

The Relation Between Galaxy Evolution and Galaxy Interaction

David Patrick O’Ryan



Physics

Department of Physics
Lancaster University

Date

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

Abstract

Dedication

Acknowledgements

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.

Interacting galaxies are smashing things.

Contents

List of Figures	ix
1 Introduction	1
1.1 WHAT IS A GALAXY?	3
1.2 GALAXY ASSEMBLY ACROSS COSMOLOGICAL TIME	4
1.3 GALAXY MORPHOLOGY & PROPERTIES	7
1.4 CATEGORISATION OF MERGERS	9
1.5 IDENTIFYING INTERACTING AND MERGING GALAXIES	9
1.6 EFFECTS OF GALAXY INTERACTION	9
1.6.1 Morphological Distortion: Tidal Features	11
1.6.2 Star Formation Enhancement	13
1.6.3 Nuclear Activation	13
1.6.4 Quenching of Galactic Systems	13
1.6.5 The Role of Environment	13
2 Creating a Large Sample of Interacting Galaxies	14
2.1 Abstract	14
2.2 INTRODUCTION	15
2.3 DATA	18
2.3.1 The <i>Hubble</i> Archives & ESA Datalabs	18
2.3.2 The Shapely Python Package	19
2.4 UTILISING A CONVOLUTIONAL NEURAL NETWORK	20
2.4.1 Zoobot	20
2.4.2 Transfer Learning	22

2.5	CREATING THE TRAINING SET	23
2.5.1	Interacting Galaxies and Galaxy Zoo	23
2.5.2	One Active Learning Cycle	25
2.6	DIAGNOSTICS	28
2.6.1	Model Performance	28
2.6.2	Duplication Removal	32
2.6.3	Bad Predictions & Removal	36
2.7	RESULTS & DISCUSSION	40
2.7.1	An Interacting Galaxy Catalogue	40
2.7.2	The Gems	43
2.7.3	Source Redshifts and Photometry	45
2.8	CONCLUSION	53
3	When do the effects of interaction happen? Interacting and disturbed galaxies in the COSMOS field	57
3.1	Abstract	57
3.2	Introduction	57
3.3	DATA: Catalogue Matching & Secondary Identification	61
3.3.1	The O’Ryan+23 Catalogue	61
3.3.2	The COSMOS2020 Catalogue	62
3.3.3	Secondary Identification	63
3.3.4	Finding Additional Systems	68
3.3.5	Redshift Effects on Visual Classification	68
3.4	METHOD: Environment, AGN and Interaction Stage	69
3.4.1	Classifying Stage of Interaction	70
3.4.2	Matching to Environment Catalogue	75
3.4.3	Classifying AGN	76
3.5	STAR FORMATION EVOLUTION WITH INTERACTION STAGE	77
3.5.1	Controlling for Interaction Stage	77
3.5.2	Projected Separation and Star Formation Enhancement . .	84
3.5.3	Controlling for Galactic Environment	87
3.6	Nuclear Activity with Interaction Stage	89
3.7	Discussion	94

3.7.1	Interaction Stage and Projected Separation	94
3.7.2	Interaction Stage and AGN	97
3.8	Conclusions	98
4	Galaxy Zoo: Mergers & Constraining Interaction	102
4.1	Abstract	102
4.2	INTRODUCTION	103
4.3	DATA	106
4.3.1	Galaxy Zoo: Mergers sample	106
4.3.2	Galaxy Zoo Mergers: Constraining Interaction	109
4.3.3	Observation Preparation	111
4.4	METHODS	112
4.4.1	Simulations: APySPAM	112
4.4.1.1	Stellar Population Evolution	113
4.4.1.2	Star Formation	115
4.4.2	Flux Distribution	117
4.4.2.1	Impact on Computation Time	118
4.4.3	Defining the Likelihood Function	120
4.4.3.1	Bayes Theorem	120
4.4.3.2	Simplifying the Prior	121
4.4.3.3	Simplifying the Likelihood Function	122
4.4.4	Architecture for Exploring Parameter Space: EMCEE	124
4.5	RESULTS & DISCUSSION	125
4.5.1	Testing on a Single System: Arp 240	125
4.5.2	Diagnostics of Pipeline	132
4.5.3	Inputting 3D Information	135
4.5.4	Recovering the Galaxy Zoo: Mergers Results	137
4.5.5	Applying to Observations	140
4.5.6	Limitations	143
4.5.6.1	Resolution & Depth Effects	143
4.5.6.2	Computational Expense	145
4.6	CONCLUSIONS & FUTURE WORK	146

5 Conclusion	148
5.1 SOFTWARE FOR STATISTICAL CONSTRAINT	148
5.2 CREATING LARGE SAMPLES WITH MACHINE LEARNING	148
5.3 APPLYING LARGE CATALOGUES: PROSPECTS	148
5.4 APPLYING LARGE CATALOGUES: CURRENT LIMITATIONS	148
5.5 FUTURE WORK	148
Appendix A MCMC Best Fit Values	149
A.1 Galaxy Zoo: Mergers Best Fit Parameters	149
Appendix B Model Diagnostics & Further Identified Objects	150
B.1 Further Model Diagnostics	150
B.2 Examples of Sources with 3-Band Information	151
B.3 Unknown Objects	151
B.4 Acknowledging PIs	156
References	157

List of Figures

- | | | |
|-----|---|---|
| 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. | 6 |
| 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. | 8 |

LIST OF FIGURES

LIST OF FIGURES

LIST OF FIGURES

LIST OF FIGURES

- 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. 38
- 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. 39

LIST OF FIGURES

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 fea- tures 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 <i>F814W HST</i> filter.	41
2.10 Sky Distribution of our catalogue, with marked positions of well known deep surveys conducted by <i>HST</i> . <i>HST</i> 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.	43
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 primar- ily trained on systems at these redshifts. There are fifteen sources with a reported $z > 5$.	48
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 dif- ferent surveys this data is from.	49

- 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. 52
- 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. There is a large difference between the measures of LePhare and EAzY here, and therefore we opt to only use EAzY measures of the SFR through this work. 63

LIST OF FIGURES

- 3.2 An example of each visual classification made on the cross matched sample. From left to right these are the secondary could not be identified, the primary had a clear secondary, the primary could not be reliably classified as an interacting galaxy, the redshift was null and the incorrect primary was identified in OR+23. 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. 65

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. 66

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. 67

3.5 Redshift vs Mass distribution for each stage of interaction we have defined. This is important as we use tidal distortion and the existence as tidal features as a fundamental for our classification methodology. As shown here, the distribution of systems across redshift and mass is consistent for all stages in our sample. Therefore, we are likely not affected by this in our analysis. 69

LIST OF FIGURES

3.6 Examples of the four stages we split our interacting galaxy sample into. Stage 1: A close pair with confirmed redshift matching. Stage 2: Two distinct systems interacting with tidal features forming. Stage 3: A tidally disturbed system with no secondary present, likely at apocentre. Stage 4: A galaxy with multiple cores while highly disturbed. At the final stage before coalescence.	71
3.7 The projected separations of the confirmed galaxy pairs in our sample. This confers with other works the definition of our different stages. Stage 1 can be at any projected separation, however, we have visually confirmed that these sources are not morphologically disturbed. Stage 2 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 stage 2. Finally, stage 3 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 stage 2 and stage 3 in projected separation, as their visual classification is also dependent on the system size.	72

LIST OF FIGURES

3.8	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. <i>Top left</i> : The stellar mass and star formation rate of Stage 1 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. <i>Top right</i> : Stage 2 of the interaction, where the two interactors are close to pericentre. The star forming sequence remains, but the red sequence is reduced significantly. <i>Bottom left</i> : Stage 3 of the interaction, where the interactors are close to apocentre or escaped. Here, we see the almost complete disappearance of the red sequence. <i>Bottom Right</i> : Stage 4 of the interaction, where the two systems are close to coalescence or have merged. The red sequence of galaxies has completely disappeared.	78
3.9	The stellar mass distribution across the four stages. Each bin is weighted based on the counts in the smallest sub-sample in stage. This is stage 4 of interaction.	80
3.10	SFR distribution weighted by mass across each stage. This weighting is based on our sample of stage 4.	80
3.11	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 3.8. 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.	83

LIST OF FIGURES

LIST OF FIGURES

3.15 The distribution of environment classifications through our sample. This is only for sources with a $\log M_* / M_\odot \geq 9.6$. However, from this subsample, we see that there is no trend with environment and that the distribution is random throughout each stage. Therefore, the effects of increasing SFR with interaction stage is not due to environmental effects.	89
3.16 The distribution of AGN through stage with SFR and stellar mass. We do not find any major changes and evolution with AGN and SFG position through interaction stage.	91
3.17 The change in AGN fraction with stage. 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 point are found by bootstrapping. We do not include any sources of which we have no information. We find a sudden drop in AGN fraction from stage 1 to stage 2, followed by a rapid increase from stage 3 to 4.	91
3.18 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 interesting structure within the AGN sample. Here, there appears to be a double peak in the number of counts of AGN. The first peak, between a projected separation of 0 and 25kpc is primarily from stage 2 galaxies while the second peak of 40 to 65kpc is of stage 3 galaxies in our sample. As shown in Figure 3.17, we expect the fraction of AGN to be similar in these two stages while they are at very different times in the dynamical timescale of the interaction. The third peak, from 75 to 125kpc is only from stage 1 and stage 3 galaxies.	93
4.1 Cutouts from SDSS of the sixty two interacting systems that we test our pipeline on. For the best fit simulations see Figure 6 of Holincheck et al. (2016a).	108

4.2	The example system used to test our pipeline: 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 pipeline can fit. These tidal features are the two tidal tails formed in the interaction and the tidal bridge linking the two systems. <i>Left:</i> The prepared observation image of the Arp 240 system created from SDSS DR16 observations. <i>Right:</i> The best fit simulation image as found by Holincheck et al. (2016a) and the first test image used in our pipeline. The different in scale and orientation are discussed below.	123
4.3	Corner plot showing the constraints made on all thirteen parameters we are exploring. Contours show the 0.5, 1, 1.5 and 2-sigma levels of the constraint. this corresponds to containing 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 find the full corner plots of each system, find them at the results website for this paper.	127
4.4	Same as Figure 4.3, but reduced to only matching parameters. . .	128
4.5	Simulations from the areas of parameter space that lay within the 0.5σ of our constraints. <i>Top:</i> The best fit simulations from this parameter space. <i>Bottom:</i> The worst fit simulations from this parameter space. Our pipeline has 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 wildly different tidal features in the areas of high probability. Therefore, identifying specific systems with the sought after tidal features remains needing to be a manual process.	131

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 parameter 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, z-velocity and ϕ_2 parameters show significant 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.	134
4.7 The constraints on the 3D velocity and z-position when including velocity in our MCMC pipeline. To add this extra information, we ran our best fit simulation of Arp 240 and summed the LOS velocity (z-velocity) in each pixel of our image. Comparing between here and Figure 4.5, we can see that we achieve complete constraint of the z-velocity of the interacting galaxy. We also significantly improve the constraint on the z-position parameter as the pipeline is able to distinguish which side the secondary galaxy is of the primary.	136
4.8 Mass corner plots for our best and worst fits from the pipeline. This was judged by the FWHM of the marginalised posteriors. <i>Top:</i> Our top three best fits. <i>Bottom:</i> Our worst fits. Even with our worst fits here, the masses are very well constrained.	137

LIST OF FIGURES

4.9	Reduced corner plot of the constraints made on the observational image of Arp 240. As shown, there is significantly more 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, although is less obvious in the two θ parameters. We are also unable to make conclusive constraints on the velocity parameters with the observations.	142
B.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 \approx 50% of interacting galaxies. This also shown in the precision recall curve where our recall is \approx 50%.	152
B.2	The F1 score found during the diagnostics of the model used in this work. The F1 score is a measure combining 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.	153
B.3	Example of six interacting systems in the catalogue with full 3-band imagery.	154
B.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.	155

Relevant Publications by the Author

Chapter 2

- Publication

Chapter 1

Introduction

In the context of large scale structure of the universe, galaxies represent a fundamental building block of matter. Through cosmic time galaxies evolve (Springel et al., 2005a), this evolution leading to the local universe as we see it today. A key part galaxy evolution is that of mutual interactions between them. Mutual interaction has a number of influences on galaxies, from morphological distortion, to mass transfer, to increases in star formation, to merging where the two (or more) galaxies coalesce into a single system. Our best cosmological model, Λ -Cold Dark Matter, dictates that matter - ergo, galaxies - assembled hierarchically (White & Frenk, 1991; White & Rees, 1978). 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 disks involved and the formation of distinct tidal features. Since the existence of such features being related to interaction was hypothesised to proved (Toomre & Toomre, 1972a) they are the best morphological indication that two galaxies are interacting. Often, we attempt to only classify interacting galaxies from their morphology. To truly know if two or more systems are interacting, we require redshift information to ascertain their 3D distances from each other. However, spectroscopic information across the many million close galaxy pairs we know about is limited.

Mutual interaction has further effects on galaxies than simply morphological disturbance, although these remain poorly understood. It is widely debated about the role of interaction in nuclear ignition and enhancement in star formation of each system. While the former may not happen at all, the latter is generally accepted to occur while the specific strength and impact on the system remains highly debated. If such processes occur strongly, they can completely change the system that we are observing. Interaction has also been proved to have an impact on the gas reserves in a galaxy, often leading to the complete quenching of the system post-interaction. However, interaction has also been shown to rejuvenate the galaxies involved, and bring them out of a quenched state and into a star forming state. This is highly dependent on the specifics of the interaction. Finally, the role of environment in galaxy interaction is also a highly researched area.

