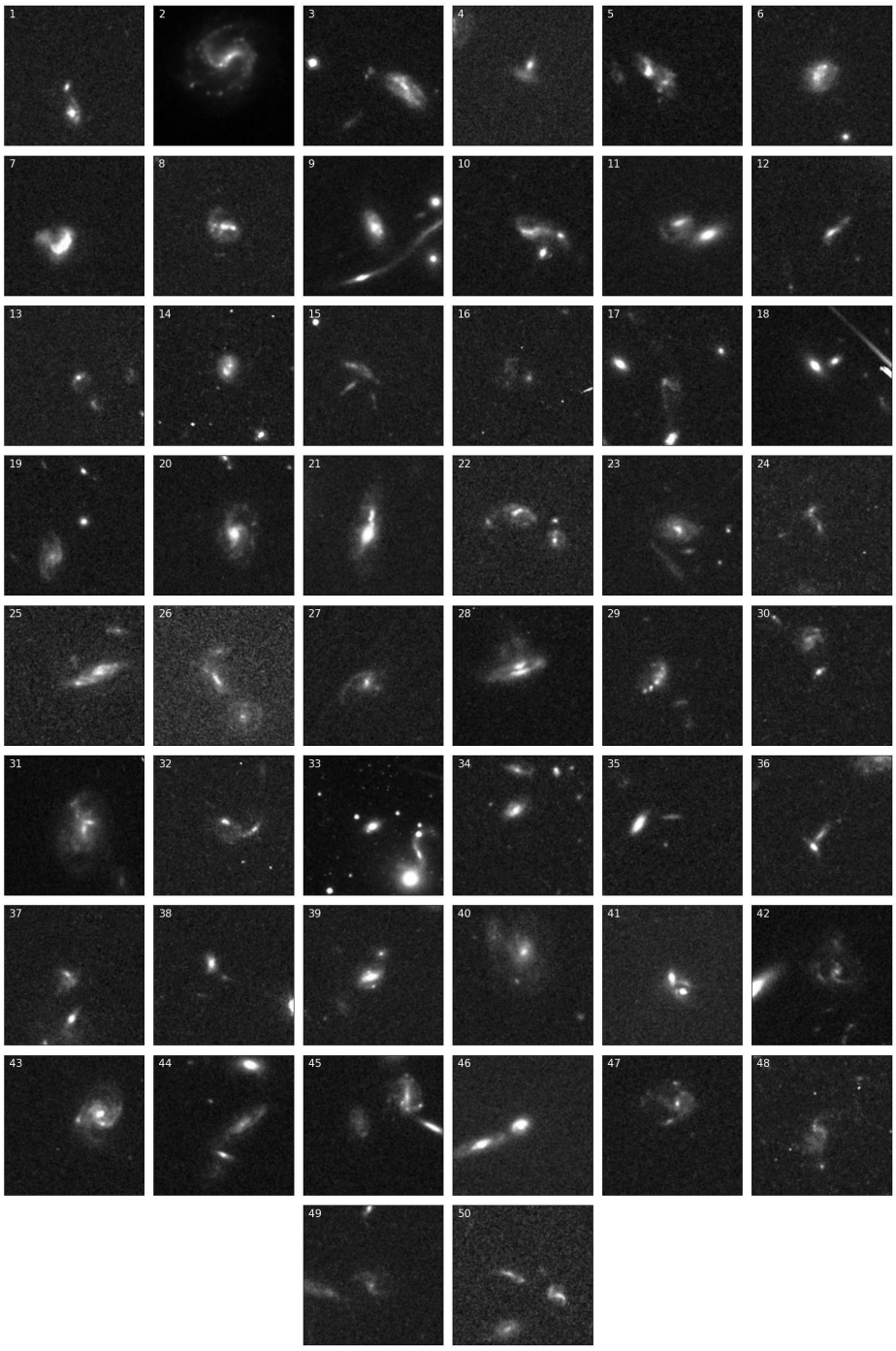

Figure 2.9: An example of 50 of the final interacting systems found with `Zoobot`. These were selected randomly from the de-duplicated and de-contaminated 21,926 sources. Each of these examples have extended tidal features and distortion. Not all of the final interacting systems have two galaxies within them (for example, image 2), but are clearly very disturbed by a tidal event. These were kept in as they would form a large part of the interacting galaxy population and would be flagged as disturbed or interacting in Galaxy Zoo. Each of these images is a 1-colour image using the $F814W$ *HST* filter.

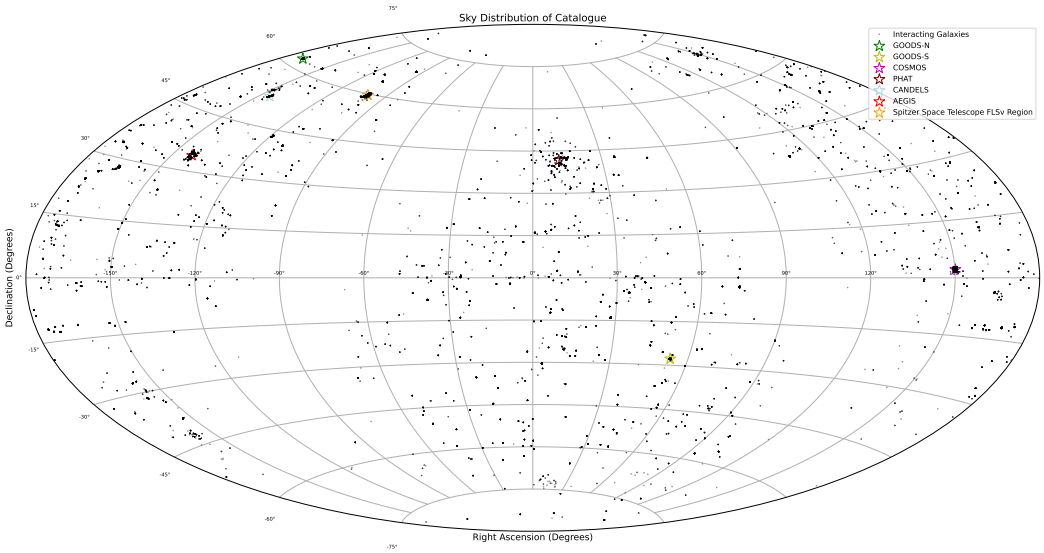

Figure 2.10: Sky Distribution of our catalogue, with marked positions of well known deep surveys conducted by *HST*. *HST* is able to observe almost the entire sky and therefore the interacting galaxies are scattered throughout. Large clusters of sources are found in the locations of surveys. This shows that often our sources are in the background of larger surveys and observations.

Figure 2.9. We also bootstrap the final catalogue as an estimate of contamination remaining. As described in Section 2.6.3, the final step of contamination removal was visual inspection by DOR of the 27,720 candidate interacting systems to remove the remaining 5,794 contaminants from the final catalogue. Visual inspection by a single expert at this scale is not perfect. We extract random sources from the catalogue in batches of 500 and manually re-classify them again. This bootstrapping reveals that ≈3% of our interacting system in the final catalogue remains contamination.

2.7.2 The Gems

By conducting a visual inspection of the 27,720 candidate systems we were able to directly identify many other objects of astrophysical interest. As `Zoobot` was trained to highly predict objects with irregular morphologies, we also find many other astrophysical objects with strange morphologies which may be of interest to the community. We call these sources of contamination gems. We make 16 subcategories of these: active galactic nuclei (AGN)/quasars, submillimetre galaxies,

Image No.	SourceID	RA (deg)	Dec (deg)	Interaction Prediction	References	Status
1	4001014298177	261.292845	37.162387	0.983999	No entry	Unreferenced
2	4001444190958	183.527536	33.183451	0.998016	[1994PASP..106..646K]	Referenced
3	4000809226818	93.960150	-57.813401	0.982266	[2019ApJ...878...66C]	Referenced
4	4553390202	73.581297	2.903528	0.968280	No entry	Unreferenced
5	4000907600174	259.037474	59.657617	0.999978	No entry	Unreferenced
6	4575187799	150.001883	2.731942	0.974649	[2007ApJS..172...99C]	Referenced
7	4000717342023	149.527791	2.126945	0.993912	[2007ApJS..172...99C]	Referenced
8	4001174802281	28.593114	-59.643515	0.982890	No entry	Unreferenced
9	4182689774	186.709991	21.835419	0.973232	[2016ApJS..224....1R, 2011ApJS..193....8B]	Referenced
10	4000958398690	186.719496	23.961225	0.999288	No entry	Unreferenced
11	4266881925	344.730228	-34.799824	1.000000	No entry	Unreferenced
12	4001084105393	150.128198	2.623949	0.982739	[2018ApJ...858...77H, 2007ApJS..172...99C]	Referenced
13	4000961670486	345.337556	-38.985521	0.954961	No entry	Unreferenced
14	4000719687395	338.173538	31.189718	0.974724	No entry	Unreferenced
15	4001435343326	331.771500	-27.826175	0.986885	No entry	Unreferenced
16	4001268932937	8.856781	-20.271978	0.986329	No entry	Unreferenced
17	4651336656	149.836709	2.141702	0.984389	[2007ApJS..172...99C]	Referenced
18	4000877021787	116.211231	39.462563	0.979178	No entry	Unreferenced
19	4000878525229	149.834893	2.516816	0.963694	[2007ApJS..172...99C, 2009ApJS..184..218L]	Referenced
20	6000290755870	186.774907	23.866311	0.981961	No entry	Unreferenced
21	4000806637434	210.253419	2.854869	0.960790	No entry	Unreferenced
22	4001215753971	135.898809	50.487130	0.998386	No entry	Unreferenced
23	4000813961830	163.678042	-12.776815	0.958405	[2005ApJ...630..206F]	Referenced
24	4001200639012	54.037618	-45.170026	0.991404	No entry	Unreferenced
25	4000921402261	150.417634	2.313781	0.990775	[2018ApJ...858...77H, 2012ApJ...753..121K]	Referenced
26	4001224732336	337.217339	-58.444885	0.955972	No entry	Unreferenced
27	4000781402752	216.968619	34.575819	0.974076	No entry	Unreferenced
28	4001283017901	120.202582	36.058927	0.994169	[2016ApJS..224....1R]	Referenced
29	4000833486119	116.260049	39.457642	0.971092	No entry	Unreferenced
30	4000949659908	146.342493	68.730869	0.961113	No entry	Unreferenced
31	4000982920478	53.084832	-27.765379	0.983472	[2010A&A...512A..12B]	Referenced
32	4001189505548	192.492491	2.436292	0.992574	No entry	Unreferenced
33	4001060882070	89.700725	-73.049783	0.962839	No entry	Unreferenced
34	4000889750512	151.176470	41.214096	0.962205	No entry	Unreferenced
35	6000322363510	53.149367	-27.823945	0.963889	[2016ApJ...830...51S]	Referenced
36	4000722901091	28.257843	-13.928090	0.982778	No entry	Unreferenced
37	6000198293960	264.488431	60.101798	0.986865	No entry	Unreferenced
38	4001095660911	258.587670	59.970358	0.955193	No entry	Unreferenced
39	4000972775076	330.960020	18.796346	0.989131	No entry	Unreferenced
40	4001132466571	126.545810	26.456196	0.997077	No entry	Unreferenced
41	4000933395648	312.810365	2.288410	0.976252	No entry	Unreferenced
42	4000932940918	218.066960	32.997228	0.990737	No entry	Unreferenced
43	4001048433104	93.880689	-57.754746	0.957755	No entry	Unreferenced
44	4001039919651	53.111470	-27.673717	0.994424	[2011ApJ...743..146C]	Referenced
45	4001282607544	333.765783	-14.006097	0.999520	No entry	Unreferenced
46	4000922341052	260.723839	58.849293	0.995477	No entry	Unreferenced
47	4000731518210	194.869144	14.146223	0.994651	No entry	Unreferenced
48	4001082523786	311.703084	-12.869002	0.976454	No entry	Unreferenced
49	4000767041112	149.784518	2.172233	0.991335	[2007ApJS..172...99C]	Referenced
50	4001024667142	150.661685	1.718587	0.967865	[2007ApJS..172...99C]	Referenced

Table 2.1: An example of the format of the final catalogue for the 50 example images presented in this paper.

galaxy groups, high redshift galaxies, jellyfish galaxies, galaxy jets, gravitational lenses/lensing galaxies, Lyman-α Emitters, overlapping galaxies, edge on proto-planetary disks, radio halos, ringed galaxies, supernova remnants, transitional young stellar objects, young stellar clusters and unknown objects.

Each sub-category has been defined by checking Simbad and VizieR for references within a $5''$ radius of each source and using the astrophysical literature for a definition of the source. DOR classified any unreferenced objects by morphological similarity to other defined objects. The platforms ESASky[1](Merín et al., 2017), NASA Extragalactic Database (NED) and the Sloan Digital Sky Survey were also used to investigate any unreferenced objects. ESASky was of paramount importance as we could investigate many objects across a range of wavelengths with many instruments.

The only objects which were classified by other means than visual morphology were AGN/quasars, submillimetre galaxies and the six unknown objects. We attempt to confirm the unreferenced AGN/quasar as candidates by investigating the source in Chandra or XMM-Newton for hard or soft X-Ray emission. The submillimetre candidates were also investigated using Herschel or Planck measurements. If there was a positive signal in their positions, they were classified as such. Further work will be needed to confirm these classification.

The final category which required further inspection was that of the unknown objects. These are objects which have unusual morphology which mark them out from the rest of the sample, but no references associated with them in Simbad or VizieR. They also did not appear in NED, meaning they could not be confirmed to be galaxies. These objects are shown in appendix A.3.

Table 2.2 shows a breakdown of the total number of objects found and the number of which were referenced or unreferenced. We have released catalogues of each sub-category in the same format as that of the main catalogue without the interaction prediction column. Each of these catalogues can also be found at the same Zenodo link.

[1]ESASky: `https://sky.esa.int/`

Category	Total Found	Referenced	Unreferenced
AGN/Quasars	35	21	14
Submillimetre Galaxies	11	8	3
Galaxy Groups	6	6	0
High Redshift Galaxies	10	7	3
Jellyfish Galaxies	18	5	13
Galaxy Jets	25	10	15
Gravitational Lenses/Lensing Galaxies	189	64	125
Lyman-Alpha Emitters	1	1	0
Overlapping Galaxies	221	92	129
Edge-on Protoplanetary Disks	9	2	7
Radio Halos	1	1	0
Ringed Galaxies	6	1	5
Supernova Remnants	4	3	1
Transitional Young Stellar Objects	2	1	1
Unknown Objects	6	0	6
Young Stellar Clusters	2	1	1

Table 2.2: A breakdown of gems found in the visual inspection stage of contamination. Each gem category has been classified based on the references associated with each object.

2.7.3 Source Redshifts and Photometry

We investigate the redshift distribution and photometric properties of sources in our catalogue. We extract all sources with pre-existing data, querying Simbad, VizieR, the HSC via the Milkulski Archive for Space Telescopes (MAST) and NED. Our queries use a 5″ search radius within the Python package `AstroQuery`. The existing data from each of these databases has undergone heterogeneous selection and analysis procedures by the various studies we extract them from; we do not try to reconcile these here. Rather than a detailed physical analysis of these sources, our priority in this subsection is to highlight how to explore and use this catalogue, as well as any difficulties which may arise.

Of the 21,926 interacting systems in our high-confidence sample, 3,037 of the 7,522 referenced sources have a measured redshift. Figure 2.11 shows the redshift distribution of this subset of our catalogue. 42.5% of the sources have a redshift $z \leq 0.5$, 45.1% have a redshift $0.5 < z < 1$ and 12.4% have a redshift $z > 1$. In fact, a small fraction (15) of these sources are found to be at z$\geq$ 5. Upon investi-

gation of these sources two of their redshifts have been measured photometrically, while the remaining 13 sources did not have the method of measurement recorded in the archive. Therefore, this finding of very high redshift interacting galaxies are uncertain at best.

It is important to note that the small sample with redshift information is affected by the selection biases of the combined studies publishing these values, and therefore the distribution may not be representative of the full sample. In addition, above redshift $z = 1$ the $F814W$ filter begins to only capture rest-frame UV flux, and therefore $z > 1$ galaxies with low star formation rates are more likely to fall below the flux limits of our detection images. Sampling only the rest-frame UV also changes a galaxy's observed brightness and morphology (e.g., Ferreira et al., 2022) – the latter being how `Zoobot` identifies interacting galaxies. For example, tidal features whose initial starburst has faded may be undetected; conversely, a single galaxy with irregular star-forming clumps may appear to be multiple interacting galaxies, which we noted as a particular source of contamination during the visual inspection stage. High-redshift interacting galaxies that are detected initially by `Zoobot` but have unusual morphologies compared to $z \sim 1$ sources may be removed during prediction (Section 2.4), given that finetuning is based primarily on the $z \lesssim 1$ imagery of Galaxy Zoo: *Hubble*. Therefore, the currently measured redshift distribution in Figure 2.11 is likely due to some combination of selection bias and training bias.

Figure 2.12 shows the basic parameter space sampled by the sub-sample of the catalogue with existing photometry and redshifts. We show the distributions of redshift with the measured apparent $F814W$ magnitude and the calculated absolute $F814W$ magnitude. The faintest objects are, as expected, observed at approximately the limiting magnitude of the deepest observations in our catalogue. Other observations have brighter limits; those wishing to select a uniform or volume-limited sample from our catalogue must consider the variable flux limits across the sample.

We finally focus on sources from our high-confidence sample that have multi-band photometry, focusing on commonly-observed filters. By construction, 100% of the sample has $F814W$ measurements, with 45% of the catalogue having

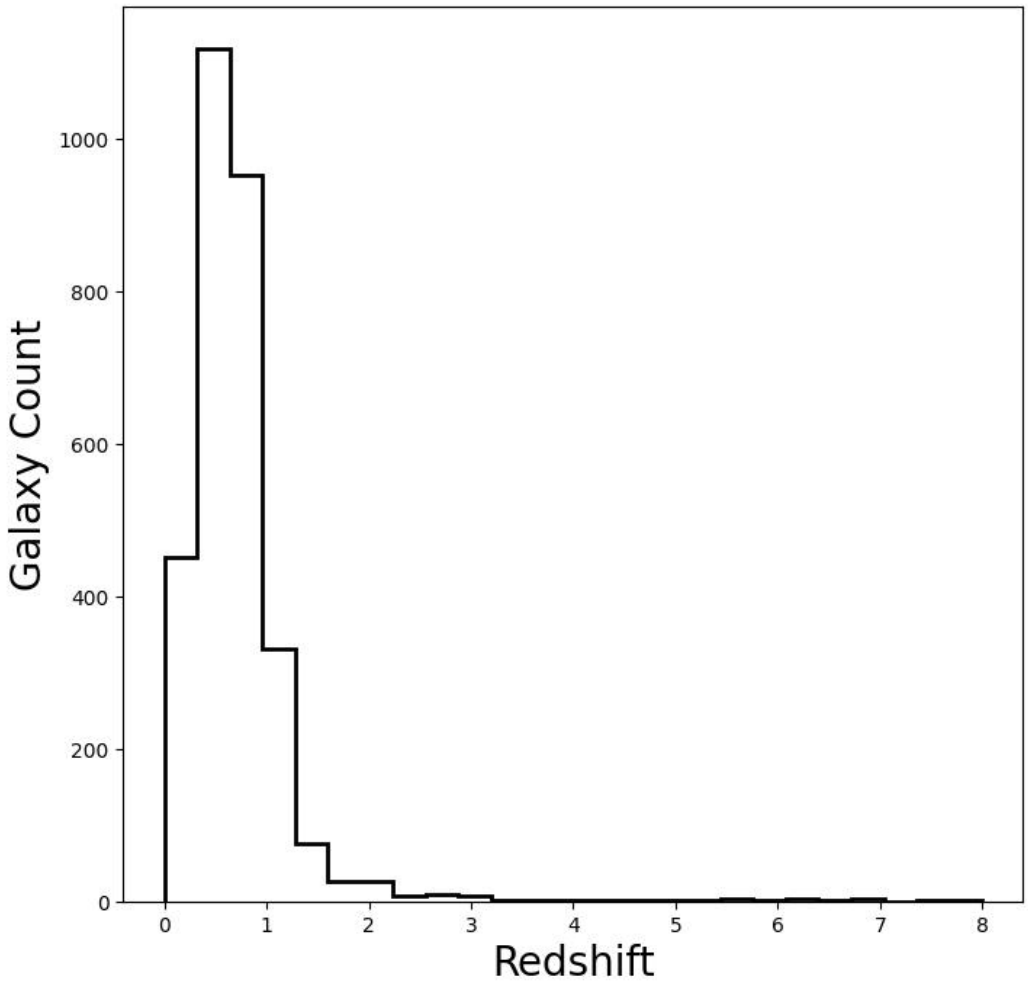

Figure 2.11: The redshift distribution of a subsample of our catalogue. Of the 7,583 referenced systems, 3,037 of them had redshift measurements in the NED, MAST or Simbad. This redshift distribution shows that our model confidently predicted interacting systems primarily for $z < 1$ systems. This was anticipated, as the model was primarily trained on systems at these redshifts. There are fifteen sources with a reported $z > 5$.

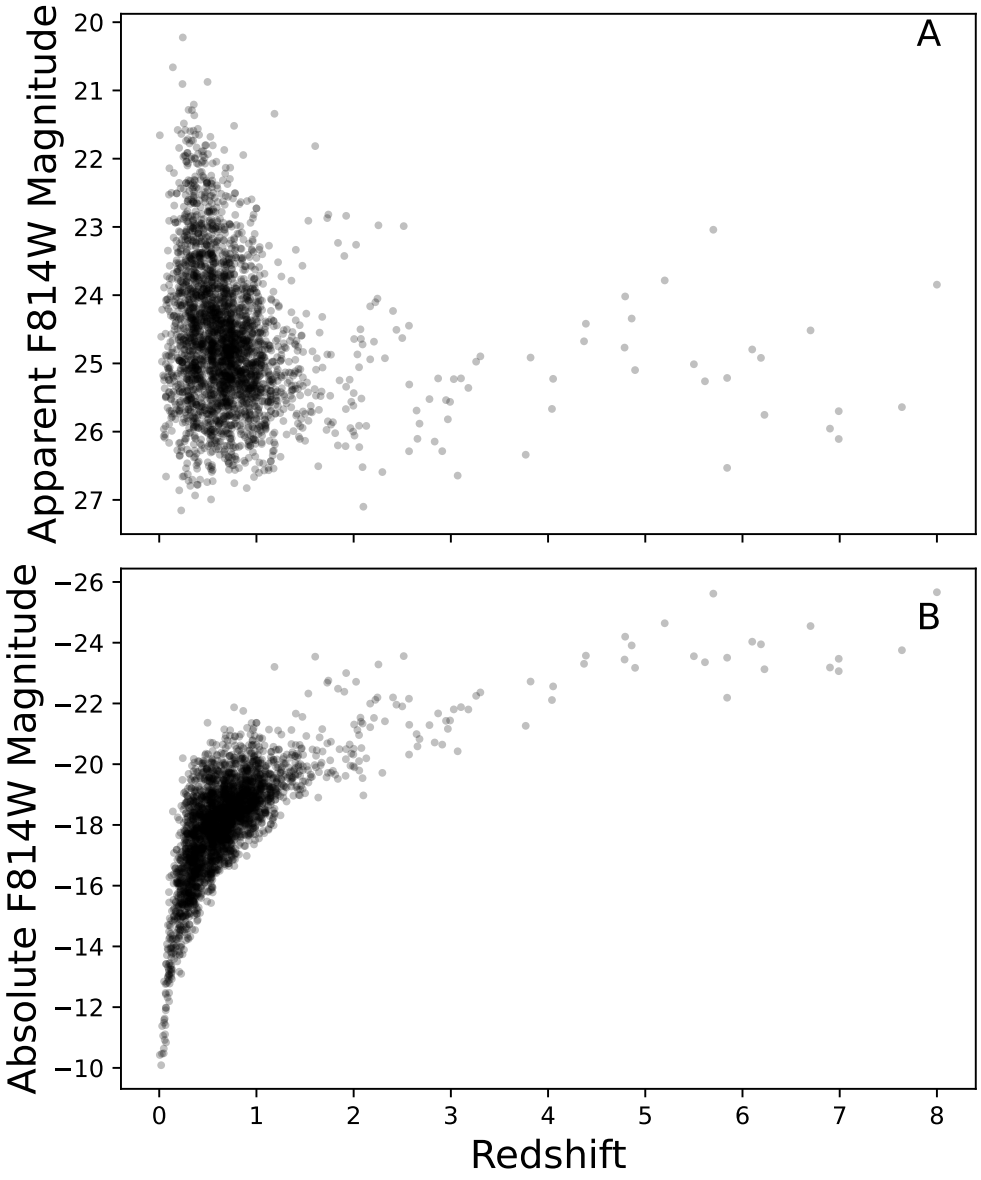

Figure 2.12: The distribution of redshift with magnitude for all sources with available data. This shows the parameter space we are sampling in this catalogue. Panel A shows that the majority of our sources are dim, background sources at low redshift. Panel B shows the faintest objects we find are at the limiting magnitudes of the different surveys this data is from.

Filter (s)	Sources Covered
F814W	100%
F606W + F814W	45.0%
F475W + F814W	11.0%
F475W + F606W + F814W	6.1%

Table 2.3: Percent of sources in the final catalogue which have observations in the relevant *Hubble* filter.

$F606W$ and only 11% having measured fluxes in $F475W$. Table 2.3 summarizes the filter coverage of our catalogue. 6.1% (1336 sources) have complete 3-band photometric information in the HSC. We use these to create examples of colour images from the catalogue (using the algorithm of Lupton et al., 2004). We used a scaling factor $Q = 2$ and $\alpha = 0.75$, with ($F814W$, $F606W$, $F475W$) as RGB channels and multiplicative factors of (1.25, 0.95, 2). The resultant images are shown in Appendix A.2.

We extract the measured magnitudes of the $F606W$ and $F814W$ filters, giving us two-band photometry for 9,876 sources. Cross referencing with each source that had a redshift yields 2,993 sources from our catalogue. We calculate the colour of each source and plot it against the absolute magnitude in the $F814W$ filter. Figure 2.13 shows the resulting colour-magnitude distribution in Panel A. The resultant distribution is very hard to interpret due to the high scatter of the sources. We extrapolate from this panel that there is little contamination from sources other than galaxies. If levels of contamination were high we would expect a second locus of sources with a very different colour-magnitude distribution.

Plotting the colour-magnitude distribution in this way captures a wide range of rest-frame wavelengths in the observed filters, which is the primary reason that panel A of Figure 2.13 is hard to interpret. In this first-look study, we do not have full spectral energy distributions (SEDs) of most sources, so K-correction of individual colours within this sample would involve assuming a template SED for each galaxy. Given that a high fraction of galaxies in our sample of mergers may deviate from standard SED templates, we wish to avoid this method. Instead, we choose redshift ranges within which to examine subsamples, such that the observed $F606W$ and $F814W$ bands cover consistent rest-frame colours within that subsample. Figure 2.13B shows only sources with $z < 0.18$, within which

the observed filters can be taken to be approximately rest-frame filters, which we define as at least 50% of the flux captured in the observed band being emitted at rest-frame wavelengths covered by that band. At $0.24 < z < 0.56$, the observed $F606W$ filter captures at least 50% rest-frame $F475W$ flux, and the observed $F814W$ filter captures at least 50% rest-frame $F606W$ flux, so Figure 2.13C is approximately a rest-frame $F475W - F606W$ vs $F606W$ plot. At $0.62 < z < 1$, Figure 2.13D is approximately a rest-frame NUV-Blue plot ($F336W - F475W$ vs $F475W$).

The galaxies in Panel B are observed in approximately the rest frame $F606W$ and $F814W$ filters. Nearly all are blue systems (by general definitions at various redshifts, *e.g.*, Kauffmann et al., 2003; Whitaker et al., 2012; Schawinski et al., 2014). This is expected for interacting systems with enough gas to fuel a starburst. The lack of many red systems is due to few gas-poor ("dry") interactions in the (relatively) local volume (López-Sanjuan et al., 2009). In Figure 2.13C, the $F606W$ and $F814W$ filters are still detecting rest-frame optical ($F475W$ and $F606W$) emission, and we find a much broader population. There are both blue and red interacting systems, with the redder mergers occurring in more luminous (likely higher mass) systems, broadly consistent with expectations (van Dokkum, 2005; Lotz et al., 2008b). The rest-frame filters approximately captured in Panel D ($F336W$ and $F475W$) sample emission across the 4000 Å break. Sensitivity to NUV means this panel effectively splits systems according to very recent star formation history (Schawinski et al., 2014; Smethurst et al., 2015). There is a significant spread in colour, with equivalent red and blue systems. We, therefore, find many young blue systems undergoing star formation and bright brighter, elliptical, massive systems also undergoing interaction in this bin.

This initial examination of the subsample of systems with easily retrievable redshifts has revealed that the interacting galaxies in the sample broadly agree with previous studies of colours in merging systems. This demonstrates the underlying promise of the catalogue. A detailed study is beyond the scope of this work, but there is considerable potential for new astrophysical insights using this high-confidence catalogue with nearly an order of magnitude more sources than those previously published.

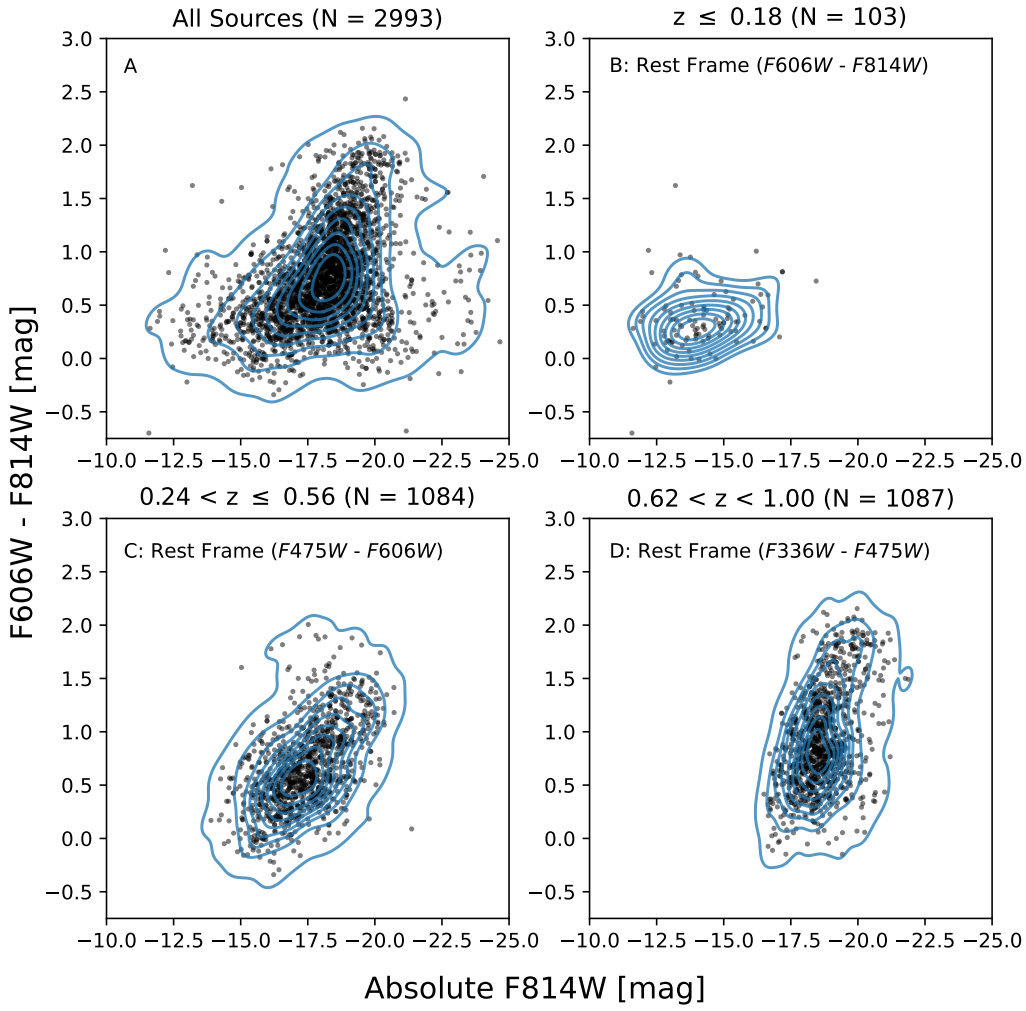

Figure 2.13: The colour-magnitude distribution of sources with a redshift measurement associated. Panel A shows the distribution of all galaxies, without controlling for redshift or dust extinction. The remaining panels then split these sources into distinct redshift bins where the $F606W$ and $F814W$ filters are observing in different rest frames. Panel B shows the colour-magnitude distribution in the local universe, where the rest frame observations are $F606W$ and $F814W$ flux. This bin reveals a blue population. Panel C shows the redshift bin where at 50% - 100% of observed $F606W$ and $F814W$ flux is rest frame $F475W$ and $F606W$ flux. This bin reveals a larger distribution of interacting galaxies, with a dominating population of blue systems and a minor population of red systems. Panel D shows the redshift bin where 50% to 100% of observed $F606W$ and $F814W$ flux is rest frame $F336W$ and $F475W$ flux. These filter bands are very sensitive to star formation, and reveal a broad distribution in colour of red and blue systems.

2.8 Conclusion

We present a large, pure catalogue of 21,926 interacting galaxy systems found from the *Hubble* Source Catalogue. This catalogue is a factor of six larger than previous works. Each interacting system was found using the European Space Agency's new platform ESA Datalabs, which allowed us to directly apply an advanced CNN - `Zoobot` - to the entire *Hubble* science archive. This corresponds to predicting over 126 million sources. The compiled catalogue has a contamination rate of ≈3% as found by bootstrapping. Table 2.1 shows an example of 50 entries in our new catalogue, Figure 2.9 showing the corresponding images. The new catalogue and all corresponding images can be downloaded from Zenodo: doi:10.5281/zenodo.7684876.

Each of our interacting galaxies were given a prediction score ≥0.95 by `Zoobot`, with such a conservative score chosen to limit contamination and maintain purity in the catalogue. Contamination was removed by applying cuts in representation space (shown by Figure 2.8) and visual inspection. Upon visual inspection, many contaminating images were found to be objects of other astrophysical interest. These have been compiled into separate catalogues, and Table 2.2 shows a breakdown of the objects found. These sub-catalogues have been released alongside our interacting galaxy catalogue. With the priority of purity in this catalogue creation, we will aim in future work to use it in the statistical analysis of interacting galaxies and begin linking the underlying parameters of interaction to the complex physical processes that occur in them. A secondary purpose of this catalogue is to serve as a training set for future models which may wish to search for interacting or merging galaxies.

With the use of ESA Datalabs, this project was conducted quickly. The entire process, from creating the source cutouts, to training `Zoobot`, to making predictions on 126 million sources took three months to complete. Using conventional methods, such as `AstroQuery` or TAP services, downloading the data would have likely taken on this timescale. By bringing the user to the data, rather than vice versa, catalogues of a similar size - and many times larger than previous catalogues - of many different objects can be created quickly.

None of the the interacting systems in this work are 'new'; every one of them exists in the background of large scale *HST* surveys and observations since their release. However, the method to directly search for them has been impractical until the release ESA Datalabs. By directly applying machine learning to existing astrophysical data repositories, a new method to creating significantly larger catalogues has been achieved.

This shows the importance of archival work, and the power that ESA Datalabs will bring to the field of astronomy. ESA Datalabs is expected to be released in Q3 and with it, the ability for large scale exploration of archival data. It will be released with introductory tutorials, step-by-step guides and different Python environments for ease of use for different telescopes and instruments the ESA is involved in. It will have a full cluster of GPUs at its disposal and a storage capability in the range of hundreds of Terabytes. In future, this entire project - from training set creation to predictions - could be conducted on ESA Datalabs.

Such a setup as ESA Datalabs also allows the creation of large observational catalogues, comparable to that we create from cosmological simulations. This is incredibly important to further constraining already existing results. In the current period of astronomy where large survey instruments are awaiting first light, or the beginning of future telescopes is uncertain, the ability to get ever more information out of the archives is paramount.

Acknowledgements

DOR gratefully acknowledges the support from European Space Agencies Visitor Archival Research program, and hosting at the European Space Astronomy Centre. DOR thanks Bruno Merín for supervising this project and Sarah Kendrew for aiding its creation. This project was conducted as part of DORs PhD program supported by the UK Science and Technology Facilities Council (STFC) under grant reference ST/T506205/1. BDS acknowledges support through a UK Research and Innovation Future Leaders Fellowship [grant number MR/T044136/1]. ILG acknowledges support from an STFC PhD studentship [grant number ST/T506205/1] and from the Faculty of Science and Technology

at Lancaster University. MW gratefully acknowledges support from the UK Alan Turing Institute under grant reference EP/V030302/1. MRT acknowledges the support from an STFC PhD studentship [grant number ST/V506795/1] and from the Faculty of Science and Technology at Lancaster University.

Much of the intense computation was conducted at the High End Computing facility at Lancaster University. This publication uses data generated via the Zooniverse.org platform, and the unending enthusiasm of citizen scientists and volunteers in classifying galaxies. We also thank the many PIs who's archival data we have used to create this catalogue. All data containing astrophysical objects of interest found in this work are public on MAST: 10.17909/wfke-n133.

This research made use of many open-source Python packages and scientific computing systems. These included `Matplotlib` (Hunter, 2007), `scikit-learn` (Pedregosa et al., 2012), `scikit-image` (van der Walt et al., 2014), `Pandas` (McKinney, 2010), `Shapely` (Gillies et al., 2007), `UMAP` (McInnes et al., 2018) and `numpy` (Harris et al., 2020). This work also extensively used the community-driven Python package Astropy (Astropy Collaboration et al., 2018a). `Zoobot` utilises the underlying code `Tensorflow` (Abadi et al., 2016) Python package.

This project used data from the *Hubble* Space Telescope and stored in the archives at the European Space Astronomy Centre. These observations are obtained from the Space Telescope Science Institute, which is operated by the Association of Universities for Research in Astronomy, Inc, under NASA contract NAS 5-26555. All sources were found using v3.1 of the *Hubble* source catalogue (Whitmore et al., 2016) and accessed using the ESA Datalabs science platform. ESA Datalabs is directly connected to the ESA *Hubble* Science Archive. This study makes use of data from AEGIS, a multiwavelength sky survey conducted with the Chandra, GALEX, Hubble, Keck, CFHT, MMT, Subaru, Palomar, Spitzer, VLA, and other telescopes and supported in part by the NSF, NASA, and the STFC.

DOR would like to thank those in the ESA Traineeship program cohort of 2022. They created a wholly welcoming environment and space of support. A

special thanks must go to Karolin Frohnapfel and Emma Vellard for much technical discussion. Finally, DOR would like to acknowledge Aurélien Verdier.

Chapter 3

When do the effects of interaction happen? Interacting and disturbed galaxies in the COSMOS field

3.1 INTRODUCTION

To fully map out the effect of galaxy interaction, we must understand its impact on galaxies through the entire merger history and dynamical timescale of the interaction. As each observation only provide snapshots of different parts of the dynamical timescale across different merger histories, it is imperative we have large samples to capture the full history. Previous works focused on simulations find sudden increases in star formation occurs early in the dynamical timescale of the interaction (Cox et al., 2008; Rodríguez Montero et al., 2019), however, confirming this definitive link observationally remains elusive (Ren et al., 2023). The same studies could also be done for nuclear activation as well as quenching of interacting systems (Ellison et al., 2011; Goulding et al., 2018; Steffen et al., 2023).

Achieving such study in the dynamical history of an interaction with observations is more difficult. While we cannot directly measure the full dynamical timescale for individual systems, we can create large samples of interacting galaxies which are representative over the dynamical history. The first interacting galaxy sample (the Arp (1966) catalogue) contained only 318 systems, while more recent samples have been created either by machine learning classification (Pearson et al., 2019a; Suelves et al., 2023), visual classification by citizen scientists (Darg et al., 2010a) or by photometric parameterisation (Lotz et al., 2004; Nevin et al., 2023). With larger samples, we can approximate different parts of the dynamical time them into their constituent stages. However, each sample has always been plagued by contamination and a loss of statistical significance when broken down into stages. It is notoriously difficult, from morphology alone, to select interacting or merging galaxies.

Often, the stage of an interaction is approximated by measuring different parameters with the projected separation between two systems. Early works showed that there was significant star formation enhancement (SFE) in galaxy pairs with small separations (Lambas et al., 2003; Ellison et al., 2008; Shah et al., 2022). A connection was also found between the projected separation in galaxy pairs and the AGN fraction where, once again, the AGN fraction appears to increase with decreasing projected separation (Rogers et al., 2009; Ellison et al., 2011; Gordon et al., 2017; Silva et al., 2021). However, using projected separation as a proxy for the stage in dynamical time of an interacting history overlooks which part of the dynamical history the interaction is at. For instance if a galaxy pair is found to have small projected separation, without visually confirming the morphology, it is difficult to ascertain if the galaxies are just approaching each other or have just passed each other. Extended morphological tracers can help break this degeneracy, which is critical to our understanding of the effects of interaction.

In this Chapter, we split our large sample of interacting galaxies into four specific stages based on their morphology. Each stage is designed to capture different parts of the dynamical time of an interaction. Each interacting or disturbed galaxy is visually classified into its stage, which ranges from galaxies being close

pairs, to overlapping and disturbed, to disturbed and distinct to finally coalescing. This follows the classification methods of developing algorithms for staging of interaction(Bottrell et al., 2019; Chang et al., 2022).

We focus here on the subset of our catalogue in the Cosmic Evolutionary Survey (COSMOS) survey[1]. This provides us with ancillary data for a subset of the catalogue described in Chapter 2. This ancillary data contains many galactic parameters of interest and we explore how they evolve with our defined stages of interaction. We primarily focus on the evolution of the stellar masses, star formation rates and AGN classification from various COSMOS catalogues. We use the photometric redshifts available from COSMOS to confirm a set of interacting galaxies and close pairs to draw on trends with projected separation.