The primary focus of this thesis is on the role of tidal distortion in formation of tidal features for the purposes of system identification and constraint. However, in this process, we must also keep an account of the other underlying processes occurring during interaction. Fully understanding them, and modelling them if necessary, is key to accurately inferring and linking galactic parameters to physical processes. The list of unresolved questions about galaxy interaction is long and complex. The focus of this thesis is in providing the tools to understand galaxy interaction and answer some of the fundamental questions regarding it. These questions include:

- What do we require to improve our understanding of interaction?
- How do we better identify interacting galaxies?
- What is the relationship between numerous galactic processes and galactic parameters?
- From existing samples and studies, how do we better constrain our understanding on the effects of interaction?

To fully answer these questions about galaxy interaction, we require three things: first, the tools to find and create large statistically significant samples 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 to better constrain these relationships. In this thesis, detail two pipelines which achieve these goals. First, in chapter 2, I will describe a pipeline for creating the largest interacting galaxy catalogue to date from combining morphological classification with novel methods of data extraction and analysis. I will then use this catalogue, in chapter 3, with cross matches to existing ancillary catalogues and make my own inferences about the relationship between galactic parameters and underlying processes. However, this will also show the limitations of only using observations and the requirements for constraint by simulations. Finally, in chapter 5, I describe a pipeline that will conduct such constraint in the context of Bayesian statistics and galactic parameterisation. I 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. However, for now, it would be prudent to begin this discussion by exploring the definition of a galaxy, and outlining the major work already conducted into linking galaxy interaction 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 be considering. A galaxy is an extended system of stars, gas and dust which all orbits around a supermassive black hole at its centre. All of this baryonic matter makes up the luminous, or visible, matter of the galaxy. This matter has two distinct components: the galactic bulge and the galactic disk. The galactic bulge lies at the centre of the system. It is an approximately spherical component which extends out to about a third of the radius of the galaxy. Galaxies will always have a bulge component, although it might be so small that the galaxy appears to only have a disk. The bulge is often made of an older population of stars, with high peculiar motions in their orbit. The galactic disk is then a very thin, extended region which orbits around the galactic bulge. This is where the majority of the galaxy's gas and dust reside. There are many younger stars throughout it as well,

and this extends out to a radii many times that of the galactic bulge. Both of these are embedded in a dark matter halo.

The dark matter halo is many times the radius of the luminous components of the galaxy. This halo is also many times the mass of baryonic components of the galaxy, and makes up the majority of the mass of a galaxy. We approximate each of these components with different potential models. For the galactic bulge, we utilise a Plummer sphere (Plummer, 1911). For the galactic disk, we use an exponential disk model. For the dark matter halo, we utilise a cored Navarro-Frenk-White model (Navarro et al., 1997).

For much of the history of extragalactic astrophysics, the idea that galaxies could interact and merge was thought to be very improbable. The small radii of the luminous matter in galaxies, and their calculated cross sections, led astronomers to believe that the probability of an interaction was close to 0 and that galaxies were, simply, island universes (Hubble, 1926). However, with the development of the idea of the dark matter halo, with its radius many times that of the luminous matter, it was realised that the probability of galactic systems interacting and merging was within the realm of possibility. And, therefore, 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 one of the most well studied concepts 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 to form the universe that we see today. Studies at high redshifts reveal that galaxy structure is significantly different when compared to today. Initially, galaxies were very small systems that formed from gravitational instability throughout the early cosmos (Lacey & Cole, 1993). These instabilities were generated by the physical processes of the very early universe brought about by the small scale

1.2 GALAXY ASSEMBLY ACROSS COSMOLOGICAL TIME

of the universe. From these initial density perturbations arose small systems which accreted from the gas and dust throughout the early universe. These early galactic systems were much smaller than those we see today. Their morphology was also very peculiar, and they were forming stars at a much higher rate than galaxies in the local universe. While these systems were accreting much of their matter to gain mass, the universe was expanding. At this time, with the universe at a much smaller scale, the rate of galaxies interacting and merging was much higher than today (Hopkins et al., 2010; Lotz et al., 2011). 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. From $z = 1$, the once massive galaxies, with rapid star formation, have transformed into bulge-dominated galaxies containing super massive black holes (Brown et al., 2007). In fact, as we go to lower redshifts, we find that the majority of the stellar mass is contained within these systems. These galaxies are dominated by old stellar components and have little ongoing star formation (Bell et al., 2004; Hogg et al., 2002). Thus, showing that mergers were the primary drivers of star formation in the past.

The morphology we see now, that formed out of this period of very intense merging and interaction, was first officially defined in Hubble (1936). The famous ‘Hubble tuning fork’, was once thought of as a map of galaxy evolution, is now known to be one of galaxy morphological classification. Figure 1.2 shows this breakdown of morphological classification, from spheroidal systems on the left to disk spiral or disk barred galaxies shown on the right. What was once thought of as a layout of galaxies with age, we now understand as an indication of the evolutionary history of individual systems. From our description here of merger rates, we realise that the systems on the left of the tuning fork are actually those with intense merger histories and that are far older than those on the right. These are spheroidal systems, also called early-type galaxies contain little gas, have almost no star formation and appear very red in our colour-colour charts. Those systems on the right are slightly younger than the early-type systems but

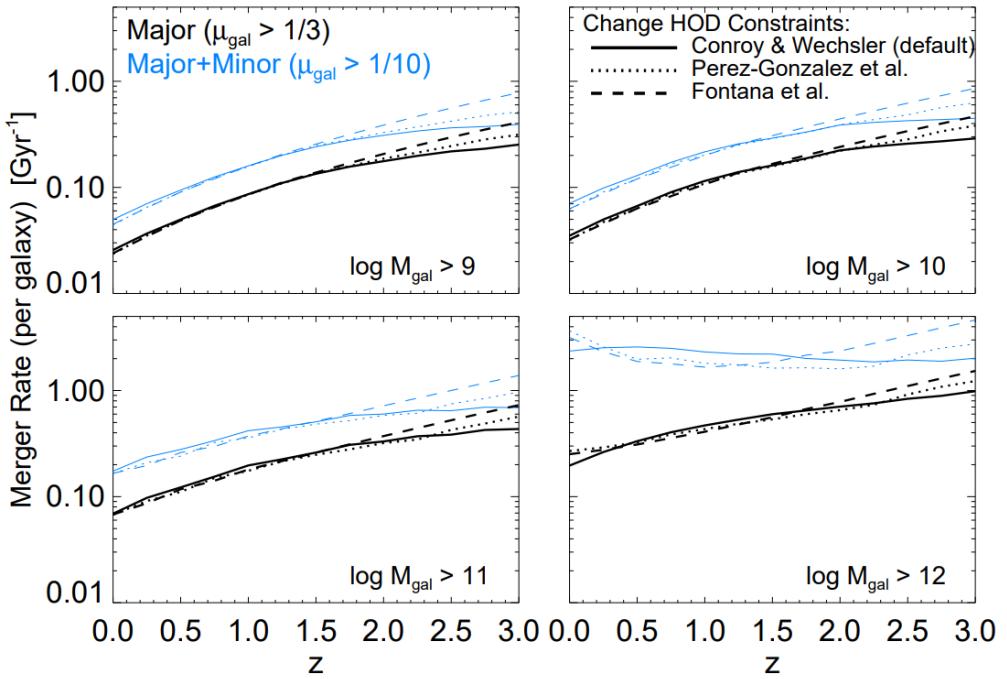


Figure 1.1: The clearly declining merger rates from $z = 3$ to $z = 0$. This is Figure 3 of Hopkins et al. (2010). This study looked at the history of a simulated set of galaxies of various stellar mass, and investigated the decrease in the merger fraction as a function of redshift and baryonic mass. 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.

have a less intense merger history, especially in the last few Gyrs (). While they could have This leads them to with plentiful gas, lots of star formation and a blue appearance.

Thus, as the merger rate has decreased through cosmic time, the survival of disk galaxies has increased. We, therefore, see two distinct populations of galaxies which can be easily defined by their morphology and colour which directly relates to their own evolutionary and merger history. One is of galaxies with a history of intense interactions with galaxies of equivalent mass, which lead to the complete destruction of their disks, and the other is of galaxies with a less intense merger history, with systems of much smaller mass than their own. These interactions and mergers serve to feed the gas within the galactic disks and increase their mass, while preserving the disk component of the system. In this thesis, I will not focus on merger rates any further than already explored. However, this whole discussion feeds into how we best identify mergers and interacting galaxies. I will, therefore, briefly discuss the morphology of galaxies and how it links into their properties. This will put things into perspective before I discuss how these systems are completely disrupted by merging and interaction.

1.3 GALAXY MORPHOLOGY & PROPERTIES

Galaxy morphology, and its change with time, specifically is the study and measure of a galaxies shape and features so that we can make different assumptions about it. There are many examples of morphological studies of galaxies. One of the largest to date, which was also a citizen science project, was that of the Galaxy Zoo collaboration first published in Lintott et al. (2008b).

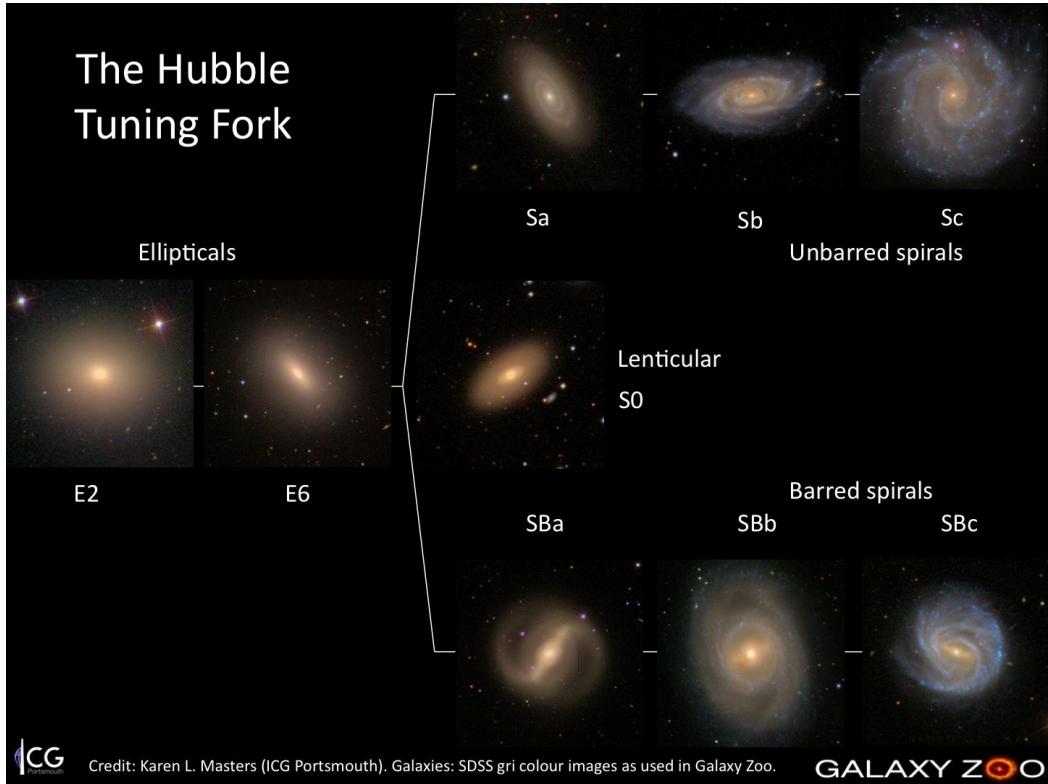


Figure 1.2: The Hubble tuning fork. Credit goes to the Galaxy Zoo collaboration (and specifically Karen Masters) for the creation of this figure. What was once thought of as an evolutionary track for galaxies, is now a widely adopted classification system for them based on morphology. On the left, we 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.

1.4 CATEGORISATION OF MERGERS

1.5 IDENTIFYING INTERACTING AND MERGING GALAXIES

1.6 EFFECTS OF GALAXY INTERACTION

The idea of two or more galaxies interacting and the effect that it would have had upon galaxy formation and evolution was thought to be minute (?). This assumption was not entirely without merit for the time. By calculating the chance of a galactic encounter from the galactic cross section measured via the galactic radius of stellar material and comparing it to the vast distances between galaxies, it seemed highly improbable that two systems would collide. However, with updated cosmological theories and the introduction of a model where the stellar component of galaxy is embedded in a dark matter halo (?), the size of galaxies and their cross section increased. This dramatically increased the probability of two galaxies interacting in the universe, and with the theory that many of the tidal features found were formed by tidal interaction, a new field of study was born.

We now understand that galaxy interaction plays an important, and fundamental role in the evolution of galaxies. As stated previously, they have multiple affects upon the systems undergoing the interaction. While the specifics of these effects lies heavily on the underlying parameters and type of interaction that is occurring, the driving force underlying them all is gravitational. Therefore, the mass of the systems involved is of great importance to the resultant system we might observe. If the two systems that interact are at vastly different masses - a so-called minor interaction - then the less massive galaxy will be highly perturbed, whilst the more massive galaxy will be relatively intact. If, however, the two systems are of a relatively similar mass both systems will be highly perturbed to the point where they may even be destroyed - a so-called major interaction. A much more difficult type of interaction to observe is a micro-interaction. This is where one system has an almost insignificant mass when compared to the system

1.6 EFFECTS OF GALAXY INTERACTION

it is interacting with. In this scenario, the least massive system is always almost entirely destroyed by the more massive system. Specific criteria are often placed upon the mass ratios - the mass of one galaxy when compared to the other - to define these three categories of interaction. For a major interaction the mass ratio must be close to 1:1, in a minor interaction it must be close to 1:10 and in a micro interaction it must be less than 1:10 (?).