This Chapter is laid out as follows in Section 3.2 we briefly summarise the COSMOS catalogue, and describe the process of catalogue matching especially focused on accurate de-duplication and cross-matching between them. Section 3.3 describes the methods by which we split the interacting systems into stages, and how we identify the secondary galaxies in each pair. In Sections 3.4 and 3.5 we show our results of star formation and active galactic nuclei evolution with interaction stage and present an initial discussion. These are followed in Section 3.6 where we compare to previous works and put our results in the context of the field. Finally, Section 3.7 we make concluding remarks and discuss future work to better our constraints.

3.2 DATA: Catalogue Matching & Secondary Identification

3.2.1 The COSMOS2020 Catalogue

In the COSMOS survey, there are multiple catalogues of galactic parameters such as star formation rates (SFR), stellar masses and line emissions. We elect to specifically use the COSMOS2020 catalogue (Weaver et al., 2022). This catalogue

[1]DOI: 10.26131/IRSA178

has a wealth of ancillary information for just over 1.7 million sources in a 2 square degree area of the sky. Each source found in this area has been analysed with well known astronomical software (for our purposes, primarily LePhare (Arnouts et al., 1999; Ilbert et al., 2006), EAzY (Brammer et al., 2008) and FAST (Aird et al., 2017)). These provide estimates of the physical parameters of each source in the COSMOS2020 catalogue based on its measured broadband photometry. For our purposes, the parameters it contains are the stellar masses and the estimated star formation rates of each source. For the stellar mass we use the best fit LePhare measurement and for the star formation rates we use the best fit EAzY/FAST measurements. We compare the output masses and SFRs for our sources in Figure 3.1. As shown, there can be wide scatter and disagreement between the two algorithms. We tested different combinations of EAzY and LePhare stellar masses and SFRs to find which would be best to use in this work. We found that using the same algorithm in both parameters revealed the degeneracies in the underlying parameter space. Therefore, we elect to inter-mix the algorithms and use the best fit LePhare stellar mass and best fit EAzY SFR measurement. This does not reveal the underlying degeneracies in the discrete templates used to fit the broad band photometry and estimate these values.

Our sample interacting galaxies has been de-duplicated, but the COSMOS2020 catalogue is not specifically a de-duplicated merger catalogue. We therefore cross match between our sample and the COSMOS2020 catalogues using a position search within $10''$ of our samples coordinates. Once we have identified the nearest COSMOS2020 source for each interacting galaxy ID, we de-duplicate based on COSMOS2020 ID and redo the coordinate matching process with any duplicate matches. If no further COSMOS2020 sources were within 10" of the source, then we classify the source as not in the COSMOS2020 catalogue. We find 3,786 of the our sources exist in the COSMOS2020 catalogue.

Once matched, we remove any sources with non-physical photometric measurements for the stellar mass or star formation rates. We then further reduce our sample by only keeping sources within the a mass range of $6.5 \leq \log_{10} \mathrm{M}_*(\mathrm{M}_\odot) \leq 12.5$ and a star formation rate range of $-5 < \log_{10} \mathrm{SFR}(M_\odot yr^{-1}) \leq 3.5$. We also opt to intitute a redshift cut of $z \leq 1.2$. Beyond this redshift, we find that identification of tidal features becomes difficult due to surface brightness dimming and

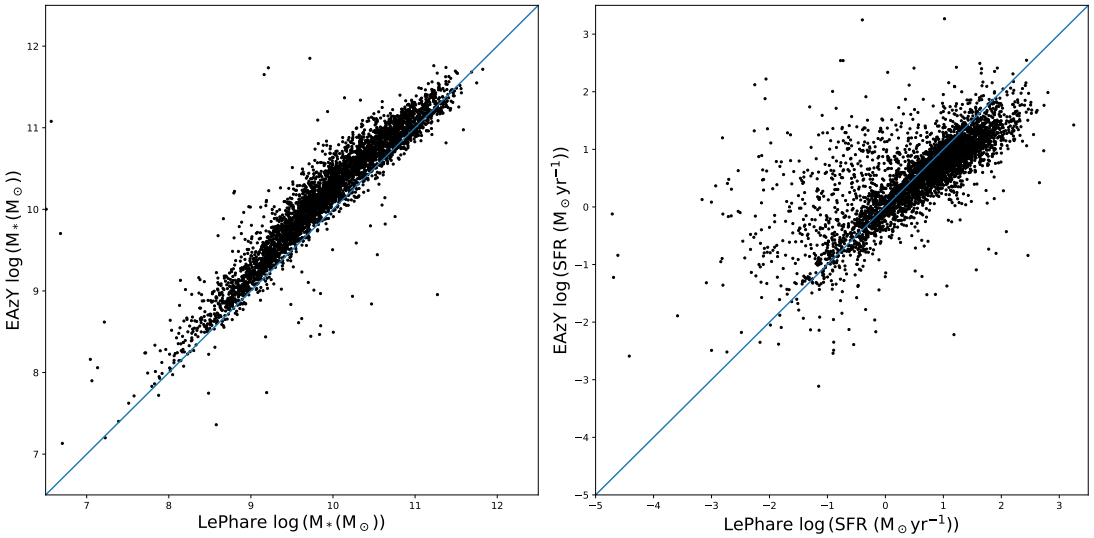

Figure 3.1: Comparison of the measures of stellar mass and SFR using either LePhare or EAzY/FAST photometric codes to calculate them. If the algorithms agreed perfectly, the sources would lie on the blue 1:1 line. *Left*: The scatter in the stellar masses between softwares. As shown, EAzY often seems to find larger stellar masses when compared to LePhare. *Right*: Scatter in SFRs when measured with LePhare or EAzY.

we risk mis-identifying the stages of the interacting galaxies. This also matches the redshift cut applied to the environment catalogue we cross match with in Section 3.3.2. Applying these cuts reduces our sample size to 3,689 interacting galaxies.

3.2.2 Secondary Identification

As the catalogue described in Chapter 2 only contains source coordinates and IDs, we must also manually identify the secondaries of many of our interacting systems. To find the secondaries, we apply three steps. First, a cutout surrounding each source was created. These cutouts were from the COSMOS cutout service, selecting HST-ACS tiles in the $F814W$ filter. Each cutout had a 30" radius (corresponding to 1001 $\times$ 1001 pixels). The original cutout from Chapter 2 was also displayed next to the enlarged cutout. We annotate each cutout with each sources COSMOS2020 ID and measured photometric redshift and error. By

annotating each cutout with the sources photometric redshift and ID, we could visually assessed each cutout and give one of four following classifications to each: system disturbed but secondary could not be identified; secondary could be identified; cannot confirm galaxy is interacting; null redshift (0 or NaN); incorrect primary assigned.

To associate a secondary galaxy for each primary, the galaxy had to be within the cutout we were visually assessing and within the recorded error of the primary photometric redshift. Using photometric redshift cutoffs in this way is often done when calculating environment parameters (e.g Baldry et al., 2006) or defining interacting galaxies by close pairs (e.g Shah et al., 2022). A null redshift is defined as one outwith our redshift limits, 0 or NaN. A minority of the cutouts we visually assessed were found to have the incorrect primary at the centre. In these cases, we record the correct primary galaxy ID and extract the ancillary data from the COSMOS2020 catalogue. We then attempt to find the secondary galaxy again for the corrected primary.

Using these definitions, we find that of the 3,689 original systems cross-matched with COSMOS2020 2,283 could not have their secondary identified, 834 had a clear secondary, 446 could not be reliably classified as an interacting galaxy, 248 had a null redshift and 149 were the incorrect primary. Figure 3.2 shows an example of each of our classifications. Each secondary we identify was added to our sample, increasing our sample size to 3,829.

While initially surprising that the majority of our systems could not have a secondary identified, we found that it was mostly due to limitations in the COSMOS catalogue or the way in which we conduct our secondary identification. Each potential secondary must have a COSMOS ID associated with it, however, when two systems are very close together and small enough they were identified under a single COSMOS ID despite being two separate systems. The same also occurred when two systems were merging or interacting. The tidal features connecting the systems or coalescing systems would only be identified under one ID in the catalogue. Figure 3.2 shows an example of two systems being close enough together that they have been identified under a single COSMOS2020 entry. Figure 3.3 shows this disparity with the different types of interaction we observe.

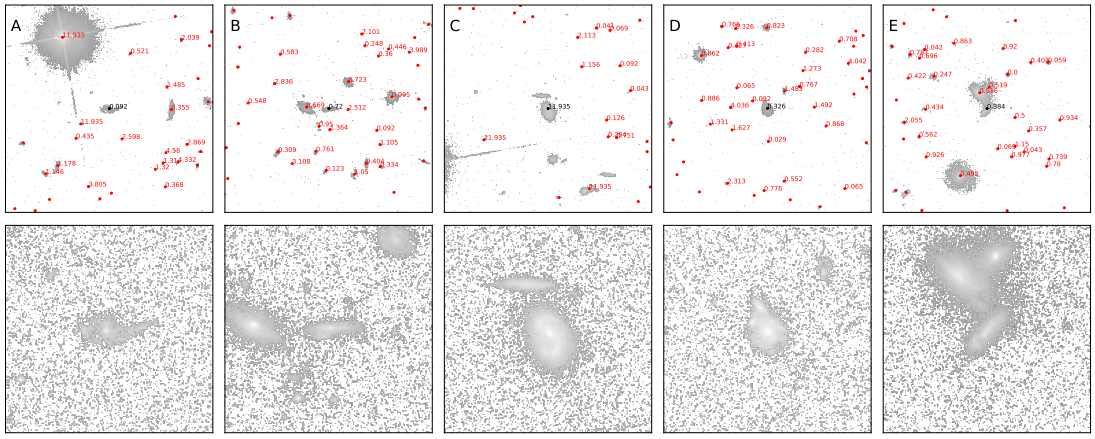

Figure 3.2: An example of each visual classification made on the cross matched sample. These are: (A) where the secondary could not be identified, (B) the primary had a clear secondary, (C) the primary could not be reliably classified as an interacting galaxy, (D) the redshift was null and (E) the incorrect primary was identified. Based on these classifications, we either add the secondary galaxies to the sample or we remove the contamination from it. These images are 30" across using the COSMOS cutout service, selecting HST/ACS tiles as the basis for the observations in the $F814W$ filter.

Those galaxies which were found to be contamination (i.e., could not be reliably classified as interacting as they showed little tidal distortion or had no neighbouring systems at a matching redshift or having a redshift of 0) were removed from our sample. Galaxies that could not be reliably classified as interacting were also removed. These systems were often overlapping but at different redshifts, or were systems with irregular morphologies of spiral arms or many clumpy galaxies. There was also many systems that were at high redshift ($z \geq 1$) where the resolution of the cutouts meant that features could not be discerned visually.

Our final classification type was that the incorrect galaxy had been identified as the priamry galaxy. This was the case for 149 systems. These were systems where the interacting galaxies were clearly in the cutout but some tidal debris or some nearby system had been cross matched from COSMOS. We reassign these systems to the correct COSMOS IDs and then take them through the secondary identification process.

By the end of this selection method, we find a sample size of 3,829 interacting galaxies. We conduct a de-duplication based on the COSMOS2020 ID,

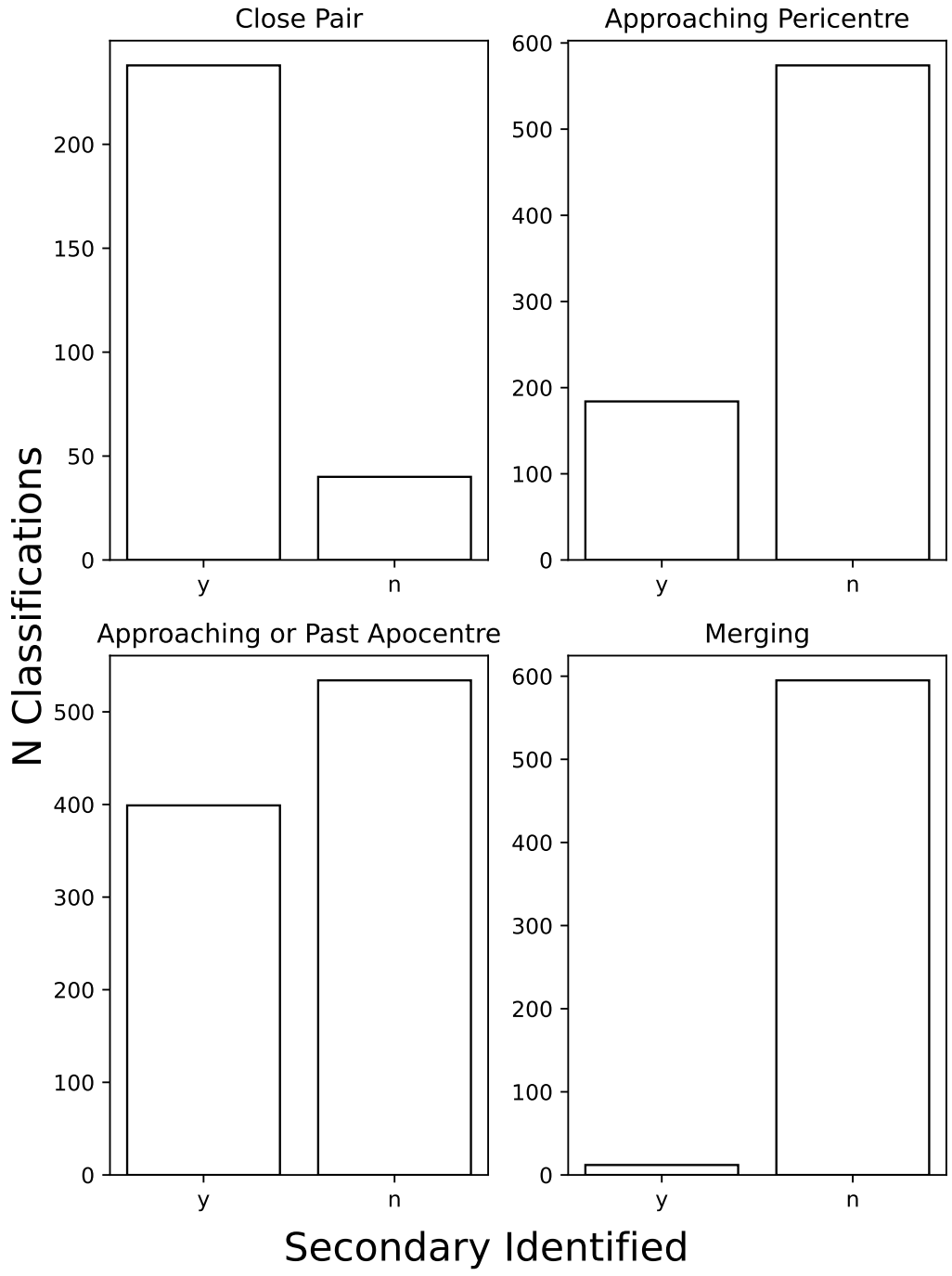

Figure 3.3: Where a secondary could be identified at different stages in the interaction. The reasoning for such disparity in secondaries identified is due to the relative distance each the secondary would be from the primary at each stage. For a close pair, we often found the secondary galaxy, but a minority of these were so close together that the entire system was given a single COSMOS ID. The same was true for those interacting systems approaching pericentre or merging. When the secondary was near apocentre, often it would be outwith the cutout we were using for visual classification.

which reduces the sample back down to 3,547 interacting galaxies. The remaining systems are each visually confirmed interacting and disturbed galaxies based on their morphology and photometric redshift as measured in the COSMOS2020 catalogue.

For each galaxy in our sample, we find and define a mass- and redshift-matched control galaxy to investigate differences in their galactic parameters. We find these control galaxies from the COSMOS2020 catalogue. All galaxies within 0.01 dex of our samples stellar mass are selected from the catalogue, and within a ± 0.01 redshift slice. We then define the control galaxy as the system furthest from our interacting galaxy. Each control galaxy is then visually confirmed to be non-interacting and have no nearby pairs. We also confirm that it does not already exist in the catalogue, and ensure there is no duplication in the control sample. Figure 3.4 shows the mass distribution of our paired sub-sample of galaxies and their control, showing a mass distribution which is approximately the same as the interacting sample. With the previously defined cuts, we find a control galaxy for all but one of our interacting galaxies.

3.2.3 Finding Additional Systems

As a result of using visual classification to find the secondary galaxy in each interaction, we were able to also confirm other interacting systems which had not been found in our catalogue. Primarily, these extra interacting galaxies are from systems which had more than two galaxies involved in the interaction. Our selection process was built to only find a primary and secondary galaxy and, therefore, we add these extra systems into our sample manually. Other interacting galaxies that were added were low redshift systems which would have appeared to completely fill the cutout of the classification process in Chapter 2. By looking at the larger COSMOS2020 cutouts, we are able to recover these galaxies and add them to this sample.

In total, we found an extra 841 interacting systems that we could add to our sample. Upon conducting a de-duplication of these with the sample already found, this was reduced to 634 interacting systems. This gives us a total sample size of 4,181 which we use through the rest of this work.

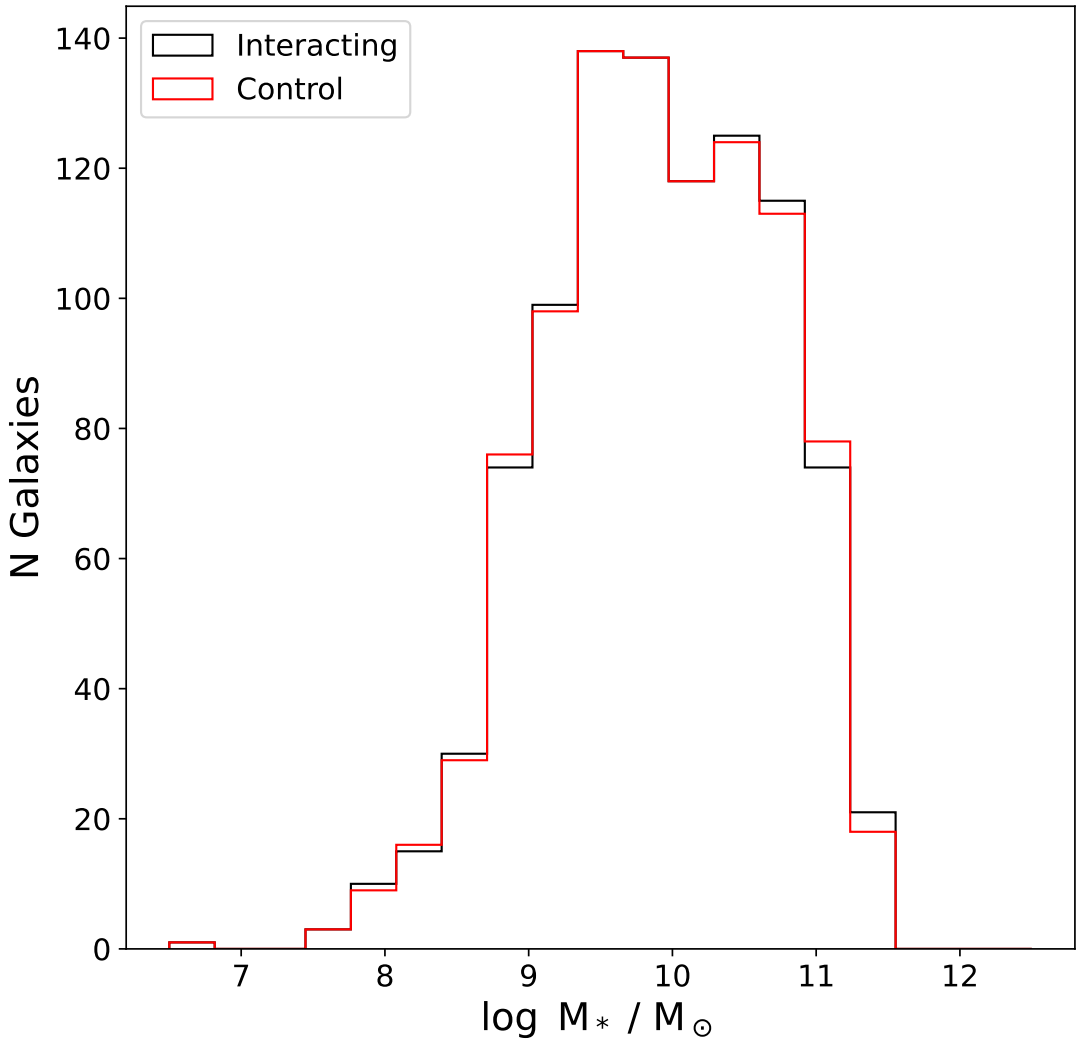

Figure 3.4: The mass distribution of the paired interacting galaxy sample and the control sample. Both primary and secondary galaxies are within this distribution. From our sample of 4,181 interacting galaxies morphologically identified, 834 are confirmed galaxy pairs. This means that the secondary in the interacting galaxy system has been identified from morphology and photometric redshifts.

3.2.4 Creating the Volume-Limited Sample

We now investigate the distribution of our sample between mass and redshift. Figure 3.5 shows the resultant distribution from our flux limited sample (what we have described thus far). ~~As shown, we find~~ we lose sensitivity to the low mass systems as the redshift of our sample increases. In order to ensure we that any changes in evolution we would observe is purely due to evolution effects, we institute a cut in the mass of the interacting systems to create a volume limited sample. We elect to use a mass cut of $\log(\frac{M}{M_\odot}) \geq 9.25$ ~~and is~~ shown by the blue dashed line in Figure 3.5. We elect to make this cut for two reasons. First, this is the lowest mass system which is still observed in our ~~smallest observed~~ sample at $z = 1.2$. ~~This is of the merging sample.~~ The second reason is that such a mass cut is close to the mass cut made in the environment catalogue we will describe later in Section 3.3.2. While in ~~this~~ Section we make a cut of $\log(\frac{M}{M_\odot}) \geq 9.6$, we find that increasing the mass cut by this amount does not make major differences to our results. Thus, we opt to lower the mass cut to increase the number of galaxies in our sample. Making such a cut reduces our flux-limited sample of 4,181 to a mass-limited sample of 3,384 interacting galaxies.

Throughout this Chapter, we will use our volume-limited sample in our analysis and when we quote the results. This gives us uniform sensitivity across our volume. However, ~~we point out at this stage that our analysis was also conducted~~ with the flux-limited sample and ~~that~~ the ~~overall~~ conclusions of this work ~~were found not to~~ change. ~~The specific results we find, their quantities and errors, do change however.~~

3.2.5 Sample Summary

Thus, to briefly sumarise this section, we have applied multiple cuts and additions to our initial sample of interacting galaxies. We have gone from an initial sample of 3,689 matched galaxies with COSMOS to a mass-limited sample of 3,384 galaxies. This mass limited sample of galaxies is the one we will use throughout this work.

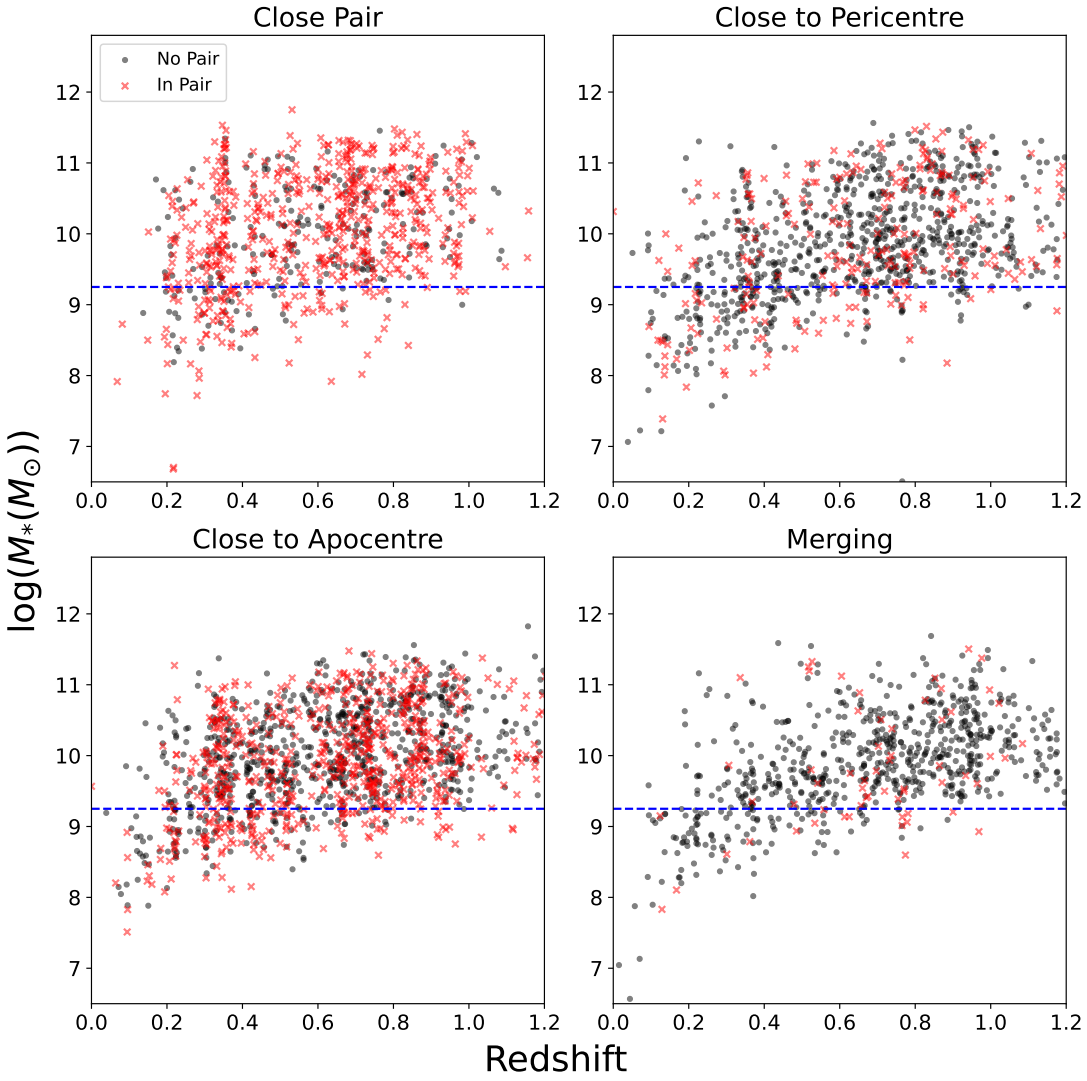

Figure 3.5: Redshift vs Mass distribution for each stage of interaction we have defined. We separate those systems with an identified secondary (in red) from those where no secondary could be found. The dashed blue line shows the mass cuts we make for our mass-limited sample and is set at $\log(\frac{M}{M_\odot}) \geq 9.25$. As shown here, the distribution of systems across redshift and mass is consistent for all stages in our sample. This is important as we use tidal distortion and the existence as tidal features as a fundamental for our classification methodology. Therefore, we are likely not affected by this in our analysis. We also find that our ability to identify pairs is primarily affected by stage and not redshift.

3.3 METHOD: Environment, AGN and Interaction Stage

A primary aim of this work is to investigate the evolution of different galactic parameters with stage of the interaction. Each stage is defined to capture a different part of the dynamical time and merger history. This stage also relates to the projected separation of galaxies with an identified secondary. Here, we fully define what we mean by interaction stage and which part of the dynamical time it covers. We also describe how we find AGN in our full sample and find measurements of the environment about each. The parameters required to calculate the AGN fraction are not found in the COSMOS2020, and we therefore must cross match with other catalogues to find the required parameters. We also describe the catalogue we use to define the environmental density about each of our sources. This is important to consider, as it is well known that measured SFRs of galaxies can be affected by environment, as well as existing biases in where interacting and merging galaxies reside. We, therefore, need to check that we have not introduced any environmental biases into our definition of interaction stage.

3.3.1 Classifying Stage of Interaction

The primary goal of this work is to find if a relation exists between a host of galactic parameters, underlying physical processes and the stage of the interaction. Each stage covers a different part of the dynamical history in an interaction and, in this case, we define it based on the morphology and projected separation between systems. We have already encountered the four stages we will investigate in Figures 3.4 and 3.5 and now define a short hand name to refer to them as well as fully describe the part of the dynamical history we are probing. These are:

- Separated: Systems which are well-separated with little to no morphological disturbance (Close pairs).
- Pericentre: Close pairs showing morphological distortion while still in a pair or show a physical connection by tidal features.

- Apocentre: Well separated pairs with morphological disturbance or isolated galaxies with clear tidal features.
- Merging: Highly disturbed systems with two or more cores within them.

Our four stage approach is not a new one, and many other works have utilised it to differentiate different parts of the dynamical history of a galaxy interaction (Chang et al., 2022; Garay-Solis et al., 2023).

Figure 3.6 shows the original source cutouts used in Chapter 2 to give an example of each stage. There are degeneracy's associated with this staging system, however. In this context, we define a degeneracy as when the interacting galaxies may be at two or more parts of the dynamical timescale and we have no way to tell without further information on the system.

The separated stage of the interaction captures the first approach of the two systems and they are observed as a galactic pair. They have no morphological disturbance yet and, in our sample, are most often those galaxies with distinct disks. At this point, we would expect no change in the underlying processes of the galaxies from their control samples. Interaction has not taken place yet, and the two systems are morphologically intact. By definition, this stage requires an identified secondary galaxy and therefore has the highest number of identified secondaries with every system having a secondary that can be visually classified. This is also the least degenerate part of the dynamical time we are sampling due to a criteria being no morphological disturbance. Therefore, it is unlikely that the galaxies would have already flown by each other.

The pericentre stage of the interaction is defined as the point where the two galaxies in the interaction are are at or just passing the pericentre of the tidal encounter. At this point, we find the beginning of morphological disturbance, the beginning of the formation of tidal features and some tidal debris. Due to the two systems having to overlap or connect via tidal features - by definition - this is the stge that suffers most from limited identification of the secondary galaxy. The COSMOS2020 catalogue often defines these two systems as a single system, and therefore, we lack information about their secondary. This stage is also highly degenerate in the context of the dynamical timescale of the interaction. Without further information, we are unable to define whether the galaxies involved at this

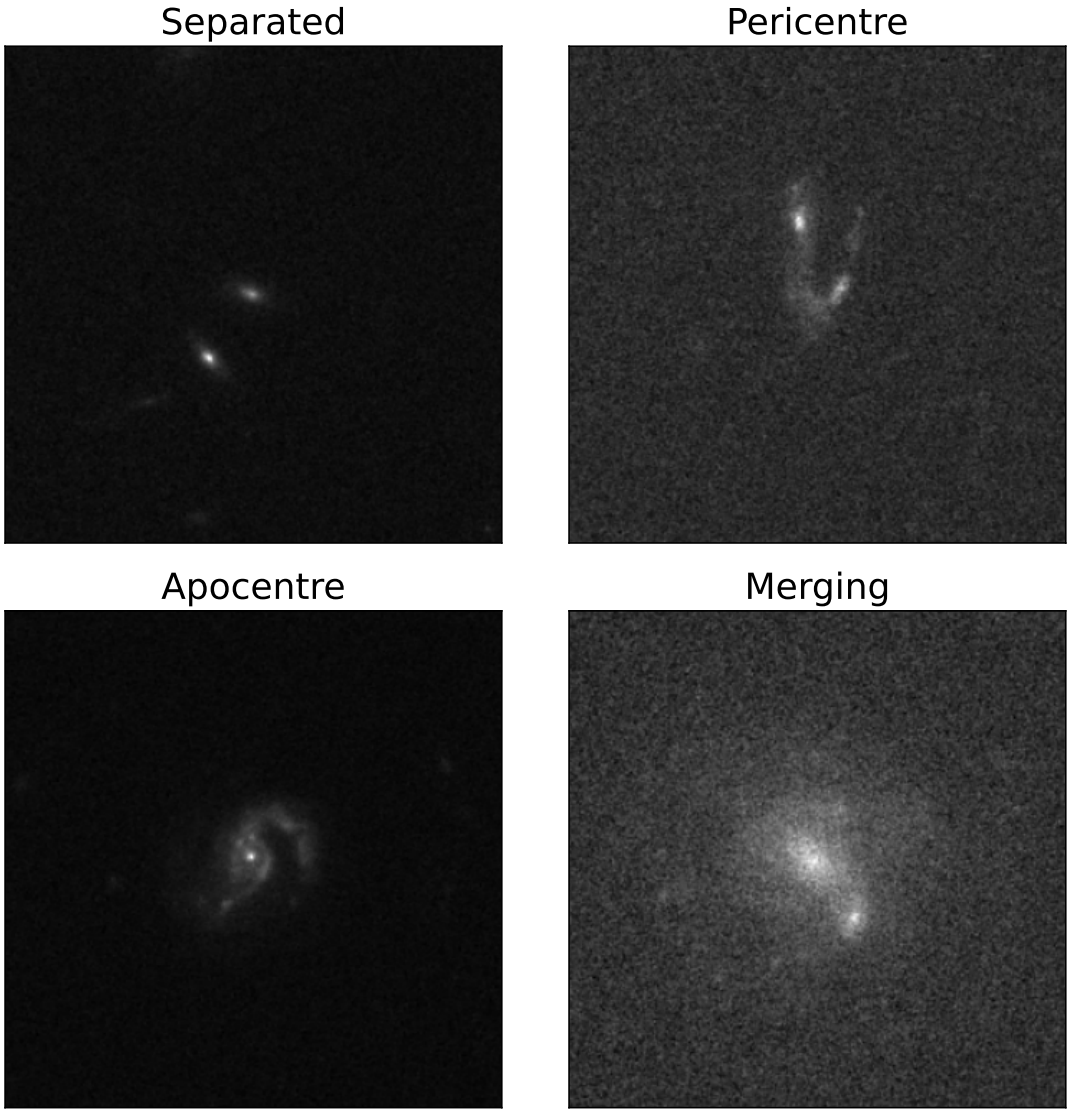

Figure 3.6: Examples of the four stages we split our interacting galaxy sample into. Separated: A close pair with confirmed redshift matching. Pericentre: Two distinct systems interacting with tidal features forming. Apocentre: A tidally disturbed system with no secondary present, likely at apocentre. Merging: A galaxy with multiple cores while highly disturbed. At the final stage before coalescence.

stage are at the first, second, third, etc passage of the tidal encounter. We also do not know if they are just passing pericentre on the first pass, or if they are approaching pericentre on more than one pass.

The apocentre stage describes those interacting systems where the two disks are fully separated and distinct from one another. They must have some morphological disturbance associated with them, but do not require a secondary galaxy to be put into this stage as if the galaxies have sufficient velocity they would escape from each other and, therefore, their secondary could be outwith our COSMOS2020 cutouts used to visually identify them. This is reflected in an even distribution of finding the primary and secondary in this stage. This stage is also defines a large part of the dynamical timescale. It spans from separating from the secondary and after pericentre, to moving out to the apocentre of the interaction (or escaping with sufficient velocity), to falling back in towards the secondary galaxy again. Without velocity information, we have no way of finding if the galaxy is moving away or moving towards its secondary.

Finally, merging stage represents the final step of a galaxy interaction. If the two galaxies do not have sufficient velocity to escape one another, then they go on to coalesce and ultimately merge. We define this stage through the severe morphological disturbance of the galaxy involved as well as the existence of a second core within it. While we attempt to capture only pre- or ongoing-coalescing systems, it is important to note that this stage is degenerate to post-merger remnants which will also be accepted by our criteria. Post-merger remnants are systems where coalescence has been completed and immediately after will be highly morphologically disturbed and difficult to distinguish from those systems with merging ongoing. At this stage, we would expect the interaction will be at its most violent with complete disruption of both galactic disks and likely increased star formation across the disk.

We have noted the degeneracy of each stage as we have described them. Thus, we can now put this fully into the context of the dynamical timescale. Over a typical interaction, we would expect the galaxy pair to move from being separated to the pericentre of the interaction in the early times of the dynamical timescale. Then, dependent on the velocity of the system, the galaxy pair will either move straight into the merging stage of classification and begin to coalesce or it will

move towards the apocentre stage. This change from pericentre to apocentre can then take many branching paths dependent on the velocity in the system. If the galaxies have sufficient velocity, the apocentre stage will be their end state until the tidal features slowly dissipate. If they do not have velocity to escape on another, the system will move from the apocentre stage back into the pericentre stage. This could then happen for many cycles until finally the galaxies enter the merging stage and coalesce. Figure 3.7 shows the branching paths that the galaxies can take through each stage. Note, these are created using the Advanced Python Stellar Animation Module restricted numerical simulation, described in Chapter 4, and are for illustrative purposes only.

More commonly, in the literature, rather than using the stage of an interaction based on morphology the projected separation of the two systems is used. To explore the difference between using our staging system and the projected separation, as well as to ensure we recover the expected relations, we measure the projected separations of our confirmed galaxy pairs. Figure 3.8 shows the stage classification with projected separation between the two pairs. We measure this by taking the average of the best fit photometric redshifts between the two galaxies, and converting their angular separation to a physical one. The most distinct projected separation ($s_{\rm proj}$) in stages is between the pericentre and apocentre stages. Here, we see that the percentre stage is dominated by systems with $s_{\rm proj}$ <35kpc while the apocentre stage is dominated $25 \leq s_{\rm proj} \leq 100$ kpc. The pericentre stage is visually classified as systems which are highly morphologically disturbed, while either being morphologically linked to each other or overlapping. Those galaxy pairs with large projected separations are pairs which are very large in angular size, while still overlapping or morphologically linked. If this criteria is not met, then the system becomes an apocentre stage interaction where the two galaxies in the pair are completely distinct. There is some overlap between the projected separations of the pericentre and apocentre stage as this is somewhat dependent on the angular size of the two systems involved in the interaction.

Figure 3.13 shows the overlap between our four stages in projected separation. It also shows the limitations in our sample of finding the secondary galaxy in the interaction. This, primarily, is at low redshift where a larger projected separation represents a larger angular separation on the sky. This is a limitation of using

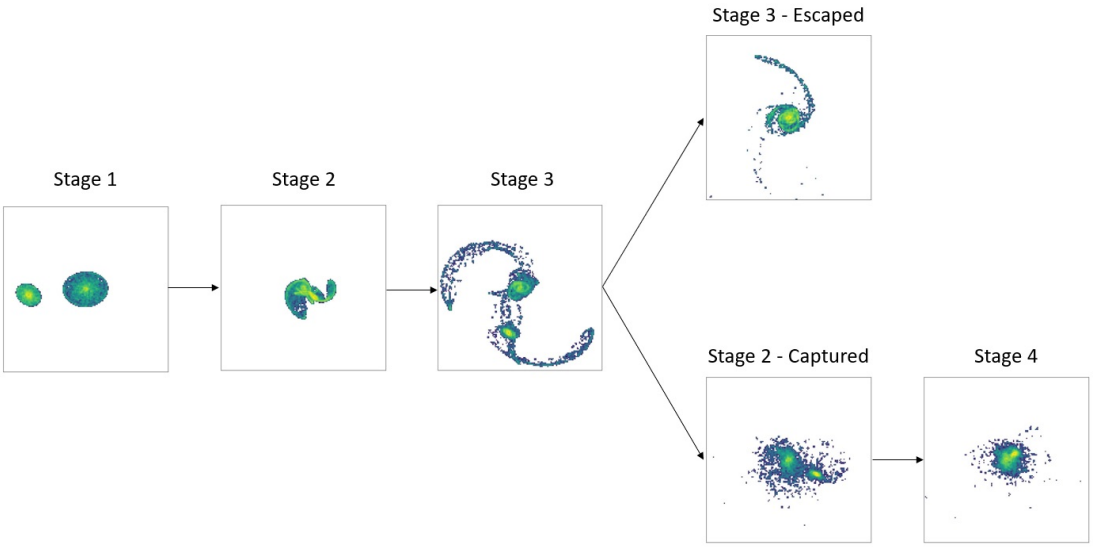

Figure 3.7: The progression through an interaction using our stage definitions. In the separated stage, we have two systems that are approaching each other but exhibit no tidal features. This is before the point of closest approach has occurred. This is followed by an intial pericentre stage: the systems are approximately at their closest approach. This stage is often mistaken, in the COSMOS catalogue, for being of only one source. Clear tidal features exist with major disturbance in the two disks. This is followed by the apocentre stage, where there are two distinct cores with clear tidal features. However, after this point, there are two outcomes to the system depending on the galactic velocities. If the secondary has the escape velocity, the system will remain an apocentre stage interaction until the tidal features dissipate (and no longer are in our sample). If they do not have enough energy, the system will return to the pericentre stage of the encounter and then begin to coalesce in the merging stage. Images are from the Advanced Python Stellar Particle Annimation Module interacting galaxy algorithm described in Chapter 4 and based on the stellar particle animation module algorithm described in Wallin et al. (2016).

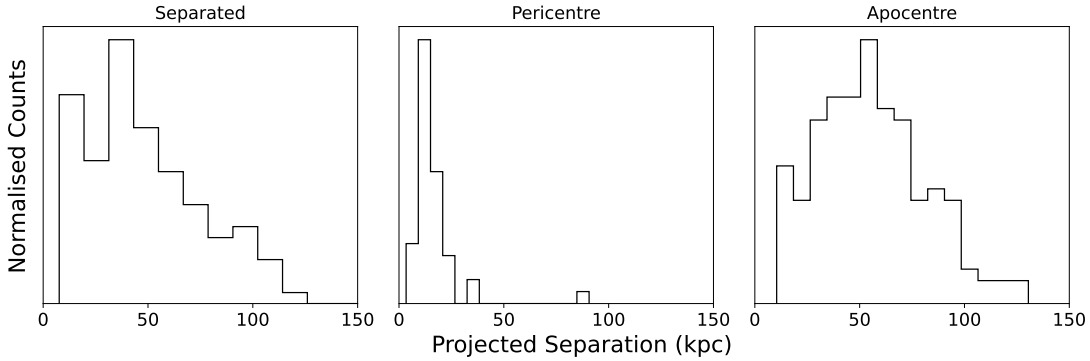

Figure 3.8: The projected separations of the confirmed galaxy pairs in our sample. This confers with other works the definition of our different stages. The separated stage can be at any projected separation, however, we have visually confirmed that these sources are not morphologically disturbed. The pericentre stage is dominated by systems with small projected separation as, by definition, they must be morphologically linked or overlapping. Those systems at larger separation are very large systems whose morphology categorise them as at the pericentre stage. Finally, apocentre stage galaxies are those which are visually confirmed to be fully morphologically separated and tidally disturbed. The bulk of these lie in a range of 25 - 100kpc in projected separation from each other. There is some overlap between the pericentre and apocentre stages in projected separation, as their visual classification is also dependent on the system size.

visual confirmation of the secondary within a cutout of limited angular size. Thus, we see at low redshift ($z < 0.2$) we only identify pairs of galaxies with projected separation below $50kpc$. Towards our limiting redshift, however, we are able to identify galaxy pairs down to a projected separation of 5kpc, as well as out to 200kpc.