However, it is also known that other aspects of the interaction also influence the outcome of the final system. Another that has a direct impact on the output morphology of the system is the velocities and orientations involved in the interaction (??). A slower interaction with a higher impact parameter (i.e. one where the point of closest approach is small) leads to a more violent interaction with highly perturbed morphology. These slower interactions often lead to the two or more systems becoming intertwined, with multiple passages and merging being the final state of the system. This often occurs due to dynamical friction in the interaction, where heating of the two disks involved and transfer of momentum in the inelastic interaction leads to the two systems spiralling into one another. Again, such an outcome is directly related to the masses of the systems involved as well as their relative velocities and their orientations.

Many underlying physical processes can also be started or enhanced due to interaction, and the effects of gravity upon the galaxies involved. If the systems involved are gas-rich - a so-called wet merger - then the star formation in the involved systems can be significantly enhanced. It has often been found that galaxies in an interaction or in the post-interaction state are starbursting (?), with their star formation rates many times what would be expected of a galaxy of their mass. This increase in star formation then leads to quenching of the system. In the same way, a system that is already quenched can undergo rejuvenation of its gas, and begin star forming after a wet merger. The increased movement in gas throughout the system can also lead to the activation of nuclear activity in the core, and completely alter what we observe of a galaxy. On the other hand, in a dry merger, no increase in star formation is expected and the discussed effects are often suppressed.

So, with these underlying processes in mind, how do they actually occur? What is happening within galaxies that leads to interaction having such a pro-

found effect upon their observed parameters. To understand this, we must go back to first principles and explore how these processes would normally occur and then extrapolate this out the extreme conditions of interaction. I will explore how these conditions relate directly to the underlying parameters of interaction, and the observational effects each result has on a galactic system.

1.6.1 Morphological Distortion: Tidal Features

The clearest example of two galactic systems interacting is the existence of tidal features. These tidal features are results of severe morphological distortion due to tidal forces an interaction. An excellent example which has been very highly studied is NGC 4038 / 4039, also known as the Antennae galaxy. This system is a major interacting pair, which are likely continuing into a merger, with long tidal tails extending far from the interacting cores. The first works attempting to simulate the outcomes of galaxy interaction focused on this system (?). The full extent of the tidal tails system was not revealed until deeper observational data became available in 2004 (?), and then could be accounted for in the simulators modelling.

This is a common problem with information on the tidal features which form in interaction. They often lie in the low surface brightness regime, at the very deepest depths that our telescopes can observe. This often makes investigating any system but the most massive, brightest systems difficult. A classic example of the limitations of our telescopes has been the case of NGC 5907, where the morphology of the stellar stream about it (likely created by a minor interaction) has been under significant scrutiny in recent years (??). With the changing morphology of the stellar stream, so to does our understanding of the dynamics of the system as well as estimates of the progenitor galaxy that formed the stellar stream.

Thus, understanding the dynamics at play in an interaction, and the underlying physics of the different tidal features which form, is of paramount importance to understanding galaxy interaction. Gravity is the driving force of the changing morphology of two systems in an interacting system. Galaxies are composed of

1.6 EFFECTS OF GALAXY INTERACTION

three components: a stellar component, a gas component and a dark matter component. In simulations, the stellar and dark matter components are treated as collisionless. Focusing on the stellar component, this is not an unfair assumption when considering the very large distances involved in an interaction to the cross section of a typical stars within each galaxy. Due to this difference, the probability of stars actually colliding is actually very small. However, the distribution of stars is often radically altered by interaction. The gravitational field of each system perturbs the ordered orbits of the stars within each system, transferring energy into them, and adding random motion to stellar orbits (?). This is an inelastic collision where as an interaction progresses, and more energy is transferred into the ordered motion of the stars, that the two galaxies merge into a single system. This final result is called a merger remnant. However, if the two systems have the required velocity, they will escape each others potential well and simply move past each other.

This transfer of energy, and disruption of the stellar disk, due to gravity is called dynamical friction. Dynamical friction acts to decelerate the stars in both galaxies, and act as a frictional drag. This force is proportional to the mass squared of the galactic system producing the perturbation. Thus, the density of stars ahead of the moving galaxy becomes less than the density behind it as the stars are decelerated from their original ordered motion. This drag acts to strip the outer regions of the galaxy of stars into tidal bridges between the systems, followed by becoming large tidal tails behind the galaxy. If the two galaxies have the velocity to escape each others potentials, these tidal tails can often coalesce themselves to form tidal dwarfs orbiting the galaxies originally involved in the interaction (?). The tidal tails can also fall onto the secondary galaxy in the interaction, and act as mass transfer between the two systems.

Thus, while the stellar component of the galaxy is often a simple calculation of the forces on each star by gravity, the same is not true of the gas component. The gas component is made up of multiple much larger clouds of gas within each galaxy that can be approximated as a fluid. If two gas clouds from different galaxies were to collide with each other - a more probable outcome than with two stars - then the shocks driven through the clouds would cause it to very quickly fragment and the formation of stars to occur (??). These newly formed stars

would then heat and ionise the surrounding gas, shocking it again when they go supernova and beginning the process all over again. The motion and position of the gas clouds within the galaxy itself will also be highly affected. Depending on the orientation of such a collision, or the external torques upon the gas cloud from the influence of the two gravitational potentials, the gas cloud may lose so much angular momentum it begins to move into the inner regions of the galaxy (?). This increase in the galactic core would then begin the process of star formation again.

This process, by which the star formation in interacting galaxies increases dramatically compared to non-interacting counterparts, is called a starburst and interaction is just one mechanism by which it can occur. A starburst is defined as when the galaxy is forming a mass of stars per year which is significantly higher when compared to what is expected from a galaxy of its mass. This is observed to occur in some, but not all, interactions (?). So, we will now discuss the fundamental process of star formation, and how the above described process can lead to a significant enhancement in star formation during an interaction or merger.

1.6.2 Star Formation Enhancement

The star formation rate (SFR) of a galaxy is governed, primarily, by its size and the total mass of the gas within it. This is not completely unexpected, as the dominant influence on star formation is the gas density. An excellent way to estimate the SFR in a galaxy is by the Kennicutt-Schmidt Law (?)

1.6.3 Nuclear Activation

1.6.4 Quenching of Galactic Systems

1.6.5 The Role of Environment

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 Λ CDM cosmology, in which structure typically assembles hierarchically (Abadi et al., 2003; De Lucia & Blaizot, 2007; Guo & White, 2008; Springel et al., 2005b). Galaxy interaction leads to highly disturbed morphologies (Hernández-Toledo et al., 2005; Toomre & Toomre, 1972b; Wallin et al., 2016a), intense starbursts (Mihos & Hernquist, 1996a; Moreno et al., 2021a; Saitoh et al., 2009a; Springel, 2000a) and, potentially, quenching of some systems (Das et al., 2022; Hani et al., 2020a; Hopkins et al., 2013a; Smethurst et al., 2018). In general, galaxies undergoing interaction are observed to have higher star formation rates than those that exist in the field (Ellison et al., 2008a; Pearson et al., 2019b; Scudder et al., 2012a). 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 their nuclear regions and igniting activity. This could be a connection with active galactic nuclei (Comerford et al., 2015a; Ellison et al., 2008a, 2011a; Li et al., 2008c). However, such a connection remains debated (Alonso et al., 2007a; Marian et al., 2020; McKernan et al., 2010). 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, 2014a,b), and examining interactions as a function of their local environments, internal properties and AGN activity (Darg et al., 2010c). These studies (and many others; for further examples, see Alonso et al., 2004; Barton et al., 2000a; Ellison et al., 2013a; Holincheck et al., 2016b; Silva et al., 2021a) 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, 1966a; Vorontsov-Velyaminov, 1959, 1977, hereafter VV). These catalogues primarily used visual inspection to identify mergers (e.g., Nair & Abraham, 2010; de Mello et al., 1997) 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. (2010a) 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 (Abd El Aziz et al., 2017; Ardizzone et al., 1996; Barchi et al., 2020; Cheng et al., 2021; Ghosh et al., 2020). 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; Bickley et al., 2021; Buck & Wolf, 2021; Jacobs et al., 2019; Walmsley et al., 2022a). Pearson et al. (2022a) 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. (2019a); Pearson et al. (2022a)).

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 Archive¹ 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., 2016a). 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 3.3 describes the HSC and all the criteria we applied to create the images we predict over. This Section also introduces ESA Datalabs²; 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 3.8 summarises our results and conclusions.

Where necessary, we use a Flat Λ CDM cosmology with $H_0 = 70 \text{ km/s/Mpc}$ and $\Omega_M = 0.3$. Hereafter in this paper, 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 tidal disturbance. A merging galaxy is the final

¹See <http://hst.esac.esa.int/ehst/>

²<https://datalabs.esa.int/>

state of these flybys, where two or more systems have coalesced to form a highly morphologically irregular system.

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., 2016a, 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., 2014) 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 \approx 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

2.4 UTILISING A CONVOLUTIONAL NEURAL NETWORK

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 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 $6\mu\text{s}$.

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; ?). `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

¹Shapely docs: <https://shapely.readthedocs.io/en/stable/manual.html>

2.4 UTILISING A CONVOLUTIONAL NEURAL NETWORK

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 $300 \times 300 \times 1$ 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 summaries 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 **Zoobots**' 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. (2022c). 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. (2022c), we only need to create a training set of $10^3 - 10^4$ interacting galaxies to achieve an accuracy of $\approx 90\%$.

2.5 CREATING THE TRAINING SET

We create a large training set of interacting galaxies following the criteria described in Section 3.3 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. (2022c), 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 Section 2.6.1. To create this large training set we use the Galaxy Zoo collaboration (initial data release described in Lintott et al., 2008a).

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., 2008a), 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, 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 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. (2022a), 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

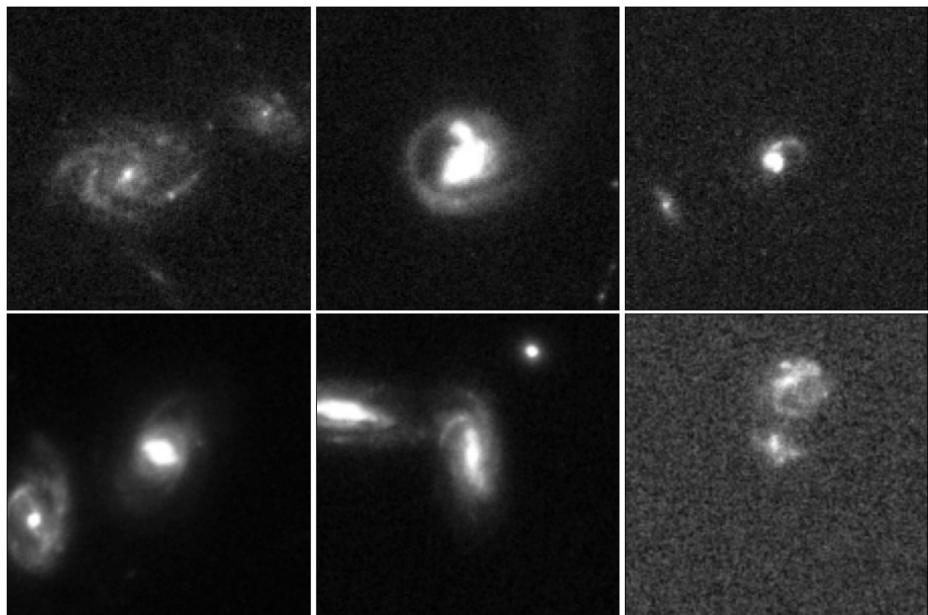


Figure 2.1: Example images of the labelled interacting galaxy systems used to train Zoobot. Each galaxy had a weighted vote fraction ≥ 0.75 in Galaxy Zoo. *Top Row:* Three examples from the Galaxy Zoo: *Hubble* project of the training set. *Bottom Row:* Three examples from the other Galaxy Zoo projects. 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.

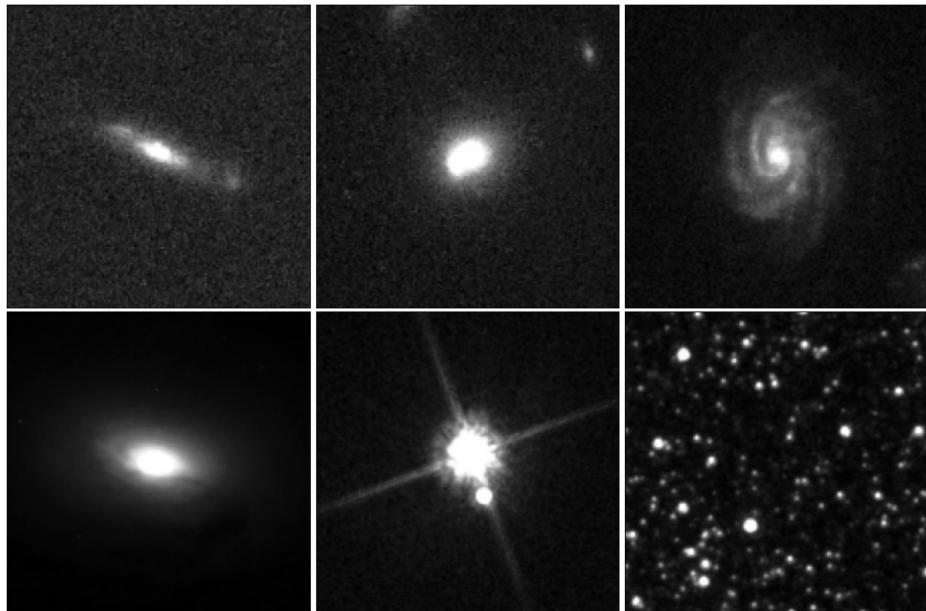


Figure 2.2: Example images of the labelled non-interacting galaxy systems used to train Zoobot. *Top Row*: Three examples from the Galaxy Zoo: *Hubble* project of the training set. *Bottom Row*: Three examples from the other Galaxy Zoo projects. These are, from right to left, Galaxy Zoo 2, Galaxy Zoo 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.