We will use the projected separation to investigate relations between AGN activity and if we observe enhancement in star formation with stage. The SFR is already present in the COSMOS2020 catalogue as measured using EAzY/FAST. To ensure that any enhancements are from interaction alone, we must ensure we have no biases in the environment distribution of our sample and identify AGN. First, we investigate the distribution of our sample with redshift and ensure that no effects we may see from stage are actually driven by redshift.

3.3.2 Matching to Environment Catalogue

It is well known that the environment has a direct impact on the observed SFR of galaxies. A galaxy in a cluster environment has, on average, a higher SFR than those in the field (Baldry et al., 2006). Thus, if any of our stage classifications are biased towards one environment or another, it could have severe consequences for our results.

There is no measure of the environmental density in the COSMOS2020 catalogue. Such a measure is often calculated in numerous ways, such as the N-nearest neighbour (Baldry et al., 2006), different Bayesian metrics (Cowan & Ivezić, 2008) or estimating it from Voronoi Tesselation (Vavilova et al., 2021). However, in this work, we use the existing environmental density catalogue produced by Darvish et al. (2017). This catalogue was created specifically for the COSMOS survey, and has a measured density for all sources with mass $\log(\frac{M}{M_\odot}) \geq 9.6$ and $z \leq 1.2$. Darvish et al. (2017) calculate not only the density, but also the density parameter δ and assign each source to a field, filament or cluster classification. For a full description of how they calculate the environment and density field see Darvish et al. (2015) and Darvish et al. (2017), but we will briefly describe it here.

To build the density field throughout the COSMOS field, Darvish et al. (2017) first construct a set of overlapping redshift slices. Within each slice, a subset of the

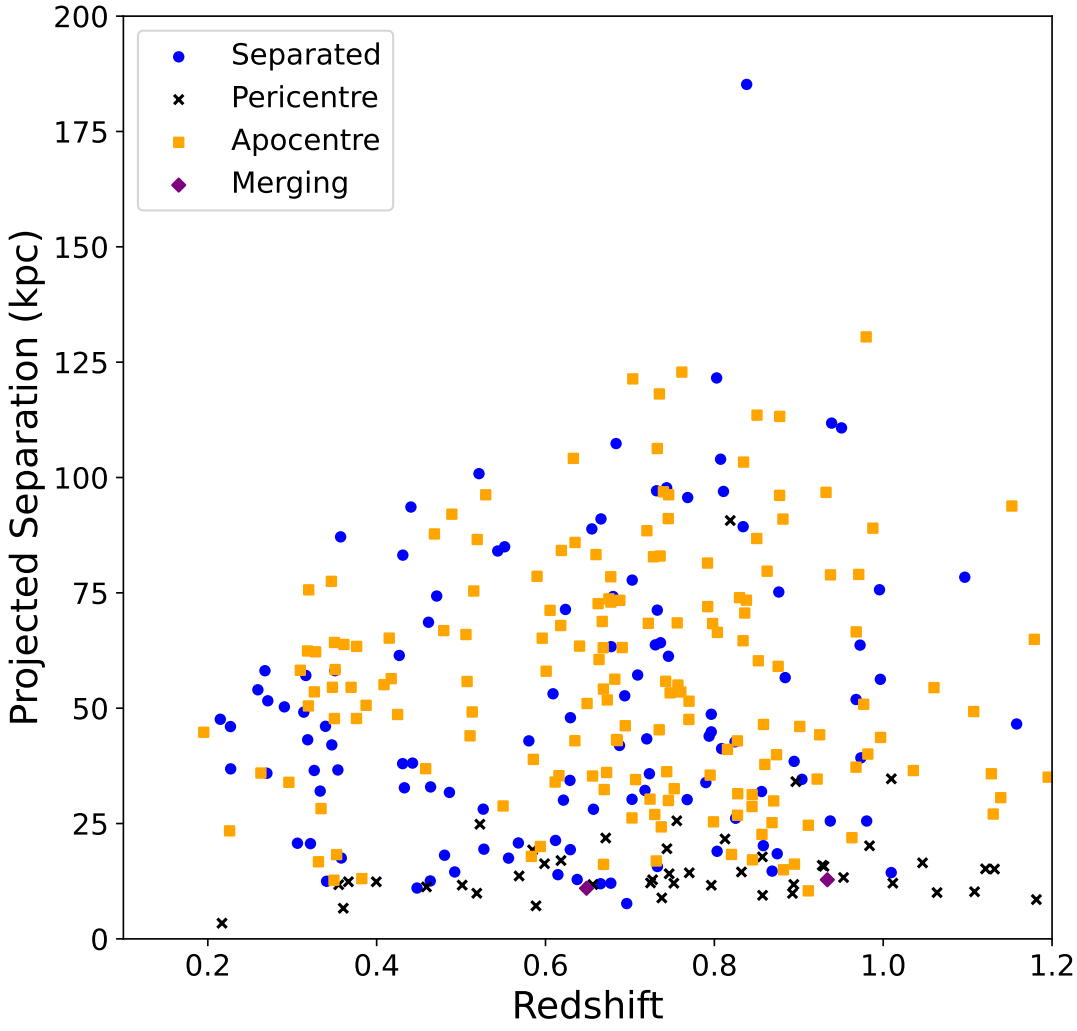

Figure 3.9: The measured distribution of projected separation with redshift. As we have made redshift cuts out to $z = 1.2$, we investigate the limitations on the projected separations we can successfully identify across our volume. We find that we can successfully identify secondary galaxies and their projected separations down to 10kpc to $z = 1.2$. This is true across defined interaction stage. We find, because of the defined size of our cutouts, the limitation of finding the secondary galaxy is at low redshift.

galaxies are selected such that the median of the probability distribution function (PDF) of their photometric redshift is within it. Then, from this subset, they calculate the weighted surface density within the redshift slice. The weighting is based upon the PDF of the photmetric redshift present within the redshift slice. These weights significantly reduce the effects of projection effects. They then apply a weighted adaptive kernel smoothing using a 2D Gaussian kernel whose width changes based on the found local density of galaxies. Once this density field is created, the density around the sample galaxies can simply be interpolated across the density field based on the angular position and the redshift slice the sample galaxy is in (based on its photometric redshift PDF).

The result of this process, and the cuts defined previously, is a catalogue of ≈45,000 galaxies with their densities accurately measured. We remove any sources which are flagged as uncertain from the catalogue (a flag existing within it) providing us with ≈39,000 sources with which to cross match the sample from Chapter 2. We apply the same cuts to our sample as applied in Darvish et al. (2017), and only consider those systems with a mass $\log(\frac{M}{M_\odot}) \geq 9.6$. To cross match with our sample, we use the COSMOS2015_ID which exists in both the COSMOS2020 catalogue and the Darvish et al. (2017) catalogue. Upon applying the mass cut to our sample, we find 2,800 matches to the Darvish et al. (2017) catalogue. Upon matching based upon the COSMOS2015 ID, we find that 628 sources in our sample do not exist in the environment catalogue. This reduces our sample to 2,172 galaxies with confirmed and reliable environment density measurements.

3.3.3 Classifying AGN

We will also investigate the effect of interaction stage on AGN activity throughout our sample. As the COSMOS2020 catalogue does not contain the relevant parameters to make this calculation, we turn to the Chandra COSMOS Legacy Survey Multiwavelength Catalogue (Marchesi et al., 2016) and the COSMOS VLA 3GHz survey (Smolčić et al., 2017; Delvecchio et al., 2017). Both of these catalogues span the entire COSMOS area, and contain detailed classifications of the sources they find. The Chandra survey spans the X-ray range of wavelengths,

and successfully identified numerous X-ray AGN. The VLA 3GHz survey is a radio survey, and we use this to find the radio AGN through our sample.

Our cross match process is very much the same as previously described. We use the previously identified COSMOS2020 source coordinates and select the nearest source with in a 10" matching radius. We first find matches of radio AGN using the VLA 3GHz survey catalogue and then search the Chandra survey. At every step, if we find a match in the relevant catalogue, we remove it from subsequent searches in other catalogues and take the first classification as the correct one.

Applying our matching criteria, we find 1,059 matches in the VLA 3GHz survey and 155 in the Chandra survey. From existing flags within the catalogues, these were split into 812 star forming galaxies and 402 AGN. We also investigate cross matching with the MPA-JHU catalogue (Kauffmann et al., 2003; Brinchmann et al., 2004; Salim et al., 2007), however, found that all matches were already represented by the VLA and Chandra surveys. We also use the COSMOS XMM-survey and, again, find no new sources to add to our sample. While the ratio of AGN to star forming galaxies in our sample seems large compared to other works, it is important to note that this is a result of limited matching between the catalogues. Of the 4,181 galaxies in our sample, only 1,214 appeared at all in either the VLA or Chandra catalogues. Any galaxy which did not appear in these catalogues, we discard as unclassified.

3.3.4 Visual Classification: Sources of Contamination

Throughout this description of our sample, ~~we have noted that~~ we have identified these interacting systems by a combination of visual classification and the best-fit photometric measurements from the COSMOS2020 catalogue. This brings with it certain limitations ~~which we will discuss in this section~~ which could bias or contaminate our sample. We identify three serious areas of bias or contamination: 1) error inherent from using only photometric redshift measurements, 2) failure of identification of tidal features at higher redshifts due to surface brightness dimming and reduction in angular size and 3) mis-identification of tidal features as disturbances caused by other processes. In this subsection, we will address each

of these points and quantify how they affect our sample. We will also discuss ~~any~~ how we mitigate these effects where appropriate.

First, there is always an inherent error in using the photometric redshift measurements of galaxies. We briefly mentioned this when describing pericentre stage systems. There, we found multiple systems which were clearly morphologically disturbed with linking tidal features, but their photometric redshifts were such that ~~they couldn't~~ possibly be interacting. ~~This leads to the question in how reliable our photometric redshifts are, and what is the likely contamination rate in our paired sample, but also in the un-paired sample where we will have missed pair identification due to incorrect photometric redshift measurements.~~ ~~Following the description in~~ Weaver et al. (2022)~~, they~~ quantify the error in ~~the~~ photometric redshift calculations as ~~at the sub-percentage level~~ for our redshift and flux range. This is true for redshifts measured using both `EAzY` and `LePhare`. In ~~fact,~~ ~~they quantify~~ the scatter in the distribution of photometric to spectroscopic redshifts ~~to be~~ $0.025(1+z)$ within our redshift and flux range. This leads to a maximum error on our photometric redshifts in this range of ± 0.06.

To investigate the affected systems, we therefore look at the redshift distribution of our paired sample. Figure 3.10 shows the scatter in our found pairs as well as the difference between them from our visual classification. During the visual classification, we instituted a cut that to be interacting, each galaxy must be within a photo-z of ± 0.1. This was to simply the visual classification at that stage. The resultant scatter in photo-zs between the primary and secondary in our paired sample is shown on the left hand panel of Figure 3.10. As expected (and by design) this follows an approximately 1-1 relation, showing that the galaxies are likely truly interacting. The right hand panel shows the distribution of the difference measured between the primary and secondary photo-zs, with the red dashed lines showing the maximum expected error on the photo-z measurement at the highest redshift of our sample. As shown, the bulk of our paired sample lies at very low redshift difference. This leads us to conclude that our pair selection process has been robust.

The second caveat we will discuss here is the failure of identifying interacting systems at high redshift due to missing the tidal features of the system. This would be a combined result of low mass systems being highly affected by surface

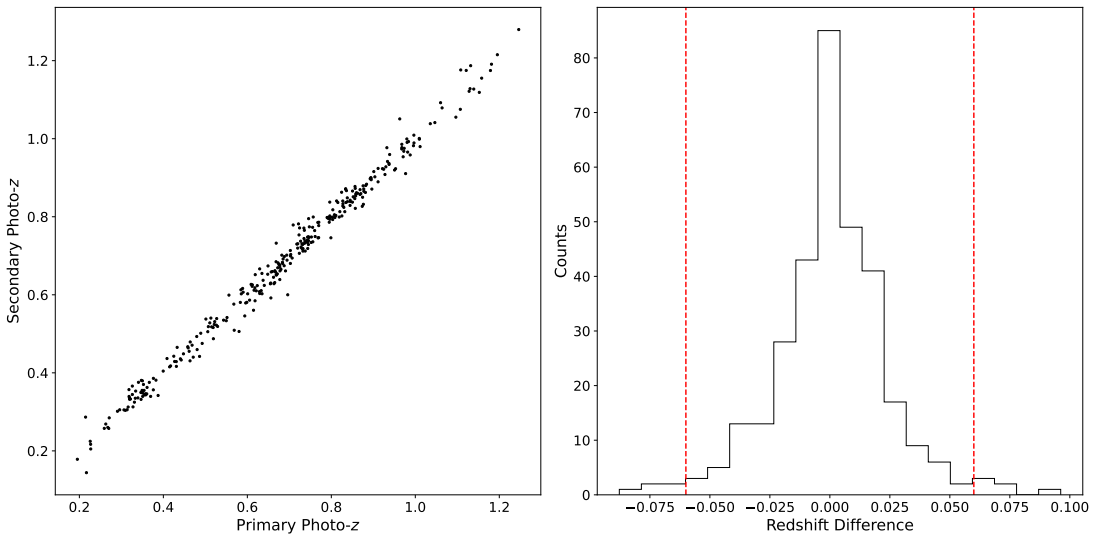

Figure 3.10: The scatter in photometric redshifts of our paired sample. Left ~~shows~~ the scatter in the primary and secondary photo-z measurements. These, approximately, follow a 1-1 relation. Right the measured difference in the primary and secondary photo-zs found in the left panel. The red dashed lines show the expected error at the maximum redshift of our sample. As shown, the bulk of the sample lies within this region.

brightness dimming of their disks and tidal features at high z and that they will have a small angular size. The former issue is rectified by our limiting of our sample to be within a range of $0 \leq z \leq 1.2$ and with our volume limit~~ation~~ of $9.25 \leq \frac{M}{M_{\odot}} \leq 12.5$. The latter is more difficult as we have a standardised cutout size no matter the redshift of the system. This means that if a system were at redshift 1.2, our pixel resolution corresponds to roughly $0.5 kpc pix^{-1}$ limiting the tidal features we would be able to identify. With such a scale we are able to identify extended tidal features away from the galactic disk such as tidal arms, but lack the resolution to those features which would be much closer to the disk. Thus, those interacting systems which have formed features such as shells, small tidal bridges or stellar streams will be missing from our samples at higher redshifts. Thus, at high redshift, we are most sensitive to apocentre and merging stage interacting systems while lacking the required resolution to pickup more intermediate tidal features at the pericentre stage.

Finally, we discuss the major caveat with our approach: that we have been

relying on visual classification to identify interacting galaxies. Besides the issue of close pairs (described in previous sections) there is that of identifying the disturbance of the sample systems is due to interaction, and not from other processes. The most obvious process which could mimic tidal disturbance is that of ram pressure stripping (RPS). RPS is an effect of the galactic environment stripping out the gas within a galaxy, and causing the formation of what appears like debris about the galactic disk. It can also cause major disturbance and irrgular morphology in the galactic disk. Thus, this could easily be a form of high contamination in the apocentre and merging stages of our sample. The environment that RPS is most ~~prelevant~~ is that of a cluster environment. Therefore, if we are highly contaminated to identifying RPS galaxies as interacting galaxies, we would see a bias in the environment of our sample towards galaxy clusters. To check this, we measure the distributions of our galaxies in each environment with stellar mass. We use the sample matched to Darvish et al. (2017) as described in Section 3.3.2.

Figure 3.11 shows the distribution of our matched samples with their density values. It is important to note that Darvish et al. (2017) has a higher mass cutoff than we have implemented in our underlying sample. Therefore, this is showing the density of all systems $\log M_*/ M_\odot \geq 9.6$. We conduct weighted KS and AD tests on the environment distributions, and confirm they are likely drawn from the same parent sample, with the resultant p-values ≈ 1 the distributions. Thus, there are no biases in our stage selection with environment, and it fair to assume that any evolution we find between interaction stages is due to interaction and not due to biases in the environment of our stages samples.

To reinforce this, Figure 3.12 shows the environment classification in our reduced stellar mass-SFR distribution. This shows that there is no ordered structure, or bias, in our sample with environment and further reinforces that the galactic environment is not responsible for effects we may find from interaction. The majority of our sample lie either within filaments or in the field. The environment which would have the most effect upon the SFRs of our sample is a cluster environment. However, the fraction of our sample within a cluster is never greater than 10%. Therefore, the lack of information here does not have a large impact on this measurement.

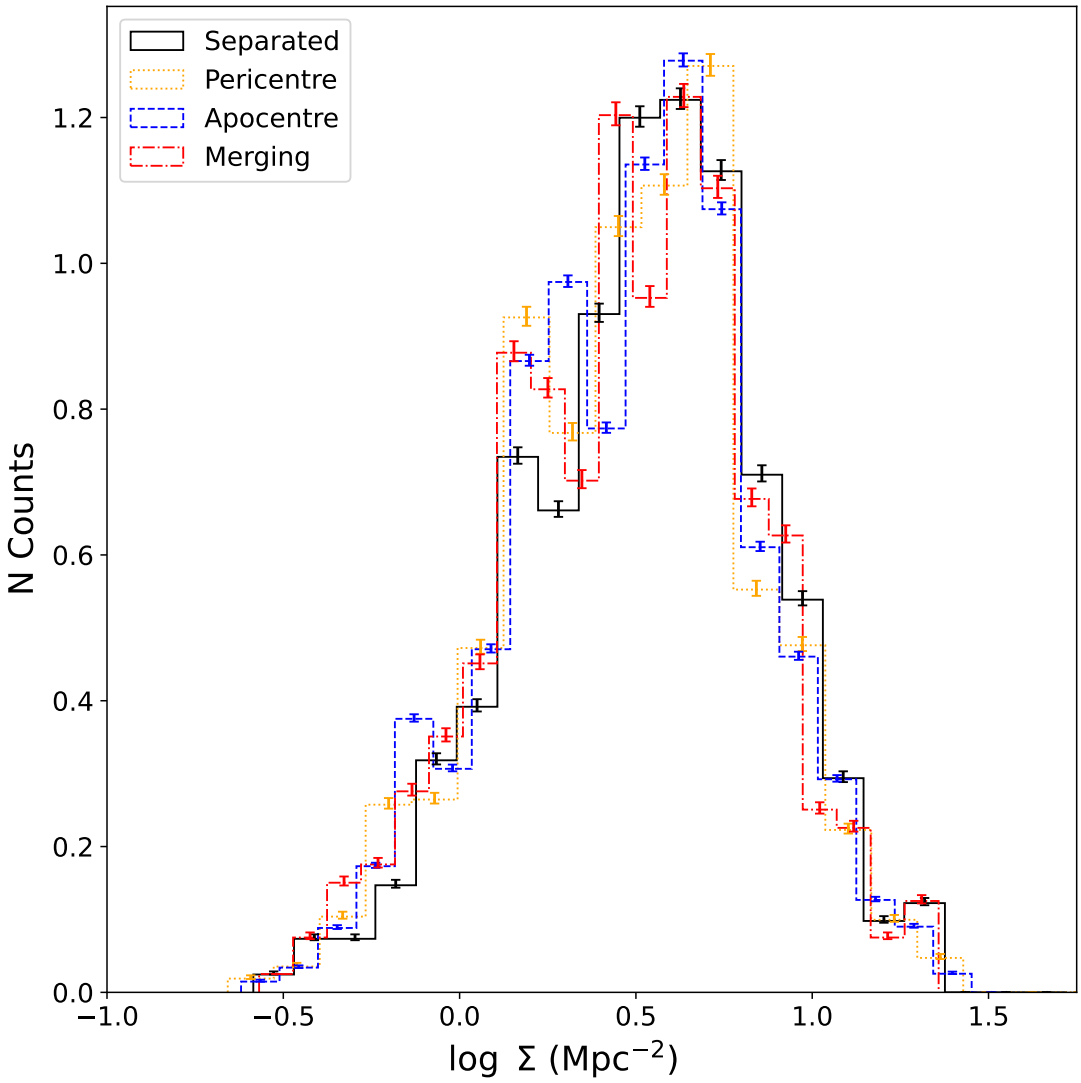

Figure 3.11: The density about each of our sources matched with the Darvish et al. (2017) catalogue. As shown, there is no existing bias in the distribution of galactic environments throughout our stages. Therefore, that we have identified primarily tidal affects and not environmental ones.

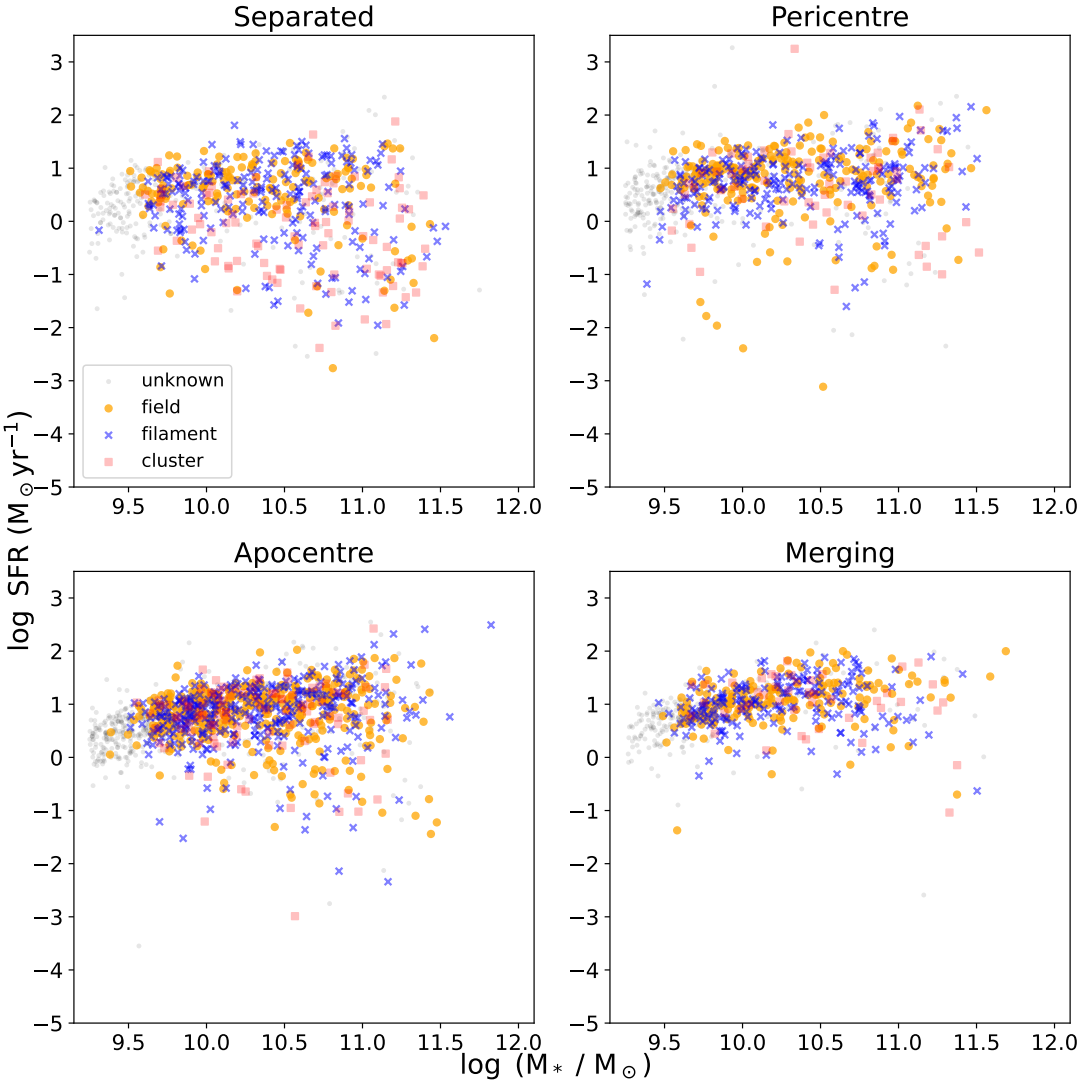

Figure 3.12: The distribution of environment classifications through our sample. This is only for sources with a logM_* / $M_\odot \geq 9.6$. However, from this subsample, we see that there is no trend with environment and that the distribution is random throughout each stage.

Another source that could be of major contamination is in our merging stage sample. We have identified the merging stage as those systems with a disturbed disk and containing more than one core. Another such system that could be recognised by this description is that of a clumpy galaxy. Clumpy galaxies are systems with regions ungergoing intense star formation. These regions appear like multiple other 'cores' ~~existing~~ throughout the galactic disk which are actually clumps of young stellar populations likely contributing to this star formation. To avoid such contamination, we apply multiple criteria to mitigate the potential impact they might have. First, our redshift range spans a sufficient range that the found fraction of clumpy galaxies declines rapidly compared to the merger rate (Adams et al., 2022). This is particularily true at $0.15 \leq z \leq 1$. Previous works have found the peak of the fraction of clumpy galaxies at $z \approx 2$ (Murata et al., 2014; Guo et al., 2018), ~~outwith~~ our redshift range. There are also two morphological distinctions that we use to our advantage to discern between merging clumpy galaxies. Clumpy galaxies often exhibit more than one clump at different radii across the galactic disk (with Adams et al., 2022, finding a mean of 3.16 clumps per galaxy), making them easy to differentiate from a second core which would be close to the centre of a perturbed disk. We therefore remove potential merging galaxies which appear to have more than one 'core' at different places of a non-perturbed disk. With the low fraction of clumpy galaxies in our redshift range and an increased number of cores compared to mergers we are confident we have separated out potential contamination by clumpy galaxies.

3.3.5 Aside: Mass Ratios in Sample

Here, we briefly discuss the mass ratios we find of our identified pairs of interacting galaxies. As stated in Section 3.2.5 we have confirmed 834 different pairs of interacting galaxies. We take this sample and break it down into its consituent stages as well as its interaction type. The interaction type is based upon the mass ratio. The mass ratio is taken between the primary and secondary galaxies. These are always defined such that the primary galaxy contains the highest stellar mass in the pair. If there is more than two galaxies involved in the interaction, we take the primary galaxy to be the galaxy with the highest stellar mass in the system.

We define three different interaction types based on the mass ratio: micro (where the mass ratio is less than 1:10), minor (where the mass ratio is between 1:10 and 1:3) and major (where the mass ratio is greater than 1:3). Figure 3.13 shows the distribution of interaction types through this subsample. We find that our sample is dominated by micro and minor interactions, although we also identify a non-negligable number of major interactions within it.

The mass ratio is known to have a direct impact on the potential enhancements we may see in SFR and in AGN fraction. While we do not focus on the mass ratio throughout this work, due to the limited pair galaxy size, knowing that we are dominated by mirco and minor interactions will inform us of the relations we expect to uncover from our analysis. This is assumes, of course, the subsample of galaxy pairs is representative. With this note on the distribution of interaction types based on mass ratio in mind, we now look at the relation between the stellar mass of our systems and the SFRs through interaction stage.

3.4 STAR FORMATION EVOLUTION WITH INTERACTION STAGE

3.4.1 Controlling for Interaction Stage

With the above samples selected, we investigate the change in multiple parameters with stage of the interaction. The four stages of interaction are designed to capture the main different parts of galaxy interaction. First, we show the results of breaking down our sample into stages with relation to the star formation and mass of our sample. We use the estimates of these parameters that exist in the COSMOS2020 catalogue itself. Then, we use our subsample of galaxy pairs to recover the relationship between projected separation and star formation enhancement (SFE). We then further break this measure down into its component stages.

Figure 3.14 shows the breakdown of stellar mass and SFR with stage. On a population scale, there is clear evolution in the star formation rate from the separated stage through to the merging stage. In the separated stage, where the

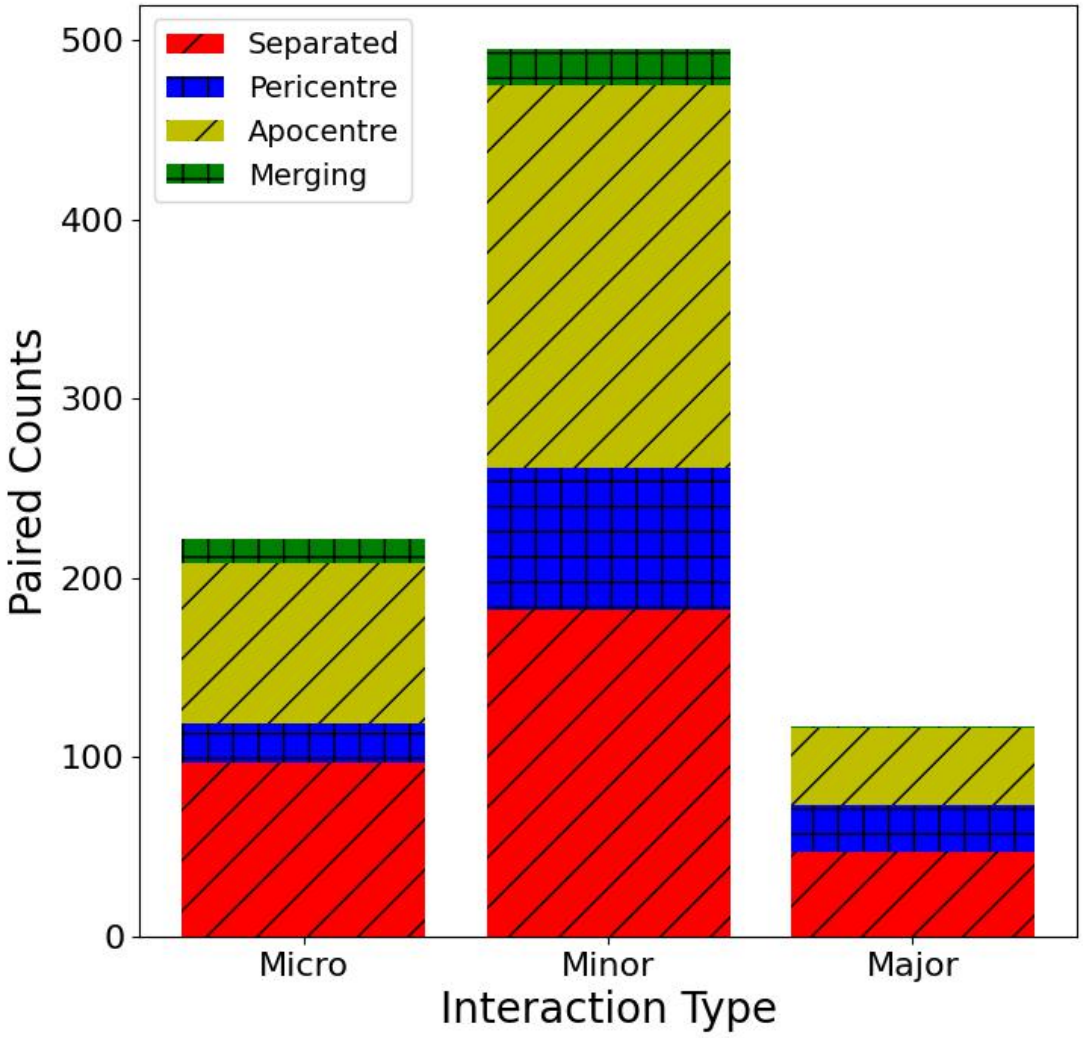

Figure 3.13: The distribution of mass ratios in our paired sample and across each stage of interaction. These are separated into our previously defined interaction stages and the interaction type (based on mass ratio). We find that, overwhelmingly, our sample is dominated by micro and minor interactions. These are where the two systems have very different stellar masses.

galaxies are distinct from one another with no clear morphological disturbance, we clearly see two populations of galaxies. These are the blue, star-forming galaxies and red sequence of galaxies. The blue contours in Figure 3.14 show increasing number density in the population into the blue cloud. In the pericentre stage, when the galaxies are actively interacting and overlapping, this red sequence remains but is highly diminished while there is no change in the blue cloud. The apocentre stage shows a similar effect, where the red sequence reduces again before finally in merging the red sequence completely disappears. Through these, the blue cloud remains highly populated and hosts the majority of galaxies in the sample.

However, this result could also be due to many other factors rather due to interaction stage. For instance, if the mass distribution of our sources evolves as well, we could simply be selecting higher mass systems as we increase stage. This would have the result of systems in the merging stage having, on average, higher SFRs than those in the separated stge and appearing like we had evolution in the star forming population with stage. Another effect that could cause this relation to appear would be our selection was highly dependent on galactic environment. It is well known that environment and star formation are closely linked, and that classifying interacting galaxies by morphology classifiers can weight up galaxies in cluster environments over galaxies in the field. In Section **??**, we control for the environment in our sample and show that this is not the case. Here, we investigate the question of the evolution of the mass distribution that could give us this result.

We can quantify the similarity of the mass distributions, and then the SFR distributions, using well known statistical tests. We opt to use two different tests: the Kolmogorov-Smirnov (KS-test; Kolmogorov, 1933) and Anderson-Darling (AD-test; Stephens, 1974). The KS-test is excellent at comparing different weighted distributions and indicating if they are drawn from the same parent sample. The AD test tests for this similarity as well, and we use both to ensure consistency and robustness in our measurements. First, we create weighted distributions of stellar mass. The weighting scheme we use balances the distributions such that each bin could be assumed to have the same number of sources within it. Therefore, any bins with fewer than a certain number of sources will be weighted up

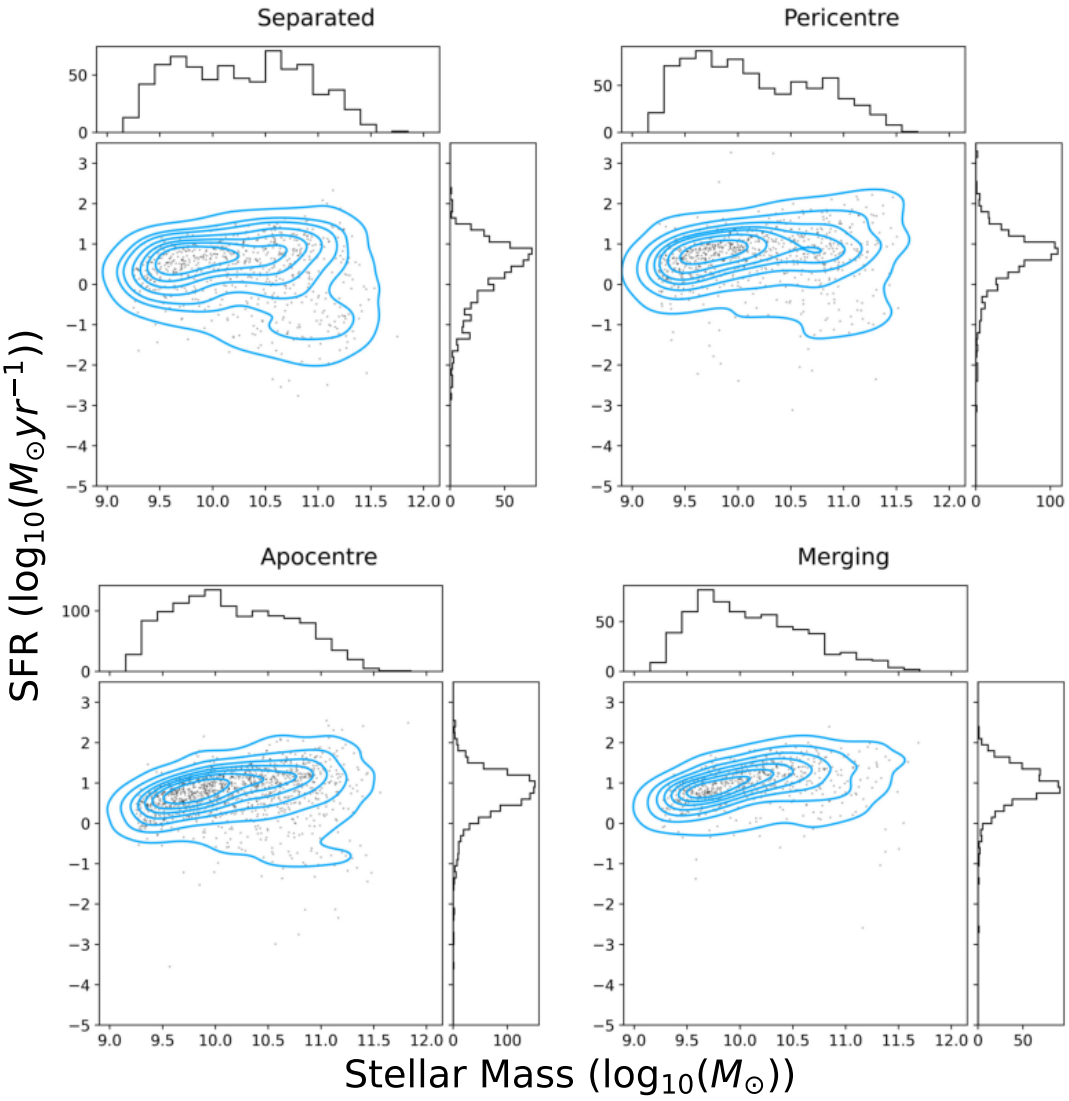

Figure 3.14: The LePhare stellar mass against the EAzY SFR across the different stages of the interaction. The blue contours are 8 levels of density of the underlying populations in each frame. *Top left*: The stellar mass and star formation rate of the separated stage of our sample. Here, the interacting galaxies are simply close pairs with little to no morphological disturbance. There are clearly two populations here: a main, star forming sequence forming the main population and a smaller red sequence. *Top right*: pericentre stage of the interaction, where the two interactors are close to pericentre. The star forming sequence remains, but the red sequence is reduced significantly. *Bottom left*: apocentre stage of the interaction, where the interactors are close to apocentre or escaped. Here, we see the almost complete disappearance of the red sequence. *Bottom Right*: merging stage of the interaction, where the two systems are close to or have coalescence. The red sequence of galaxies has completely disappeared.

while those bins with more will be weighted down. These weights applied to the mass distribution are then applied to the SFR distribution as to control for stellar mass in this distribution.

Figure 3.15 shows the weighted mass distributions through the four different stages. These appear highly similar, however to statistically measure the similarity we apply our KS- and AD-tests to them. We chain the KS and AD test through each distribution and calculate the value of the test values and a p-value. This p-value represents the probability that each distribution is drawn from the same parent distribution. For each mass distribution, we find that p-value of both tests is ≈ 1. Thus, proving that the distribution ins tellar mass though each stage is similar and drawn from the same parent sample.

Figure 3.16 shows the SFR distributions while being weighted by stellar mass. When comparing the separated and pericentre, separated and apocentre, separated and merging, pericentre and merging and apocentre and merging stages, the p-values are $\ll 0.05$. This allows us to reject the null hypothesis for these distributions and assume they are from different parent samples. However, for comparing pericentre and apocentre stages, the p-value of the KS test is ≈ 0.90. Thus, while these distributions are likely to be not identical, they are very similar in the parent sample they have been drawn from. For the same mass distribution through each stage of interaction, the star formation distribution changes from separated to pericentre stage and from apocentre to merging stage, while remaining similar from the pericentre to apocentre stage. The errors on these distributions are calculated using the methodology of Cameron (2011) for the beta distribution around the binomial population. In this approach, we assume that the error about our measurements is simply a measure of the limited samples we have of the underlying beta distribution of the full population. Thus, by providing a confidence measure of 68.3%, we are able to estimate the size of the errors using the 16% and 84% confidence intervals.

Putting this result into the context of the dynamical timescale of an interaction, it shows there are distinct points at which the SFR changes in these systems. The first is when the interacting system moves from being a close pair to actually morphologically disturbing each other in a close flyby. This difference, most likely enhancement, then persists through through to apocentre stage - where

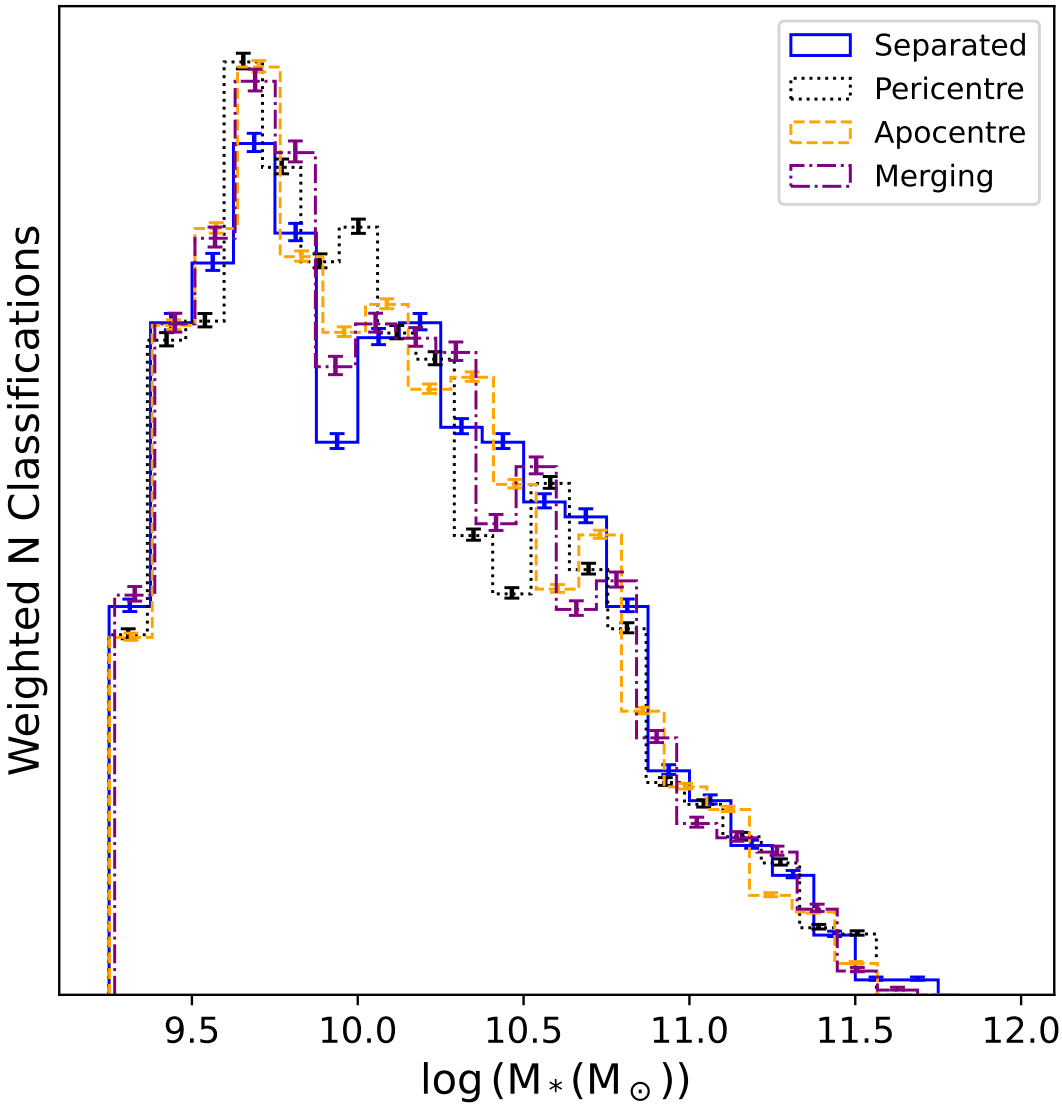

Figure 3.15: The stellar mass distribution across the four stages. Each bin is weighted based on the counts in the smallest sub-sample in stage: the merging stage of the interaction.