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 ≥ 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-modal 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 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 B.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

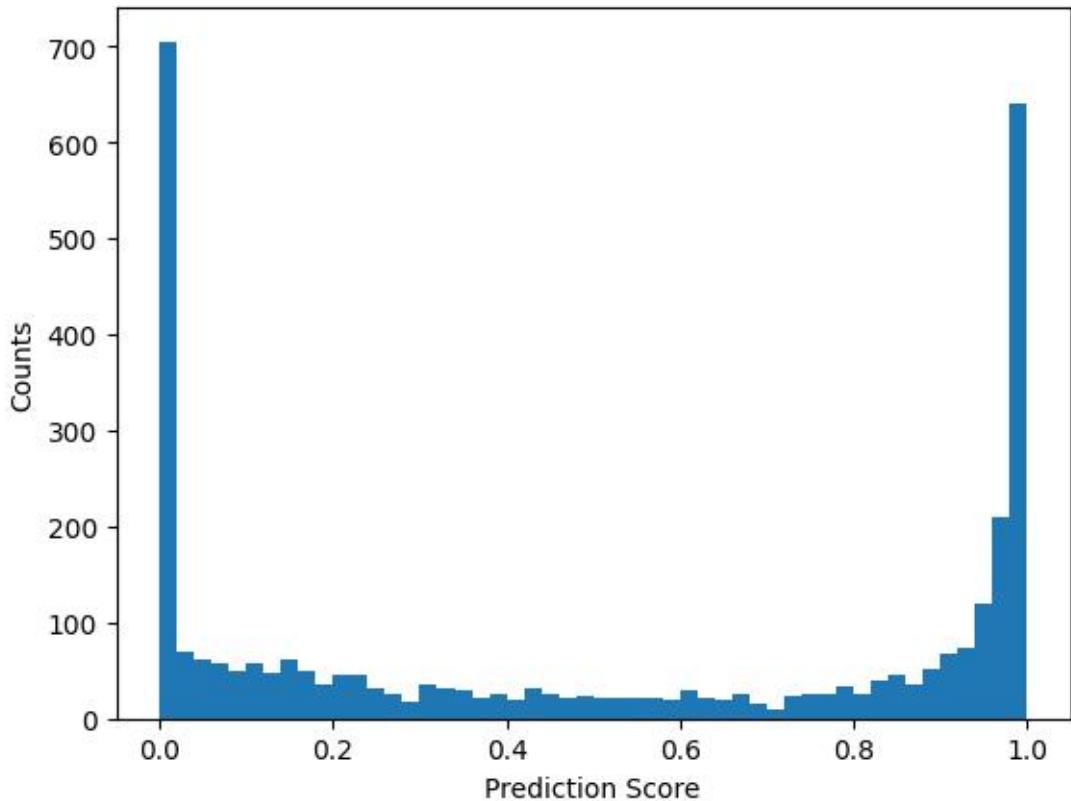


Figure 2.3: The distribution of prediction scores given to our validation set of 3,270 labelled sources set aside by Zoobot in training. These were split into 1,648 non-interacting sources and 1,622 interacting sources. As can be seen from the distribution, our model is often confident when a source 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.

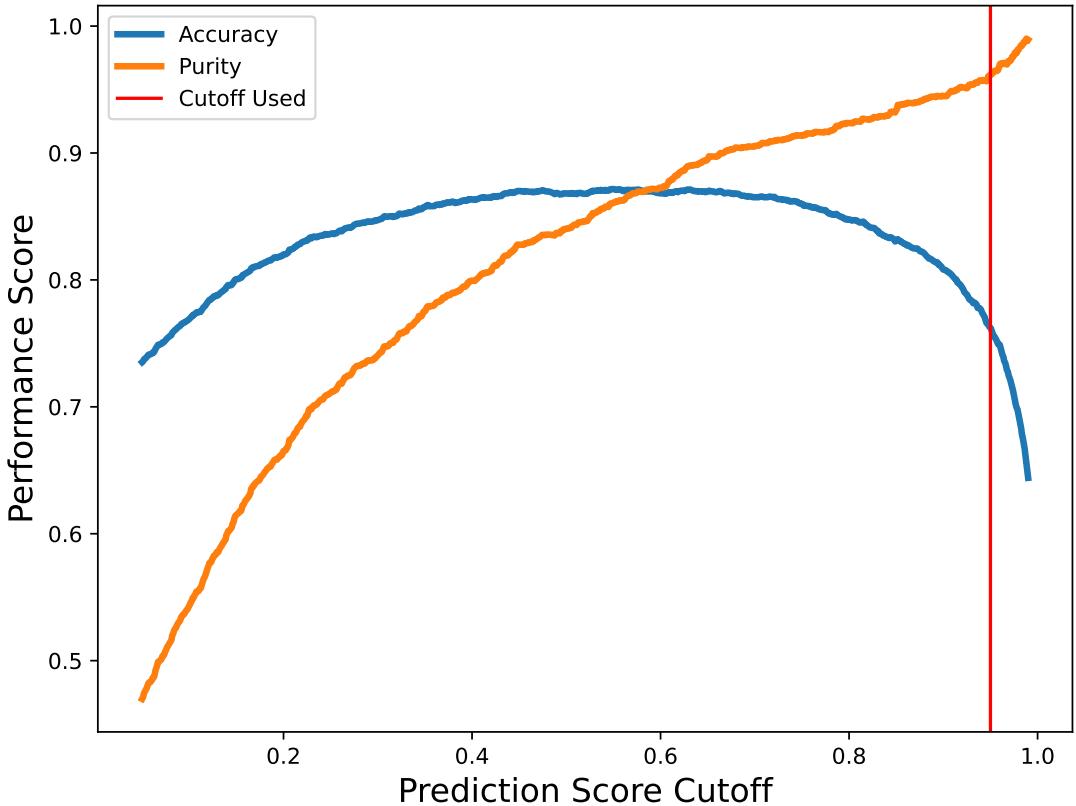


Figure 2.4: A measure of accuracy and purity against prediction score. The accuracy (in blue) is a direct measure of the number of sources Zoobot correctly predicted vs the total number of predictions made. The measure of purity (in orange) is the the number of predictions Zoobot correctly made vs the 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 B.2 shows this change with prediction cutoff in the Appendix.

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. 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, $\approx 0.2\%$ 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.

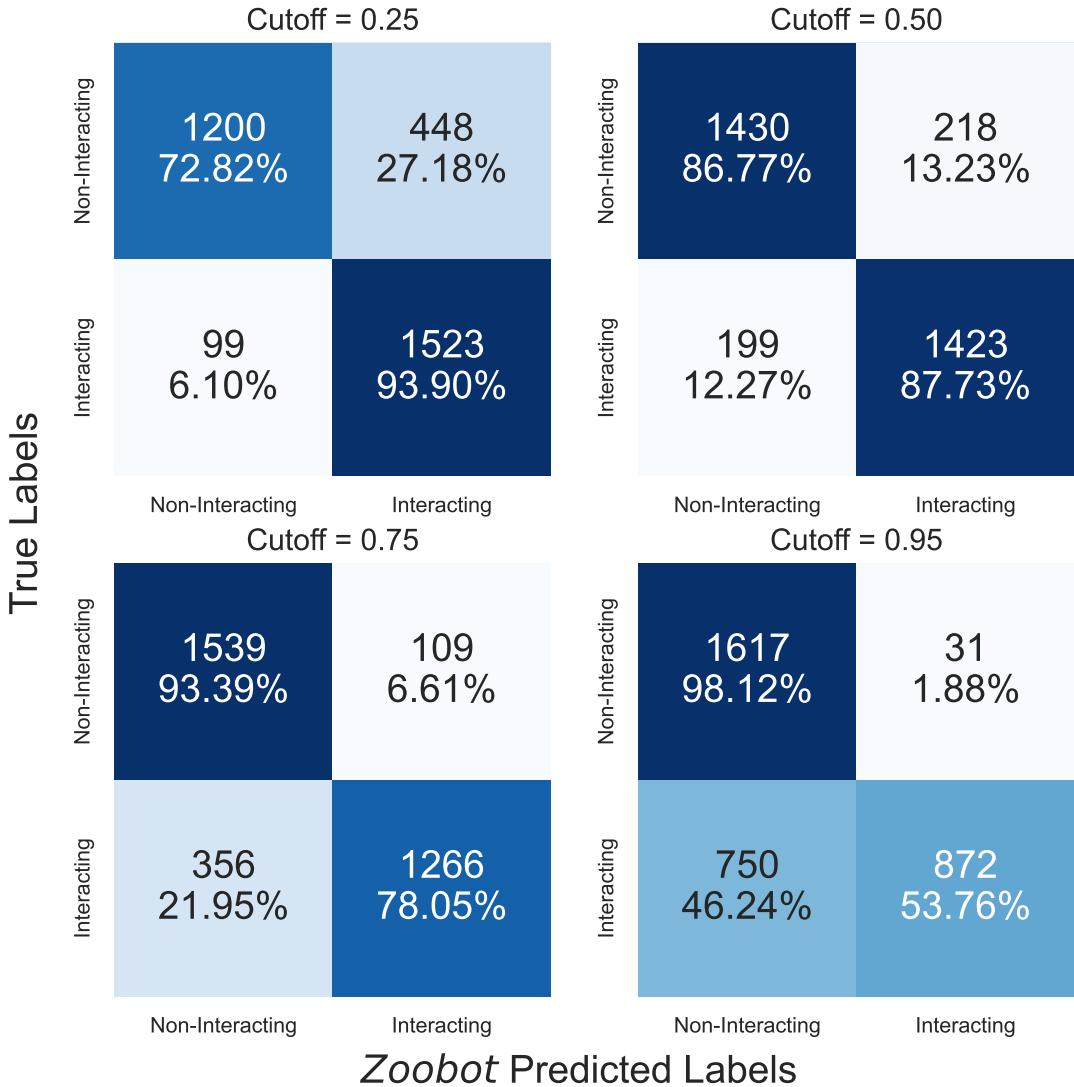


Figure 2.5: Confusion matrices of four different cutoffs of prediction score defining a binary classification of interacting galaxy or not. Confusion matrices break down our accuracy measurement into how Zoobot is misclassifying sources. At a cutoff of 0.50, the accuracy is highest at 88.2%. However, at this cutoff, $\approx 10\%$ 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.

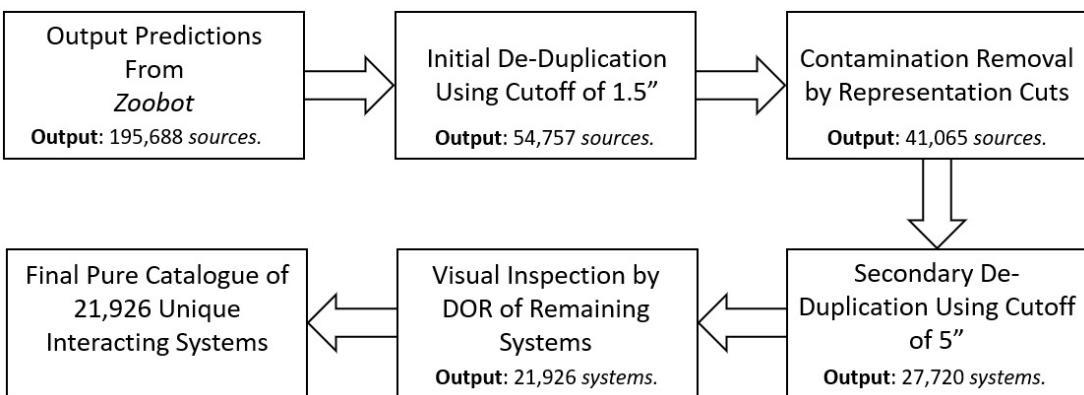


Figure 2.6: Flow diagram of our contamination and duplication removal process. De-duplication used agglomerative clustering based on sky separation. The first step of de-duplication uses a cutoff of $1.5''$. This significantly reduced duplication in the catalogue, as well as the size of the catalogue 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.