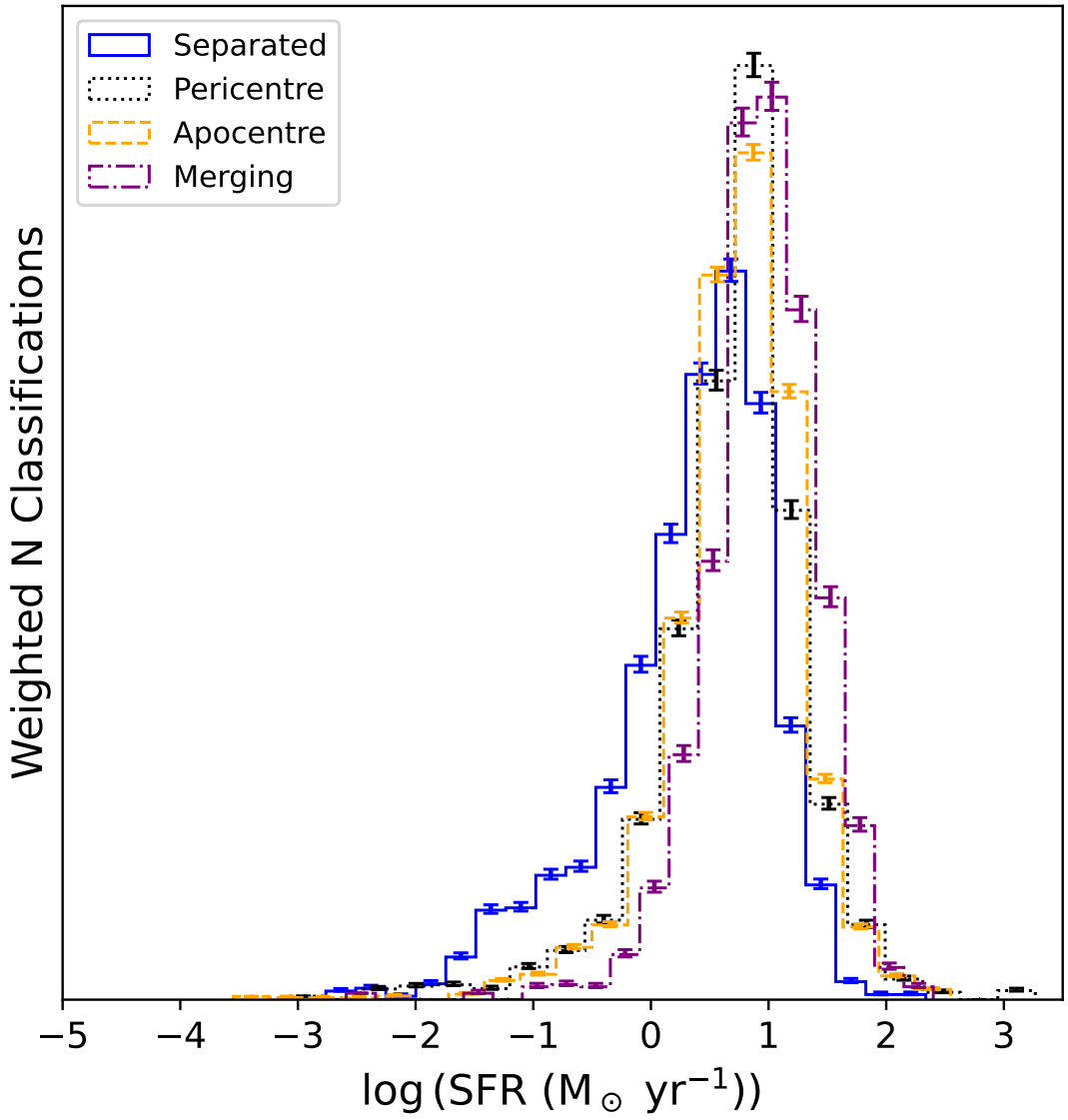

Figure 3.16: SFR distribution weighted by mass across each stage. This weighting is based on our sample of merging stage.

the galaxies remain highly disturbed but are no longer overlapping with their secondary. The SFR distribution remains approximately the same between these two, meaning the forces that drive and affect star formation remain equivalent between these two stages. Finally, the SFR changes again when we approach the merging or post-merger stage of the interaction. We can also say that this change is likely an enhancement in the SFR of the galaxies through the interaction due this change being driven by the disappearance of the red sequence through each stage.

We further examine this result by controlling for redshift and making specific classification of galaxy star formation class. While the red sequence is significantly reduced across across each stage, it is difficult to ascertain the change in the galaxies within the blue cloud. We expect that, due to an interaction, the population of starbursting and quiescent galaxies will change. Therefore, we define a star forming main sequence (SFMS) through across redshift. This allows us to classify each source based on its SFR independently of redshift. To define the star forming main sequence, we follow the example of Aird et al. (2019). There, they define the star forming main sequence as a function of galactic stellar mass and redshift as

$$\log \mathrm{SFR_{MS}}(z)[M_{\odot} yr^{-1}] = -7.6 + 0.76 \log \frac{M_*}{M_{\odot}} + 2.95 \log(1+z). \tag{3.1}$$

This finds the expected, main sequence, SFR of a galaxy at a given stellar mass, M_*, and redshift, z.

The ratio of the measured galaxy SFR and the expected main sequence SFR is then taken. We then use this fraction to classify each galaxy into distinct bins. These are:

(i) Starburst galaxies. Here, the galaxy SFR is highly elevated compared to the SFMS. We follow Aird et al. (2019) and define a cutoff of $\log \mathrm{SFR}/\mathrm{SFR_{MS}} > 0.4$.

(ii) Main sequence galaxy. The galaxy is within 0.4 dex of the SFMS and approximately has the expected SFR. Defined as $-0.4 < \log SFR/SFR_{MS} < 0.4$.

(iii) Sub-main sequence galaxy. A galaxy whose SFR is below the majority of the SFMS, but likely not quiescent. Defined as $-1.3 < \log SFR/SFR_{MS} <$ -0.4.

(iv) Quiescent (High) galaxy. A galaxy with an SFR in the top ≈50% of the quiscent galaxy population. Defined as $-2.3 < \log SFR/SFR_{MS} <$ -1.3.

(v) Quiescent (Low) galaxy. A galaxy with low SFR and very likely completely quenched. Defined as $\log SFR/SFR_{MS} <$ -2.3.

We split our sample into its different stages and apply these criteria. Figure 3.17 shows the ratio between the expected SFMS SFR and the measured SFR in the COSMOS2020. This clearly shows a large increase in galaxies classified as starburst from the separated to merging stages and a large reduction in the number of quenched systems. Figure 3.18 shows the change in fraction of the different galaxy classifications through stage, reflecting the results found in Figure 3.14. Initially, in the separated stage, we find that the majority of our galaxies lie on the SFMS or just below it. There also exists a small population of galaxies which are classified as starburst with a population of quiescent galaxies that is roughly double the starburst fraction. As we move through the interaction stage, we see that the quiescent galaxy fraction gradually decreases to the point of almost non-existence in the merging stage galaxies. The inverse is true in our starburst fraction. We find that this almost quadruples in the fraction of our sample over the course of the different stages of interaction. The fraction of galaxies on the SFMS remains dominant throughout, however, we do find the fraction of sub-MS galaxies significantly reduced. Thus, we find that, in general, the SFR of these galaxies is increasing with interaction stage (though, not in the pericentre to apocentre stage). It appears to have sub-MS and quiescent galaxies move up and join the SFMS (as it remains dominant). But, many galaxies from the SFMS are moved upwards and into a starbursting phase.

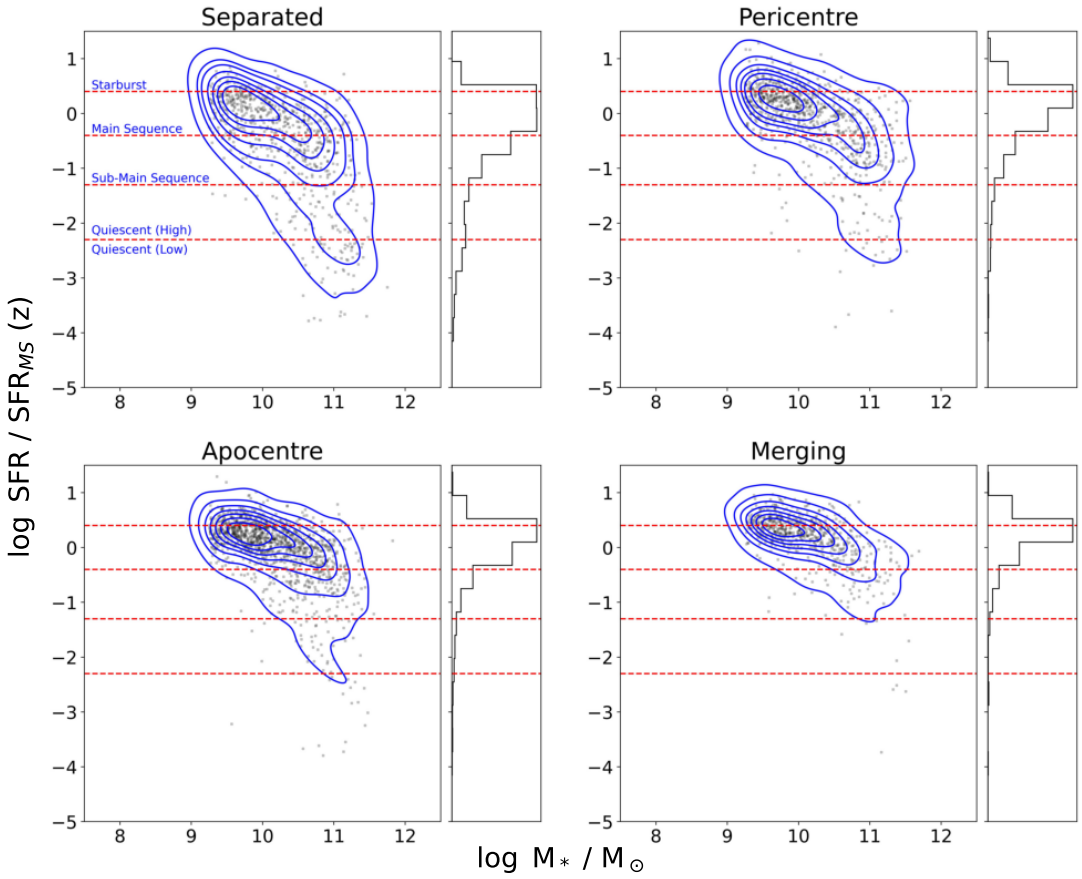

Figure 3.17: Stellar mass against the ratio of measured SFR to the expected SFR if the galaxy was on the SFMS. Black points are the individual sources, while the blue contours are as in Figure 3.14. The red dotted lines show the cutoffs for different galaxy classifications based on their SFR, with each cut off being defined by the text in blue. We find that through interaction stage, the quiescent galaxy population significantly reduces while the starburst population rapidly increases. As these cutoffs are also dependent on redshift, we find that this evolution in SFR with interaction stage is independent of redshift.

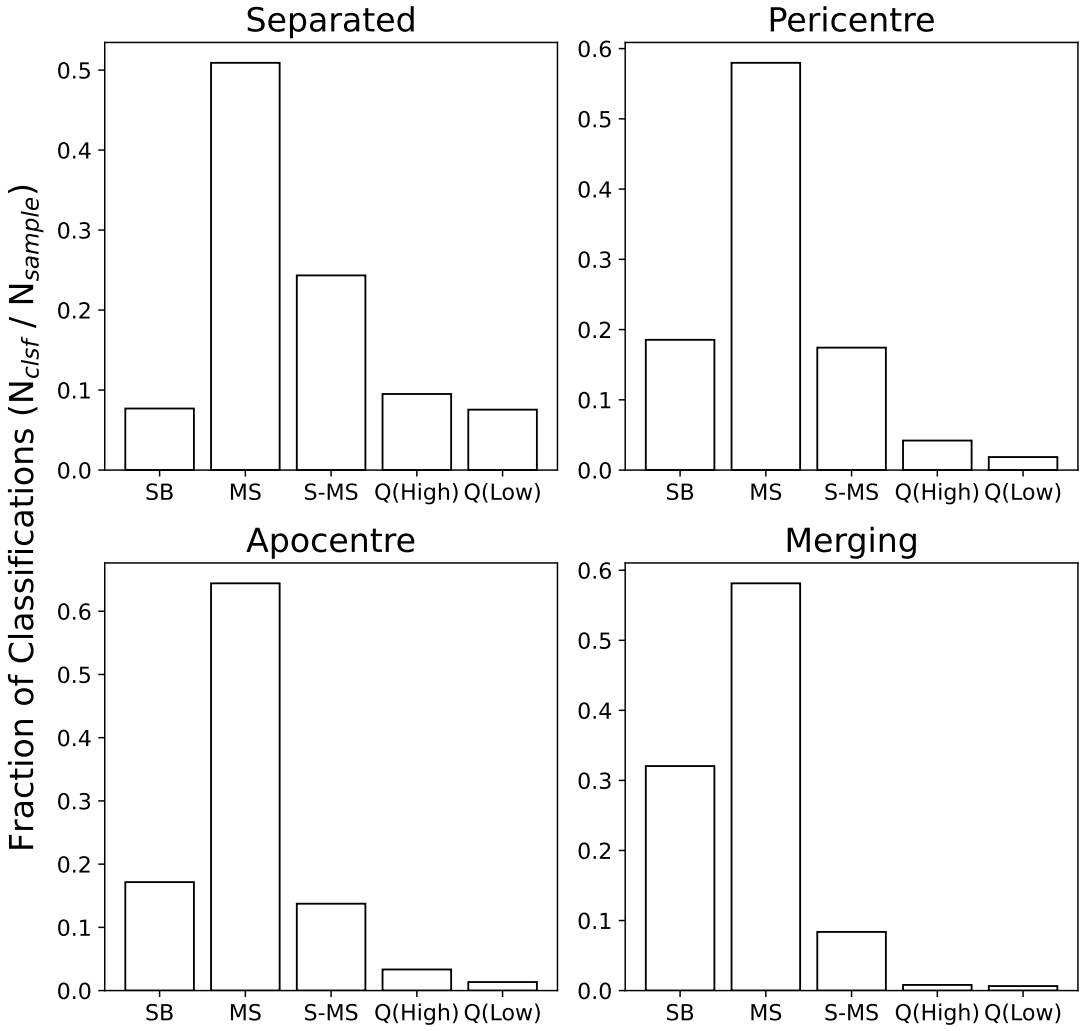

Figure 3.18: The change in fraction of different galaxy classifications from the fraction of SFR to the expected SFR on the SFMS. While galaxies on the SFMS remain dominant through each sample across interaction stage, there is significant change in the starburst and quiescent populations. The starburst galaxy population moves from being roughly half the size of the quiescent population in the separated stage to completely dominating it in the merging stage. This is occurring while the quiescent population is significantly reduced to almost non-existence in the merging stage.

It is important to note the parameter space that we are searching in these examples. We are probing interactions between galaxies of high mass, where the resultant tidal features that form would be classifiable in an image. The majority of our sample is minor interactions where one of the two systems is very highly perturbed by the interaction. In this parameter space, we would expect large increases in the SFRs. However, we do not see a significant increase in the starbursting population until we the two galaxies begin to actually coalesce. Thus, the effect of the interaction itself may be to only enhance the SFR, while at coalescence we find the dramatic starburst. This can be argued from our results of clear, statistically robust change and evolution of a galaxy's SFR with interaction stage. The SFR changes dramatically from the separated to pericentre stage and apocentre to merging stage - after the initial passage of closest approach and at the point of coalescence of the interaction.

3.4.2 Projected Separation and Star Formation Enhancement

We directly investigate the relation between the star formation enhancement (SFE) and the projected separation using our confirmed sub-sample of galaxy pairs. This sample is significantly smaller than our non-pair sample: containing 480 pairs or 960 galaxies (we have also added some pairs from the additional interacting galaxy sample). Upon sub-dividing this into different stages we find 310 separated galaxies, 146 pericentre galaxies and 498 apocentre galaxies. Only 3 merging galaxies were in our confirmed galaxy pair sub-sample, and therefore we do not attempt to make inferences about this population.

To measure the SFE of our galaxy pairs, we directly compare to the mass- and redshift-matched control sample that was also created with this sub-sample and defined in Section 3.2.2. We separate our galaxy pairs into different bins based on the projected separation between them. We find that the bulk of our sample has a projected separation between the two galaxies of $\leq$50kpc. Therefore, we sample from this region of the parameter space with high precision and smaller bin widths before we increase the bin widths at larger projected separations. We

define a cutoff that each bin must contain at least 10 counts to be included in this plot and, therefore, by changing the bin widths with projected separation we are able to maintain some level of statistical robustness. We define our bins as [0.5, 10, 20, 50, 100, 125, 150] kpc.

We create two projected separation distributions: one of interacting galaxies and the other of our control galaxies. We then take the average of each bin. By taking the average of each bin, we find the total excess of the SFR caused by interaction alone and then compare it to what we would expect from non-interacting galaxies in the same bin. Note, that the control galaxy relation to the projected separation is meaningless as they are not paired. The control galaxies are simply mass and redshift matched to each interacting galaxy and used as a reference for what SFR we would expect of it had it not been interacting. Finally, we divide the the averaged interacting SFR distribution by the averaged control SFR distribution. This provides us with a measure of the excess star formation due to interaction, and the SFE when compared to the regular population.

Figure 3.19 shows star formation enhancement between our interacting and control binned SFRs and the projected separations between each galaxy pair. On the left of the plot, we have the distribution for the full galaxy pair sample, without taking account of stage. The errors on our measurements are calculated as the standard deviation from the mean within the bin. The dashed black line represents no enhancement in SFR, as the average SFR in the interacting galaxy sample would be equal to the average SFR in the control sample. We find at very small projected separation a SFR enhancement of 1.87 which gradually decreases with projected separation down to approximately 1 at 117.5kpc.

On the right, we break the galaxy pair sample into different stages and plot out the resultant enhancement in star formation. This shows very different behaviour in enhancement dependent on the stage of the interaction. In the separated stage, we find that the star formation is very weakly enhanced, and with no overall structure. The enhancement moves around 1 through the distribution, with some enhancement being a result of low number counts in the bin. Taking into account the errors on our measurements, the true value of the enhancement could be around 1 for the entire projected separation distribution. This is not unexpected. The separated stage represents when the two galaxies are close pairs, but with

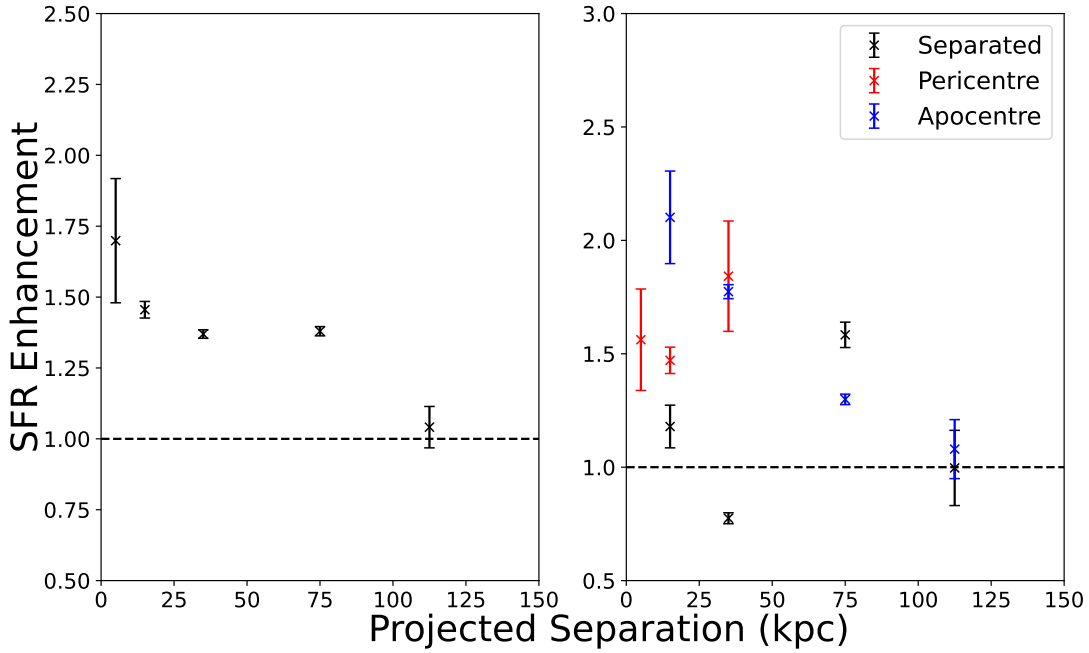

Figure 3.19: The projected separation against the star formation enhancement in average star formation at different bins of projected separation. Each bin must contain at least 10 counts to be considered. As the bulk of our galaxy pair sample is at low projected separation, we heavily sample from this region of the parameter space. The bins are: [0.5, 10, 20, 50, 100, 125] kpc. As we move to higher projected separation, the bins increase in width to maintain statistical significance in our sample. *Left*: The star formation enhancement found in the entire galaxy pair sample. As expected, we see a gradually decreasing enhancement. *Right*: As left but broken up into different stages of the interaction. Black markers are the separated stage, red the pericentre stage and blue the apocentre stage. We limited our investigation to the separated, pericentre and apocentre stages as only three galaxy pairs were identified in the merging stage. We find a generally decreasing star formation enhancement with projected separation but very different individual behaviour dependent on the stage classification.

no morphological disturbance. Therefore, we would not expect significant, if any, enhancement in the SFR based on the projected separation. This is the stage with the lowest enhancement across projected separation, remaining close to 1.

In pericentre stage, we see consistent enhancement with projected separation. It is also much larger than the separated stage, with it being approximately 1.75 up to 1.90. In fact, the measured enhancement increases across the 50kpc of projected separation. However, accounting for the errors on these measurements and the declining counts in each bin, it is likely that the enhancement does drop as the projected separation increases. We see a very large enhancement in the overlapping apocentre stage galaxies which overlap with the pericentre stage measurements. The enhancement in apocentre stage galaxies then rapidly declines as we move to higher projected separations. After 150kpc, our galaxy pair sample does not have the counts to make robust estimates of the SFE. However, it is important to note that this result has been found with a sample of 960 interacting galaxies in their pairs. No merging stage galaxies have been represented and in numerous projected separation bins the counts are small enough to have conclusion altering error bars. Therefore, it is imperative that future works, with larger sample sizes not only look at the projected separation of the systems but the morphology as well.

3.5 Nuclear Activity with Interaction Stage

We now use our AGN sample from cross matching with the AGN catalogues described in Section 3.3.3. As stated previously, we find matches of 1,361 AGN and star forming galaxies (SFGs). The breakdown of number of classifications per stage is shown in Table 3.1. Figure 3.20 shows distribution of stellar mass to SFR with stage, with the confirmed AGN and SFGs marked. We find no major changes in AGN with stellar mass or SFR with stage in our sample. Galaxies containing AGN appear throughout the starburst, the SFMS and red sequence across each stage. There is no obvious bias in stellar mass or SFR for hosting an AGN. We do, however, see a concentration of AGN in the separated stage on the red sequence at higher masses. This shifts as we increase interaction stage, with

	Separated	Pericentre	Apocentre	Merging
SFG	135	212	337	199
AGN	103	117	163	95

Table 3.1: Breakdown in number of classified AGN and SFGs per stage. With our counts so low in this sample, it is difficult to make concrete conclusions about the evolution of AGN during interaction.

the bulk of the AGN beginning to appear at lower masses in the apocentre and merging stages. Finally, in the merging stage we see AGN and SFGs distributed almost evenly across the star forming main sequence with little bias in distribution with mass. This hints at evolution with stage.

Applying KS or AD tests to these, however, reveals p-values close to 1 and therefore these distributions are consistent with being drawn from the same parent sample. However, due to the low counts of AGN and SFGs in our sample, using KS and AD tests to verify this is not optimal. Therefore, we instead investigate the global AGN fractions in our sample across different stages. We apply a weighting, as in our SFR and stellar mass distribution examples, to account for the various counts across the different stages. We then take a second weighting of the measured AGN fraction from the total size of the different samples we have. Thus, we have assumed that we have equivalent counts in each mass bin of our sample as well as approximated having the same sample size in each stage bin.

Figure 3.21 shows the changing AGN fraction with interaction stage. The dotted line shows the expected fraction of AGN in the field from Ellison et al. (2008). We find that the AGN fraction generally remains static with stage and approximately at the fraction of the field. There is a small decrease increase from the separated to pericentre stage, followed by the fraction mostly remaining unchanged until the apocentre stage. Then, in the merging stage we measure a large increase in the AGN fraction from 0.12 to 0.15, far above the expected field fraction. However, before we discuss this result further, we must point out that the large error bars on our measurements and the low number counts of confirmed AGN and SFG make this difficult to interpret. We measure the errors as described for the stellar mass and SFR distributions in Section 3.4.

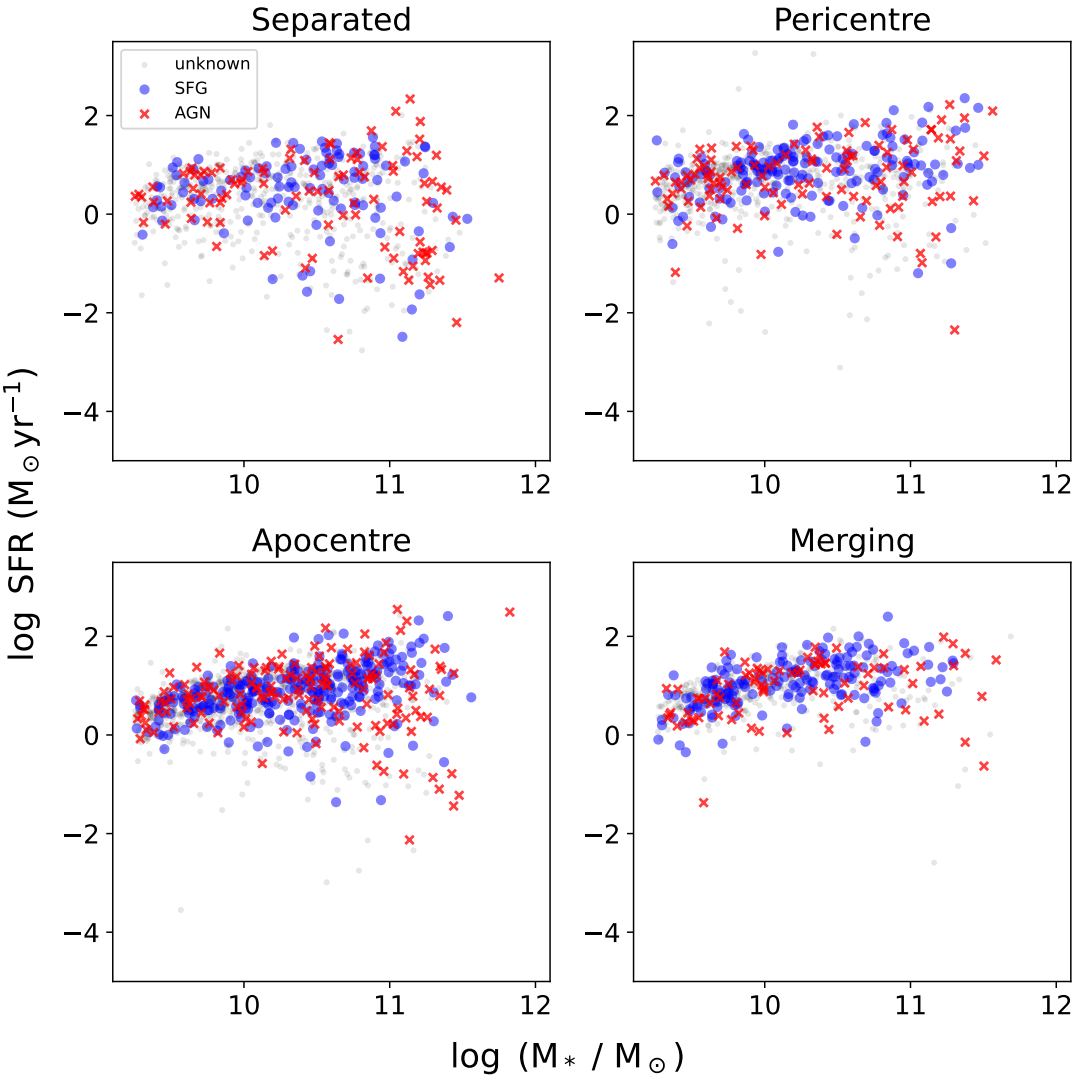

Figure 3.20: The distribution of AGN through stage with SFR and stellar mass. We find that the AGN populate every part of the SFR-M_* parameter space probed in this merger sample.

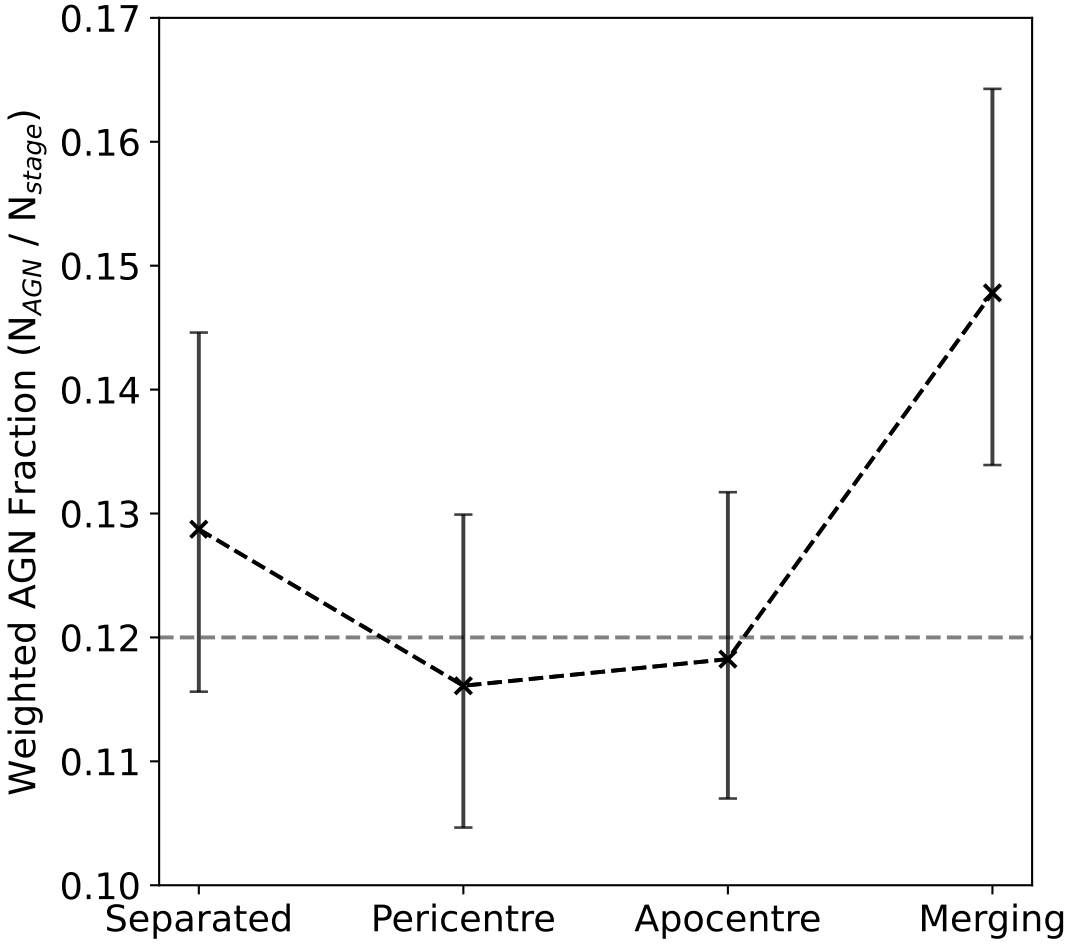

Figure 3.21: The change in AGN fraction with stage. The dotted line is the expected AGN fraction in the field (from Ellison et al. (2008)). We define the AGN fraction as the ratio of number of confirmed AGN divided by the number confirmed AGN and SFGs in the sample. This is then weighted by the relevant sizes of each subsample and the number of unclassified sources in each. Errors on each fraction are found as the confidence intervals defined via the beta distribution. These confidence intervals are calculated based on the 16% and 86% confidence interval. We do not include any sources of which we have no information. We find a sudden drop in AGN fraction from the separated to pericentre stage, followed by a rapid increase from the apocentre to merging stage.

It is also difficult to use interaction stage as a proxy for the projected separation in this context. As shown previously, the pericentre and apocentre stages have some overlap in the range of projected separations we have classified them into. Therefore, we also investigate any overlap between our confirmed pairs and the AGN fractions we see here. We find 104 AGN and 180 SFGs overlap with our confirmed pairs. This sample is far too small to breakdown further into different stages. Therefore, Figure 3.22 shows the change in density of AGN classifications and SFGs with increasing projected separation. As expected, we find that with increasing projected separation the number of confirmed AGN and SFGs decreases. In the AGN distribution, we find two contributing components in it. The first is a peak in the projected separation distribution is from 0kpc-25kpc. The only contribution here is from pericentre stage galaxies. The second peak, from a projected separation of 30kpc - 60kpc is dominated by apocentre stage galaxies, although some separated stage AGN are in this peak as well. Finally, the third peak from 85 - 125kpc is representative of a mixture of the separated and apocentre stage galaxies. None of our merging stage galaxies which had their secondaries identified overlapped with our AGN and SFG identified sample.

We find the AGN fraction only increases at the point of coalescence in interaction: the merging stage. This indicates that the mechanism driving AGN ignition in interaction is primarily at the point of the closest approach between the two systems. When we investigate the AGN number counts with respect to the projected separation of systems, we find that it gradually declines over the projected separation. However, we find two peaks; first clearly in the pericentre of the interaction where the galaxies are overlapping and secondly in the early parts of the apocentre stage. This could be evidence of a delayed AGN ignition depending on the underlying parameters of the interaction. Thus, we find that the AGN fraction is generally unchanged with stage, however, the mechanisms responsible for an increase in the fraction primarily occur when the two systems are actually merging. We find evidence of a delay of ignition, as there is an increase in fraction at a early projected separations in the apocentre stage.

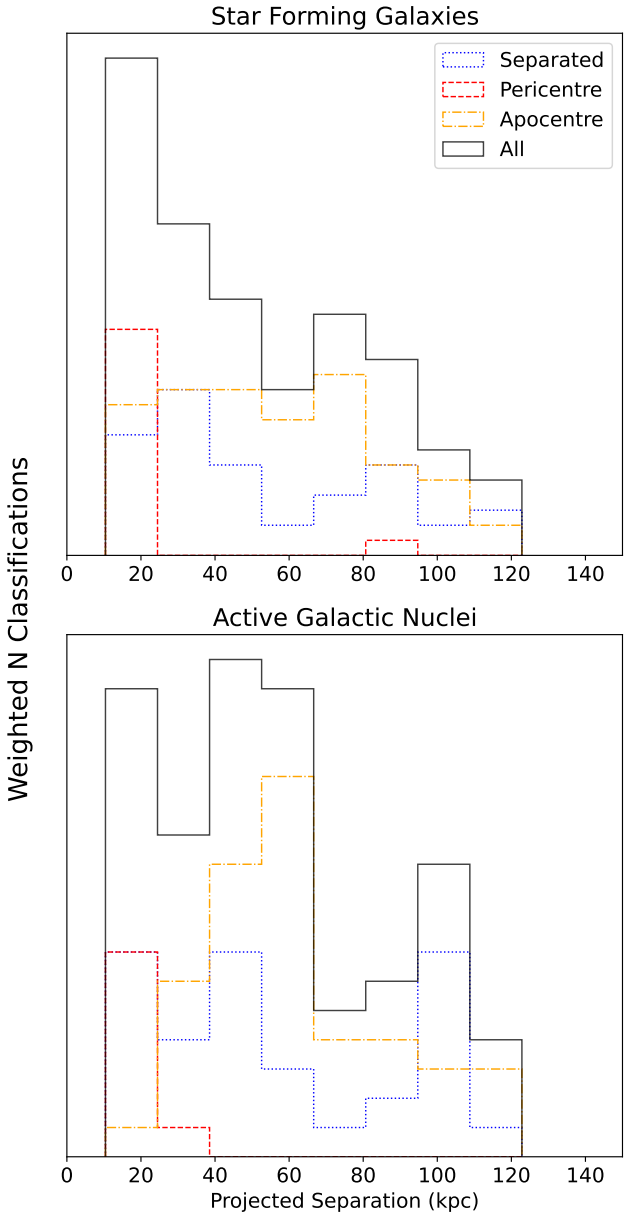

Figure 3.22: The distribution of both AGN and SFGs which have been cross matched with our confirmed galaxy pairs. While we find, as expected, the number of AGN and SFGs decreases with projected separation we find an interesting double peak in the AGN sample. The first peak, between a projected separation of 0 and 25kpc is primarily from pericentre stage galaxies while the second peak of 40 to 65kpc is of apocentre stage galaxies in our sample. As shown in Figure 3.21, we expect the fraction of AGN to be similar in these two stages despite being at different parts of the dynamical timescale in interaction. The third peak, from 75 to 125kpc is only from separated and apocentre stage galaxies.

3.6 Discussion

3.6.1 Interaction Stage and Projected Separation

Finding evolution with interaction stage is not a new idea in the field. Multiple works have found increases in SFE and SFR as a function of projected separation of pairs of galaxies (Barton et al., 2000; Ellison et al., 2008; Patton et al., 2013). Projected separation is often seen as a proxy for the point in the dynamical time of the interaction that is being measured. It is likely that, when the galaxies in the pair are closer together, they are closer to coalescence in the dynamical time and can be thought of as a linear progression in the interaction from being close pairs to coalescence. However, what this fails to capture is the larger complexity of interaction. Without morphological consideration, we are unable to tell whether galaxies at close projected separation are actually at the closest point of flying by each other or about to coalesce.

We find that throughout the different stages of interaction the SFR within the systems is increasing. Observations of interacting galaxies at various stages have found increased gas inflows into the nuclear regions of the galaxies which lead to enhancement in SFR there over their outer reaches (Barrera-Ballesteros et al., 2015). This is often confirmed with deep observations of individual systems, which capture snapshots of different parts of the dynamical timescale of the interaction (Karera et al., 2022). Our study utilises numerous snapshots of the dynamical time of interaction in attempt to build a full picture of the change in SFR. We have found here is that this process of increasing SFR and galaxies being driven to starburst begins from the pericentre of interaction. Multiple simulation works, which have the ability to model the entire dynamical history, show that this is to be expected (Di Matteo et al., 2007; Hopkins et al., 2013; Karman et al., 2015; Moreno et al., 2021). Often, these works show an initial dramatic increase in the SFR of interacting systems followed by an exponential decline through the dynamical time before increasing dramatically again at either a second passage or coalescence. Moreno et al. (2015) is a direct example of this SF history through the merger history.

Our results differ here as we do not find an exponential decrease in the SFR as move from the pericentre systems to the apocentre systems. As stated previously, observational work does support a continued increase in SFE in galaxies to projected separations of out to 80kpc - well into our defined apocentre stage systems (for further examples, see Li et al., 2008a; Scudder et al., 2012). However, it is important to note that our apocentre stage defined classification is a 'wide net' that captures many systems that may be very soon after the initial passage in the dynamical timescale (recall Figure 3.7). The criteria defining pericentre and apocentre stages are simply that the galaxies must be no longer connected or overlapping morphologically with tidal features. They must only be distinct and separate galaxies. Therefore, our found large enhancement in the apocentre stage may be from interacting systems which are only just out of the pericentre passage and not enough time has passed for the rapid decline in SFR to begin.

Nonetheless, we still find a disappearance of the red sequence in the apocentre systems which may come as unexpected to when compared to simulations. It is important to note, however, that this is not the same as saying that a large proportion of the apocentre systems are classified as starbursting galaxies. While previously mentioned simulations approximate an initial large starburst before rapid decline, we find the interacting galaxy population is pushed being quiescent / sub the star forming main sequence to being on the star forming main sequence. Therefore, it may be more likely that the impact of interaction on star formation is not to suddenly cause rapid star formation in the aforementioned starburst before declining, but rather to gradually increase a galaxy's SFR up and into the blue cloud of galaxies. This could be supported by the surprising lack of rapidly quenched post-interaction galaxies found in both observations (Weigel et al., 2017) and simulations (Hani et al., 2020; Quai et al., 2021). Thus, the impact of interaction on star formation is may not be a catastrophic increase in star formation that leads to quenching but, rather, a small increase in galaxies to use whatever gas they have into star formation.