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., 2019a; Pearson et al., 2022a). 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 generalisable 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. (2022c). We reduce the dimensionality from 1,280 to 40(as in Walmsley et al. (2022c)), 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 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

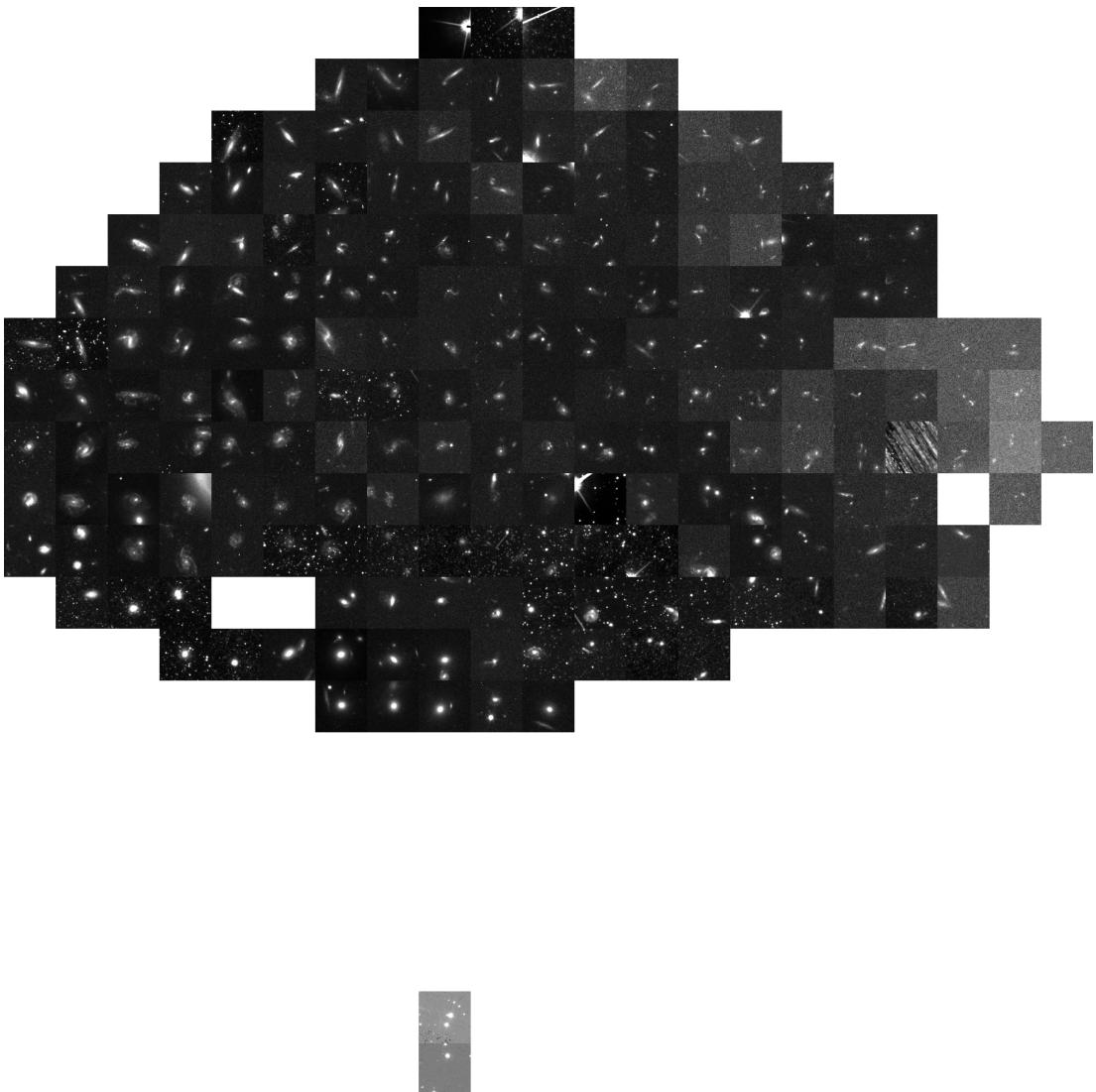


Figure 2.7: The representation distribution of 54,757 candidate interacting galaxies. This distribution is the compressed 2D representation of the 1,280 dimensional representation that *Zoobot* has learned of each image. Each image is a randomly selected one from sources within each bin in the distribution. 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.

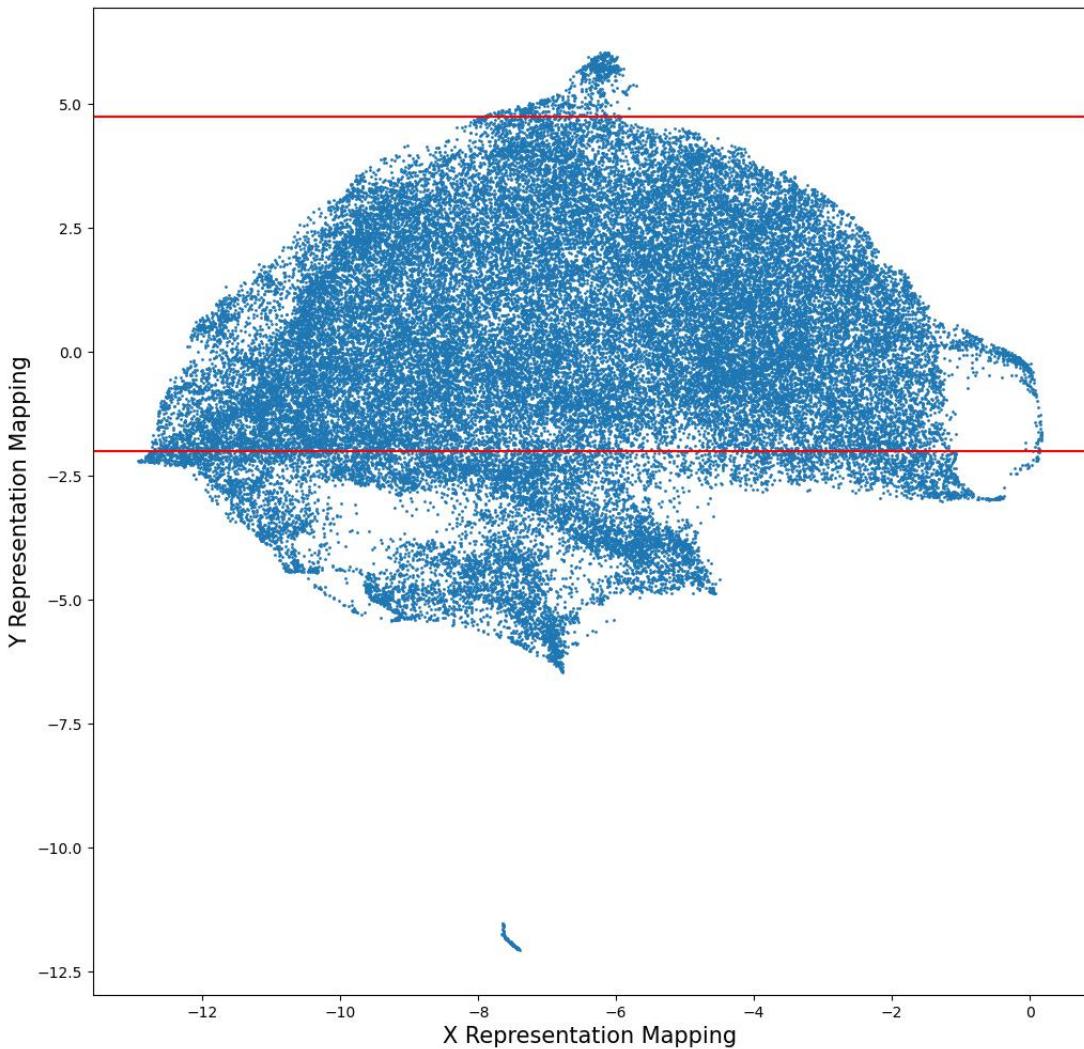


Figure 2.8: Scatter plot showing the precise distribution of each representation of sources in the remaining 54,757 sources. This is the unbinned version of Figure 2.7. The two red lines show the cutoffs utilised to remove the majority of close pairs by projection as well as the very obvious contamination 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.

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.

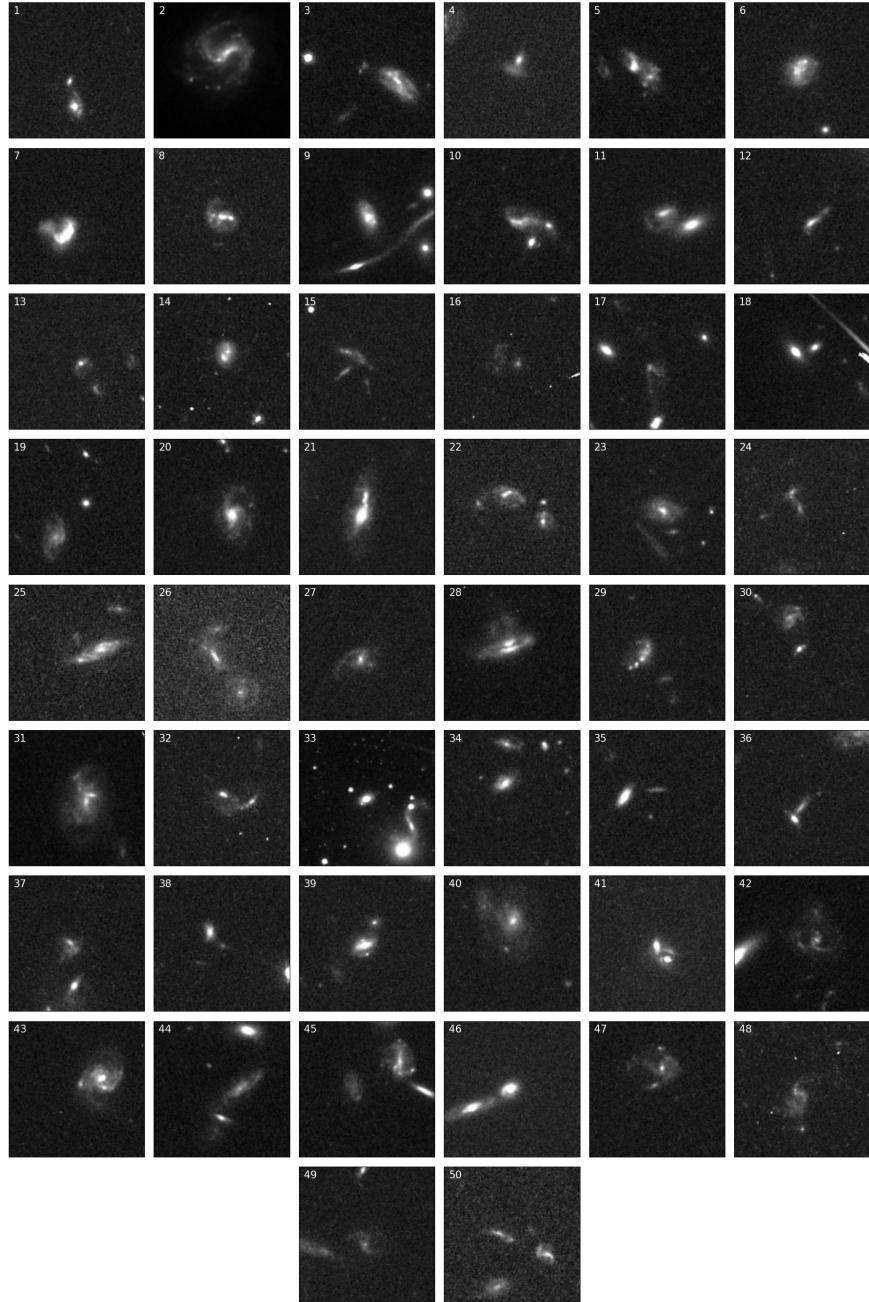


Figure 2.9: An example of 50 of the final interacting systems found with Zoobot. These were selected randomly from the de-duplicated and de-contaminated 21,926 sources. Each of these examples have extended tidal features and distortion. Not all of the final interacting systems have two galaxies within 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.

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) (?) and VizieR (?) 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 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. 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

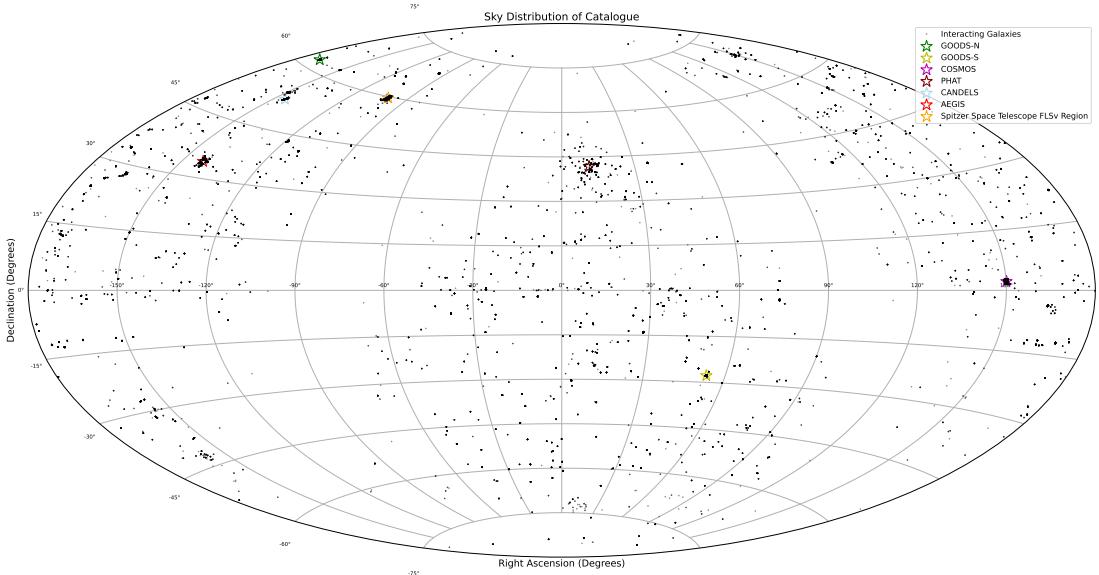


Figure 2.10: Sky Distribution of our catalogue, with marked positions of well known deep surveys conducted by *HST*. *HST* is able to observe almost the entire sky and therefore the interacting galaxies are scattered throughout. Large clusters of sources are found in the locations of surveys. This shows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

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 $\approx 3\%$ 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 sub-categories of these: active galactic nuclei (AGN)/quasars, submillimetre galaxies, galaxy groups, high redshift galaxies, jellyfish galaxies, galaxy jets, gravitational lenses/lensing galaxies, Lyman- α Emitters, overlapping galaxies, edge on protoplanetary disks, radio halos, ringed galaxies, supernova remnants, transitional

2.7 RESULTS & DISCUSSION

Image No.	SourceID	RA (deg)	Dec (deg)	Interaction Prediction	
1	4001014298177	261.292845	37.162387	0.983999	
2	4001444190958	183.527536	33.183451	0.998016	[1994]
3	4000809226818	93.960150	-57.813401	0.982266	[2011]
4	4553390202	73.581297	2.903528	0.968280	
5	4000907600174	259.037474	59.657617	0.999978	
6	4575187799	150.001883	2.731942	0.974649	[2001]
7	4000717342023	149.527791	2.126945	0.993912	[2001]
8	4001174802281	28.593114	-59.643515	0.982890	
9	4182689774	186.709991	21.835419	0.973232	[2016ApJS..224..11C]
10	4000958398690	186.719496	23.961225	0.999288	
11	4266881925	344.730228	-34.799824	1.000000	
12	4001084105393	150.128198	2.623949	0.982739	[2018ApJ...858..11C]
13	4000961670486	345.337556	-38.985521	0.954961	
14	4000719687395	338.173538	31.189718	0.974724	
15	4001435343326	331.771500	-27.826175	0.986885	
16	4001268932937	8.856781	-20.271978	0.986329	
17	4651336656	149.836709	2.141702	0.984389	[2001]
18	4000877021787	116.211231	39.462563	0.979178	
19	4000878525229	149.834893	2.516816	0.963694	[2007ApJS..172..11C]
20	6000290755870	186.774907	23.866311	0.981961	
21	4000806637434	210.253419	2.854869	0.960790	
22	4001215753971	135.898809	50.487130	0.998386	
23	4000813961830	163.678042	-12.776815	0.958405	[2001]
24	4001200639012	54.037618	-45.170026	0.991404	
25	4000921402261	150.417634	2.313781	0.990775	[2018ApJ...858..11C]
26	4001224732336	337.217339	-58.444885	0.955972	
27	4000781402752	216.968619	34.575819	0.974076	
28	4001283017901	120.202582	36.058927	0.994169	[2001]
29	4000833486119	116.260049	39.457642	0.971092	
30	4000949659908	146.342493	68.730869	0.961113	
31	4000982920478	53.084832	-27.765379	0.983472	[2010]
32	4001189505548	192.492491	2.436292	0.992574	
33	4001060882070	89.700725	-73.049783	0.962839	
34	4000889750512	151.176470	41.214096	0.962205	
35	6000322363510	53.149367	-27.823945	0.963889	[2011]
36	4000722901091	28.257843	-13.928090	0.982778	
37	6000198293960	264.488431	60.101798	0.986865	
38	4001095660911	258.587670	59.970358	0.955193	
39	4000972775076	330.960020	18.796346	0.989131	
40	4001132466571	126.545810	26.456196	0.997077	
41	4000933395648	312.810365	2.288410	0.976252	
42	4000932940918	218.066960	32.997228	0.990737	
43	4001048433104	93.880689	-57.754746	0.957755	
44	4001039919651	53.111470	-27.673717	0.994424	[2011]
45	4001282607544	333.765744	-14.006097	0.999520	
46	4000922341052	260.723839	58.849293	0.995477	
47	4000731518210	194.869144	14.146223	0.994651	
48	4001082523786	311.703084	12.869002	0.976454	

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 ESASky¹(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 B.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.

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`.

¹ESASky: <https://sky.esa.int/>

2.7 RESULTS & DISCUSSION

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.

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 investigation 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

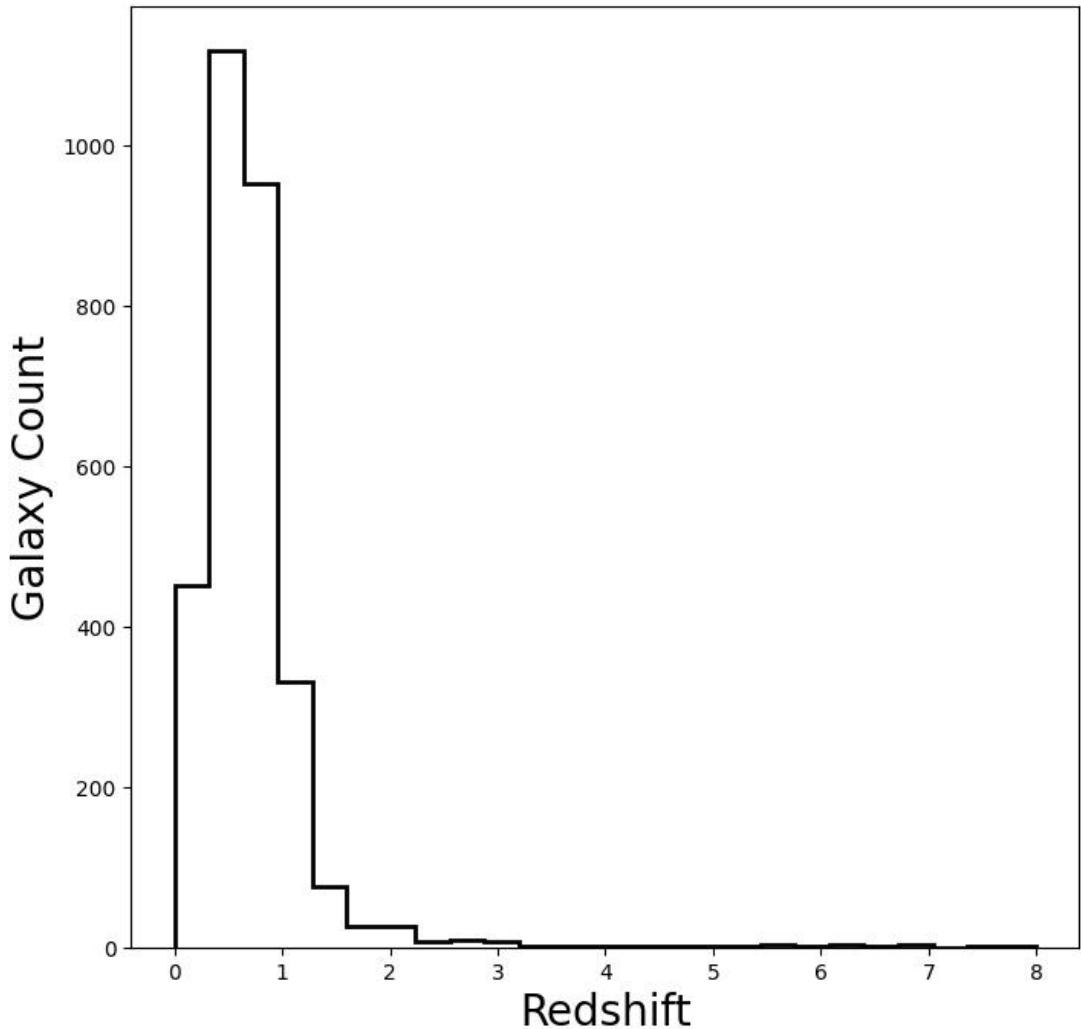


Figure 2.11: The redshift distribution of a subsample of our catalogue. Of the 7,583 referenced systems, 3,037 of them had redshift measurements in the NED, MAST or Simbad. This redshift distribution shows that our model confidently predicted interacting systems primarily for $z < 1$ systems. This was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

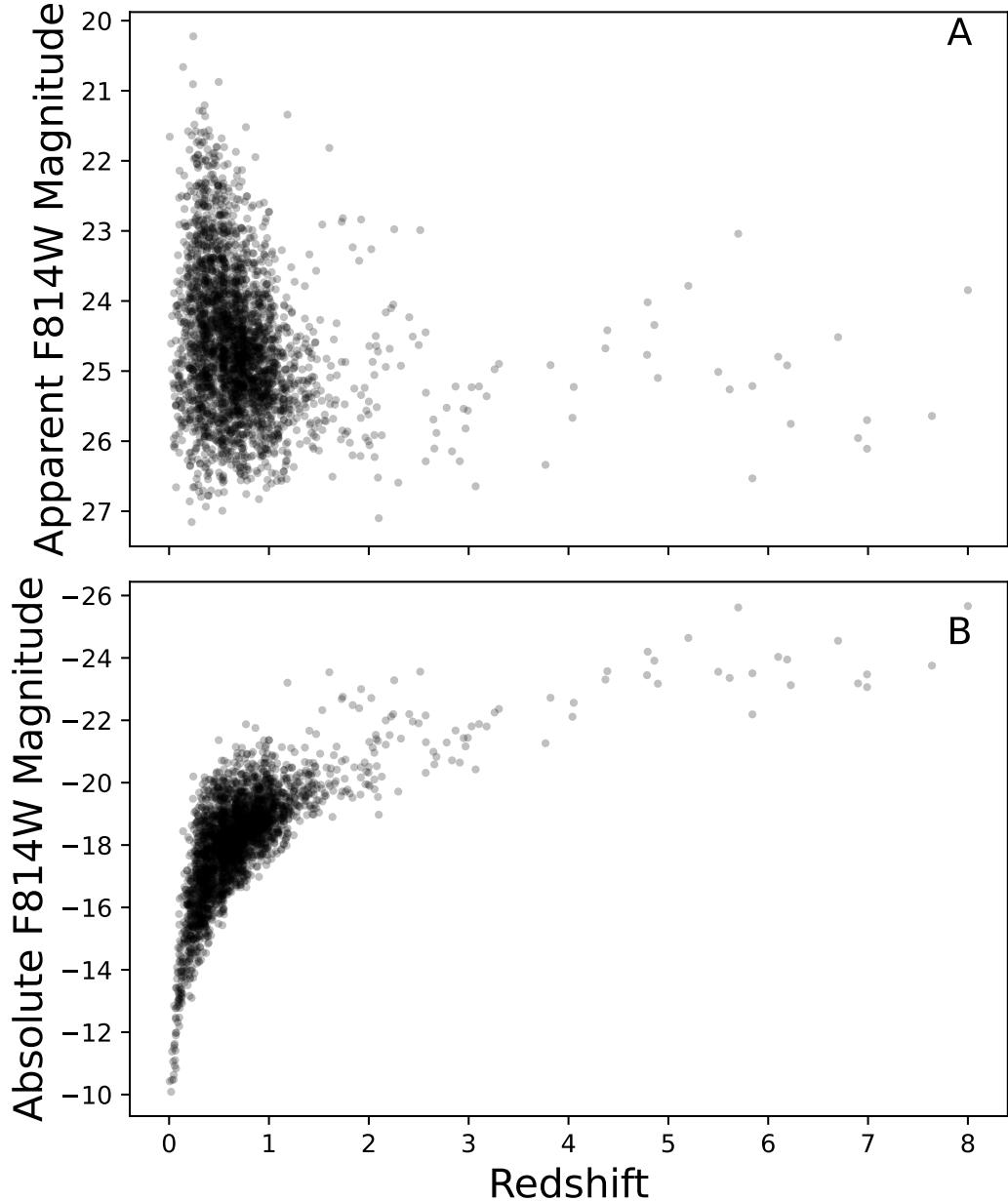


Figure 2.12: The distribution of redshift with magnitude for all sources with available data. This shows the parameter space we are sampling in this catalogue. Panel A shows that the majority of our sources are dim, background sources at low redshift. Panel B shows the faintest objects we find are at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

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 $F606W$ 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., 2004a). 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 B.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., 2003a; Schawinski et al., 2014; Whitaker et al., 2012). 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 (Lotz et al., 2008b; van Dokkum, 2005). 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,

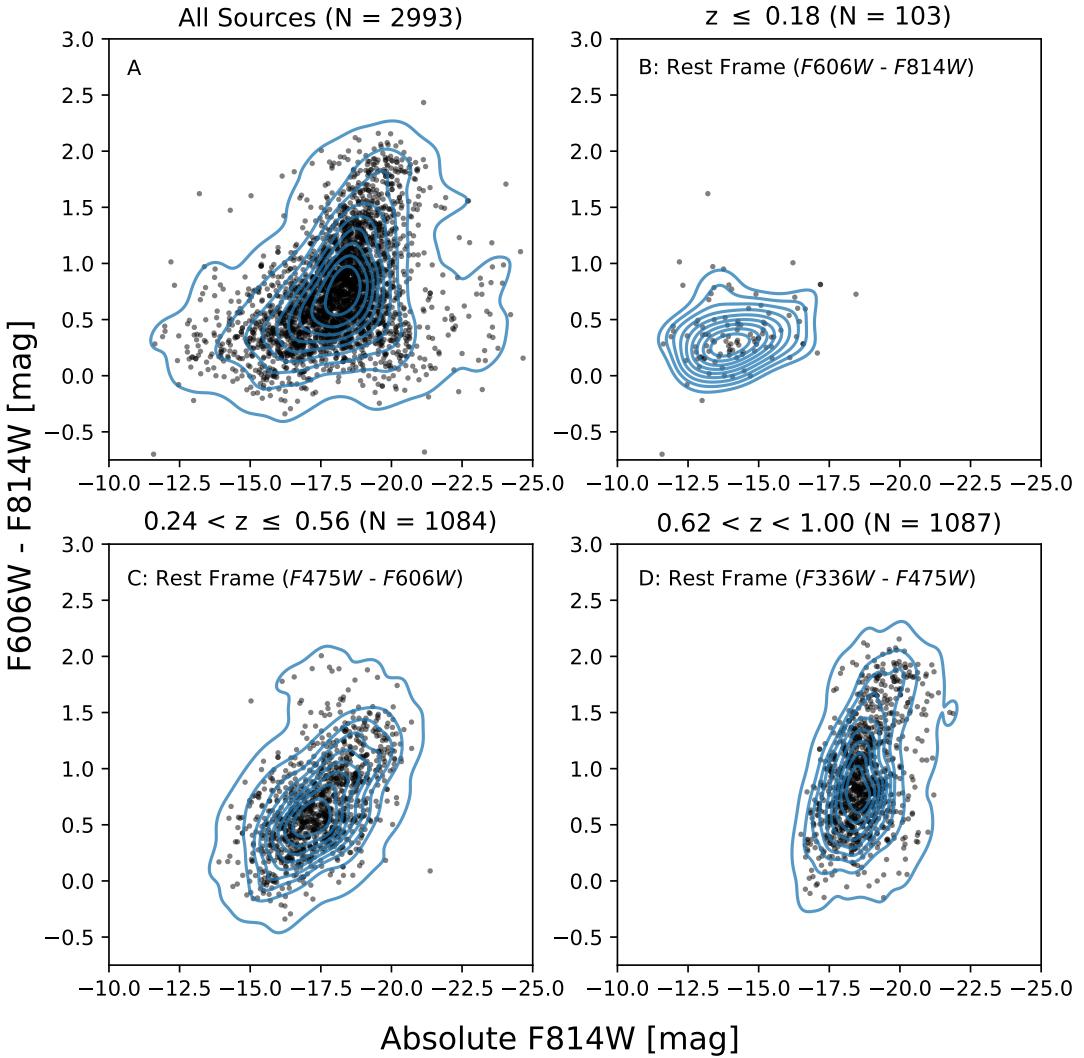


Figure 2.13: The colour-magnitude distribution of sources with a redshift measurement associated. Panel A shows the distribution of all galaxies, without controlling for redshift or dust extinction. The remaining panels then split these sources into distinct redshift bins where the $F606W$ and $F814W$ 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.