This idea can be brought further forward by considering the large increase in starbursting galaxies we find when going from pericentre / apocentre stage to the merging galaxies. The merging stage represents, in our sample, galaxies that are undergoing the final coalescence of the two systems involved. We find that the

point of final coalescence leads to a large increase in the fraction of starbursting galaxies as well as the almost complete disappearance of the quiescent and sub-main sequence fraction of galaxies. Thus, from this result we can conclude that during coalescence galaxies undergo a huge starburst and enhancement in star formation which will quickly lead to their quenching. This is also supported by works such as Ellison et al. (2022), which find that galaxies post-coalescence are 30-60 times more likely than control galaxies to have rapidly shut down star formation.

Thus, we can conclude that, for merging galaxies if gas is present within them, star formation will increase and change our classification of the galaxy. We will observe a sudden increase in the star formation rates of these systems followed by a rapid quenching as the gas is used up entirely. This differs from galaxies that move into the apocentre stage and do not merge. These apocentre stage systems will then slowly lose their enhancement over a long period of time, and return to star forming at their expected rate, whereas only moving into a merging stage galaxy will lead to a starburst which may incur rapid quenching.

Such a conclusion would also, therefore, explain the often quite large divide in the literature between whether interaction actually leads to enhancement or not. The only part of the interaction which actually causes the enhancement, and is the forking point, is that of the pericentre stage. From this point, if the galaxies then move off to the apocentre stage and escape, we will see a gradual decline in the SFE with the apocentre stage galaxy only having a minor increase in its star formation classification. Whereas, if the galaxies move into the merging stage of the interaction and coalesce, that is where we see the catastrophic results of a major starburst and the complete using and of all the gas in the systems. Thus, leading to a large fraction of quenched post-merger galaxies, but only after the initial coalescence.

3.6.2 Interaction Stage and AGN

The evolution of the AGN fraction in interacting galaxies is similar to that of the evolution of star formation. It has often been found with projected separation the AGN fraction increases. There are multiple observational works that show

this (Alonso et al., 2007; Ellison et al., 2013; Shah et al., 2020) as well as works on cosmological simulations which support these conclusions (Byrne-Mamahit et al., 2023). In simulations, the increased likelihood of nuclear ignition comes from the sudden increase in gas density in the galactic core which naturally leads to increased black hole accretion and ignition. To satisfy these conditions, actual coalescence of the two systems are required. Observations of interacting galaxies are often interesting, as they contain examples of dual AGN and individual examples (e.g. Ellison et al., 2017; Stemo et al., 2021) or investigate the increase in the AGN fraction in only the merger / post-merger stage (Gao et al., 2020).

We find peaks in the AGN fraction with projected separation, before the galaxy merging and coalescence takes place. This is supported with other works which specifically look at AGN fraction with projected separation (Ellison et al., 2011; Steffen et al., 2023). However, we specifically find that the AGN fraction increases rapidly in the merging stage and holds quite constant between stages 1 to 3. This shows, again, shows two distinct effects of interaction which can colour our interpretation of the link between AGN and interaction. It appears, from our results, that the AGN fraction is driven up during the points in the dynamical time when the inner parts of the galaxy are majorly disturbed, and not by the simple movement of gas and dust into the core during the two galaxies passing each other. There is evidence that the onset of nuclear ignition from interaction may be delayed (Ellison et al., 2011), or even flicker (Schawinski et al., 2015) after ignition. Both of these possibilities are reflected distributions of AGN fraction with projected separation.

Again, it is important to note the rather 'wide net' that our stage classification takes. There is overlap between our pericentre and apocentre stage classifications at high pericentre projected separation and low apocentre projected separations. However, these stages naturally lead onto one another. Therefore, what see in the bi-modal distribution of AGN fraction with projected separation in Figure 3.22 are two peaks. The second of these peaks is at the cross over point of pericentre to apocentre; the point at which (if the two galaxies have just flown by each other) a delayed AGN ignition could take place. This is also reflected in Figure 3.20 where we see a slight increase in AGN fraction in the apocentre stage. This

is likely from the ignition of nuclear activity taking place over a large period of time with a delay being involved from the initial flyby.

We also see a dramatic increase in AGN fraction in the merging stage systems, where coalescence is just beginning or has occurred. Multiple works show that this is as expected. Byrne-Mamahit et al. (2023) has used cosmological simulations to show that we expect a large increase in AGN fraction at coalescence, and even for sometime into the post-merger phase. This is supported by measured AGN fractions in post-merger galaxies. Here, we see that the enhancement actually begins in the pericentre stage, before the coalescence has even begun, but the actual coalescence is what sends it into overdrive. This is matches observations mentioned previously of increased gas densities in nuclear regions and cores, and that the merging stage is when this really occurs in earnest,

However, to more fully study this, we would need a larger sample of confirmed AGN and star forming galaxies from existing photometry or catalogues to make more decisive conclusions based on stage. In this work, we were limited to a small number of AGN and star forming galaxies across each stage. When compared to our paired sample, we do not have the numbers to also divide into stage again. Therefore, Figure 3.22 shows the global AGN number count with projected separation. We show that the two peaks are due to different stages of the interaction, however, these are using very low number counts and should be treated carefully.

3.7 Conclusions

In this work, we investigate the evolution of multiple parameters and processes in galaxy interaction with interaction stage. We use the O'Ryan et al. (2023) catalogue of interacting galaxies and cross match it with the COSMOS survey to gain ancillary data. This gives us a sample of 4,135 interacting galaxies of which 960 have a confirmed secondary from available photometric redshift data. We use visual morphology as well as angular separation to split our sample into four distinct stages: (1) close pair, (2) morphologically disturbed and overlapping, (3) morphologically disturbed and distinct, (4) merging. Each stage is designed to capture a different part of the dynamical timescale of the interaction. We then

cross match these samples with existing catalogues of environment and active galactic nuclei for further study.

We first split our sample of galaxies into their different stages and investigate their evolution with stellar mass and star formation rate. We conduct Kolmogorov-Smirnov and Anderson-Darling tests to show the mass distribution of our sample does not change with stage, while the star formation rate changes dramatically. This change in star formation rate from the separated to merging stages is found in the red sequence of galaxies reducing to the point of disappearance. This is further confirmed by sub-classifying each sampled stage into starbursting, main sequence and quiescent galaxies. We find that as the galaxies move from the separated to pericentre stages the fraction of starbursting galaxies increases while the quiescent galaxy fraction reduces. In the merging stage, the starbursting fraction increases dramatically while almost no quiescent galaxies exist in the sample. This implies that the mechanisms responsible for enhancement in star formation in interacting galaxies is dominant from separated to pericentre stages and in the final coalescence of the system. We find that, for all of our galaxies, some enhancement in star formation is observed.

To further investigate this change in enhancement, we investigate our subsample of galaxy pairs and compare it to a mass and redshift matched control sample. We bin our projected separations and measure the ratio between the average interacting SFR and control SFR in each bin. We find, across the whole subsample, a general increase in star formation efficiency as we move to smaller projected separation. The highest found enhancement is below 10 kpc. However, when broken down into their constituent stages, we find dramatically different behaviour in the star formation efficiency. This is best seen in the pericentre stage enhancement, which does not seem to change with projected separation, while there is a dramatic decrease in the pericentre stage enhancement from 2.5 at 10 kpc down to 1.2 at 100 kpc. This shows that just using projected separation as a proxy for stage will leave out crucial information to the underlying causes and mechanisms fueling star formation enhancement. To confirm that the effects observed here are not related to biases in the environment, we investigate its relation to our staged sample. We find that the environment is consistent between all stages, with no biases existing.

Finally, we investigate the change in AGN activity across our whole sample. We find that, on average, the AGN fraction remains constant with stage until the point of coalescence. However, when looking at project separation, we find an almost bi-modal distribution. This could be evidence for a delayed ignition in the AGN present in those systems undergoing an interaction. However, it is difficult to draw this conclusion definitively due to small number counts. We also find that the AGN fraction is highest in merging and merged galaxies.

While the results with projected separation are not unexpected, those with the different stages are. We have shown that one cannot simply use the projected separation of interacting galaxies as a proxy for the stage of the interaction. We find very different behaviour in the star forming behavior of interacting galaxies based on stage which are at in the dynamical timescale of the interaction. We find the beginning of an enhancement in star formation occurs in the pericentre stage. Then, through to the apocentre stage, the enhancement remains although with no further increase. There is then another dramatic starburst in the merging stage, where the two galaxies actually begin to undergo the coalescence processes.

Our work shows the importance of considering morphological stage when considering interaction, and that there is a fine interplay between underlying processes and dynamical timescale of an interaction. We require larger samples of correctly staged galaxies to further understand and exploit what these relations are, and how to best investigate when and where star formation and its enhancement occurs. This is also particularly true of the relation between active galactic nuclei and the interaction stage, where we are unable to have the sample size to definitively draw conclusions from our results. We also require numerical models to better identify the point in the dynamical timescale an interacting system is. This would lead to more reliable constraints on the relations we have explore in this Chapter. We present our work on building such an algorithm in the following Chapter. This algorithm is focused on constraining both the dynamical timescale of the interaction and the full set of underlying parameters needed to describe a galaxy interaction.

Chapter 4

Advanced PySPAM: An Infrastructure to Constrain Underlying Interacting Galaxy Parameters

4.1 INTRODUCTION

The precise interplay between the observed processes we have discussed in this thesis and the underlying physical parameters of the involved galaxies is difficult to quantify. Algorithms have been developed to match the underlying parameters of an interaction by directly modelling the expected final morphology. Examples include Identikit by Barnes & Hibbard (2009) or the Stellar Particle Animation Module (SPAM) by Wallin (1990). However, finding the best fit underlying parameters for more than a handful of systems is often seen as unfeasible. There are often greater than 10 different parameters which contribute to different observed morphologies. This leads the underlying parameter space to be upwards of 10-dimensional, complex and, potentially, highly degenerate. Exploring even a sub-space of this parameter space fully requires tens of thousands of models to

be run. This is also hindered by its degeneracy. For example, Smith et al. (2010) found that four of their best-fit models were able to accurately match morphology of the Arp 284 system. Therefore, it is often preferable to use large cosmological simulations (e.g. Schaye et al., 2015; Hopkins et al., 2018; Hani et al., 2020) to create large samples of synthetic interacting galaxies where analogues to the observed system can be found. This, however, limits our capability to explore observational parameter space beyond the limitations of such cosmological simulations. By their nature, cosmological box simulations have finite scope and size; meaning the rarest (and likely most fundamental) interacting systems will remain unexplored without significant further computational expense.

This limitation of directly comparing numerical simulations to observations was solved in a novel way by Holincheck et al. (2016) in the Galaxy Zoo: Mergers (GZM) project (for details on Galaxy Zoo, see Lintott et al. (2008)). GZM worked with citizen scientists to run simulations of interaction and visually compare the simulated morphology to observations. GZM studied a sample of 62 interacting galaxies from the Arp catalogue of interacting galaxies (Arp, 1966). Citizen scientists' were given a selection of simulation outputs around an observation and would select the one which appeared most alike. The simulations would then be run again with tweaked underlying parameters with the citizen scientist selecting the new the best fit output. With enough citizen scientists on the project, enough new simulations run and enough time, GZM were able to find the best fit parameters for their interacting galaxy sample of 62 across a 14-dimension parameter space; providing one of the largest, observational fully constrained interacting galaxy samples to date. In the era of the Vera C. Rubin Observatory such an approach will not be practical, with statistically significant samples of thousands of interacting galaxies potentially being produced every week, a new approach is required.

In this work, we present an automated methodology to constrain the underlying parameters of interaction. We combine a fast restricted numerical simulation code with a Markov-Chain Monte Carlo (MCMC) framework. We apply this over a 13D parameter space and constrain the probability distribution and find the likely parameter values. This approach allows us to marginalise over each parameter, provide the best fit value as well as the error on each measurement.

First, we apply this to a set of synthetic interacting galaxy observations created from a restricted N-body simulation where we have input the 12 underlying parameters. This provides us with diagnostics of our methodology. We then apply our process to three real observations of interacting galaxies and discuss how our methodology could be applied to real data, and the limitations therein. We also discuss the more general limitations of this approach, with a particular emphasis on the computational expense required, and the potential future solutions.

The layout of this paper is as follows. In section 4.2, we describe the sample of 51 galaxies we will study and our approach to observation preparation. Section **??** will summarise our simulation code and how it has been updated from previous iterations. We also describe how we build our sample of synthetic interacting galaxies and the parameter space we will explore with this simulation. A full discussion of our results will be given in section 4.5 followed by our conclusions and future work in section 4.6.

4.2 DATA

4.2.1 Sample of Major Interacting Galaxies

In this work, we test our Markov-Chain Monte Carlo (MCMC) process on idealised, mock observations of 52 major interacting systems. These mock observations are created using our numerical simulation where we have input the underlying parameter values directly. As we know what the true underlying parameters of each interaction are, we can test the accuracy of our algorithm. To create these mock observations, we use the sample of 62 interacting galaxies from the GZM project. This project, with the help of citizen scientists, approximated the best fit parameters for a set of 62 major interacting systems. We use GZM's found best fit parameters to create our set of mock observations.

Of the sixty two systems, we create synthetic images of 51 of them to act as our idealised observational sample. As we are including the flux distribution of each system, we require an accurate redshift of each system. 3 of the 62 systems do not have a redshift measurement in the NASA Extragalactic Database (NED), the Sloan Digital Sky Survey (SDSS) or Simbad databases. Therefore, we discard

these. We create each mock observation as though it was observed using SDSS. Of which, a further eight systems are not present in. Therefore, we also discard these systems. This leaves us with the GZM best-fit parameters of 51 systems with which to create our synthetic observations. The details of the real systems are displayed in Table 4.3, including the target coordinates, the SDSS ID and the redshift measurements we use. How we create our idealised, mock observations is described in Section 4.3.2 after we have described our simulation.

4.2.2 Observation Preparation

As a final test of our MCMC algorithm, we will constrain the underlying parameters of observations of a subset of the systems described in Table 4.3. Therefore, we describe how we build such observational images here. We downloaded the FITS files from SDSS Data Release 16 which contained the full system of interacting galaxies. These were then used to create smaller cutouts of the systems. The coordinates of the primary galaxy are used as the center of the image. We judged the size of the cutout by whichever size contained the full interacting system, most often between 600 - 1000 pixels. We created cutouts in each SDSS filter ($ugriz$).

We convert each observational image into counts from the native unit of nanomaggies using the conversion value in each FITS header. Each filter image was then stacked into white image by simply summing them together. We elected to do this so that we would gain higher signal-to-noise in the tidal features of the interacting systems. Then, we took these native resolution white images and block reduced them to a 100×100 thumbnail. Each cutout was visually inspected to ensure that, even with the reduced resolution, the tidal features were still clear and prominent in the image.

We then reduce the resolution of each observed image of the system to be 100×100 cutout. We take the 0.396" pixel scale of SDSS on the sky with the measured redshift to calculate its physical size. As we know the pixel-to-distance conversion for each image, the position of each secondary galaxy was then calculated from the central pixel. No attempt was made to approximate the z-position of the secondary galaxy, as this is a parameter to constrain. In

Name	SDSS ID	RA	Dec	Redshift
Arp 240	587722984435351614	204.980417	0.835278	0.02250
Arp 290	587724234257137777	30.946317	14.72365	0.01171
Arp 142	587726033843585146	144.429583	2.763056	0.02329
Arp 318	587727177926508595	32.380516	-10.158508	0.0132
Arp 256	587727178988388373	4.710417	-10.369167	0.02730
UGC 11751	587727222471131318	322.247796	11.382539	0.02909
Arp 104	587728676861051075	203.037083	62.733889	0.01082
Double Ring, Heart	587729227151704160	238.287292	54.147861	0.040
Arp 285	587731913110650988	141.040000	49.226111	0.00967
Arp 214	587732136993882121	173.145221	53.067922	0.00331
NGC 4320	587732772130652231	185.740516	10.548328	0.02668
UGC 7905	587733080814583863	190.952917	54.900278	0.01648
Arp 255	587734862680752822	148.290000	7.870000	0.04106
Arp 82	587735043609329845	122.811250	25.193056	0.01368
Arp 239	587735665840881790	205.423852	55.672324	0.02489
Arp 199	587736941981466667	214.265833	36.573333	0.01024
Arp 57	587738569246376675	199.198750	14.424444	0.048
(HWB2016)Pair 18	587738569249390718	206.209583	13.921361	0.089
Arp 247	587739153356095531	125.889478	21.342976	0.01108
Arp 241	587739407868690486	219.461958	30.481222	0.03472
Arp 313	587739505541578866	179.418333	32.285556	0.01045
Arp 107	587739646743412797	163.069583	30.065278	0.03318
Arp 294	587739647284805725	174.931624	31.920108	0.00892
Arp 172	587739707420967061	241.389583	17.597222	0.029
Arp 302	587739721376202860	224.251667	24.612222	0.03286
Arp 242	587739721900163101	191.544583	30.727222	0.02205
Arp 72	587739810496708646	236.733750	17.878333	0.01100
Arp 101	587739845393580192	241.124946	14.800192	0.026
Arp 58	587741391565422775	127.990209	19.211523	0.03722
Arp 105	587741532784361481	167.804167	28.724722	0.021
Arp 97	587741534400217110	181.439583	31.068889	0.02305
Arp 305	587741602030026825	179.655833	27.490833	0.004
Arp 106	587741722819493915	183.902522	28.173576	0.02199
NGC 2802/3	587741817851674654	139.172619	18.963463	0.02914
Arp 301	587741829658181698	167.470000	24.259722	0.02059
Arp 89	587742010583941189	130.665852	14.285624	0.00687
Arp 87	587742014353702970	175.185000	22.437778	0.02373
Arp 191	587742571610243080	166.834167	18.431111	0.02739
Arp 237	587745402001817662	141.933458	12.286750	0.02899
Arp 238	588011124116422756	198.886667	62.126944	0.03106
MCG +09-20-082	588013383816904792	181.161667	52.956111	0.078
Arp 297	588017604696408086	221.330417	38.761389	0.0298
NGC 5753/5	588017604696408195	221.328663	38.805889	0.01374
Arp 173	588017702948962343	222.869434	9.328297	0.028
Arp 84	588017978901528612	209.649167	37.438889	0.01158
UGC 10650	588018055130710322	255.060770	23.106346	0.00986
Arp 112	758874299603222717	0.368333	31.437778	0.024
Arp 274	587736523764234706	218.786250	5.356389	0.02800

total, there are 12 parameters that we must constrain over our mock and real observation images. These parameters are detailed in the following section. The observation image preparations were made using the Astropy Python package (Astropy Collaboration et al., 2013b, 2018b).

4.3 SIMULATING GALAXY INTERACTION

The aim of this work is to constrain the underlying parameters of interacting systems. To do this, we require a method of mapping underlying parameters to our observed systems morphology and calculating how likely it is they are representative. We do this by inputting our set of underlying interaction parameters into a restricted three-body simulation which predicts the morphology and flux distribution of the interacting system with such parameters. In this section, we describe our simulation code, its previous iterations and how we add in additional flux information.

4.3.1 APySPAM

4.3.1.1 Restricted Three-Body Simulation

The simulation code we opt to use is the Java Stellar Particle Animation Module (JSPAM) (Wallin et al., 2016). For an in-depth description of the underling code, we direct the reader to Wallin (1990); Wallin et al. (2016) but also describe it here. JSPAM is a restricted three-body code focused on recreating the morphology of interacting systems. It approximates the interaction as two massive bodies each being orbited by a set of massless test particles. The simulation calculates the gravitational potential of the two massive bodies and the resultant forces upon the massless test particles and uses this to predict the particles position and velocity at different timesteps. The size and number of time steps can be input by the user. The simulation is computationally efficient, approximating the morphology of an interacting system using thousands of particles in seconds on a regular work PC at a reasonable time resolution. The primary particle integrator is a fourth-order Runge-Kutta which applies backwards integration to calculate the position

and velocity of the massive bodies from t=0 to the time set by the user. The test particles are then added to the simulation, and forward integration is conducted to find their position and velocity in each timestep through the trajectory of the massive bodies. In this way, a flyby of the two galaxies is simulated with the test particles representing the extended morphology of the two galaxies.

The user has the option to choose a N-body approximation or a softened point mass approximation. Each of these slightly changes the way that the integrator calculates the forces on each particle. Therefore, there will be slight discrepancies between the final morphology of systems interacting with the same underlying parameters but different force approximations. We elect to utilise the softened point mass approximation, as this has improved computational efficiency over the higher accuracy of N-body approximation (for more on this see Wallin et al., 2016).

The base code of JSPAM was purely a morphology matching code. Attempting to extract underlying parameters of interaction from morphology matching alone has been shown in prior works to be very difficult (e.g. Barnes & Hibbard, 2009). JSPAM itself has been used in genetic algorithms to find the best fit parameters of different systems (e.g West et al., 2023). However, this lacks the exploration of parameter space and the quantification of uncertainties we aim to achieve in our approach. Using both morphology and flux matching between simulations and observations improves accuracy of recovered parameters (Miller & van Dokkum, 2021) and we have therefore enhanced the original JSPAM algorithm with the ability to model population evolution with star formation/star bursts to approximate the flux distribution of the interacting system. We have created this enhanced version in Python 3.7.4 (hence, we shall refer to this algorithm as Advanced Python Stellar Particle Animation Module, APySPAM).

Thus, to approximate the flux distribution of the galaxies, we calculate luminosity of each particle while minimising the computational cost. We incorporate the evolution of the underlying stellar populations of each galaxy over the time taken in the interaction. Also incorporating formation of new stellar populations in potential starbursts due to the interaction, and approximating the impact this will have on each galaxys' flux distribution. To preserve the computational efficiency, we build a semi-analytic model and define a global stellar population to

assign a luminosity to each particle based on its assigned mass. In this way, we 'paint on the stars' of each galaxy without having to directly model stellar evolution and star formation. This process is detailed in the following two subsections.

4.3.1.2 Stellar Population Evolution

To model the underlying stellar populations, we utilise a Bruzual & Charlot (2003) (BC03) simple stellar population. These contain SEDs generated from flux libraries from a Chabrier (2003) initial mass function. We set this stellar population model to have a delayed exponentially declining star formation rate (Johansson et al., 2009; Pacifici et al., 2013; Simha et al., 2014; Boquien et al., 2019). However, to capture any star formation due star formation enhancement in the interaction, we increase the star formation based on the conditions in the interaction. We assume an e-folding time for star formation of $\tau = 1.7$Gyrs. This value is chosen as it matches field, massive galaxies found in the literature (e.g Peng et al., 2010b) and used in hydrodynamical simulations (e.g. Jeon et al., 2022). Our simulation outputs a spectrum normalised to $1\mathrm{M}_{\odot}$ with each particle assumed to be the same age as the galaxy.

This spectrum must then be scaled to the stellar mass present at each particle. We follow the prescription as stated in Wallin et al. (2016) to distribute the mass between the three components of the galaxy, and then to each test particle. The test galaxy in the JSPAM simulation has its mass distribution as $M_{\mathrm{bulge}} = 0.05M_{\mathrm{galaxy}}$, $M_{\mathrm{disk}} = 0.14M_{\mathrm{galaxy}}$ and $M_{\mathrm{halo}} = 0.81M_{\mathrm{galaxy}}$. We take the total mass assigned to the galaxy (input by the user) and divide it into these three components. We then assume that the bulge and disk masses are fully baryonic, with the remaining mass being non-baryonic dark metter.

We, then, further divide the calculated baryonic mass into two components: stellar and gas. The total stellar mass of each particle will be used to scale our final output SEDs from our model while total gas mass of each particle is used to calculate the star formation. The gas and stellar mass to be distributed to the particles is then defined by a gas fraction parameter that the user can alter. By default, this value is 0.15 for both the primary and secondary galaxies.

We assume the initial ages of the galaxies (both 10Gyrs by default) at initialisation and calculate the final age of the SEDs based on the number of time units the user wishes to run the simulation. This age is then used to extract the normalised spectra from the BC03 templates. These output SEDs are then convolved with given telescope filters of the users choice and integrated, giving a colour flux value to each particle. However, this process only gives the final flux values at each particle of the initial stellar population. During the interaction, we assume that the galaxy begins to form new stars at a rate significantly higher than is modelled in the BC03 templates. Therefore, we account for this by modelling newly created stellar populations as the interaction progresses.

To incorporate the increase in star formation in our simulations, we manually enhance the expected star formation rate through the simulated interaction. In high-resolution simulations which also model gas, starbursts occur naturally (Saitoh et al., 2009). However, in our simulations we must approximate this behaviour in a semi-analytic fashion. The change in star formation is heavily dependent on the mass ratio and the kinematics of the interaction and, therefore, we implement an enhancement parameter based on these parameters. We calculate the excess star formation due to the interaction compared to that already expected in the SED, and distribute this to each particle based on the initial gas mass. Approaching the problem in this semi-analytic way is similar to what is done in the CIGALE (Boquien et al., 2019) algorithm. Here, we detail this enhancement parameter.

At any given timestep the total stellar mass is given by

$$SFR_{\text{enhancement}} = \beta(\frac{t}{\tau^2})\exp(-\frac{t}{\tau})M_{\text{baryonic}}M_{\odot}\text{yr}^{-1}. \tag{4.1}$$

Here, τ is the e-folding time of star formation, $M_{baryonic}$ is the baryonic mass of the galaxy and t is the age of the galaxy at the given timestep. This is the star formation at any time given by the BC03 template. However, we add the β parameter: our enhancement value. This is a dimensionless value given by

$$\beta = M_{\text{ratio}}D^2_{\text{ratio}}. \tag{4.2}$$

M_{ratio} is the mass ratio between the galaxy being enhanced and the galaxy causing the interaction. D_{ratio} is the ratio of each galactic radius to the distance each galaxy is apart. In a system with a high mass galaxy interacting with a low mass galaxy, the high mass galaxy will have relatively little star formation enhancement while the less massive galaxy will have significant enhancement. This is similarly true for the ratio of the radius and separation. If the galaxies are interacting in such a way that the distance of closest approach is less than each galactic radius, then this ratio will rapidly increase above one; enhancing star formation further. This represents a significantly more violent interaction. It is important to note, however, that this has significantly less impact on strengthening star formation than that of the mass ratios.

These parameters successfully reflect the findings of the current astrophysical literature, where mass ratio has a significantly higher role on star formation enhancement than impact parameter (Barton Gillespie et al., 2003; Lotz et al., 2008a; Li et al., 2008a). We base our semi-analytic approach on the star formation histories found in a range of high resolution N-body simulations (Mihos & Hernquist, 1996; Springel, 2000; Rodríguez Montero et al., 2019) which measure the change in star formation directly from the Kennicutt-Schmidt (Kennicutt, 1998) relation. These simulations directly model the star forming gas through the interaction, measuring the change its evolution which we are able to approximate. Thus, we achieve an accuracy comparable to directly-modelled simulations at a fraction of the computational cost.

The output of Equation 4.1 is global star formation of each interacting galaxies at any given timestep. However, the aim of our models is to be able to match the flux distribution across the entire galaxy (especially the tidal features) to any observation that the code is given. Therefore, we must distribute the star formation throughout the particles. To keep computational efficiency, this is done by utilising weights which have been assigned to each particle. These weights are based on the ratio of the gas mass of the particle which has been assigned at initialisation to the total gas mass of the galaxy. So, to find the star formation

rate of a single particle at any given timestep, the following equation is applied,

$$SFR_{\text{Particle}} = \frac{M_{\text{gas,Particle}}}{M_{\text{gas,Galaxy}}} SFR_{\text{Galaxy}}. \tag{4.3}$$

After every time step, the gas within each particle is reduced by the mass converted into stars. Once this drops below a user defined value, the particle is cut off from star formation and is considered quenched. Currently, each particle is assigned equivalent gas mass at initialisation of the simulation. Therefore, when a particle is quenched in this example, every particle in the galaxy will also be quenched. The user can define a gas distribution model, which will lead to different particles being quenched at different times.

There are limitations to this approximation. We assume that each galaxy is a disk galaxy prior to the interaction when we assign gas masses to each particle. We assume that all of the gas mass assigned can be used in star formation, i.e. all the gas is cold molecular gas. We make no account of gas ionisation or the turbulence in the ISM that likely occurs during these interactions. We also assume that the disruption occurring to the test particles represents what would occur to the gas disk of galaxies within an interaction. However, we find that with these assumptions, the output star formation histories mimic those simulations which directly calculate these values.

Upon finding the star formation rate at the position of each particle, we convert this into a stellar mass formed through the timestep taken. We then compare this formed mass to the expected mass formed in the initial underlying stellar population. If the new mass formed is so low that it would be captured by the underlying population, we do not add this mass. If excess stellar mass has been formed, we assign an SED to it and its age is recorded. Once the simulation is completed, each new stellar population age is used to extract the relevant BC03 SED and multiplied by the total mass of the new stellar population. We then stack all of the SEDs together. This gives us the total extra emission we expect from the stars formed during the starburst throughout the simulation. This is then added to the initial stellar population emission defined at the beginning of the simulation. This gives us the total SED of each particle throughout the simulation. We convert this to total output flux of each particle.

4.3.1.3 Extending Flux Distribution

One remaining challenge in the simulations is that we are attempting to model full interacting systems with a number of particles significantly less than the number of pixels in each image. This is to maximise computational efficiency. The resulting effect is that large gaps can appear in the tidal features that form or within the disks themselves as the interaction progresses. To mitigate this effect, we calculate the flux at each particle position and then use the procedure described below to distribute it through each pixel of our image. This results in more realistic images of galaxies compared to just binning the particle flux based on position.

First, we calculate the flux at each particle described in the previous sections. We take each particle SED, and convolve it with the filter(s) of the user's choice. These convolved SEDs are then integrated to give a value of the flux in counts at each particle. We then create a grid of pixels and calculate the physical distances between each of their centres. We then calculate the contribution of flux from each particle to each pixel centre. Once the 2D flux distribution has been found across the pixel grid, we then use the measured redshift of the galaxies to find the measured fluxes as if the system had been observed.

The result of this is a well distributed galaxy image where there are no empty spaces in the tidal features nor in the disk. However, it does have a limitation when particles are not within the galaxy. This can be because they have been flung out to different parts of the image during the interaction. This has the effect of any particles in isolation being smeared into seemingly larger orbiting systems to the interaction. When doing our pixel matching, this can lead to much lower, unrepresentative χ^2 values being calculated despite excellent reproductions of the primary tidal features very well. As a result, we set the value of any particle with no neighbour within 5×5 pixels to zero.

The process described above successfully creates continuous, synthetic images of a galaxys' flux distribution with a limited number of particles in the simulation. However, this approach does impose two limitations on our constraining methodology. These limitations are related to the existence of tidal features at the very limits of detectability of our telescopes. The first is if a galaxy has a

'hidden' tidal feature that hasn't been detected. In this example, the correct underlying parameters would lead to the formation of the tidal feature that would be measured in the synthetic image. When compared to the observational image, this would be identified as a mis-match between the morphologies of the observed and simulated images. It is important to be aware of all the tidal features within the image and understand the limiting flux of all images in a sample. Here, we have selected major mergers in SDSS with very clear tidal features which are far from the low surface brightness regime. Thus, we do not encounter this issue here.

The second is with tidal features at the very limit of our detectability, where a single isolated particle could be representative of the tidal feature but we now remove it with our 5×5 neighbour criteria. Primarily, in testing, we find that this is a lesser problem that the first. The 5×5 criteria is a very lenient one, and often such tidal features are close enough to the galactic disk in our images that they remain in the image. They often form continuous features which then incur the first limitation. This can be rectified by selecting lower resolution and getting more flux into larger pixels.

4.3.1.4 Impact on Computation Time

The new algorithms to calculate flux have been added to the original JSPAM code while preserving computational efficiency. The choice to create an interaction constraining code which uses flux distribution rather than morphology matching is due to the prior difficulties of using such a method. Therefore, the introduction of extra algorithms which require extra computation time has been necessary. The runtimes of JSPAM and APySPAM are shown in table 4.2. As shown here, even with our extra flux calculations, for reasonable particle counts our new code APySPAM outperforms JSPAM by at least a factor of four. When we have translated JSPAM into Python, we have also re-written the underlying code to take full advantage of Numpy and Python's speed with vectorisation over for loops. As shown, the computational efficiency impact only becomes noticeable at very low particle number, where the overheads of Python's vectorisation is comparable to the base runtime of a for loop.

N Particles	JSPAM (s)	APySPAM (s)
10	0.062	0.250
100	0.45	0.338
1000	4.22	1.090
2500	10.535	2.392
5000	21.104	4.458
10000	42.796	8.625

Table 4.2: Timing comparison between the original JSPAM code (as used in Galaxy Zoo: Mergers) and the advanced version of PySPAM we are using here. These timings were taken using Python 3.7.4 on an Intel(R) Core i7-8665U CPU. Our version of APySPAM significantly outperforms that of the original JSPAM by many times, even with the added architecture of approximating the flux distribution. This is because in our re-write of the underlying simulation code we take advantage of Python's efficiency with vectorisation and array multiplication over that of for loops. These tests were performed by running the simulation for seven hundred steps fifty times and then taking the average run time of each iteration.

To explore the full parameter space we must run APySPAM many thousands of times. Therefore, we need to use the simulation specified with the fastest runtime possible for the smallest tradeoff in resolution of the tidal features. We elect to use 2,500 particles throughout our run. This is still relatively fast, taking approximately 2 seconds, but also maintains high resolution of the tidal features. This is still five times faster than using the original JSPAM code with this many particles.

By using the flux distribution method described in Section 4.3.1.3, we mitigate the effect of lower particle compared to pixel number. This is at the cost, however, the adding the largest computational overhead. We find that this part of the algorithm is much more sensitive to image size than particle number. The timing calculations shown in Table 4.2 use an output image size of 100×100. When increased to 500×500, the computation time for 2,500 particles increased to over 30s. Therefore, it is imperative that the user keeps this in mind when selecting cutout size.

4.3.2 Creating Test Images

With the descriptions of our simulation completed, we now describe how we build our idealised observations. We use the APySPAM three-body algorithm described in Section 4.3.1.1 to create these images. These images are measured in counts, as if observed in the SDSS. We utilise the best fit parameters which were found in the GZM project, and re-create their best fit images for each named interacting system. The parameters to create these images are shown in Table 4.3. We run our APySPAM with 20,000 particles and a high time resolution of 0.057Myrs per timestep. The resultant images are shown in Figure 4.1. Each image is a white, created by the stacking of *ugriz* filters of SDSS. We find the original SDSS observations of each system, and extract the conversion from nanomaggies (native SDSS flux unit) to counts and apply them to the native standard units our simulation outputs the flux distribution in. By stacking, we increase the signal in the tidal features, as well as other points in the disk.

Each test image is centered on the 'primary' galaxy, with the xyz-position of the secondary galaxy being used to calculate the size of the cutout. The primary galaxy is the galaxy with most mass in the pair, as defined in H16. These images were then reduced from their native resolution to 100×100 cutouts. This image size was found to be the best compromise between detail in the tidal features. As we will be matching the flux distribution of each system, it requires the redshift of the interacting system. As stated previously, the redshifts for our sample were found from the NASA Extragalactic Database. The full range redshift range of our sample is $0.003 < z < 0.113$.

4.4 CONSTRAINING INTERACTION

We have now described how we will create our simulation images to compare to our test images and, later, some observational data. This comparison acts as a way to map the set of underlying parameters to an output morphology that we can then constrain by comparing to observations. However, we need a framework to conduct this comparison and to explore the parameter space of our interactions. To do this, we use an Markov-Chain Monte Carlo (MCMC) algorithm. In this

Names	x(kpc)	y(kpc)	z(kpc)	v_x(kms^{-1})	v_y(kms^{-1})	v_z(kms^{-1})	log $M_{T,1}$($M_\odot$)	log $M_{T,2}$($M_\odot$)	R_1(kpc)	R_2(kpc)	ϕ_1	ϕ_2	Θ_1	Θ_2
Arp 240	-149.08	-68.71	231.50	-260.83	-504.46	794.60	12.54	12.47	44.69	63.75	301.00	35.50	310.81	321.99
Arp 290	16.22	-25.02	-31.02	-31.22	-67.22	-4.94	10.71	10.56	14.11	6.58	12.97	135.12	97.54	305.01
Arp 142	-7.40	-21.57	0.90	12.44	-173.23	152.27	11.04	11.19	19.16	5.01	48.79	43.80	50.07	135.06
Arp 318	14.10	2.70	39.98	0.00	-71.83	196.50	10.56	11.01	9.87	9.21	6.21	95.00	200.00	289.23
Arp 256	-6.12	-28.88	29.37	-194.60	-173.47	132.02	11.12	11.10	14.11	8.16	97.79	60.52	144.60	216.50
UGC 11751	10.00	-31.91	24.23	-31.46	-39.20	34.96	10.28	10.19	12.22	11.34	15.00	87.50	203.48	22.36
HEART	-1.89	5.25	50.36	-21.98	94.78	680.87	10.89	10.99	4.58	6.18	152.78	195.63	147.23	138.43
NGC 4320	-1.58	2.14	-3.10	-339.22	-30.34	101.96	10.90	11.05	7.07	2.57	117.88	147.04	326.46	65.76
Arp 255	-18.50	21.49	17.61	-40.64	-89.59	126.67	11.36	10.64	15.20	11.73	140.38	-4.88	317.52	221.37
Arp 82	-7.30	-21.75	-9.39	68.90	-90.15	7.84	11.17	10.87	5.90	2.89	26.65	49.48	342.60	232.84
Arp 199	-4.10	2.85	11.17	218.16	176.37	84.61	10.25	10.05	4.55	4.63	96.61	49.83	152.65	109.12
Arp 57	-69.44	-52.30	128.05	-173.38	-160.84	440.18	11.15	10.99	17.70	9.77	176.94	71.08	337.85	335.44
Arp 247	7.14	-10.80	-41.04	26.09	-53.22	-43.00	10.87	10.55	7.17	3.33	55.45	45.65	319.41	335.29
Arp 241	-4.54	-8.23	-0.55	-74.39	79.62	0.00	10.20	10.37	3.72	5.17	96.08	107.20	227.48	229.44
Arp 107	-29.43	33.42	-9.48	0.20	19.43	46.26	10.99	10.86	16.49	6.55	82.03	271.03	191.08	296.27
Arp 294	-4.61	13.67	-62.09	-122.55	-52.72	-25.10	10.63	10.82	8.69	8.20	70.38	178.66	59.54	109.24
Arp 172	-6.62	-18.38	-30.98	-42.56	-33.24	-52.22	11.26	11.12	8.81	8.09	213.74	67.31	203.50	241.01
Arp 302	2.76	-25.61	74.51	-196.00	-305.20	97.39	10.75	10.32	14.16	9.52	6.24	36.33	283.48	345.34
Arp 242	-6.47	-13.90	-26.85	4.19	60.44	-3.14	10.52	10.74	10.69	5.86	7.45	57.97	85.88	212.75
Arp 72	16.78	-8.81	26.86	8.24	-23.74	171.19	10.12	9.98	9.18	2.59	6.80	90.49	224.26	228.89
Arp 58	-28.91	-25.10	-15.37	20.13	-158.10	11.60	11.32	10.55	19.35	2.85	173.29	45.58	44.63	230.34
Arp 105	-1.53	-31.87	80.03	37.05	-3.92	265.62	10.77	11.45	9.23	11.88	-34.38	141.75	306.34	0.00
Arp 106	-6.39	-12.55	-28.98	-73.92	1.21	-185.26	11.12	10.30	8.98	2.94	108.93	39.06	319.92	160.84
NGC 2802	15.95	15.68	24.54	169.82	53.15	263.10	10.75	10.36	13.78	7.52	108.04	30.31	327.91	146.78
Arp 301	-17.28	-4.91	-19.08	-314.96	25.19	-3.15	11.05	11.17	10.24	16.68	9.19	78.00	160.99	252.67
Arp 89	-14.81	-9.89	-36.98	21.05	-81.46	-20.52	10.71	10.49	12.05	4.66	149.19	110.69	78.26	70.43
Arp 87	-3.46	26.92	-38.84	-42.32	45.27	-106.64	10.16	10.23	8.08	5.42	116.69	65.78	32.05	249.08
Arp 191	9.03	-5.50	-47.81	-2.88	34.73	-272.22	10.87	11.09	7.37	8.22	101.88	53.06	2.24	216.89
Arp 237	-8.24	1.92	1.88	85.13	-141.79	48.60	11.04	10.65	7.85	5.18	86.86	55.83	348.72	148.42
Arp 173	11.63	-26.45	-54.92	75.32	5.69	-139.47	12.04	10.56	6.95	6.87	83.54	329.68	310.01	317.61
Arp 84	7.61	23.25	2.26	-24.13	44.84	26.11	10.67	10.06	17.24	4.08	11.06	65.56	111.80	290.68
UGC 10650	-1.64	-16.00	-26.56	43.91	30.74	5.69	10.59	10.61	8.50	6.96	73.19	-63.69	346.58	187.83
Arp 112	-9.47	9.57	4.03	-170.75	-32.95	71.45	10.96	10.98	4.37	4.09	97.25	-4.75	194.53	35.78
Arp 274	19.50	-4.72	55.56	226.62	154.11	-150.67	10.75	10.24	17.75	12.69	126.88	51.38	296.59	232.55
Arp 146	18.05	10.10	9.61	72.89	85.41	73.70	10.70	10.78	8.83	12.91	72.50	48.81	111.80	67.08
Arp 143	-3.91	-16.09	24.25	53.06	-21.88	84.27	10.62	10.15	7.35	7.75	45.23	32.38	236.54	61.02
Arp 70	9.24	-30.53	-29.87	42.03	-128.45	-0.41	10.78	10.67	13.72	7.80	149.90	136.19	50.40	38.01
Arp 218	35.55	-20.53	3.84	74.29	-59.10	-25.77	10.74	10.12	11.07	6.38	148.30	49.41	231.50	261.45
Violin Clef	-15.21	-33.86	-16.08	-15.95	-70.31	-186.56	10.59	10.78	10.47	7.86	142.49	65.29	258.15	238.65
Arp 104	0.89	-9.61	-9.34	32.76	-26.66	-53.61	10.94	10.68	1.44	0.88	78.09	87.64	36.70	347.77
Arp 285	-16.66	21.33	25.37	-84.45	86.28	14.76	9.20	10.50	5.01	5.71	51.19	-35.00	295.16	249.32
Arp 214	-43.09	13.54	0.93	-57.28	31.96	-21.48	10.07	10.01	8.01	6.31	135.70	354.86	140.82	138.14
UGC 7905	1.32	-9.53	16.47	-112.54	-80.94	67.17	10.31	11.07	4.19	5.03	14.93	68.72	103.69	239.11
Arp 239	-12.19	4.02	21.45	-104.37	16.96	58.77	10.97	10.27	7.45	2.21	58.82	94.31	235.38	335.40
PAIR18	-12.44	32.72	-18.36	14.70	-38.05	-156.67	10.96	11.36	11.02	16.62	153.57	43.84	168.63	326.13
Arp 313	-13.42	11.35	-39.17	-142.02	-61.52	-303.92	10.71	10.44	4.64	11.10	7.92	42.80	115.72	304.23
Arp 101	14.52	-39.46	-33.44	27.22	-132.95	-143.42	10.71	10.78	6.60	4.45	45.10	358.74	62.34	207.08
Arp 97	-0.57	27.13	4.51	-40.93	31.86	45.10	9.98	10.09	8.14	3.62	53.18	146.65	46.34	24.76
Arp 305	44.52	54.29	-1.86	50.93	115.06	12.07	10.53	10.27	10.51	7.03	45.81	166.88	286.21	234.78
Arp 181	-3.11	-1.81	-7.72	-199.27	-63.32	-201.94	10.69	10.94	1.56	1.73	96.00	65.69	127.45	221.37
MCG 09-02-082	8.67	-5.60	-16.64	197.24	16.52	-228.72	11.10	10.77	2.97	3.49	85.43	97.51	317.60	4.75
Arp 297	11.55	-2.79	-35.87	38.79	-62.45	20.98	10.84	10.48	10.76	5.24	51.44	136.06	199.01	243.73
NGC 5753	-24.85	-34.94	-21.60	-59.57	-86.24	-5.50	10.64	10.40	4.62	5.77	141.06	110.25	138.63	26.83

Table 4.3: The best fit parameters found the GZM project to represent each of the named interacting systems. We use these best fit parameters to create idealised images of each interacting system using APySPAM. Each of these parameters are the final parameters of the galaxy at the point of the observation. The positional and velocity vectors are of the secondary galaxy in the frame of the primary. The masses are the total mass of the system. The radii is that of the disk initialised to create the final morphology. Finally, the four orientation parameters are with respect to the y- and z- axis and allows the disk to be rotated in 3D with respect to the sky. The resultant simulation cutouts are shown in Figure 4.1.

Figure 4.1: Our mock observations of each interacting system. These act as an idealised observation, which is completely noiseless. We also know the underlying parameters which formed these outputs. Therefore, this forms an excellent test-set to see if we can recover the underlying parameters when searching over a large parameter space.

section, we first describe the parameter space we are going to explore, followed by our MCMC algorithm, how we quantify the similarity between simulation and observation images and how we find the parameters which are best fit to our input images.

4.4.1 The Parameter Space of Interaction

In this work, we aim to explore the underlying parameter space of interaction and find the best fit parameters which describe an input observation or test image. The parameters we will be constraining are simply the required parameters in the APySPAM three-body simulation algorithm. To function, APySPAM requires 15 different parameters. These are: the six positional and velocity vectors of the secondary galaxy. APySPAM runs in the frame of reference of the primary galaxy at its centre. This reduces the size of the parameter space to explore. It also requires the total mass, the radius and two orientation parameters tp correctly align the inclination of the disk to the sky for each galaxy. It also requires the redshift measurement of the system to correctly estimate the flux distribution of the system. However, some of these parameters we can find before we begin exploring the underlying parameter space.

The first two parameters we can find manually first is the projected 2D position of the secondary from the primary. These can be measured from the observational image of the system by converting between pixel and physical coordinates in the image. We take this position as the centre of the secondary disk. The third parameter that is required is the redshift of the system. The redshift not only directly impacts the flux distribution of the input image, but its scale and resolution. Therefore, we assume that this parameter is known and provided. Work is currently being conducted to making this a free parameter that can also be constrained based on the interacting systems flux distribution and size.

Thus, we require 12 different parameters to recreate an interacting systems morphology and, therefore, are the 12 parameters we will constrain over. A description of these parameters and the size of the parameter space explored are shown Table 4.4. We choose the limits of the parameter space based on the maximum values of the parameters we used to build our test images in Table

Parameter	Description	Conversion(Spec. Units)	Parameter Range(Spec. Unit
z-position	Secondary z-position	15kpc	-300kpc - 300kpc
v_x,v_y,v_z	Secondary velocities	169.34kms[1]	-1693kms^{-1} - 1693kms^{-1}
M_1,M_2	Total Masses of Galaxy	$10^{11}M_\odot$	$1 \times 10^9 M_\odot$ - $4 \times 10^{12} M_\odot$
R_1,R_2	Radii of Galaxies	15kpc	0.15kpc - 150kpc
ϕ_1, ϕ_2	Y-axis orientation	deg	0° - 360°
θ_1, θ_2	Z-axis orientation	deg	0° - 360°
t	Time of Min. Separation	57.7Myr	0Myr - 500Myr

Table 4.4: The thirteen parameters used in both JSPAM and APySPAM to recreate an interaction. Each of these parameters must be found to consider an interaction constrained. The Parameter column shows how each parameter will be described throughout the rest of this paper. The third column then gives the conversion required to go from simulation units to SI units.

4.3. Table 4.4 provides the size of the parameter space as well as the conversion between the simulation units and physical units. This is primarily for reference for those who would wish to use our simulation to their own datasets. The conversions are found from Wallin (1990).

4.4.2 Defining the Likelihood Function

4.4.2.1 MCMC & Bayes Theorem

We now need a framework by which to explore our parameter space, and identify areas where the parameter set successfully represent the input image. We combine APySPAM with a MCMC methodology in order to fully explore the underlying parameter space. In an MCMC, a set of walkers are created and are then moved through parameter space in an ensemble. Each walker position is a set of parameters in our 13D parameter space. At each point, we calculate the likelihood that the output simulation is representative of the input interacting system. We then compares the likelihood between the old and new position, and if the likelihood is higher at the new position the walker moves there. If not, the walker remains in place and makes another attempt to find a higher likelihood. Thus, the walkers form their own chain of steps which gradually move towards the areas of highest likelihood. In our case, the likelihood is a measurement of

the similarity in flux distribution between a simulated image with parameters of the walker position and an observed image of unknown underlying parameters. Therefore, the walkers are moving from a set of underlying parameters that poorly describe the observed system to a set of underlying parameters which describe the observed system well.

We use the well known the Python package EMCEE (Foreman-Mackey et al., 2013) for our MCMC. This is an ensemble MCMC package with numerous predefined moves and algorithms to make getting to the area of high likelihood more efficient. We construct contours of walker mass and use these to calculate the errors and probability distribution of our best fit measurement; i.e. we can construct a posterior for each of our parameters. For full details of EMCEE and the different modes that it can use, see the extensive readthedocs[1]; but here we will briefly state the hyper parameters that we use.

For each observed image, an ensemble of six hundred walkers was initialised which would explore a total chain length of 7500 steps. Following the advice in the documentation regarding potentially complex and multi-model parameter spaces we utilised two different walker move proposals in our algorithm. These were the Differential Evolution (DE) Move (Nelson et al., 2014) and the Snooker Differential Evolution (DES) Move (ter Braak & Vrugt, 2008). An identical version of our setup can be found on GitHub[2]. Here, a user can download our setup to reproduce our results, or to update the model for their own purposes.

We define a likelihood function to compare the input images to our simulation outputs. By Bayes Theorem, the probability that a set of underlying parameters which produced a simulation also describe the observed image follows Equation 4.4,

$$P(H_i|D_{obs}, C) = P(H_i|C)\frac{P(D_{obs}|H_i, C)}{P(D_{obs}|C)}). \tag{4.4}$$

P(H_i—D_{obs},C) is the probability that some hypothesised set of underlying parameters, H_i, successfully describes some observational data, D_{obs}, under some prior constraints, C. Applying this to our hypothesis, H_i, allows us to utilise the prior knowledge that we have about the interacting system in question and can be used

[1] `https://emcee.readthedocs.io/en/stable/`
[2] `https://github.com/AstroORyan`

to put constraints on the parameter spaces we explore to shorten computation time. This is described by the expression P(H_i—C). This is multiplied by the likelihood that the observation is defined by the hypothesised parameters given the constrains,P(D_{obs}—H_i,C), all divided by a normalisation constant, P(D_{obs}—C).

4.4.2.2 Simplifying the Prior

In order to simplify this expression, we make assumptions about the underlying parameter space to increase efficiency and simply our computations. We first assume uniform priors for each of our 13 parameters. Therefore, we define a range of parameter values that if a walker moves outwith, we set the probability immediately to zero. The ranges we allow for each parameter are specified in Table 4.4. These ranges can be tweaked, or a different prior function defined, by the user. Here, we elect to set the priors part of Equation 4.4 to one.

We improve efficiency in our code further by adding to the prior based on the likelihood that tidal features will form in any given interaction. This is defined by a filter parameter, γ, and is fully described in Holincheck et al. (2016, where it is called β but we call it γ here to not be confused with our star formation enhancement parameter of Equation 4.2):

$$\gamma_{min} = \frac{M_1 + M_2}{r_{min}^2 V_{r_{min}}}. \tag{4.5}$$

Here, r_{min} is the closest approach distance, $V_{r_{min}}$ is the relative velocity at the time of closest approach and M_1 and M_2 are the primary and secondary masses, respectively. This parameter is designed to capture two important quantities: the mutual gravitational attraction and the inverse of the closest approach velocity. Each of which is important for the resultant gravitational distortion of the interacting system. By maximising the total mass of the system, while minimising the distance of closest approach and maximising the time of closest approach (i.e. minimising $V_{r_{min}}$) we would expect stronger tidal distortion.

In our case, we use it to inform our prior as the MCMC continues. This significantly enhances the computational efficiency and pushes the walker ensemble

to areas of high likelihood quickly. In each step of the MCMC, running the simulation itself is the highest computational cost, so we calculate γ first and then make a decision on whether to run the simulation. This decision is based on an exponentially declining probability dependent on the value of γ. This probability, or prior, is defined as

$$C = \begin{cases} \exp(0.5\frac{\gamma}{\gamma_{min}}), & \text{if } \gamma < 0.5 \\ 0, & \text{if } \gamma \geq 0.5 \end{cases} \tag{4.6}$$

Here, γ is a user defined cutoff, 0.5 in our case. Taking the log of this, we can directly add it to prior. If the prior is initially calculated above 100, we do not run the simulation and move the walker to a new set of parameters.

4.4.2.3 Simplifying the Likelihood Function

To further simplify the likelihood function, we can assume our probability distribution is Gaussian. This is a reasonable assumption to make as a starting point for our constraining attempts. However, as will be described in section 4.5 this is found to not always hold true; particularly for the orientations of the system. However, making this assumption allows us to utilise the following Equation to compare our mock observations to our observed data;

$$P(D_{obs}|H_i, C) = (2\pi\sigma_j^2)\exp(-\frac{1}{2\sigma^2}\sum_{j=1}^{n}(x_j - \mu)^2) \times C. \tag{4.7}$$

Here σ_j is the uncertainty in the observed image, x_j is our mock observation and μ is the observed image. The above expression can be simplified further by noting that the expression in the exponential function is just a half of the χ^2 difference between the observational image and the mock observational image. This is the same χ^2 function that is used in the code GALFIT (Peng et al., 2002). Where χ^2 is given by;

$$\chi^2 = \frac{1}{N - n_{dof}}\sum_{0}^{N_x}\sum_{0}^{N_y}\frac{(p_{x,y} - q_{x,y})^2}{\sigma_{x,y}^2}. \tag{4.8}$$

Here, N is the number of pixels in the observed image, which has the number of degrees of freedom subtracted from it, n_{dof}. $p_{x,y}$ and $q_{x,y}$ are the flux values of the (xth, yth) pixel in the observed and simulated images respectively. $\sigma_{x,y}$ is the σ value of the (xth,yth) pixel; this is the uncertainty in the observed images pixel value and follows the $\sigma_{x,y}$-image definition from GALFIT (Peng et al., 2002; Peng et al., 2010a). This is then summed over all pixels in the image, giving us a single χ^2 value between each simulated image and observed image.

Finally, to help computation time the log is taken of our likelihood function. This leaves us with a final expression that a given set of parameters describing a simulation image also describe the observed image input into the algorithm,

$$\log_{10}(P(H_i|D_{obs}, C)) = \log_{10}(L) = -\frac{\chi^2}{2} + \log_{10}(C). \tag{4.9}$$

This is used at every step of our MCMC chain, with a simulation have to be run for each.

Thus, we now have a method by which to quickly predict the resultant morphology and flux distribution that the user would observe in a set of given filters. We also have a means of exploring the likely parameter space of an interaction, given an existing observation. To test this methodology, we first apply it to an idealised scenario where we know the true underlying parameters that formed the system.

4.5 RESULTS & DISCUSSION

We now apply our MCMC algorithm to our set of synthetic observations. To reduce the number of repetitive figures in this paper, we focus here on a specific system as a representative sample, and publish the results of all samples online. They are presented online[1]. As these input images are created from APySPAM, they provide us with a set of synthetic images of interacting pairs with no noise and where the underlying parameters are well known. Initially, we explore the results of running this on the synthetic image of Arp 240. We discuss how the

[1]All results are found here: `Link_to_results`

constraints could be improved, and discuss improvements on our constraints with extra, 3D information. We then plot the results for the entire dataset, and discuss trends we see in the parameters we recover.

We then apply our methodology to a trio of real, observed images of interacting galaxys'. We select the three which had the tightest constraints from our simulated dataset. We apply our MCMC to only a subset as the computational expense is significantly increased to make constraints on the observational systems and for our MCMC to reach convergence. We compare our found best fit values with those of our synthetic system, and discuss the difference between applying this to best fit simulations and observations. Finally, we describe the applicability of our approach to other systems; keeping an emphasis on those in the low surface brightness regime and the limitations this would introduce.

4.5.1 Testing on a Synthetic Image

We apply our MCMC to a single image from our mock dataset described in Section 4.3.2. We elect to use this system as it is is composed of two clear and distinct disks, with a tidal bridge connecting them and tidal tails forming on the opposite side. The tidal features lie in the high surface brightness regime, and have an inclination close to 0. The observation and synthetic image are shown in Figure 4.2. There are clear differences between the morphology of our mock image and the observation of Arp 240. Therefore, testing on the synthetic image is not the same as constraining the Arp 240 system. However, the synthetic image is an ideal scenario to test our methodology on. We elect to test our MCMC on the Arp 240 system first, as it is one of the more massive systems in our sample. Thus, the tidal features contain more flux and should be easier for our MCMC to constrain.

We constrain the synthetic Arp 240 image by running our simulation with 2,500 particles with 600 walkers and 7,500 steps in the MCMC. An example of our full results is shown in Figure 4.3. This corner plot is created using the Corner Python Package (Foreman-Mackey, 2016). However, displaying our results as is shown in Figure 4.3 will be difficult as we will be discussing multiple different systems throughout this section. This larger corner plot also uses a lot of space

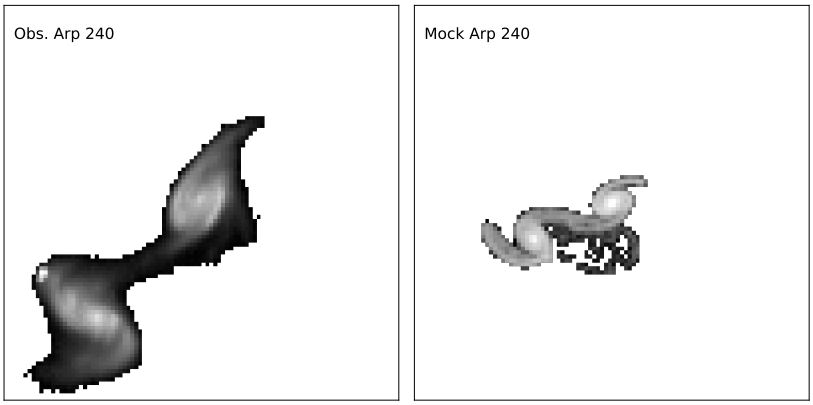

Figure 4.2: The example system used to test our MCMC: the Arp 240 interacting system. This system is considered an easy one to constrain. It is composed of two clearly distinct galaxies, with strong tidal features that our MCMC can match. These tidal features are the two tidal tails formed in the interaction and the tidal bridge linking the two systems. *Left*: The prepared observation image of the Arp 240 system created from SDSS DR16 observations. *Right*: The best fit simulation image as found by Holincheck et al. (2016) and the first test image used in our pipeline. The different in scale and orientation are discussed below.

displaying corner plots and contours of parameters we do not expect to correlate. Therefore, we will present and discuss our results using reduced corner plots as shown in Figure 4.5.

Figure 4.5 shows the constraints we have found on each parameter from our MCMC. The golden lines then show the true values used to create the synthetic image, and are those described in Table 4.3 for Arp 240. The contour levels correspond to 11.8%, 39.3%, 67.5% and 86.4% of the samples across all walker chains (these are default values), indicating the confidence level (CI). For reference, we have set up the corner plots here such that the contours containing 68% of the walkers roughly corresponds to 1σ[1]. However, this assumes a Guassian probability distribution, which is not strictly true for the results we find here. Figure 4.5 shows that each of the truth values is within a 67.5% CI of the probability distribution with the exception of the z-position and time.

We get excellent constraints on the masses of the two galaxies. This system is the most massive system in our sample, and therefore we expect the mass at

[1]Full explanation in Corner plot Docs: `https://corner.readthedocs.io/en/latest/pages/sigmas/`

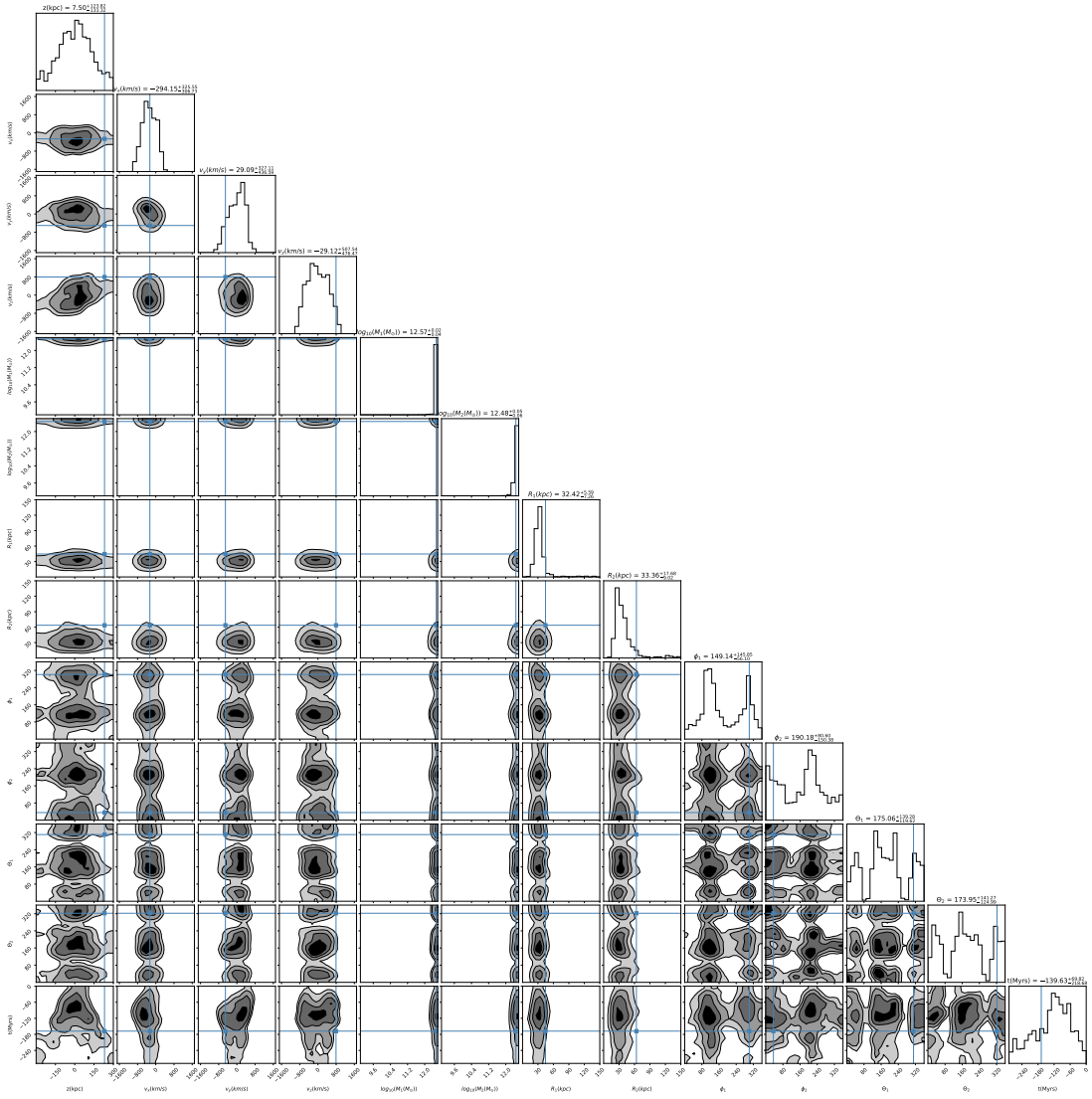

Figure 4.3: Corner plot showing the constraints made on all thirteen parameters we are exploring. Each contour level contains 11.8%, 39.3%, 67.5% and 86.4% of the samples across all walker chains. Displaying our results using the full corner plot is difficult in a paper because of the high dimensional results that we obtain. Therefore, we elect to show all remaining results in this paper as reduced corner plots like Figure 4.5. We elect to put the parameters which are most likely to correlate together in different corner plots. To view the full corner plots of each system, find them at the results website for this paper.

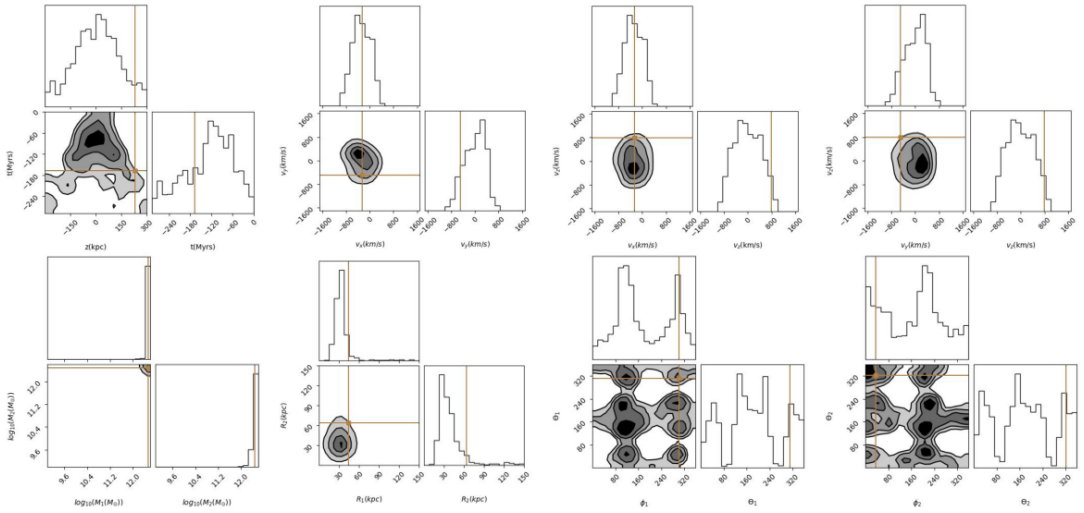

Figure 4.4: Same as Figure 4.3, but reduced to only corresponding parameters.

the very limits of the range we are exploring. We find that the MCMC first converges on the two masses of the disks, and are the easiest for us to constrain. We provide the algorithm with the secondary 2D position, and therefore it only has to fit the flux distribution of the inner disk correctly to get good constraints on the galactic masses. The largest gains in likelihood maximisation come from matching the flux distribution in the inner parts of the primary and secondary. While the formation of the tidal features depends on the mass ratio, they also depend heavily on the orientation of the interaction as well as the relative sizes.

The relative sizes of the disks is important for the constraining the tidal features as it is the outer parts of the disk that are sheared off to form them. The primary radius is constrained very well, and is primarily expected to be smaller than the true value of this system. This is, again, due to the χ^2 nature of calculating the distance between the two images. The likelihood, on average, is lower with some pixels within the galaxy being missed than a pixel containing the central disk in it when it should not. We therefore have a bias effect where the peak of probability drifts to just below the true value of the radius. However, the true value of the radius remains within 86.4% CI of the found contours.

The recovery of the radius of the secondary is worse than that of the primary. We find the peak in the probability distribution is far lower than the true value.

This is from a limitation of our simulation as well. While the output simulations of the MCMC are always centred on the primary, the secondary galaxy central position is not so certain. Due to the backwards integration and trajectory calculated here, the secondary does not always end in the exact same image bin in the output simulation image. Therefore, the secondary disk likely moves slightly per simulation. This change transfers further uncertainty in the secondary disk size, and a preference for the secondary to be smaller to improve the likelihood calculation. The size of the secondary disk also contributes much more to the tidal debris formed in the interaction. The simulation is conducted in the frame of the primary, leading to the 3D velocity of the secondary being higher than that of the primary. Therefore, the particles in the secondary are much more likely to be ejected during the encounter than the simply orbiting primary particles. This adds further uncertainty to the secondary radius. However, once again, the best fit values found are within 84.6% CI of the true value from the base simulation.

There will also be some uncertainty involved in this measurement from the formation of the tidal features. The flux distribution of the tidal features that form are inter-dependent on multiple parameters; primarily the mass ratio, the size ratio and the orientation of the galaxies in the interaction. The significant degeneracy in the orientation constraints undoubtedly has some effect on the fitness of the resultant system. Due to a lack of three dimensional information, our algorithm cannot discern which way the galaxy is rotating or which way the tidal features should be orientated in the line-of-sight. Therefore, degeneracy at $\pm 180^\circ$ of the true parameter values for ϕ and $\pm 180^\circ$ for θ. Figure 4.5 shows this with several different peaks in both measurements of ϕ and θ. However, we recover the true parameters in one of the peaks of the marginalised posterior for each orientation parameter.

It is important reiterate here that ϕ is the orientation of the galactic disk with respect to the y-plane while θ is the orientation with respect to the z-plane. With 3-dimensional information, such as the direction of rotation of the disks or the line of sight (LOS) velocities of the tidal features, we would be able to resolve the degeneracy in the ϕ parameter. However, The source of the degeneracy in θ has a different source. The tidal features can form in the opposite direction from the mock observation and still be found to have high likelihood. This is a result

of our likelihood being based on flux matching on pixels. There is no knowledge provided of the direction the tidal features should be moving or forming, only if the pixels contain the correct flux. Therefore, this gives a significant degeneracy in the θ parameter. Therefore, while the disk can be flipped in the z-direction and still match the observation, it can also be flipped in the y-direction as well. While the degeneracy in ϕ can be solved with velocity information, the θ orientation will require information on the rotation of the disk itself. Having rotational information will allow us to constrain the bulk motion within the galaxy and the direction the inner disk should be rotating as well. These two pieces of 3D information will remove the 4-fold degeneracy in each parameter.

The lack of 3D information also affects our constraints in the z-direction: the z-position and the z-velocity. As seen in Figure 4.5, the finds the peak in the posterior distribution for both parameters as much lower than they truly are. First, the z-position is difficult to constrain as we lack 3D information. The simulation is run in the reference plane of the primary galaxy, therefore the secondary can only be behind or in front of iy. There will be change in the flux of the secondary based on whether it is in front or behind of the primary galaxy. However, this change in flux is completely dominated by the distance due to the redshift of the galaxy. Therefore, there is little to no observable change in the absolute values of flux unless the true value of the z-position was very large.

The constraint of the z-velocity remains far from the true parameters due to similar reasons as above. The true value lies at the very edge of the probability distribution found, at approximately 86.4%. This would, again, be rectified readily by introducing velocity information into our constraints. The simulation works so that the secondary galaxy is always at the same x and y position that is defined by the user. Therefore, an output simulation with a positive or negative LOS velocity will have the same flux distribution. Hence, this constraint simply peaks about zero for the z-velocity.

Our MCMC works significantly better, however, with the x- and y-velocity of the secondary galaxy. We find the the truth value of this is within 67.8% CI of our found distribution. The velocity values are directly related to the strength of the interaction, and therefore indirectly relate to the tidal features which form.

Thus, our MCMC is informed by the flux distribution of the resultant simulated image and gives an excellent constraint on the parameters.

Finally, we discuss attempting to constrain the time of the interaction. We show the time since closest approach, not the total interaction time. The underlying simulation utilises backwards integration to calculate the trajectory of the interaction. Therefore, the time we input into the simulation simply tells it how far back in said trajectory to put the secondary galaxy. Therefore, the same interaction will occur whether we input -10 time units or -100 time units. The algorithm will require more computation time to calculate the particle positions in the lead up to the interaction. It is important to note that the total integration time will only affect the output system when we make it too small. In other words, if the secondary starts after the point of closest separation or at closest separation, our simulation breaks down and gives nonphysical results.

We calculate the time of closest approach for all of our walker steps and then present this as a measure of the time posterior distribution. Our measured value is significantly smaller than the true value of our best fit simulation, although it does lie within the region of 86.4% CI. This parameter is highly dependent on the velocity and position constraints that we have made, and these are skewed to significantly smaller values than the truth. Therefore, it is unsurprising that our time of closest approach value is also found to be much smaller.

To fully put this result into context, we explore the simulations that lie in the areas of highest probability within these posteriors. To select which simulations to present, we take those walkers that were within the 11.8% CI throughout each walker chain and take the top 5 as an illustration here. Our MCMC is not able to precisely reproduce the input image. It is often able to reproduce the tidal features of the primary as well as the tidal bridge connecting the two systems. This appears to be where the MCMC has centered the posterior upon. The likely reason for this is actually due to the filtering parameter that we use on the simulation. The Arp 240 simulation lies in an unlikely area of tidal features to form - $\gamma = 0.259$ - and, therefore, we update our prior to make the true result appear less likely. However, the γ parameter remains a necessity in our MCMC. Without the ability to filter the simulations quickly, we begin to approach

Figure 4.5: Simulations from the areas of parameter space that lay within the 11.8% CI of our constraints. *First Row*: The synthetic image we are attempting to constrain. *Second Row*: The best fit simulations from this parameter space. *Bottom*: The worst fit simulations from this parameter space. Our MCMC found those parameters which cause the formation of the correct tidal features, as well as the tidal bridge connecting the two systems. However, it has been unable to fully identify the tidal features of the secondary. There is also a lot of noise in this posterior distribution, with many systems with different tidal features in the areas of high probability. Therefore, identifying specific systems with the sought after tidal features requires manual intervention.

very large requirements of computational expense. Therefore, for this particular example the γ parameter is a hindrance.

In the surrounding area of probability space, however, we find some interesting results. Changing each parameter by small amounts based on the posteriors of each parameter space leads to variation in the simulation outputs and tidal features. Due to the disks being well defined and aligned, this can often lead to them being weighted highly. Therefore, the question remains, how would one use this code to find their best fit simulation and actually make constraints using it? This algorithm is best used as an indication of where in parameter space the true parameters lie in recreating the tidal features observed in an observation.

This reduces the size of parameter space to explore dramatically, and could be an indication of where to search with more accurate simulation models for true interacting galaxy parameters.

Overall, from our example of Arp 240, we are able to recover nearly all the true values of the input simulation to within an 86.4% CI of the true parameters. The only missing parameter is the time since the flyby. There is significant degeneracy in constraining the orientations of this interaction, but this is expected. While our results appear like they have converged in the MCMC, we will also describe the diagnostics with which to prove this.

4.5.2 Diagnostics of Pipeline

It is important to ensure our results are reliable by using diagnostics to investigate the MCMC chains. We investigate three different diagnostics of our MCMC run. First, we check that they have truly converged with the Geweke diagnostic. The Geweke diagnostic is a Z-test of equality of means where the autocorrelation in the flattened samples is taken into account as the standard error is measured. We compare the means of the first 10% and last 50% of each chain in each parameter, and require the resultant Z-score to be < 1 for convergence. We use the Geweke diagnostic as written in the ChainConsumer (Hinton, 2016) Python package. For this case, every parameter passes this convergence test, with the exception of the orientation parameters (although, this is likely the result of converging on multiple best-fit values).

Figure 4.5 clearly shows that each orientation has incredibly complex, multi-model structure in the parameter space. We have a 2- or 4-fold degeneracy in the output models due to three dimensional information not being available. This disrupts our Gewke diagnostic measure, which is looking for a single peak in parameter space. Therefore, by folding the parameter space over and only exploring over 0° - 180°, we achieve a single peak in parameter space. These results then pass the Geweke diagnostic test.

The second diagnostic we check is that the walkers have fully explored parameter space and that we have removed enough of the steps at the beginning of the run to consider the MCMC burnt-in. Once again, using ChainConsumer,

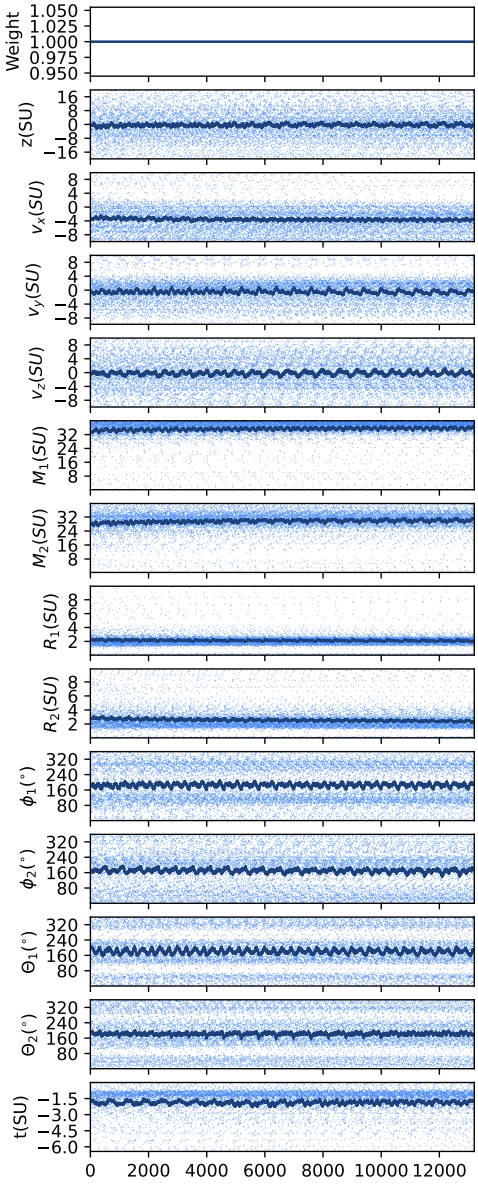

Figure 4.6: Steps taken by each walker in our MCMC chain to constrain the Arp 240 best fit simulation. Note, the y-scales here do not extend over the full parmaeter space for some galaxies, and only show where the walkers have stepped after the burn-in phase. The deeper the blue, the more walkers have stepped at that point. This figure shows that our MCMC has successfully burnt in and very quickly goes to high areas of probability for parameter space. They then oscillate around the best fit values while searching the remaining parameter space. The z-position, z-velocity and ϕ_2 parameters show dignificant uncertainty as the walkers move around the entire parameter space. In the $\phi_1, \theta_1, \theta_2$, we can see the two fold degeneracy form very early on and then the walkers do not explore across them at any point.

we can plot out each walker step throughout parameter space. Figure 4.6 shows the flattened walker chains through parameter space. We have removed the first 200 steps of each walker chain before thinning the chain and flattening it. By flattening we have taken each walker chain and combined them into one chain for every parameter. Discarding the first 200 steps has been enough for the burn in of the MCMC, as the walkers have already moved mostly through parameter space and are centering on a central value of high probability. The structure in the orientation parameters is also interesting. The degeneracy structure in the parameter space has already formed by the end of the burn-in and then the walkers move within those areas of high probability. This is for two reasons. First, they affect the flux distribution of the disks primarily in if they are face-on or edge-on. The MCMC quickly converges on face-on systems in our case. Second, when they are in that degenerate space, the slight changes in inclination of the disks given does not change the likelihood calculation significantly enough to reduce this degenerate space further. Hence, the degenerate areas are very large with very flat areas of probability at their peak.

We have tested resolving this problem by running further steps in our MCMC to achieve convergence naturally within the parameter space. We find that increasing the number of steps does improve convergence on the orientation parameters, but at a much larger computational cost and without substantial improvement on the other parameters. Therefore, we elect to fold our resultant degenerate solutions into a smaller parameter space. This achieves convergence, and gives us excellent estimates on the orientation for these systems.

A second solution to this problem would be involving velocity information into our models. Knowing the bulk motion of the tidal features would allow us to constrain the tidal features based on which way they were rotating. This would eliminate part of the degenerate space. However, as stated previously, little spectroscopic data exists of our sample of interacting galaxies. Therefore, we run this test using the our synthetic Arp 240 image and including the LOS velocity of the particles and creating a velocity grid to constrain over.

4.5.3 Inputting 3D Information

Very few of the systems described in Table 4.3 have associated integral field spectroscopy (IFU) data in order to get LOS velocities to incorporate 3D information into our fitting process. This is because they are very large in the field-of-view of many instruments, and therefore we can only achieve measurements of their inner disks and not their tidal features. We, therefore, map the velocity distribution of our synthetic Arp 240 image. This map is created by summing the z-velocities of each particle in the bin and then create a total LOS velocity map for which to compare to simulations. This has little impact on the measurements of mass, size and x- and y-velocity measurements. However, it completely changes our measurements of the z-position and time of interaction as well as improves our constraints on all three velocity vectors dramatically.

Figure 4.7 shows the new measurements of the constraint on the z-position, y- and z- velocities. The constraints on each of these parameters is significantly improved. This is with the same number of MCMC walkers and steps as without 3D information. For the y- and z-velocities, the constraints are improved to the point where we completely recover the true underlying parameter values within 39.3% CI. With the z-position, we also gain significant constraint. We are able to recover which side of the primary the secondary lies, with a sharp drop in the marginalised posterior over the negative part of the z-position parameter space. We also significantly improve our constraint on the time of the interaction, and approximately recover the true parameter. We are able to constrain the final velocity expected in the observation, and therefore, means the galaxy cannot be moving too fast at in the final timestep. This has the effect of shifting the expected time of contact back significantly.

Interestingly, we lose a significant amount of constraint on the relative sizes of the two systems. Adding the velocity information in the way we have likely increases the required size of the disks so the LOS velocities match within them. Therefore, more particles are stripped from the secondary galaxy and cause further uncertainty in the final image. We also find the removal of some of the degeneracy in the orientation parameters, as expected. It is important to note

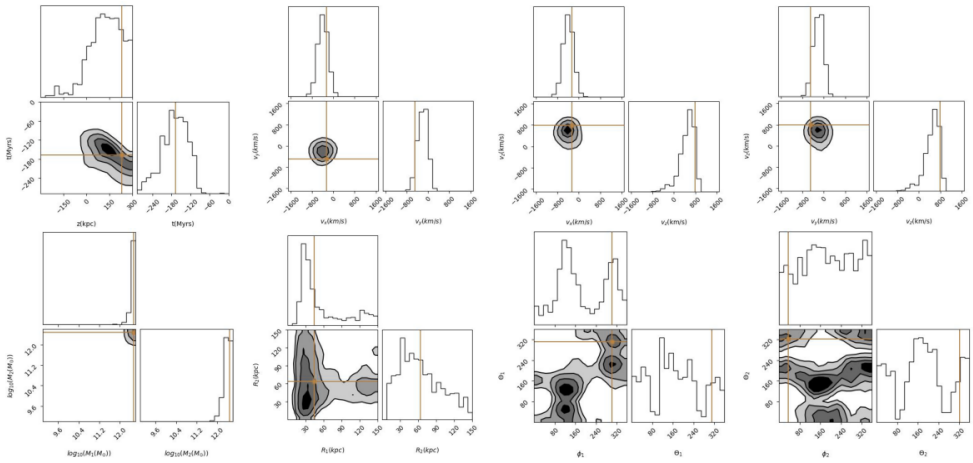

Figure 4.7: The reduced corner plot of the synthetic Arp 240 system with a velocity map also used for constrained. To add this extra information, we created a mock velocity map of our Arp 240 synthetic image and summed the LOS velocity (z-velocity) in each pixel of our image. Comparing between here and Figure 4.5, we see we make different constraints on the z-position and the time as well as resolve one of the degeneracys in the orientation space.

that, in this example, we have only used the LOS velocities to achieve this improvement in constraint. To better get constraint on orientation, we likely need further velocity and 3D information of the system. For instance estimates of the internal rotation, and accurate measures of the bulk motion of the tidal features in the galaxy. Currently, the simulation is not able to accurately reproduce this outwith simply assuming circular velocities at different radii from each galactic centre. Finally, the mass constraints remain the same as was found without velocity information as they are dependent primarily on the flux distribution and dominated by the inner disk of the system.

This shows the improvement that adding velocity information to our MCMC could bring, and how far interacting galaxy simulation and constraint can go once we incorporate IFU spectroscopy over more systems. In the sample we are using here, of the most massive, large, major interacting systems, only three have got any IFU data. This data is from the MaNGA (Bundy et al., 2015) IFU spectrograph, whose field of view is only able to capture the central disk of these systems. While having velocity information on even a subset of the pixels of the galaxy would improve constraints, to significantly improve them we will need this

of the full tidal features of the systems. IFUs with larger fields of view are soon to come online, such as WEAVE (Dalton et al., 2014).

4.5.4 Running on Full Idealised Sample

With constraints being ascertained on nearly all parameters of our synthetic Arp 240 image, we then applied our MCMC to the remaining 50 synthetic images. The reader is invited to see the resultant corner and reduced corner plots of each system on the website. Here, we will detail describe our best and worst three corner plots, and then discuss the trends throughout applying the MCMC to the dataset. First, our ability to constrain is highly dependent upon the stage of interaction. Our tightest constraints, the ones with narrowest posteriors, were on those interacting galaxies which were only just past the point of closest approach. I.e., they were the systems which had highly distinct tidal features and disks were fully separate. Our best three fits were those of synthetic images of Arp 172, Arp 240 and Arp 290. The 'best fits' have been judged by those with the smallest FWHM of their marginalised probability distribution in mass. The reduced corner plots are shown in here in Figures 4.5, 4.8 and 4.9. Each of these systems is clearly in stage 2 or 3 of the interaction, where tidal features have formed clearly, and there are two distinct disks.