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.

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 $\approx 3\%$ 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

DO R gratefully acknowledges the support from European Space Agencies Visitor Archival Research program, and hosting at the European Space Astronomy Centre. DO R thanks Bruno Merín for supervising this project and Sarah Kendrew for aiding its creation. This project was conducted as part of DO R's 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

2.8 CONCLUSION

5-26555. All sources were found using v3.1 of the *Hubble* source catalogue (Whitmore et al., 2016a) 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.

For the purpose of open access, the authors have applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising.

DOOR 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, DOOR 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 Abstract

Abstract will go here.

3.2 Introduction

Galaxy interaction and merging is of fundamental importance to our current theories of galaxy formation and evolution. Our current cosmological theory, Λ CDM dictates that structure is expected to form and assemble hierarchically (Moreno et al., 2013; Springel et al., 2001; White & Rees, 1978). Thus, understanding the underlying processes that occur in galaxy interaction and merging are paramount to understanding cosmological theories. It is well known that many underlying

processes are triggered or enhanced due to galaxy interaction. The clearest result of interaction is the disruption of galactic morphology, the acceleration of disks to spheroidal systems and the formation of tidal features (González-García & Balcells, 2005; Hopkins et al., 2009a; Theys & Spiegel, 1977; Toomre & Toomre, 1972c), increases in star formation rates when compared to the general galaxy population (Barnes & Hernquist, 1991; Berrier et al., 2006; Knapen & Cisternas, 2015; Scott & Kaviraj, 2014) and, potentially, their rapid quenching (Das et al., 2023; Hopkins et al., 2013b). There is also evidence to suggest that interaction plays a role in the ignition of active galactic nuclei (AGN) (Comerford et al., 2015b; Ellison et al., 2011b; Hernández-Toledo et al., 2023). Therefore, with interaction having such an affect on the galaxies we see - over a long time period - understanding galaxy interaction is fundamental to understanding galaxy evolution.

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. The most obvious place to start with mapping the full dynamical histories is with simulations. Previous works find sudden increases in star formation occurs early in the dynamical timescale of the interaction (Cox et al., 2008; Rodríguez Montero et al., 2019a), 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., 2011b; 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 (1966b) 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., 2010b) 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 (Ellison et al., 2008b; 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 (Ellison et al., 2011b; Gordon et al., 2017; Rogers et al., 2009; Silva et al., 2021b). 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 just passed each other. The differences between these being paramount to understanding the effects of interaction on the galaxies involved.

In this work, we split a large sample of interacting galaxies into four specific stages based from 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., 2019b; Chang et al., 2022).

We elect to use the newly created interacting galaxy catalogue created by O’Ryan et al. (2023) (hereafter, OR23) which found a sample of 21,926 interacting galaxy systems. This is the largest visually classified interacting galaxy catalogue to date. This sample was created using the new ESA Datalabs¹ platform and the convolutional neural network machine **Zoobot** (Walmsley et al., 2022b, 2023). This classified all sources in the Hubble Source Catalogue (Whitmore

¹ESA Datalabs: [link](#)

et al., 2016b) from the Hubble Space Telescope (HST) archives into interacting or non-interacting. Once this was done, a rigorous process of de-duplication, contamination removal and visual inspection was conducted.

We cross match this catalogue with the Cosmic Evolutionary Survey (COSMOS) survey¹. This provides us with ancillary data for a subset of the OR23 catalogue. 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 utilise the photometric redshifts available from COSMOS to confirm a set of interacting galaxies and close pairs to draw on trends with projected separation.

This paper is laid out as follows in Section 3.3 we explain the OR23 and COSMOS catalogues, and describe the process of catalogue matching especially focused on accurate de-duplication and cross-matching between the different catalogues. Section 3.4 describes the methods by which we split the interacting systems into stages, and how we identify the secondary galaxies in each pair. In Sections 3.5 and 3.6 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 3.7 where we compare to previous works and put our results in the context of the field. Finally, Section 3.8 we make concluding remarks and discuss future work to better our constraints.

Where necessary, we use a Flat Λ CDM cosmology with $H_0 = 70 \text{ km/s/Mpc}$ and $\Omega_M = 0.3$. Hereafter in this paper, 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 tidal disturbance. A merging galaxy is the final state of these flybys, where two or more systems have or are about to coalesce to form a morphologically irregular system.

¹DOI: 10.26131/IRSA178

3.3 DATA: Catalogue Matching & Secondary Identification

3.3.1 The O’Ryan+23 Catalogue

The interacting galaxy catalogue we draw from is the OR23 catalogue of interacting galaxies. These systems are classified based on their visual morphology and, while the full process of its creation are detailed there, we will briefly describe it here. The morphological classifications were made by a convolutional neural network (CNN) - *Zoobot* - which searched all final product data in the *Hubble* Legacy Archive. While a CNN primarily classifies images by extracting features from them *Zoobot* differs by extracting features from an image and reducing them to generalised representations. By learning the generalise representation of different galaxy morphologies, *Zoobot* is incredibly flexible and easy to transfer learn on different tasks and datasets. This is further reinforced by training it on citizen scientist responses to the Galaxy Zoo: DECaLS (Walmsley et al., 2022b) and using it to make classifications on multiple morphology questions at once. In OR23, using a small dataset of citizen science-classified interacting galaxies from Galaxy Zoo (Lintott et al., 2008b) *Zoobot* was finetuned to *Hubble* data.

Cutouts of each source were created as single-band *F814W* images at 150×150 pixels across. Source coordinates were found in the *Hubble* Source Catalogue (Whitmore et al., 2016b). This corresponded to just over 92 million cutouts created and classified. Due to well known issues of classifying interacting galaxies by morphology alone (such as close pair contamination) an extremely conservative prediction score from *Zoobot* was required to make a positive classification. This resulted in a very pure, but highly incomplete catalogue of 21,926 interacting galaxy systems. This is the largest catalogue of interacting galaxies directly found to date.

The OR23 catalogue only provided the coordinates and *Hubble* Source catalogue ID for each source. Therefore, we must either calculate our own ancillary data or cross match with existing catalogues containing it. We elect to cross match the OR23 catalogue with the numerous catalogues of the COSMOS sur-

vey. The overall survey has numerous sub-catalogues ranging from calculations of galactic physical parameters to predictions on AGN classifications.

3.3.2 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 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 most important parameters these calculate are the stellar masses and the estimated star formation rates of each source. We opt to simply use one of these two measurements in our analysis. For the stellar mass we use the LePhare measurement and for the star formation rates we use the 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, therefore, elect to only use one software for each parameter and not to match the stellar masses and SFRs between them.

We cross match between the OR23 catalogue with the COSMOS2020 catalogue by matching coordinates between them. We search the COSMOS2020 catalogue for any sources within $10''$ of the OR23 coordinates and make a list of any COSMOS2020 sources. If more than one COSMOS2020 source is found within this radius we take the source with the smallest angular separation from the OR23 coordinates as the correct source. Once we have identified one COSMOS2020 source for every OR23 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 OR23 source, then we the OR23 source not in the COSMOS2020 catalogue. We find 3,786 of the OR23 sources exist in the COSMOS2020 catalogue.

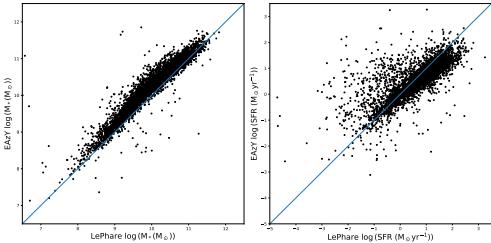


Figure 3.1: Comparison of the measures of stellar mass and SFR using either LePhare or EAzY/FAST photometric codes to calculate them. If the algorithms agreed perfectly, the sources would lie on the blue 1:1 line. *Left:* The scatter in the stellar masses between softwares. As shown, EAzY often seems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

Once matched, we remove any sources with unphysical photometric measurements for the stellar mass or star formation rates. We then further reduce our sample by only keeping sources within a mass range of $6.5 \leq \log_{10} M_*(M_\odot) \leq 12.5$ and a star formation rate range of $-5 < \log_{10} \text{SFR}(M_\odot \text{yr}^{-1}) \leq 3.5$. To also nullify the effect of surface brightness dimming that may occur when attempting to classify tidal features of interacting galaxies, we institute a redshift cut of $z \leq 1.2$. This also makes our sample match the catalogue that will be later used to investigate the incidence of our sample of interacting galaxies with environment (detailed in Section 3.4.2). Applying these cuts reduces our sample size to 3,689 interacting galaxies.

3.3.3 Secondary Identification

As the OR23 catalogue 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 OR23 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 1001×1001 pixels). The original OR23 cutout was also displayed next to the enlarged cutout. We annotate each cutout with each sources COSMOS2020 ID

3.3 DATA: Catalogue Matching & Secondary Identification

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; incorrect primary assigned.

To classify a secondary galaxy for our primary one, it 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 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 priamry 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 were very close together and small enough they would be 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.

3.3 DATA: Catalogue Matching & Secondary Identification

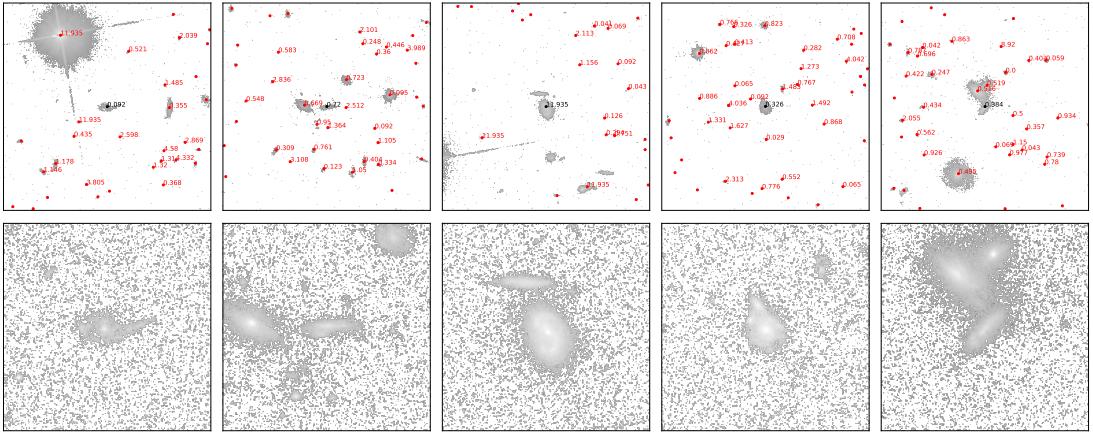


Figure 3.2: An example of each visual classification made on the cross matched sample. From left to right these are the secondary could not be identified, the primary had a clear secondary, the primary could not be reliably classified as an interacting galaxy, the redshift was null and the incorrect primary was identified in OR+23. 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 or having a null redshift) were removed from our sample. Galaxies that could not be reliably classified as interacting as they showed little tidal distortion or had no neighbouring systems at a matching redshift 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 primary 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 pipeline, we find a sample size of 3,829 interacting galaxies. We conduct a de-duplication based on the COSMOS2020 ID, which reduces the

3.3 DATA: Catalogue Matching & Secondary Identification

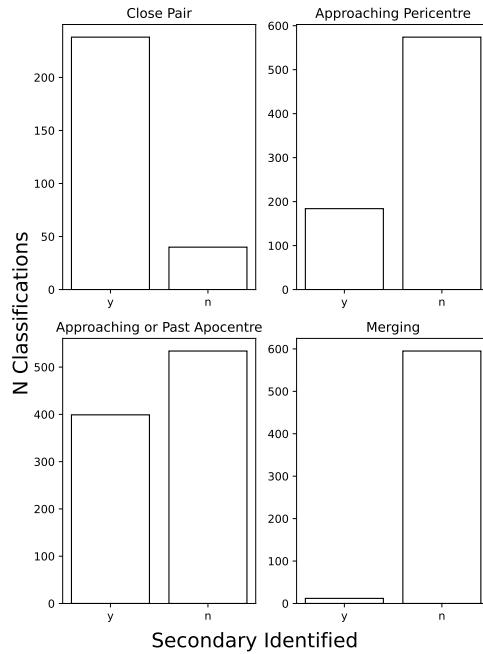


Figure 3.3: Where a secondary could be identified at different stages in the interaction. The reasoning for such disparity in secondaries identified is due to the relative distance each the secondary would be from the primary at each stage. For a close pair, we often found the secondary galaxy, but a 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.