The worst fits we achieved were of those systems which were close to the merging stage. They were systems where the two cores were close to coalescence, very little tidal features were visible or the position of the secondary galaxy was very unclear. Our worst three fits and examples of this were Heart, NGC4320 and Arp 57. The reduced corner plots are shown in Figures 4.10, 4.11 and 4.12. Each of these systems represent the three limitations, respectively. First, is the issue with constraining a stage 4 system. If two cores are close to coalescence, the flux distribution will appear similar to two overlapping disks. Therefore, our χ^2 calculation will lead to a maximised likelihood which is equivalent between systems with little to no interaction and a merger with multiple flybys undergoing coalescence. This would be improved by including velocity information, where the velocity distribution will be very different between two overlapping disks and coalescing galaxies. Our process is also not yet designed to account for multiple

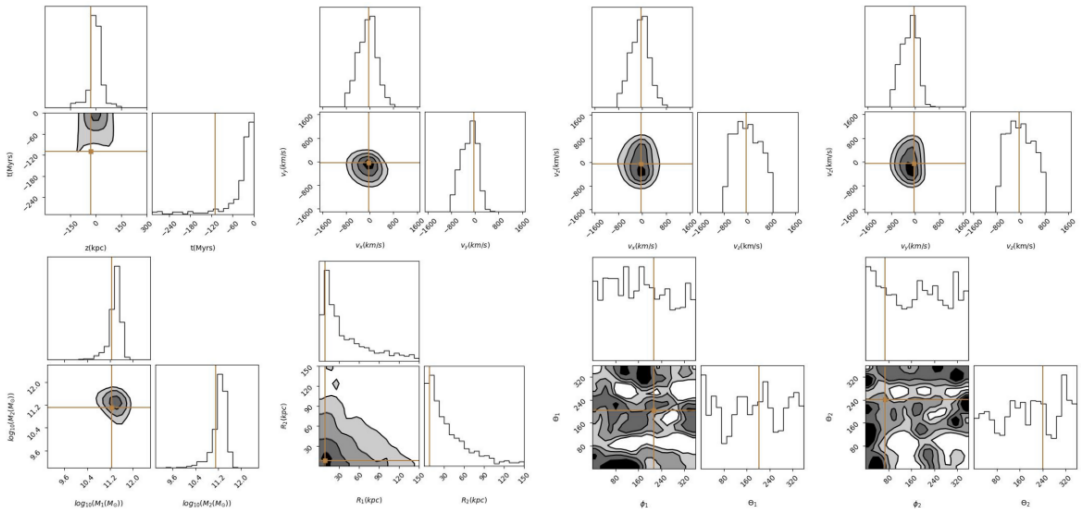

Figure 4.8: Second example of our best fits from our methodology is representative of the simulated image of Arp 172, a stage 3 system. Gold lines represent the true values.

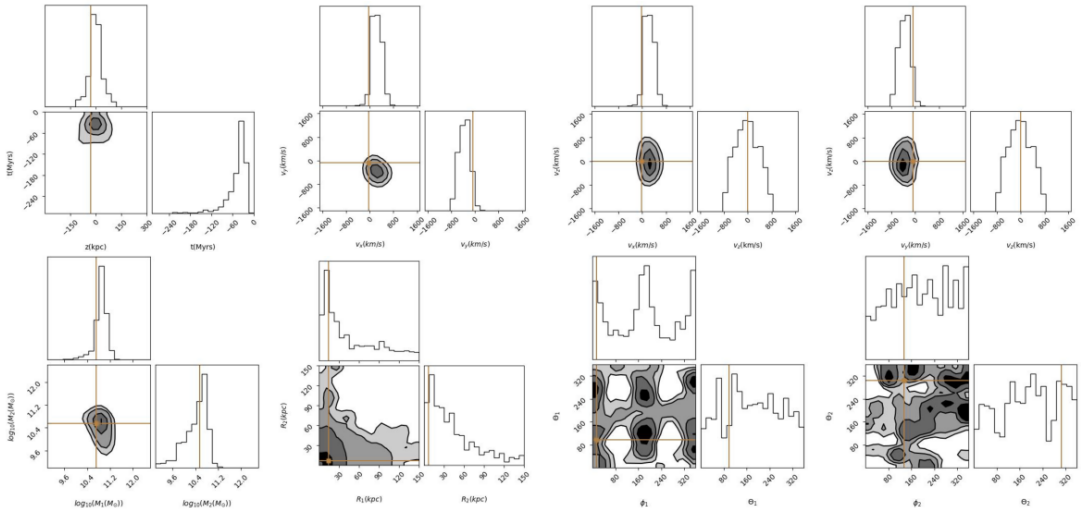

Figure 4.9: Third example of our best fits from our methodology is representative of the simulated image of Arp 290, a stage 2 system.

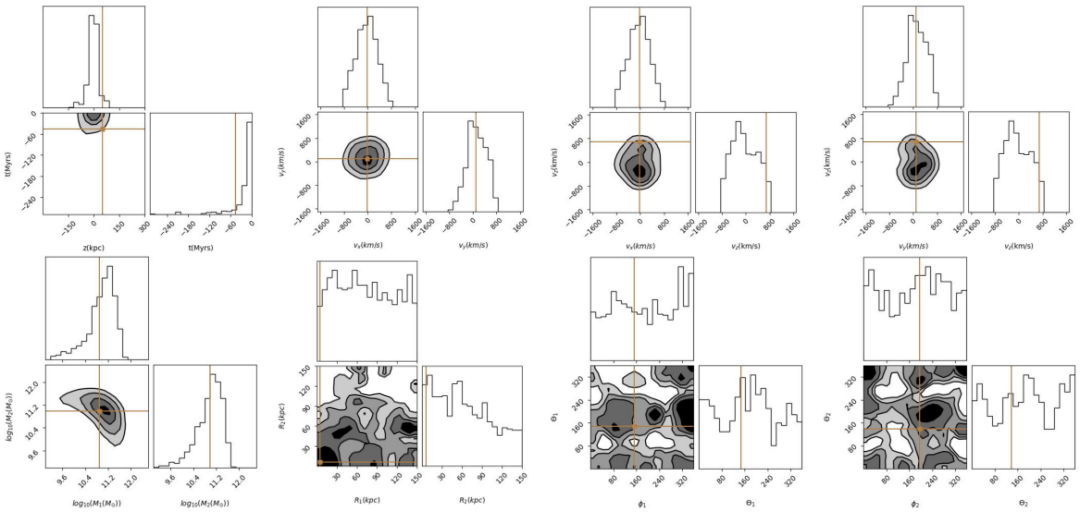

Figure 4.10: First example of our worst fits. This shows our constraints of the simulated image of the Heart system, a stage 4 system.

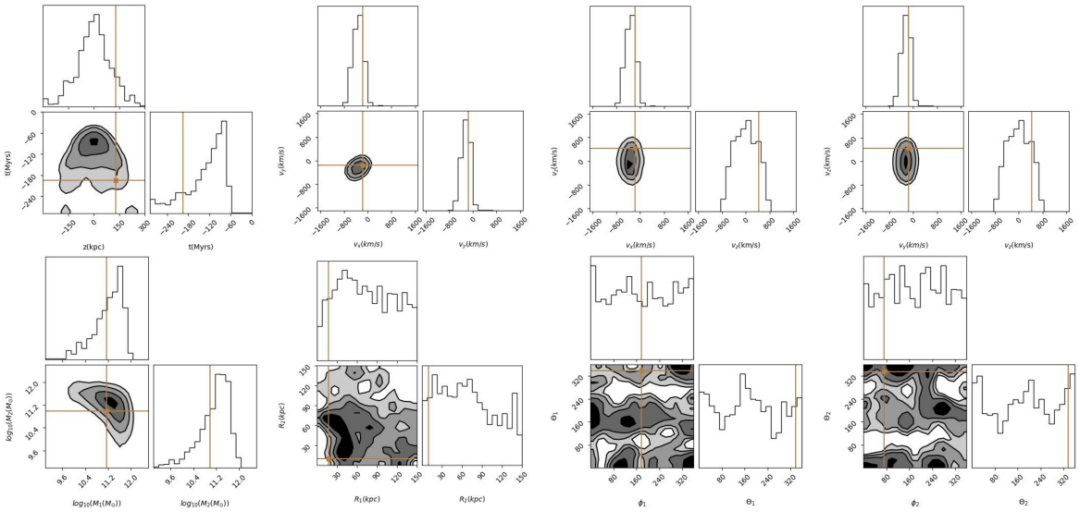

Figure 4.11: Second example of our worst fits. This shows our constraints on the simulated image of NGC 4320, a stage 4 system.

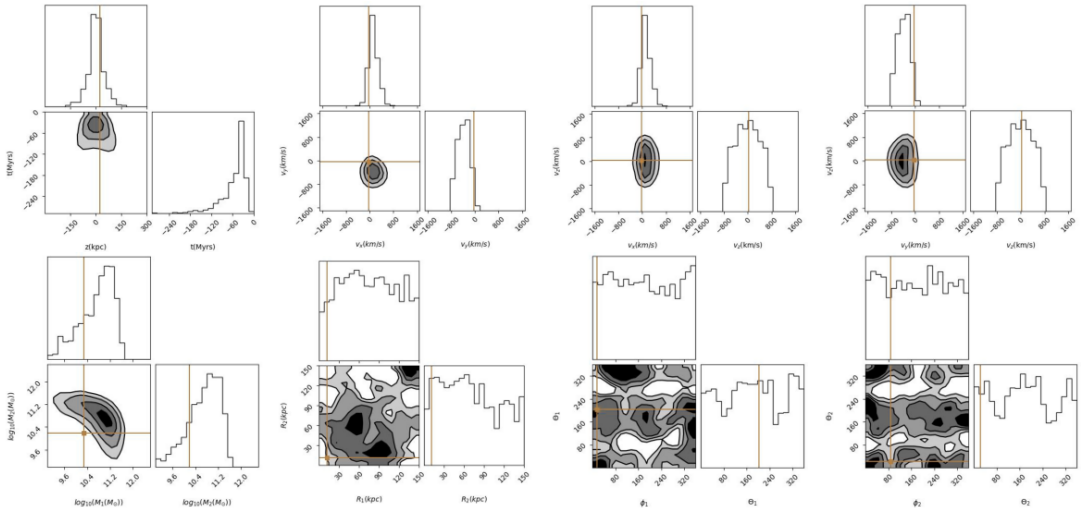

Figure 4.12: Third example of our worst fits. This shows our constraints on the simulated image of UGC 11751, a stage 3 system.

flybys, which is likely to have occurred in a merging system. To make accurate constraints on many parameters, we require clear and distinct tidal features. UGC 11751 is a system where we have reduced the resolution so much that we lose spatial resolution in the flux distribution of the tidal features. Therefore, we insert a significantly higher uncertainty into our constraints. Finally, our MCMC requires a secondary galaxy to make reasonable constraints on the underlying parameters. The example of NGC 4320 is a final stage merger where a large tidal feature has formed as a result of the coalescence.

Throughout every system we put constraint on, the parameter we are able to get reliable constraints on are the masses of the systems. This is shown above as even our worst fit systems still hold excellent constraints, with the values confined to small areas of parameter space. However, our worst constraints lie in the orientation and time since the interaction. We show the relation between the true parameter used to create the synthetic image and our recovered best fit parameter with errors included. Figure 4.13 shows this distribution. We define the best fit value from our MCMC as the 50th percentile of our walker distribution. This is represented by each point in Figure 4.13. The upper and lower error bars are calculated from the difference between the 50th percentile value and the 84th and 16th percentile values, respectively. The red lines in each

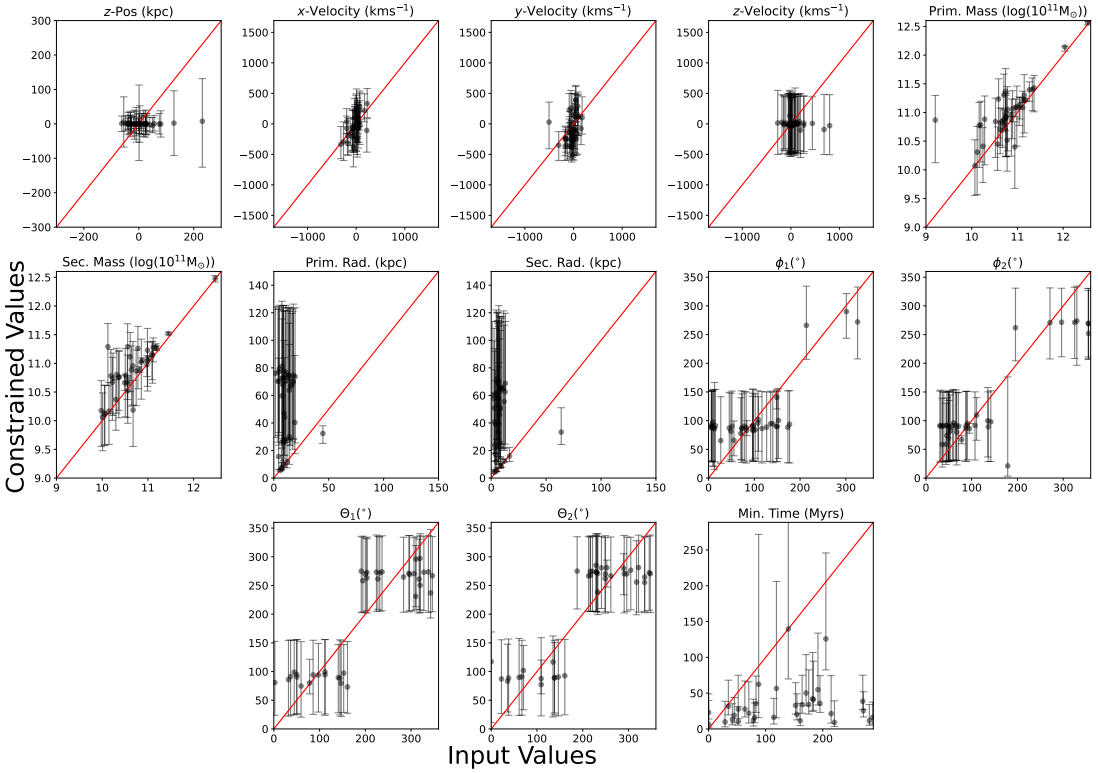

Figure 4.13: The distribution of parameter values used to create our synthetic interaction images and the best-fit values we recover from our algorithm. The constrained values are taken as the 50th percentile of our walker distribution. Our lower and upper errors are then calculated as the difference from our best found value and the 16th and 84th percentiles, respectively. This gives an idea of the extent of our walker distribution in 1D. The red lines show a 1-1 relation and where our points would lie if the 50th percentile represented the exact input value found. Parameter names and unts are found in the title of each subplot.

subplot represent a 1-to-1 relation between the true parameter value and the recovered value.

Figure 4.13 clearly shows where our MCMC has made excellent constraints and where it has not. The best constraints are on the masses of the interacting systems, where each best-fit value approximately follow the 1-to-1 relation. Where there is some discrepancy, the true values are within our errors. There is a trend here, however, that the total mass of the system is over estimated when compared to the true value. We also find excellent constraints on the majority of the positional and velocity vectors of the interaction, where the 50th

percentile value approximates the true value. However, in these parameters we have explored a very wide parameter space when the majority of the true values are clustered around the centre of the parameter space. We will require further investigation with interacting systems more representative of the full parameter space we have explored.

We also have reasonable constraints on our orientation parameters. As we often have two areas of high probability due to degeneracy, we opt to only plot the 50th percentile and errors of the high probability space near the input value. This would not be possible with an observational system with unknown parameters. Therefore, the user will either need to make a choice of what area of parameter space to explore, or input some prior on the orientation / inclination of the system. Once we have accounted for the degeneracy, we find that our 50th percentile value is representative of the true value, with the true value being within error of it.

There are three parameters remaining where we find poor constraints or struggle to recover the true parameter value in general. The first are the radii of the two galaxies. We find that these are often over-estimated. Upon inspection of many of the output simulations, we find this is likely an effect of the distribution of particles and that we are using idealised synthetic images with no noise in the image. The way in which the particles are distributed is with an exponentially declining disk. This is achieved by creating a disk of particles which are sectioned into rings. Each ring has an exponentially declining probability of particles appearing within it based on the radius from the centre. Therefore, we find the outer rings of the galaxy are diffuse of particles, and do not dominate the likelihood function at high radius. This, in turn, means that galaxies with large angular size can be calculated as likely if the inner parts of the galaxy, with high particle numbers, match the flux of the inner disk. This also explains the over-estimate of the mass of each system. There are likely less particles than expected still within the galactic disk, and therefore more mass (ergo, luminosity) is required to keep the flux matched in the disk.

We also find we often under-estimate the time since the closest point of the encounter - i.e., the time of the passage - by a large margin. While those synthetic images whose underlying parameters had their point of passage very recently we recover very well, those at larger values we fail to. This is due to the likelihood

function finding equivalent, high probability values between systems with correct tidal features and simple disks which match the inner disk of the interacting system. The former, tidal feature formation, requires a high β parameter with low velocity and low impact distance. However, if the system is fast and does not form tidal features, but ends the simulation in the right place with the correct flux, the calculation will be dominated by the inner disk. Thus, we find a radius that covers the inner parts of the galactic disk successfully, with little disturbance that matches the flux of a large part of the galaxy. Therefore, the β parameter is of the utmost importance to filter out these galaxies and give weight to higher t_{min} values.

Aside from the points that have already been raised regarding our MCMC constraining each parameter, a further problem was that the likelihood calculation is dominated by contributions from the mass and velocity parameters. Therefore, to continue improvements on other parameters further walkers and a longer chain must be run, incurring a cost of far more computation time. For the purposes of this work, we show the full degeneracy, limitations and difficulties of our model and what is to be expected in the most general case. Many of these limitations and problems can be resolved from tighter priors with more information regarding the systems being investigated.

The true power of this approach will be when investigating large populations of interacting galaxies and combining and marginalising over many different posteriors. When combining the parameter spaces of many different systems, it will be possible to identify those areas of parameter space which lead to the formation of certain features across populations of interacting galaxies. Our method will allow more intense simulations to sample smaller parameter spaces and be more efficient when finding systems with specific features. However, how to combine the posteriors is somewhat up for debate. We have ensured that the parameter spaces we are exploring are equivalent in size and that the prior is equal across parameter space. This introduces the question of the β parameter in filtering simulations which, indirectly, changes the prior based on the trajectory of the interacting system. To conduct this combination, we would recommend that the β parameter was not utilised when building the different posteriors if the increased computational cost is viable.

However, all of our results so far have been in the best-case scenario of a noiseless best fit simulation. The translation from simulation to observation in previous algorithms such as these is never an easy one. We, therefore, use our top three best fit systems here (Arp 172, Arp 240 and Arp 290) and apply our pipeline to their reduced observations.

4.5.5 Applying to Observations

We apply our pipeline to the reduced observational data of Arp 172, Arp 240 and Arp 290. Cutouts were created as described in Section 4.2.2. We reiterate here that the cutout resolution is reduced from its native resolution to cutouts of 100 $\times$ 100 pixels. Before we input the images into the pipeline, we find the central pixel of the secondary galaxy and convert this into a physical x- and y-position. We also find the total size physical size of the cutout in kpc, convert it to simulation units and provide this to the pipeline. We find that the physical size of the cutouts are significantly different from those of our synthetic dataset. As an example, we create the cutout of Arp 240 at 600 $\times$ 600 pixels at native resolution. With SDSS data, this corresponds to a physical size of 111.50kpc $\times$ 111.50kpc at the redshift of Arp 240. The synthetic image we previously used of Arp 240 is 785.35kpc $\times$ 785.35kpc. Figure 4.2 clearly shows a very different scaling between the observation and simulation, however, it is not enough to be seven times zoomed in on the system. We also find that the secondary positions are very different between the observation image and that used previously. As a result, the parameters we will find for constraining the observation will likely be very different.

We apply our MCMC to our three best fit simulations. We only investigate these three systems as we find the computational expense is significantly higher when constraining the observations compared to the best fit simulations. We find the reasons for these are two fold. First, we must run each walker for twice the number of steps than when constraining the synthetic images to reach convergence. This is from increased noise contributions in the observation images. Second, due to the smaller scale size of our images, the defined γ parameter is much stronger than with the synthetic images. Both of these reasons for higher

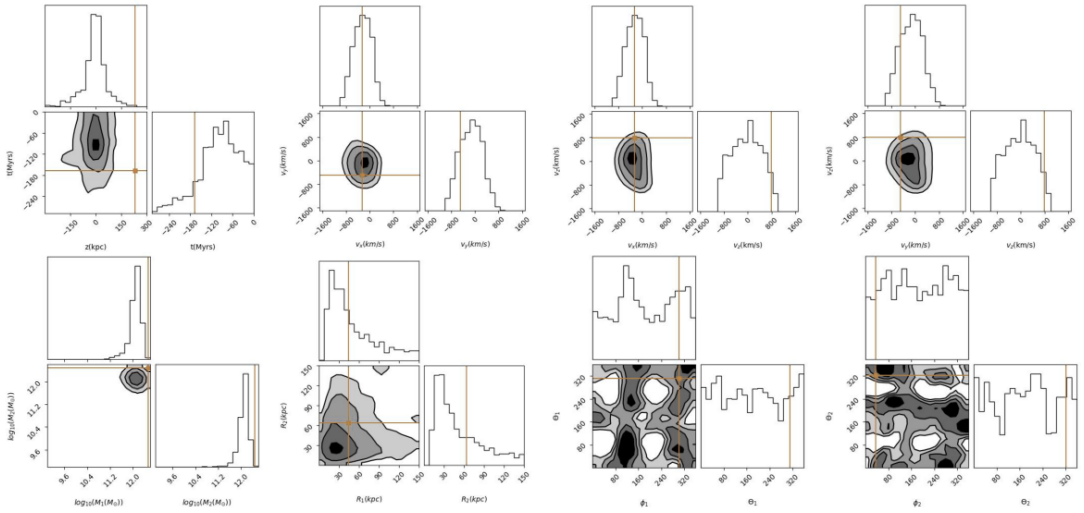

Figure 4.14: Reduced corner plot of the constraints made on the observational image of Arp 240. As shown, there is significant additional uncertainty in these measurements than those of the best fit simulation. However, we are able to make constraints on almost all parameters in the sample. The degeneracy in the orientations remains. The gold lines represent the true values we used in to create the synthetic image.

computational expense are to be expected. The increase in the γ parameter leads to us filtering out less candidate systems and running the base simulation code more often. This directly translates into a higher runtime for our MCMC.

Figure 4.14 shows the reduced corner plot for the observed Arp 240 system. The constraints on the velocity parameter have significantly worsened when compared to the constrains on the best fit simulation. They are also in a different area of parameter space when compared. The velocity and spatial parameters are the most likely to be affected by the change in scale between the two input images. As the secondary galaxy position has been completely altered the best fit trajectory of the interaction has also completely changed. The secondary is significantly closer to the primary, therefore meaning the secondary velocity must be significantly slower than previously. Due to the increase in noise in the observational image, and the extent of the tidal features of the primary and secondary galaxies, the constraint has also significantly weakened.

However, for the remainder of the parameters the level of constraint remains

similar even with the change in the best fit values. We retain the two- and four-fold degeneracy in the orientation parameters. This follows on that the pipeline outright rejects disks which are edge on and loses accuracy when attempting to pin point the precise inclination of each galaxy. We are unable to make any constraint on the θ parameters, while the degeneracy of the ϕ parameters is still visible. The radius parameter is well constrained at 11.8% CI, although with large tails in the probability space for both the primary and secondary galaxies. These tails are significantly larger here than when using the best fit simulation as the input. This is due to the observation image having no hard cutoff of the edge of the galaxy, and simply moving into more noise. The outer edges of our simulated galaxies also moves to very low signal, and therefore, there is lots of uncertainty surrounding the true radius of the galaxy.

We, again, make excellent constraints on the mass parameters. The accuracy in the flux distribution of the primary and secondary disk is what dominates here, and therefore, we find this value incredibly quickly. The found value is comparable to the best fit simulation, although slightly lower. We also compare our constrained values to what is found in the literature. Using the criteria described previously, we recover total masses of $M_1(10^{11}M_\odot) = 7.97 - 15.9$ and $M_2(10^{11}M_\odot) = 8.1 - 16.2$ which correspond to stellar masses between $M_{1,*}(10^{11}M_\odot) = 0.94 - 1.89$ and $M_{2,*}(10^{11}M_\odot) = 1.0 - 1.9$ in our simulations. These are slightly higher than literature values, with a recent work quoting the stellar mass as $M_{1,*}(10^{11}M_\odot) = 0.94$ and $M_{2,*}10^{11}M_\odot = 1.05$ (He et al., 2020). Thus, while our 50th percentile value is higher than the measured values, we do recover them within error.

This is true in the Arp 172 and Arp 290 observational constraints as well. The synthetic image we have used is at a different scale when compared to the the SDSS observations. We therefore end up with similar constraints to the different parameters, however with significant variation in the spatial parameters. The uncertainties in the radius parameters are significantly increased, while the constraint on the θ parameters is completely lost. In fact, for the same number of steps compared to constraining the best fit parameters, our MCMC fails the Geweke diagnostic. Therefore, due to the added noise running longer chains is imperative which drives up computational expense.

This leads us onto the limitations of this process to constraining galaxy interaction. While it has been very successful for many of the underlying parameters, there are many different parameters which must be provided to the pipeline so it is able to make those constraints. We discuss the limitations of applying this methodology to large interacting galaxy datasets, and how these can be offset in the short term but a long term solution is still required.

4.5.6 Limitations

While this methodology does show a lot of promising in being able to constrain underlying parameters across populations of interacting galaxies, it is important to note that at its base is a restricted numerical interaction code with limited resolution. There may be some interacting systems that simply cannot be modelled realistically using three-body approximations, and may therefore cause a skew in results. However, there are also some more subtle limitations which may affect how a user wishes to use this MCMC approach.

4.5.6.1 Resolution & Depth Effects

One of the fundamental parameters that must be provided for this methodology to work is the redshift of the system. This is used to calculate the distance to the system and then converted to an assumed resolution for the scale of the image. If this is input incorrectly the simulation images will be created at an incorrect resolution, and the MCMC will likely not reach convergence. This is also true of providing and calculating correctly an approximation of the position of the secondary galaxy. Significant computation power can be wasted if these parameters are incorrect, and will give spurious results. The reduction in resolution of the images is also an important limitation. This toy model needs small, thumbnail size images with a reasonable number of degrees of freedom as to reach convergence. Therefore, input images can only be of limited resolution which will affect the quality of the constraints we can actually make on different systems.

The redshift is also used in scaling the fluxes calculated for each particle in the simulation. The base SED is calculated for a 1 $M_{\odot}$ system at a distance of

10pc. This is then scaled to the mass assigned to a particle and then put at the distance calculated from the redshift. If the redshift is incorrect (or even slightly off) then this could lead to a bad fit for the mass of the two galaxies in the system. A flag exists in the algorithm to give it the freedom to slightly vary the redshift (and therefore the distance and resolution) by 0.001 with each step. It will then attempt to fit the redshift and resolution of the system. Note, however, this is untested and currently significantly increases computational expense.

Depth effects also play a significant role in the ability of our algorithm to fit a system. For this work, we used primarily synthetic images, meaning we were not at risk of this. However, when using observational imaging this could become a significant problem. Our observational examples were of major interacting systems which resided in the high surface brightness regime, where we retained the full extent of the tidal features and could provide better fitting. However, for systems that lie closer to the low surface brightness regime, our algorithm will become less efficient and require more computational expense to make effective constraints.

This also has the opposite effect in terms of tidal features formed. For example, if a system is being fit but at the true parameters a tidal feature exists which has not been detected due to its low surface brightness then our MCMC pipeline will not be able to converge on the true values. The algorithm would instead converge on those parameters where the disks were at the correct flux rather than getting the tidal features correct. Therefore, when exploring the parameter space of interacting systems with our pipeline, it is imperative that the full structure is within detection of the observing instrument.

There must be a trade off between the resolution of our simulation and the observations. In this work, we used cutouts of 100×100 pixels. This was so that we could still get consistent system outputs from our simulations using only 2500 particles. If lots of pixels are used, then using a low number of particles can lead to many 'unphysical' output simulations. I.e. these are simulation systems where the disks will have large holes in them where there haven't been enough particles to fill the disk. To mitigate the effect of potentially using low particle numbers, we distribute the flux as described in Section 4.3.1.3, which is an imperfect solution.

It is recommended for any user to make a balanced trade off between resolution and computational expense.

4.5.6.2 Computational Expense

The main drawback of this methodology is that of computational expense required to run such an MCMC over a the full parameter space. In the case of this work, the simulation was set up with 2500 particles on a High End Computer Cluster with 50 CPUS. Each simulation took approximately two seconds, with the full sample run taking approximately 30 days to complete. Each galaxy was given 600 walkers to move through parameter space with each walker moving 7,500 steps. The highest memory requirement of any system was 6GB. Therefore, the runtime is very high but the memory required for it is approximately 122MB per core.

This methodology was successful in constraining 51 different interacting galaxy synthetic images. Achieving the same in prior work projects, such as Galaxy Zoo: Mergers, took months to complete. We have reduced the runtime required of it to 30 days on a powerful High End Computer Cluster. However, this is not the solution to the large scale dataset problems that we will be seeing when LSST comes online. If we are to run this methodology on a much larger galaxy sample - such as thousands of galaxies - then we will need to find ways to significantly improve the runtime of this method.

4.6 CONCLUSIONS & FUTURE WORK

In this work, we have introduced a new algorithm based on a MCMC framework to apply constraints on the underlying parameters of interacting galaxies. We have made these constraints, and explored the underlying parameter space, through directly comparing input images to output simulations. Thus, we have mapped the underlying parameters that define interaction to an output morphology that we compare for similarity with a synthetic or observation image of an interaction. We have introduced an updated version of the restricted three body simulation JSPAM and modified it to calculate and match the flux distribution of interacting galaxies based on thirteen underlying parameters. This updated

algorithm, APySPAM, calculates the flux distribution by assigning particles with a spectral energy distribution and accounting for star formation throughout the interaction.

To test our MCMC, we applied it to a set of synthetic images of interacting systems created from the parameters found in the Galaxy Zoo: Merger project for a set of real interacting systems. We were able to recover the true parameters used to create these images, with associated error measurements on each. We have explored the specific results of a single synthetic image: that of the Arp 240 system, and shown the corner plots and full posteriors of each parameter. We have took the 50th percentile value as the best fit value for our synthetic dataset, and used the 16th and 84th percentile to define the errors on our measurements. We then presented the distribution of our recovered underlying parameters and the true values, describing the general trends in our methodology as well as its limitations.

Every parameter of our thirteen was recoverable, however there was significant degeneracy in the orientation angles of the two galactic disks. This was expected as works such as Smith et al. (2010) found significant degeneracy between systems, but not a direct cause. In this work we found, without taking account of kinematic information, multiple orientation angles can recreate observed tidal features as well as limitations in the pixel matching method of constraint. When tested on observational data of our best fit synthetic systems, we were able to retain the excellent constraints on nine of the thirteen parameters, but lost any constraint on the orientations. However, we were able to maintain a tight constraint on the masses of the two systems.

There are limitations to this method that any users must be aware of. First, the core simulation of this process is a restricted three-body code with a highly optimised flux distribution calculation and approximation. It is probable that there are systems that this approach can not constrain, or may give nonphysical results for. The required reduction in the parameter space of the images, such as reducing the degrees of freedom and limiting individual pixel resolution in favour of computation time, also may lead to constraints only on the disks of the galaxies and not the tidal features themselves. There is also limited time and spatial resolution within the simulation itself. When using true observations, they

had to be artificially scaled up so our simulation could maintain the resolution to have constraint. In terms of the temporal and spatial parameters, this scale up can be accounted for. However, in terms of the masses of the flux distribution, this can be significantly hindered. Therefore, a user must be careful when choosing the image scales to use when attempting to constrain individual systems.

We have demonstrated the power of this methodology, and that it is capable of recovering the underlying parameters of interaction in idealised examples. The next, serious test of this MCMC approach will be to use it on a set of observed interacting systems, where the underlying parameters are known to properly diagnose the results. However, completely constrained interacting galaxy samples are only of a small scale and few exist. We will apply our MCMC to the observed dataset of our 51 interacting systems in SDSS.

This is an initial step towards developing the methodology to serve the field as large scale surveys - such as LSST - come online. Large scale automation where we can constrain interaction, make diagnostics and estimates about parameter space are not only important for inferences about individual datasets but for the interacting galaxy population as a whole. The main limitation of this method is the computational expense associated with it. To run this on 51 systems, the total run time was thirty days; an unusable timescale when being applied to interacting galaxy samples in the thousands. Therefore, future works with this algorithm and methodology will also be focused on increased computational efficiency and working on larger interacting galaxy datasets.

The bottleneck for computational efficiency lies in the time spent running the APySPAM simulation. While incredibly cheap individually, this is exceptionally expensive when having to be run in an MCMC chain. With the growing power of GPUs, and their ability to run numerical simulations significantly more efficiently than CPUs, a pathway to solve this limitation may already exist. This, combined with further developments in methodologies such as simulation based inference, machine learning and gaussian processes could begin to reduce computation time.

Chapter 5

SUMMARY AND FUTURE WORK

5.1 SUMMARY

This work has investigated the role of galaxy interaction with respect to galaxy evolution, shown a new and novel way of creating large interacting galaxy samples and presented the initial pilot of a pipeline to definitively link the parameters of galaxies to the tidal features that form and their underlying processes. The novel process of creating a large catalogue of interacting galaxies was detailed in Chapter 2. By using newly developed data-access architecture with the newly developed Bayesian CNN, `Zoobot`, the largest catalogue of interacting galaxies to date was created. Along with this, we demonstrated how the new data-access architecture, ESA Datalabs, can be used to explore archival observations like never before. To make concrete links between galaxy evolution and interaction in the local volume, this catalogue was cross-matched to catalogues of ancillary data. This was done using the catalogues created from deep observations of the COSMOS field. This provided us with a sample of 4,135 interacting galaxies to explore.

In Chapter 4 we conducted this exploration of our cross-matched sample in the context of interaction stage. We investigated whether the often conflicting theo-

ries about galaxy interaction were a result of not accounting for interaction stage based on observed morphology. We confirmed that interaction stage does have a significant impact on the underlying processes and enhancements that result from interaction. It was immediately found that the SFR increases dramatically through stage - from close pair to merger. This was demonstrated by measuring the distributions of stellar mass and SFR through stage and comparing them. We found that for distributions of identical mass, the distribution of SFR changes completely. This is in the form of the disappearence of the red sequence, as our samples SFR was enhanced. We compared this with existing works which utilise the projected separation between systems as an approximation of the stage of interaction in the two systems. We found a complete disconnect with the projected separation and stage, noting that conflicting results can emerge from only looking at the projected separation without properly accounting for the interaction stage from the morphology.

This degeneracy in the projected separation was particularily true of the change in the fraction of active galactic nuclei with stage, a topic the literature is particularily divided upon. Two modes of activation were found, one at stage 2 and one at stage 3. Thus, this was not only evidence for AGN flickering, but also that there is some delay in the activation of the AGN which is difficult to account for when only using projected separation. While this methodology shows how we can use ancillary data to connect the underlying parameters of galaxies to interaction, and fundamental processes we attempted to conduct this in a more general way and to bring in linking to tidal feature formation.

Chapter 4 saw the development of an algorithm to find the underlying parameters the galaxys' involved in interaction purely from their flux distribution. We combined a fast, efficient restricted numerical simulation with a MCMC methodology and Bayesian statistics and placed constraints on the underlying parameters of 38 simulated interacting systems. While our uncertainties increased when applied to observational data, constraints were nonetheless able to be made upon them. The creation of this algorithm, APySPAM, paves the way to apply this to large interacting galaxy datasets like the one created in chapter 2. However, the limitations in the runtime of the algorithm; taking approximately 15-20 hours

per interacting galaxy, leads this to not be feasible. Therefore, methods of further improvement in efficiency of the underling numerical simulation must be explored. There is particular promise in this regard with the development of numerical simulations on GPUs, and the massive acceleration they provide to such projects.

We have created a large interacting galaxy dataset and demonstrated its capability in the context of how the progression of a galaxy changes the underlying properties of galaxies. We have introduced a method by which further examination of these systems could be explored, although further advances in its efficiency must be made for this approach to be viable. Potential methods, plus descriptions of future works in Chapters 2 and 3 are discussed in the final subsection of this thesis.

5.2 FUTURE WORK

5.2.1 Catalogues of Galaxy Morphology with Ancillary Data

Chapter 2 demonstrated that new data access architecture is now ready for us to conduct source classification at a scale rarely seen to this point. We can now directly access millions of sources across multiple filters, instruments and observatories to yield unprecedented sample sizes. These can be combined with novel machine learning algorithms requiring small training set sizes to fine-tune making them versitile and accurate across many different observatories. Since the publication of the paper underlaying Chapter 2, the `Zoobot` algorithm has been updated many times. Not only is it now trained upon *HST* data, but part of the representations of galaxies it learns is if a galaxy is interacting or not. By redoing the work conducted in Chapter 2, one could classify galactic morphologies across the Galaxy Zoo workflow rather than just an individual question.

The fundamental component of this which makes it possible is the ESA Datalabs platform. This platform has also been updated to provide access to many different observatory's archives within it (such as JWST and Euclid) and also has

access to much larger storage spaces and GPUs. Therefore, the entire work of Chapter 2 can be done on the platform with no requirement for data transfer (in our case, data had to be moved from ESA Datalabs to the Lancaster computer cluster for classification: taking the bulk of the project time). ESA Datalabs now allows us to conduct to a project similar to the one outlined in Chapter 2 with much greater efficiency. Not only could this classification be applied for interacting vs non-interacting, but we can also begin to consider creating larger catalogues of much broader morpholgoy classifications. The recent Galaxy Zoo: DESI release contained 9.7 million galaxies with full morphology classification. Applying this to just the sources created in Chapter 2 would find 126 million morphology classifications - over a factor of 10 greater.

We can also take this further and not limit the opportunity here to only to only large catalogues of morphology classification. With the all-sky photometry that will be available from Euclid, or survey scale photometry that will be available from ground based observatories like the Legacy Survey of Space and Time telescope, it will be possible to get broad-band photometry for millions of overlapping sources. By applying well-known astrophysical software such as FAST, EAzY or LePhare, it will be able to estimate many galactic parameters. These include the stellar masses, the photometric redshifts or, in some cases, the precense of AGN. By creating such large morphology catalogues, combined with this ancillary data, we will be able to robustly link different galaxy parameters to their physical morphologies. This can then be expanded out to include linking to the source environments across the sky.

The development of algorithms like those proposed and tested in Chapter 4 will also provide excellent methods for constraining the underlying parameters of sources. While Chapter 4 focused on such an algorithm in the context of interacting galaxies, algorithms such as GALFIT are able to find morphological parameters purely from comparison to flux distributions and images. Even without further broad band photometry from other observatories, we will be able to explore the underlying parameter spaces of galaxy's and link this to their morphologies using these data access architectures.

5.2.2 Constraining Interacting Galaxy Parameters

In Chapter 4, we made direct links between the flux distribution of interacting galaxies and the parameters of those galaxies. We made constraints and revealed degeneracies across multiple parameters. However, the limitation of this approach was the computational expense: taking between 15-20 hours per system with the associated run time cost for this. For the large scale surveys or the catalogues such as those created in Chapter 2, this efficiency is not enough. Making constraints on the entire sample selected in Chapter 2 would take between 37 and 50 years. However, there are multiple development routes that could be taken in future to bosst computational efficiency signifcantly. The first is to take full advantage of the next generation of GPUs. Many cases of accelerating numerical simulations have been written about and analysed, particularily from the GPU developer NVIDIA. These have demonstrated striking improvements in efficiency up to a factor of 3, especially in the context of numerical models of fluid dynamics (recent examples include Mantas et al., 2016; Costa et al., 2021). Thus, reworking the code in this way would remove this limitation and have constraint being made in a matter of minutes: much more feasible for applying to large scale samples.

A second approach would be to move away from the direct method of MCMC and into utilising simulation based inference (SBI), sometimes called likelihood free inference. Much work has been conducted into SBI, and its massive increases in efficiency when constraining over large and complex parameter spaces (for an excellent describtion of likelihood free inference, see Jeffrey et al., 2021). In this context, rather than running an MCMC and directly comparing simulation outputs to an observational image, a machine learning algorithm trained on simulation outputs is used to explore the parameter space. By running an initial set of simulations and feeding this into a rudimentary CNN, a low dimensional representative vector can be created of each image. This is then applied to the observational image. The CNN then approximates the posterior through parameter space, achieving constraints comparable to a direct MCMC exploration. However, while this approach has been very successful for small parameter spaces defining 1D cases, applying to cases of images - especially one defined by a 13D parameter space presented here - is in its infancy.

Once this is evolved, however, it would succeed in SBI significant improvements in the efficiency of our constraining algorithms. We could take advantage of the amortisation of SBI, and be able to apply our initially trained network and posterior distribution to many different systems instead of having to explore parameter space directly every time. This has the effect of front-loading the computational expense of our constraining process, running thousands to millions of simulations across parameter space initially to train the neural network and then being able to use the model in multiple cases, albeit with some caveats. Amortisation of the trained model across parameter space would be very sensitive to many different fundamenal parts of our analysis process. For instance, changing the position of the secondary galaxy, the changing resolution of the images, the number of particles in the simulation, etc. Changing any of these would mean amortisation is not valid, and the model would have to be retrained - re-introducing the problem of computational efficiency.

However, if these cavearts can be resolved, this would open the way for large-scale application of SBI and our direct inferencing methodology to large samples. These will be our next steps in the development of this algorithm, and to then make it widely accessible to the community, with the application of our sample of interacting galaxies from Chapter 2. We will then be able to apply the process as conducted in Chapter 3, and link the effects of interaction to further underlying parameters than the stellar mass, interaction stage and SFR.

Appendix A

Model Diagnostics & Further Identified Objects

A.1 Further Model Diagnostics

In Section 2.6 we present diagnostic properties of our model. These include the accuracy measurements, purity measurements as well as confusion matrices at different cutoffs of our model. Here, we present the Receiver Operating Characteristic (ROC) curves, the precision-recall (PR) curves, and measures of true and false positive rates vs the cutoff threshold.

Figure A.1 shows the ROC and PR curves of the final `Zoobot` model we applied to the the *Hubble* archives. The ROC shows the rate of change of finding true positives and false positives with changing cutoff. The PR curve shows the changes of precision against recall. Precision is the ratio of true positives (interacting galaxies correctly predicted as so) to the sum of true and false positives (non-interacting galaxies incorrectly predicted as interacting). The recall is then the ratio of true positives to the sum of true positives and false negatives (interacting galaxies that have been misclassified as non-interacting). The red crosses in both plots shows how the model was behaving when we use a cutoff of 0.95.

These are both as expected. Both curves show that the model behaves well, and are much better than a random classifier (which would have a 1:1 relation).

The ROC plot shows that we are minimising our false positive rate when using a prediction score cutoff of 0.95. However, we are misclassifying approximately 50% of interacting galaxies as non-interacting galaxies. The contamination rate in our final catalogue (False Positives rate) will be very low (close to zero in this ideal validation set). The PR curve shows a similar result. Here, we are operating with a high precision (finding a pure catalogue) while keeping our recall minimal.

We also present the changing F1 score for the model used in this work, shown in Figure A.2. The F1 score is twice the ratio of precision multiplied by recall upon precision summed to recall. This combines our measure of accuracy and purity into a single metric. The cutoff we use in this work is at the point where the F1 score has began to decline. This is because we are beginning to lose recall rapidly, but gaining significantly in precision. As discussed in Section 2.6, this was an acceptable trade off in this work for a very large, pure interacting galaxy catalogue.

A.2 Examples of Sources with 3-Band Information

Of the full catalogue of 21,926 interacting systems, only 1336 of them had got all 3-band information. Six examples are shown in Figure A.3. These were created using the Lupton et al. (2004) algorithm, with a scaling factor $Q = 2$ and $\alpha = 0.75$, with ($F814W$, $F606W$, $F475W$) as RGB channels and multiplicative factors of (1.25, 0.95, 2).

A.3 Unknown Objects

From the final catalogue, there were six sources which we could not visually identify. These objects were also not referenced anywhere in the astrophysical literature. $F814W$ cutouts of the six objects are shown in Figure A.4. Their Source IDs are shown in the upper left of each image, and a separate catalogue

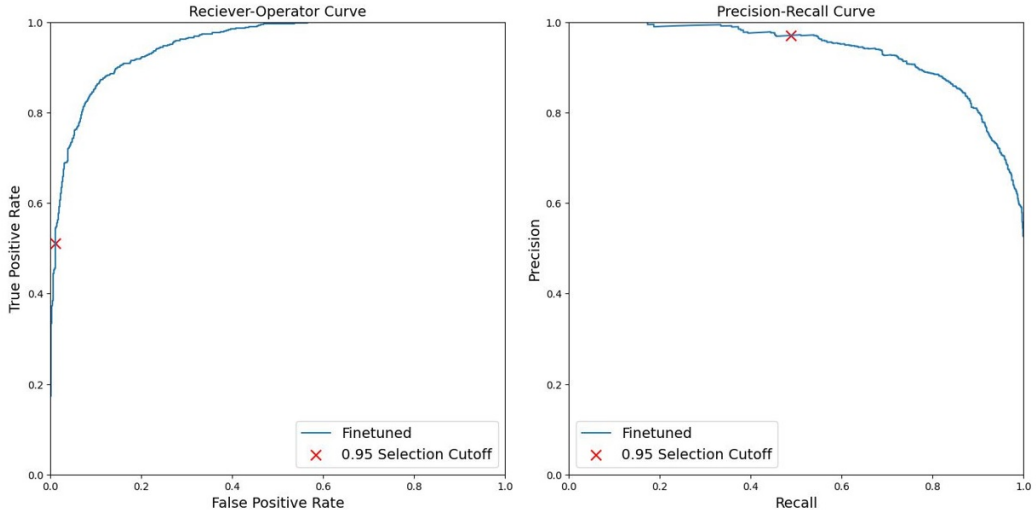

Figure A.1: The Receiver-Operator and Precision-Recall Curve for the `Zoobot` model that was used to explore the Hubble archives. The blue curves are the measured curves. These curves measure the relevant rates or characteristics based on the changing cutoff applied to how `Zoobot` defines an interacting galaxy. The red crosses are where the prediction score cutoff is for this work. We can see in the Reciever-Operator Curve that the prediction score cutoff we use would have an incredibly low false positive rate, while it would be misclassifying ≈50% of interacting galaxies. This also shown in the precision recall curve where our recall is ≈50%.

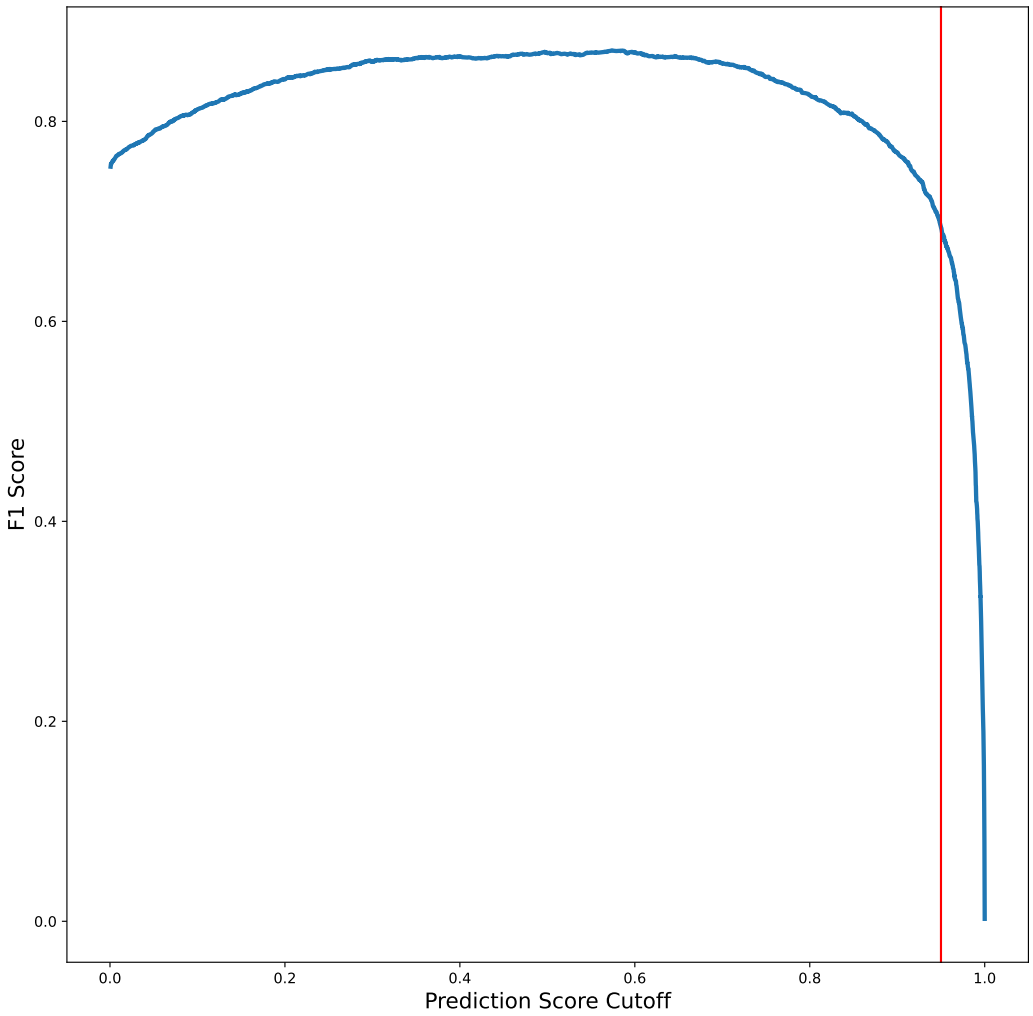

Figure A.2: The F1 score found during the diagnostics of the model used in this work. The F1 score is a measure combing the measure of accuracy and purity into one metric. The cutoff we use is at the point where the F1 score begins to rapidly decline. This point is shown by the red vertical line.

Figure A.3: Example of six interacting systems in the catalogue with full 3-band imagery.

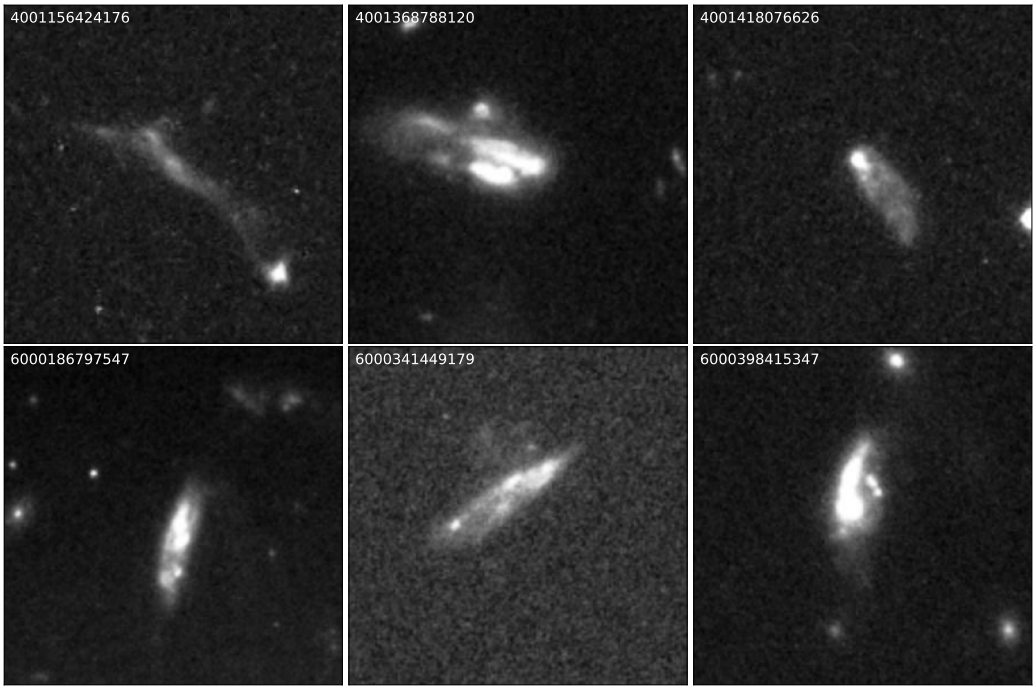

Figure A.4: The six unknown systems found in this work. These have no reference in Simbad or in NED, and their morphology could not be classified by the authors. Investigation into these six objects are presented to the community, with the authors hoping that future work and investigation of them can be conducted by them.

has been released of these with all other objects. This catalogue can be found at the data release on Zenodo.

Four of the six objects (40001156424176, 4001368788120, 4001418076626 and 6000398415347) have a bright central source, followed by a low-surface brightness tail. Initially, it was assumed that these were solar system objects such as comets. This, however, could not be confirmed. The first of these four sources is also thought to potentially be a highly disruped system with a significantly elongated tidal feature. The final two unknown sources (6000186797547 and 6000341449179) have no clear central source, though there is extended structure to them. These are likely to be highly irregular galaxies, but no confirmation could be found.

These objects are released to the community for identification and investigation, as the authors cannot find definitive agreement on what they are.

Proposal ID	Observation ID	Observation Date	DOI	references
8183	hst_8183_54_acs_wfc_f814w_j59l54	18/07/2002	https://doi.org/10.5270/esa-88k8vcj	
9075	hst_9075_2a_acs_wfc_f814w_j6fl2a	24/07/2002	https://doi.org/10.5270/esa-gsxhb4b	
9351	hst_9351_11_acs_wfc_f814w_j8d211	31/03/2003	https://doi.org/10.5270/esa-5lba8bo	
9361	hst_9361_03_acs_wfc_f814w_j8d503	22/07/2003	https://doi.org/10.5270/esa-ecmnqgh	
9363	hst_9363_09_acs_wfc_f814w_j8d809	02/07/2002	https://doi.org/10.5270/esa-ethtec5	
9367	hst_9367_02_acs_wfc_f814w_j8ds02	10/06/2003	https://doi.org/10.5270/esa-3j404ll	
9373	hst_9373_02_acs_wfc_f814w_j6la02	05/07/2002	https://doi.org/10.5270/esa-ztsq94u	Rejkuba et al. (2005)
9376	hst_9376_02_acs_wfc_f814w_j8e302	13/07/2002	https://doi.org/10.5270/esa-h90iavd	Keel et al. (2006)
9381	hst_9381_02_acs_wfc_f814w_j8fu02	13/03/2003	https://doi.org/10.5270/esa-vlapyea	
9400	hst_9400_04_acs_wfc_f814w_j6kx04	29/05/2003	https://doi.org/10.5270/esa-39rnout	
9403	hst_9403_02_acs_wfc_f814w_j8fp02	09/07/2002	https://doi.org/10.5270/esa-k5mv9ct	
9405	hst_9405_6k_acs_wfc_f814w_j8iy6k	22/05/2003	https://doi.org/10.5270/esa-zy9phm1	
9409	hst_9409_03_acs_wfc_f814w_j6n203	29/06/2003	https://doi.org/10.5270/esa-vjngw7r	Goudfrooij et al. (2004)
9411	hst_9411_09_acs_wfc_f814w_j8dl09	11/02/2003	https://doi.org/10.5270/esa-debpiln	
9427	hst_9427_13_acs_wfc_f814w_j6m613	21/10/2002	https://doi.org/10.5270/esa-bw1b97v	
9438	hst_9438_01_acs_wfc_f814w_j6me01	16/01/2003	https://doi.org/10.5270/esa-e5eaam5	Gregg & West (2017)
9450	hst_9450_02_acs_wfc_f814w_j8d402	25/08/2002	https://doi.org/10.5270/esa-9ttmykz	York et al. (2005)
9453	hst_9453_02_acs_wfc_f814w_j8f802	03/12/2002	https://doi.org/10.5270/esa-1xvyjfy	Brown et al. (2003)
9454	hst_9454_11_acs_wfc_f814w_j8ff11	23/03/2003	https://doi.org/10.5270/esa-xsdowj9	

Table A.1: Twenty example of the accompanying data table of observations used.

A.4 Acknowledging PIs

In the final section of this work, we wish to acknowledge all of the PIs whose observations we have used. A machine readable table containing the proposal IDs, the DOIs and the references (if provided/found) is presented with this work. Table A.1 shows the first twenty observations used in this work and is an example of this table.

References

Abadi M. G., Navarro J. F., Steinmetz M., Eke V. R., 2003, ApJ, 591, 499

Abadi M., et al., 2016, arXiv e-prints, p. arXiv:1605.08695

Abd El Aziz M., Selim I. M., Xiong S., 2017, Scientific Reports, 7, 4463

Ackermann S., Schawinski K., Zhang C., Weigel A. K., Turp M. D., 2018, MNRAS, 479, 415

Adams D., Mehta V., Dickinson H., Scarlata C., Fortson L., Kruk S., Simmons B., Lintott C., 2022, ApJ, 931, 16

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

Aird J., Coil A. L., Georgakakis A., 2019, MNRAS, 484, 4360

Alonso M. S., Tissera P. B., Coldwell G., Lambas D. G., 2004, MNRAS, 352, 1081

Alonso M. S., Lambas D. G., Tissera P., Coldwell G., 2007, MNRAS, 375, 1017

Ardizzone E., Di Gesù V., Maccarone M. C., 1996, Vistas in Astronomy, 40, 401

Arnouts S., Cristiani S., Moscardini L., Matarrese S., Lucchin F., Fontana A., Giallongo E., 1999, MNRAS, 310, 540

Arp H., 1966, ApJS, 14, 1

Arp H. C., Madore B., 1987, A catalogue of southern peculiar galaxies and associations

Astropy Collaboration et al., 2013a, A&A, 558, A33

Astropy Collaboration et al., 2013b, A&A, 558, A33

Astropy Collaboration et al., 2018a, AJ, 156, 123

Astropy Collaboration et al., 2018b, AJ, 156, 123

Atkinson A. M., Abraham R. G., Ferguson A. M. N., 2013, ApJ, 765, 28

Avila R. J., Hack W., Cara M., Borncamp D., Mack J., Smith L., Ubeda L., 2015, in Taylor A. R., Rosolowsky E., eds, Astronomical Society of the Pacific Conference Series Vol. 495, Astronomical Data Analysis Software an Systems XXIV (ADASS XXIV). p. 281 (`arXiv:1411.5605`), doi:10.48550/arXiv.1411.5605

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

Balogh M. L., Baldry I. K., Nichol R., Miller C., Bower R., Glazebrook K., 2004, ApJL, 615, L101

Barchi P. H., et al., 2020, Astronomy and Computing, 30, 100334

Barnes J. E., 1992, ApJ, 393, 484

Barnes J. E., Hernquist L. E., 1991, ApJL, 370, L65

Barnes J. E., Hibbard J. E., 2009, AJ, 137, 3071

Baron D., et al., 2018, MNRAS, 480, 3993

Barrera-Ballesteros J. K., et al., 2015, A&A, 579, A45

Barton Gillespie E., Geller M. J., Kenyon S. J., 2003, ApJ, 582, 668

Barton E. J., Geller M. J., Kenyon S. J., 2000, ApJ, 530, 660

Baugh C. M., Cole S., Frenk C. S., 1996, MNRAS, 283, 1361

Beckmann V., Shrader C. R., 2012, Active Galactic Nuclei

Bell E. F., et al., 2004, ApJ, 608, 752

Bell E. F., et al., 2006, ApJ, 640, 241

Bell E. F., Monachesi A., Harmsen B., de Jong R. S., Bailin J., Radburn-Smith D. J., D'Souza R., Holwerda B. W., 2017, ApJL, 837, L8

Bergvall N., Laurikainen E., Aalto S., 2003, A&A, 405, 31

Berrier J. C., Bullock J. S., Barton E. J., Guenther H. D., Zentner A. R., Wechsler R. H., 2006, ApJ, 652, 56

Bickley R. W., et al., 2021, MNRAS, 504, 372

Bluck A. F. L., Conselice C. J., Bouwens R. J., Daddi E., Dickinson M., Papovich C., Yan H., 2009, MNRAS, 394, L51

Blumenthal K. A., et al., 2020, MNRAS, 492, 2075

Bolatto A. D., et al., 2013, Nature, 499, 450

Boquien M., Burgarella D., Roehlly Y., Buat V., Ciesla L., Corre D., Inoue A. K., Salas H., 2019, A&A, 622, A103

Bottrell C., et al., 2019, MNRAS, 490, 5390

Bournaud F., Jog C. J., Combes F., 2005, A&A, 437, 69

Bournaud F., Jog C. J., Combes F., 2007, A&A, 476, 1179

Bower R. G., Lucey J. R., Ellis R. S., 1992, MNRAS, 254, 589

Brammer G. B., van Dokkum P. G., Coppi P., 2008, ApJ, 686, 1503

Brinchmann J., Charlot S., White S. D. M., Tremonti C., Kauffmann G., Heckman T., Brinkmann J., 2004, MNRAS, 351, 1151

Brown T. M., Ferguson H. C., Smith E., Kimble R. A., Sweigart A. V., Renzini A., Rich R. M., VandenBerg D. A., 2003, ApJL, 592, L17

Brown M. J. I., Dey A., Jannuzi B. T., Brand K., Benson A. J., Brodwin M., Croton D. J., Eisenhardt P. R., 2007, ApJ, 654, 858

Bruzual G., Charlot S., 2003, MNRAS, 344, 1000

Buck T., Wolf S., 2021, arXiv e-prints, p. arXiv:2111.01154

Bundy K., Ellis R. S., Conselice C. J., 2005, ApJ, 625, 621

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

Bushouse H. A., 1987, ApJ, 320, 49

Byrne-Mamahit S., Hani M. H., Ellison S. L., Quai S., Patton D. R., 2023, MNRAS, 519, 4966

Cameron E., 2011, PASA, 28, 128

Chabrier G., 2003, PASP, 115, 763

Chang Y.-Y., Lin L., Pan H.-A., Lin C.-A., Hsieh B.-C., Bottrell C., Wang P.-W., 2022, ApJ, 937, 97

Cheng T.-Y., Huertas-Company M., Conselice C. J., Aragón-Salamanca A., Robertson B. E., Ramachandra N., 2021, MNRAS, 503, 4446

Cheung E., et al., 2016, Nature, 533, 504

Cicone C., et al., 2014, A&A, 562, A21

Ćiprijanović A., Snyder G. F., Nord B., Peek J. E. G., 2020, Astronomy and Computing, 32, 100390

Comerford J. M., Pooley D., Barrows R. S., Greene J. E., Zakamska N. L., Madejski G. M., Cooper M. C., 2015, ApJ, 806, 219

Conselice C. J., Arnold J., 2009, MNRAS, 397, 208

Costa P., Phillips E., Brandt L., Fatica M., 2021, Computers & Mathematics with Applications, 81, 502

Cowan N. B., Ivezić Ž., 2008, ApJL, 674, L13

Cox T. J., Jonsson P., Somerville R. S., Primack J. R., Dekel A., 2008, MNRAS, 384, 386

Dalcanton J. J., et al., 2012, ApJS, 200, 18

Dalton G., et al., 2014, in Ramsay S. K., McLean I. S., Takami H., eds, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series Vol. 9147, Ground-based and Airborne Instrumentation for Astronomy V. p. 91470L (arXiv:1412.0843), doi:10.1117/12.2055132

Darg D. W., et al., 2010a, MNRAS, 401, 1043

Darg D. W., et al., 2010b, MNRAS, 401, 1552

Darvish B., Mobasher B., Sobral D., Scoville N., Aragon-Calvo M., 2015, ApJ, 805, 121

Darvish B., Mobasher B., Martin D. C., Sobral D., Scoville N., Stroe A., Hemmati S., Kartaltepe J., 2017, ApJ, 837, 16

Das A., Pandey B., Sarkar S., 2023a, Research in Astronomy and Astrophysics, 23, 095026

Das A., Pandey B., Sarkar S., 2023b, Research in Astronomy and Astrophysics, 23, 115018

De Lucia G., Blaizot J., 2007, MNRAS, 375, 2

De Lucia G., Springel V., White S. D. M., Croton D., Kauffmann G., 2006, MNRAS, 366, 499

Delvecchio I., et al., 2017, A&A, 602, A3

Dey A., et al., 2019, AJ, 157, 168

Di Matteo P., Combes F., Melchior A. L., Semelin B., 2007, A&A, 468, 61

Eisenstein D. J., et al., 2003, ApJ, 585, 694

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

Ellison S. L., Patton D. R., Mendel J. T., Scudder J. M., 2011, MNRAS, 418, 2043

Ellison S. L., Mendel J. T., Patton D. R., Scudder J. M., 2013, MNRAS, 435, 3627

Ellison S. L., Secrest N. J., Mendel J. T., Satyapal S., Simard L., 2017, MNRAS, 470, L49

Ellison S. L., et al., 2022, MNRAS, 517, L92

Elmegreen B. G., Elmegreen D. M., 2005, ApJ, 627, 632

Emsellem E., et al., 2011, MNRAS, 414, 888

Faber S. M., Gallagher J. S., 1976, ApJ, 204, 365

Fakhouri O., Ma C.-P., Boylan-Kolchin M., 2010, MNRAS, 406, 2267

Ferreira L., et al., 2022, ApJL, 938, L2

Foreman-Mackey D., 2016, The Journal of Open Source Software, 1, 24

Foreman-Mackey D., Hogg D. W., Lang D., Goodman J., 2013, PASP, 125, 306

Gabor J. M., Davé R., Finlator K., Oppenheimer B. D., 2010, MNRAS, 407, 749

Gadotti D. A., Kauffmann G., 2009, MNRAS, 399, 621

Gao F., et al., 2020, A&A, 637, A94

Garay-Solis Y., Barrera-Ballesteros J. K., Colombo D., Sánchez S. F., Lugo-Aranda A. Z., Villanueva V., Wong T., Bolatto A. D., 2023, ApJ, 952, 122

Geach J. E., et al., 2018, ApJL, 864, L1

Gebhardt K., et al., 2001, AJ, 122, 2469

Ghosh A., Urry C. M., Wang Z., Schawinski K., Turp D., Powell M. C., 2020, ApJ, 895, 112

Giavalisco M., et al., 2004, ApJL, 600, L93

Gillies S., et al., 2007, Shapely: manipulation and analysis of geometric objects, `https://github.com/Toblerity/Shapely`

Goldreich P., Lynden-Bell D., 1965, MNRAS, 130, 97

González-García A. C., Balcells M., 2005, MNRAS, 357, 753

Gordon Y. A., et al., 2017, MNRAS, 465, 2671

Goudfrooij P., Gilmore D., Whitmore B. C., Schweizer F., 2004, ApJL, 613, L121

Goulding A. D., et al., 2018, PASJ, 70, S37

Gregg M., West M., 2017, in Early stages of Galaxy Cluster Formation. p. 13, doi:10.5281/zenodo.831767

Guo Q., White S. D. M., 2008, MNRAS, 384, 2

Guo Y., et al., 2018, ApJ, 853, 108

Hani M. H., Gosain H., Ellison S. L., Patton D. R., Torrey P., 2020, MNRAS, 493, 3716

Harris C. R., et al., 2020, Nature, 585, 357

He H., Wilson C. D., Sliwa K., Iono D., Saito T., 2020, MNRAS, 496, 5243

He C., Xu C. K., Domingue D., Cao C., Huang J.-s., 2022, ApJS, 261, 34

Helou G., Madore B. F., Schmitz M., Bicay M. D., Wu X., Bennett J., 1991, in Albrecht M. A., Egret D., eds, Astrophysics and Space Science Library Vol. 171, Databases and On-line Data in Astronomy. pp 89–106, doi:10.1007/978-94-011-3250-3˙10

Hernández-Toledo H. M., Avila-Reese V., Conselice C. J., Puerari I., 2005, AJ, 129, 682

Hernández-Toledo H. M., Cortes-Suárez E., Vázquez-Mata J. A., Nevin R., Ávila-Reese V., Ibarra-Medel H., Negrete C. A., 2023, MNRAS, 523, 4164

Hernquist L., 1990, ApJ, 356, 359

Hinton S. R., 2016, The Journal of Open Source Software, 1, 00045

Hogg D. W., et al., 2002, AJ, 124, 646

Holincheck A. J., et al., 2016, MNRAS, 459, 720

Hopkins P. F., Hernquist L., Cox T. J., Kereš D., 2008, ApJS, 175, 356

Hopkins P. F., et al., 2009a, MNRAS, 397, 802

Hopkins P. F., Cox T. J., Younger J. D., Hernquist L., 2009b, ApJ, 691, 1168

Hopkins P. F., et al., 2010, ApJ, 715, 202

Hopkins P. F., Cox T. J., Hernquist L., Narayanan D., Hayward C. C., Murray N., 2013, MNRAS, 430, 1901

Hopkins P. F., et al., 2018, MNRAS, 480, 800

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

Hubble E. P., 1936, Realm of the Nebulae

Hunter J. D., 2007, Computing in Science and Engineering, 9, 90

Ilbert O., et al., 2006, A&A, 457, 841

Jackson R. A., Kaviraj S., Martin G., Devriendt J. E. G., Noakes-Kettel E. A., Silk J., Ogle P., Dubois Y., 2022, MNRAS, 511, 607

Jacobs C., et al., 2019, ApJS, 243, 17

Jeffrey N., Alsing J., Lanusse F., 2021, MNRAS, 501, 954

Jeon S., et al., 2022, ApJ, 941, 5

Johansson P. H., Naab T., Burkert A., 2009, ApJ, 690, 802

Karera P., Drissen L., Martel H., Iglesias-Páramo J., Vilchez J. M., Duc P.-A., Plana H., 2022, MNRAS, 514, 2769

Karman W., Macciò A. V., Kannan R., Moster B. P., Somerville R. S., 2015, MNRAS, 452, 2984

Kauffmann G., et al., 2003, MNRAS, 341, 33

Kaviraj S., 2014a, MNRAS, 437, L41

Kaviraj S., 2014b, MNRAS, 440, 2944

Kaviraj S., Devriendt J., Dubois Y., Slyz A., Welker C., Pichon C., Peirani S., Le Borgne D., 2015, MNRAS, 452, 2845

Keel W. C., White Raymond E. I., Owen F. N., Ledlow M. J., 2006, AJ, 132, 2233

Keel W. C., et al., 2022, AJ, 163, 150

Kennicutt Robert C. J., 1998, ApJ, 498, 541

Kingma D. P., Ba J., 2014, arXiv e-prints, p. arXiv:1412.6980

Knapen J. H., Cisternas M., 2015, ApJL, 807, L16

Knapen J. H., Cisternas M., Querejeta M., 2015, MNRAS, 454, 1742

Kolmogorov A., 1933, Giorn Dell'inst Ital Degli Att, 4, 89

Koss M., Mushotzky R., Treister E., Veilleux S., Vasudevan R., Trippe M., 2012, ApJL, 746, L22

Lacey C., Cole S., 1993, MNRAS, 262, 627

Lambas D. G., Tissera P. B., Alonso M. S., Coldwell G., 2003, MNRAS, 346, 1189

Li C., Kauffmann G., Heckman T. M., Jing Y. P., White S. D. M., 2008a, MNRAS, 385, 1903

Li C., Kauffmann G., Heckman T. M., White S. D. M., Jing Y. P., 2008b, MNRAS, 385, 1915

Li Y. A., Ho L. C., Shangguan J., 2023, ApJ, 953, 91

Lintott C. J., et al., 2008, MNRAS, 389, 1179

López-Sanjuan C., Balcells M., Pérez-González P. G., Barro G., García-Dabó C. E., Gallego J., Zamorano J., 2009, A&A, 501, 505

López-Sanjuan C., et al., 2013, A&A, 553, A78

Lotz J. M., Primack J., Madau P., 2004, AJ, 128, 163

Lotz J. M., Jonsson P., Cox T. J., Primack J. R., 2008a, MNRAS, 391, 1137

Lotz J. M., et al., 2008b, ApJ, 672, 177

Lotz J. M., Jonsson P., Cox T. J., Croton D., Primack J. R., Somerville R. S., Stewart K., 2011, ApJ, 742, 103

Lupton R., Blanton M. R., Fekete G., Hogg D. W., O'Mullane W., Szalay A., Wherry N., 2004, PASP, 116, 133

Lynds R., Toomre A., 1976, ApJ, 209, 382

Malin D. F., Carter D., 1983, ApJ, 274, 534

Mantas J. M., De la Asunción M., Castro M. J., 2016, An Introduction to GPU Computing for Numerical Simulation. Springer International Publishing, Cham, pp 219–251, doi:10.1007/978-3-319-32146-2˙5, https://doi.org/10.1007/978-3-319-32146-2_5

Marchesi S., et al., 2016, ApJ, 817, 34

Marian V., et al., 2020, ApJ, 904, 79

Martig M., Bournaud F., 2008, MNRAS, 385, L38

Martig M., Bournaud F., Croton D. J., Dekel A., Teyssier R., 2012, ApJ, 756, 26

Martin G., Kaviraj S., Devriendt J. E. G., Dubois Y., Laigle C., Pichon C., 2017, MNRAS, 472, L50

Masters K. L., et al., 2010, MNRAS, 405, 783

McInnes L., Healy J., Melville J., 2018, arXiv e-prints, p. arXiv:1802.03426

McKernan B., Ford K. E. S., Reynolds C. S., 2010, MNRAS, 407, 2399

McKinney W., 2010, `https://conference.scipy.org/proceedings/scipy2010/pdfs/mckinney.pdf`

Merín B., et al., 2017, arXiv e-prints, p. arXiv:1712.04114

Mihos J. C., Hernquist L., 1994, ApJL, 425, L13

Mihos J. C., Hernquist L., 1996, ApJ, 464, 641

Mihos J. C., Richstone D. O., Bothun G. D., 1992, ApJ, 400, 153

Miller T. B., van Dokkum P., 2021, ApJ, 923, 124

Morales-Vargas A., et al., 2020, MNRAS, 499, 4370

Moreno J., Bluck A. F. L., Ellison S. L., Patton D. R., Torrey P., Moster B. P., 2013, MNRAS, 436, 1765

Moreno J., Torrey P., Ellison S. L., Patton D. R., Bluck A. F. L., Bansal G., Hernquist L., 2015, MNRAS, 448, 1107

Moreno J., et al., 2019, MNRAS, 485, 1320

Moreno J., et al., 2021, MNRAS, 503, 3113

Morganti R., Garrett M. A., Chapman S., Baan W., Helou G., Soifer T., 2004, A&A, 424, 371

Murata K. L., et al., 2014, ApJ, 786, 15

Nair P. B., Abraham R. G., 2010, ApJS, 186, 427

Nelson B., Ford E. B., Payne M. J., 2014, ApJS, 210, 11

Netzer H., 2015, ARA&A, 53, 365

Nevin R., Blecha L., Comerford J., Simon J., Terrazas B. A., Barrows R. S., Vázquez-Mata J. A., 2023, MNRAS, 522, 1

Nielsen F., 2016, Hierarchical Clustering. pp 195–211, doi:10.1007/978-3-319-21903-5˙8

O'Ryan D., et al., 2023, ApJ, 948, 40

O'Shea K., Nash R., 2015, arXiv e-prints, p. arXiv:1511.08458

Ochsenbein F., Bauer P., Marcout J., 2000, A&AS, 143, 23

Pacifici C., Kassin S. A., Weiner B., Charlot S., Gardner J. P., 2013, ApJL, 762, L15

Patton D. R., Ellison S. L., Simard L., McConnachie A. W., Mendel J. T., 2011, MNRAS, 412, 591

Patton D. R., Torrey P., Ellison S. L., Mendel J. T., Scudder J. M., 2013, MNRAS, 433, L59

Patton D. R., et al., 2020, MNRAS, 494, 4969

Pawlik M. M., et al., 2018, MNRAS, 477, 1708

Pearson W. J., Wang L., Trayford J. W., Petrillo C. E., van der Tak F. F. S., 2019a, A&A, 626, A49

Pearson W. J., et al., 2019b, A&A, 631, A51

Pearson W. J., et al., 2022, A&A, 661, A52

Pedregosa F., et al., 2012, arXiv e-prints, p. arXiv:1201.0490

Peng C. Y., Ho L. C., Impey C. D., Rix H.-W., 2002,] 10.1086/340952

Peng C. Y., Ho L. C., Impey C. D., Rix H.-W., 2010a, AJ, 139, 2097

Peng Y.-j., et al., 2010b, ApJ, 721, 193

Petsch H. P., Theis C., 2008, Astronomische Nachrichten, 329, 1046

Pierce C. M., et al., 2007, ApJL, 660, L19

Quai S., Hani M. H., Ellison S. L., Patton D. R., Woo J., 2021, MNRAS, 504, 1888

Quinn P. J., 1984, ApJ, 279, 596

Quirk W. J., 1972, ApJL, 176, L9

Rejkuba M., Greggio L., Harris W. E., Harris G. L. H., Peng E. W., 2005, ApJ, 631, 262

Ren J., Li N., Liu F. S., Cui Q., Fu M., Zheng X. Z., 2023, ApJ, 958, 96

Renaud F., Segovia Otero Á., Agertz O., 2022, MNRAS, 516, 4922

Rodriguez-Gomez V., et al., 2015, MNRAS, 449, 49

Rodríguez Montero F., Davé R., Wild V., Anglés-Alcázar D., Narayanan D., 2019, MNRAS, 490, 2139

Rogers B., Ferreras I., Kaviraj S., Pasquali A., Sarzi M., 2009, MNRAS, 399, 2172

Ross D., Lim J., Lin R., et al. 2008, Int J Comput Vis, p. 125–141

Saitoh T. R., Daisaka H., Kokubo E., Makino J., Okamoto T., Tomisaka K., Wada K., Yoshida N., 2009, PASJ, 61, 481

Salim S., et al., 2007, ApJS, 173, 267

Salo H., Laurikainen E., 1993, ApJ, 410, 586

Sánchez S. F., et al., 2004, ApJ, 614, 586

Satyapal S., Ellison S. L., McAlpine W., Hickox R. C., Patton D. R., Mendel J. T., 2014, MNRAS, 441, 1297

Schawinski K., et al., 2009, MNRAS, 396, 818

Schawinski K., et al., 2014, MNRAS, 440, 889

Schawinski K., Koss M., Berney S., Sartori L. F., 2015, MNRAS, 451, 2517

Schaye J., et al., 2015, MNRAS, 446, 521

Schiminovich D., et al., 2007, ApJS, 173, 315

Schmidt M., 1959, ApJ, 129, 243

Schweizer F., Seitzer P., 1992, AJ, 104, 1039

Scott C., Kaviraj S., 2014, MNRAS, 437, 2137

Scoville N., et al., 2007, ApJS, 172, 1

Scudder J. M., Ellison S. L., Torrey P., Patton D. R., Mendel J. T., 2012, MNRAS, 426, 549

Shah E. A., et al., 2020, ApJ, 904, 107

Shah E. A., et al., 2022, ApJ, 940, 4

Silva A., Marchesini D., Silverman J. D., Martis N., Iono D., Espada D., Skelton R., 2021, ApJ, 909, 124

Simha V., Weinberg D. H., Conroy C., Dave R., Fardal M., Katz N., Oppenheimer B. D., 2014, arXiv e-prints, p. arXiv:1404.0402

Simmons B. D., et al., 2017, MNRAS, 464, 4420

Smethurst R. J., et al., 2015, MNRAS, 450, 435

Smethurst R. J., et al., 2018, MNRAS, 473, 2679

Smethurst R. J., et al., 2022, MNRAS, 510, 4126

Smith B. J., et al., 2010, in Smith B., Higdon J., Higdon S., Bastian N., eds, Astronomical Society of the Pacific Conference Series Vol. 423, Galaxy Wars: Stellar Populations and Star Formation in Interacting Galaxies. p. 227 (arXiv:0908.3478), doi:10.48550/arXiv.0908.3478

Smolčić V., et al., 2017, A&A, 602, A6

Sparre M., Whittingham J., Damle M., Hani M. H., Richter P., Ellison S. L., Pfrommer C., Vogelsberger M., 2022, MNRAS, 509, 2720

Springel V., 2000, MNRAS, 312, 859

Springel V., White S. D. M., Tormen G., Kauffmann G., 2001, MNRAS, 328, 726

Springel V., et al., 2005, Nature, 435, 629

Steffen J. L., et al., 2023, ApJ, 942, 107

Stemo A., Comerford J. M., Barrows R. S., Stern D., Assef R. J., Griffith R. L., Schechter A., 2021, ApJ, 923, 36

Stephens M. A., 1974, Journal of the American Statistical Association, 69, 730

Strateva I., et al., 2001, AJ, 122, 1861

Suelves L. E., Pearson W. J., Pollo A., 2023, A&A, 669, A141

Tanaka M., et al., 2023, PASJ, 75, 986

Theys J. C., Spiegel E. A., 1977, ApJ, 212, 616

Toomre A., 1977, in Tinsley B. M., Larson Richard B. Gehret D. C., eds, Evolution of Galaxies and Stellar Populations. p. 401

Toomre A., Toomre J., 1972, ApJ, 178, 623

Vavilova I., Elyiv A., Dobrycheva D., Melnyk O., 2021, in Zelinka I., Brescia M., Baron D., eds, , Vol. 39, Intelligent Astrophysics. pp 57–79, doi:10.1007/978-3-030-65867-0˙3

Villforth C., et al., 2014, MNRAS, 439, 3342

Violino G., Ellison S. L., Sargent M., Coppin K. E. K., Scudder J. M., Mendel T. J., Saintonge A., 2018, MNRAS, 476, 2591

Vorontsov-Velyaminov B. A., 1959, Atlas and Catalog of Interacting Galaxies, p. 0

Vorontsov-Velyaminov B. A., 1977, A&AS, 28, 1

Wallin J. F., 1990, The Astronomical Journal, 100, 1477

Wallin J. F., Holincheck A. J., Harvey A., 2016, Astronomy and Computing, 16, 26

Walmsley M., et al., 2022a, MNRAS, 509, 3966

Walmsley M., et al., 2022b, MNRAS, 513, 1581

Walmsley M., et al., 2023, The Journal of Open Source Software, 8, 5312

Weaver J. R., et al., 2022, ApJS, 258, 11

Weigel A. K., et al., 2017, ApJ, 845, 145

Wenger M., et al., 2000, A&AS, 143, 9

West G., Ogden M., Wallin J. F., 2023, Astronomy and Computing, 42, 100691

Whitaker K. E., Kriek M., van Dokkum P. G., Bezanson R., Brammer G., Franx M., Labbé I., 2012, ApJ, 745, 179

White S. D. M., 1978, MNRAS, 184, 185

White S. D. M., Frenk C. S., 1991, ApJ, 379, 52

White S. D. M., Rees M. J., 1978, MNRAS, 183, 341

Whitmore B. C., et al., 2016, AJ, 151, 134

Willett K. W., et al., 2013, MNRAS, 435, 2835

Willett K. W., et al., 2017, MNRAS, 464, 4176

Woods D. F., Geller M. J., Barton E. J., 2006, AJ, 132, 197

Xu C. K., Zhao Y., Scoville N., Capak P., Drory N., Gao Y., 2012, ApJ, 747, 85

Xu C. K., Lisenfeld U., Gao Y., Renaud F., 2021, ApJ, 918, 55

Yoon Y., Park C., Chung H., Lane R. R., 2022, ApJ, 925, 168

York T., Jackson N., Browne I. W. A., Wucknitz O., Skelton J. E., 2005, MNRAS, 357, 124

Zhong Y., Inoue A. K., Yamanaka S., Yamada T., 2022, ApJ, 925, 157

de Mello D. F., Infante L., Menanteau F., 1997, ApJS, 108, 99

ter Braak C. J. F., Vrugt J. A., 2008, Statistics and Computing, 18

van Dokkum P. G., 2005, AJ, 130, 2647

van Dokkum P., et al., 2018, Nature, 555, 629

van Dokkum P., et al., 2019, ApJL, 883, L32

van der Walt S., et al., 2014, PeerJ, 2, e453