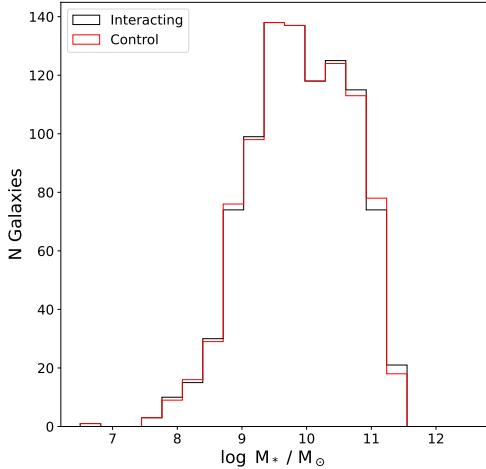


Figure 3.4: The mass distribution of the paired interacting galaxy sample and the control sample. Both primary and secondary galaxies are within this distribution. From our sample of 4,181 interacting galaxies morphologically identified, 834 are confirmed galaxy pairs. This means that the secondary in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

sample back down to 3,547 interacting galaxies. This shows that many of the secondaries we had identified for the different systems were already found in the OR23 catalogue. 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 sub-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.

3.3.4 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 the OR23 catalogue. Primarily, these extra interacting galaxies are from systems which had more than one galaxy involved in the interaction. Our pipeline was built to only find a primary and secondary galaxy and, therefore, we add these extra systems through another pipeline. Other interacting galaxies that were added were low redshift systems which would have appeared very large in the small OR23 cutout. 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.

3.3.5 Redshift Effects on Visual Classification

There are two limits when using visual morphology to find interacting galaxies. The first is the limitation of being able to characterise tidal features with increasing redshift. As the redshift increases, surface brightness dimming of the systems may lead to a lack of visual confirmation in the tidal features of different objects. Thus, if we are heavily affected by this in our redshift range ($0.0 \leq z \leq 1.2$), we would see distinct signs of it in our selection function of our mergers. For instance, if all of our stage 4 classifications were actually high redshift and high mass stage 3 galaxies, we would observe this in our selection function. Therefore, Figure 3.5 shows our the distribution of stellar mass across our redshift range for each defined stage in our interaction. As shown, the distribution of masses across redshift that we find in our sample is reasonably consistent with each other.

The second issue that could affect our sample is that of contamination by close pairs by projection effects rather than by galaxies which are physically close together. For the majority of these systems, we are limited to only having their photometric redshift from the COSMOS2020 catalogue and not their well-constrained spectroscopic redshift. As we are conducting visual classification, this

3.4 METHOD: Environment, AGN and Interaction Stage

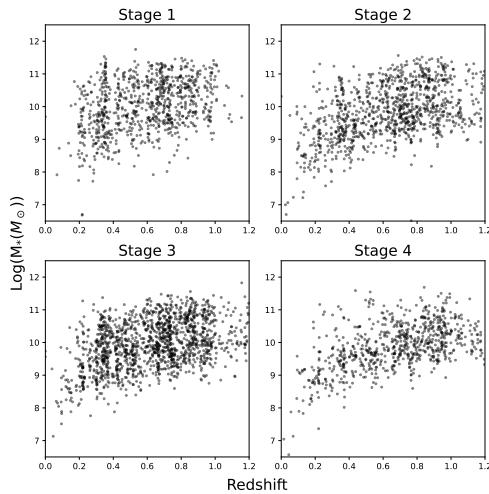


Figure 3.5: Redshift vs Mass distribution for each stage of interaction we have defined. This is important as we use tidal distortion and the existence as tidal features as a fundamental for our classification methodology. As shown here, the distribution of systems across redshift and mass is consistent for all stages in our sample. Therefore, we are likely not affected by this in our analysis.

issue is rather limited. The criteria on all of our sources besides stage 1 is that they must be morphologically disturbed. Each source is confirmed to have some kind of tidal feature $f=$ present in the image, even if a secondary is not present in the image. There are, always, of course other galaxies in each image and there is no guarantee that the calculated photometric redshift is correct. We often find galaxies which are clearly interacting or paired, but with the photometric wildly different. In these cases, we defer to the redshift and only record one of the galaxies in our sample. This approach, however, means we must be careful to investigate the effects of environment on our galaxies as disturbance by, say, ram pressure stripping can appear like tidal features.

3.4 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 dif-

ferent 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.

3.4.1 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 define four dinstinct stages of interaction:

- Stage 1: Systems which are well-separated with little to no morphological disturbance (Close pairs).
- Stage 2: Close pairs showing morphological distortion while still in a pair or show a physical connection by tidal features.
- Stage 3: Well separated pairs with morphological disturbance or isolated galaxies with clear tidal features.
- Stage 4: Highly disturbed systems with two or more cores within them.

Figure 3.6 shows the original OR23 source cutouts of an example galaxy in each stage. There are degeneracies associated with this staging system, however. In

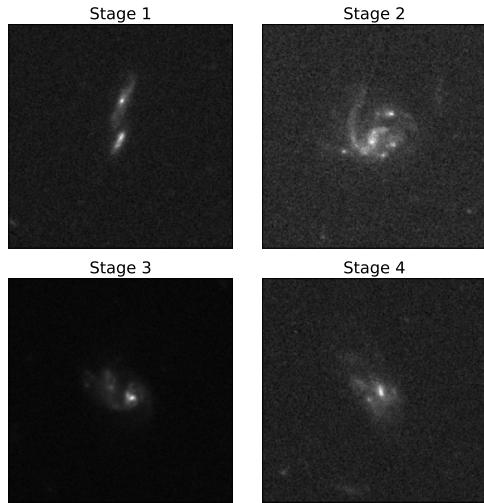


Figure 3.6: Examples of the four stages we split our interacting galaxy sample into. Stage 1: A close pair with confirmed redshift matching. Stage 2: Two distinct systems interacting with tidal features forming. Stage 3: A tidally disturbed system with no secondary present, likely at apocentre. Stage 4: A galaxy with multiple cores while highly disturbed. At the final stage before coalescence.

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.

Stage 1 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.

Stage 2 of the interaction is defined as the point where the two galaxies in the interaction 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

3.4 METHOD: Environment, AGN and Interaction Stage

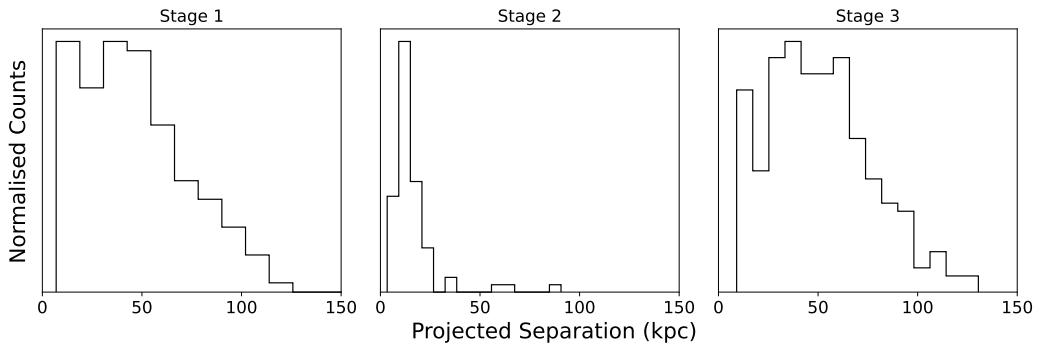


Figure 3.7: The projected separations of the confirmed galaxy pairs in our sample. This confers with other works the definition of our different stages. Stage 1 can be at any projected separation, however, we have visually confirmed that these sources are not morphologically disturbed. Stage 2 is dominated 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 stage 2. Finally, stage 3 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 stage 2 and stage 3 in projected separation, as their visual classification is also dependent on the system size.

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 stage 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 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.

Stage 3 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

3.4 METHOD: Environment, AGN and Interaction Stage

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, stage 4 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 it's most violent with complete disruption of both galactic disks and likely increased star formation across the disk.

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). 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 always move from stage 1 to stage 2 in the early times of the dynamical timescale. Then, dependent on the velocity of the system, the galaxy pair will either move straight into stage 4 of classification and begin to merge or it will move into stage 3. This change from stage 2 into stage 3 can then take many branching paths dependent on the elocity in the system. If the galaxies have sufficient velocity, stage 3 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 stage 3 back into stage 2. This could then happen for many cycles until finally the galaxies enter stage 4 and coalesce. Figure ?? shows the branching

3.4 METHOD: Environment, AGN and Interaction Stage

paths that the galaxies can take through each stage. Note, these are created using the PySPAM restricted numerical simulation (O’Ryan et al. in prep.) and are for illustrative purposes only.

More commonly, in the literature, rather than using the stage of an interaction based on morphology we use the projected separation of the two systems. 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 3.7 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_{proj}) in stages is between stage 2 and stage 3. Here, we see that stage 2 is dominated by systems with $s_{proj} < 35\text{kpc}$ while stage 3 is dominated $25 \leq s_{proj} \leq 100\text{ kpc}$. Stage 2 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 a stage 3 interaction where the two galaxies in the pair are completely distinct. There is some overlap between the projected separations of stage 2 and stage 3 as this is somewhat dependent on the angular size of the two systems involved in the interaction.

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 from both EAzY/FAST and LePhare. To ensure that any enhancements are from interaction alone, we must ensure we have no bias’ in the environment distribution of our sample and identify AGN. We, therefore, cross match to two further catalogues COSMOS catalogues containing the required information.

3.4.2 Matching to Environment Catalogue

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 utilise 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 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 photometric 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 $\approx 45,000$ 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 $\approx 39,000$ sources with which to cross match the OR23 sample. 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 simply utilise 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 that we are matching 2,800 sources 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.

3.4.3 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 (Delvecchio et al., 2017; Smolčić 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 conduct a radial search of $10''$. If more than one source is within this radius, we match on the closest source. 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 classifications in the VLA 3GHz survey and 155 in the Chandra survey. These were split into 812 star forming galaxies and 402 AGN. We also investigate cross matching with the MPA-JHU catalogue (Brinchmann et al., 2004; Kauffmann et al., 2003b; Salim et al., 2007), however, found that all matches were already represented by the VLA and Chandra surveys. We also utilise 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.

3.5 STAR FORMATION EVOLUTION WITH INTERACTION STAGE

3.5.1 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 utilise the estimates of these parameters that exist in the COSMOS2020 catalogue itself. Then, we utilise 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 it's component stages.

Figure 3.8 shows the breakdown of stellar mass and SFR with stage. On a population scale, there is clear evolution in the star formation rate from stage 1 through to stage 4. In stage 1, where the 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 3.8 show increasing number density in the population into the blue cloud. In stage 2, 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. Stage 3 shows a similar effect, where the red sequence reduces again before finally in stage 4 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

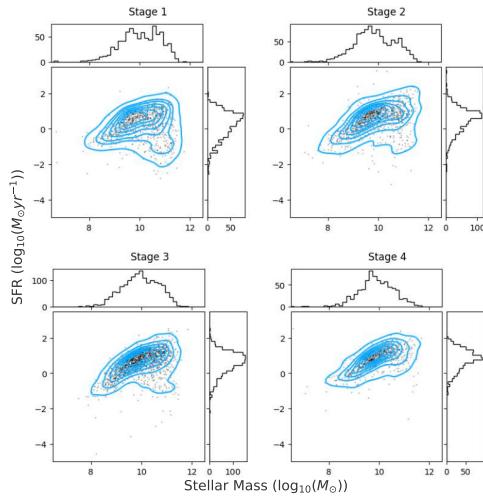


Figure 3.8: The LePhare stellar mass against the EAzY SFR across the different stages of the interaction. The blue contours are 8 levels of density of the underlying populations in each frame. *Top left*: The stellar mass and star formation rate of Stage 1 of our sample. Here, the interacting galaxies 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*: Stage 2 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*: Stage 3 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*: Stage 4 of the interaction, where the two systems are close to coalescence or have merged. The red sequence of galaxies has completely disappeared.

3.5 STAR FORMATION EVOLUTION WITH INTERACTION STAGE

well, we could simply be selecting higher mass systems as we increase stage. This would have the result of systems in stage 4 having, on average, higher SFRs than those in stage 1 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 3.5.3, 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 utilise both to ensure consistency and robustness in our measurements. First, we create weighted distributions of stellar mass. The weighting scheme we utilise 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 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 3.9 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 approximately one. Thus, proving that the distribution in stellar mass though each stage is very similar, and no stellar mass evolution is occurring. Figure 3.10 shows the SFR distributions while being weighted by stellar mass. When comparing stages 1 and 2, 1 and 3, 1 and 4, 2 and 4 and 3

3.5 STAR FORMATION EVOLUTION WITH INTERACTION STAGE

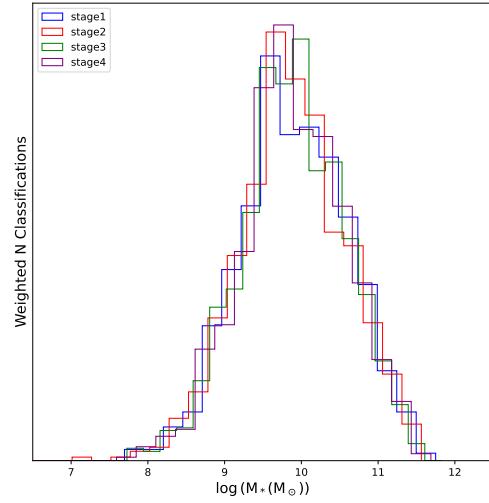


Figure 3.9: The stellar mass distribution across the four stages. Each bin is weighted based on the counts in the smallest sub-sample in stage. This is stage 4 of interaction.

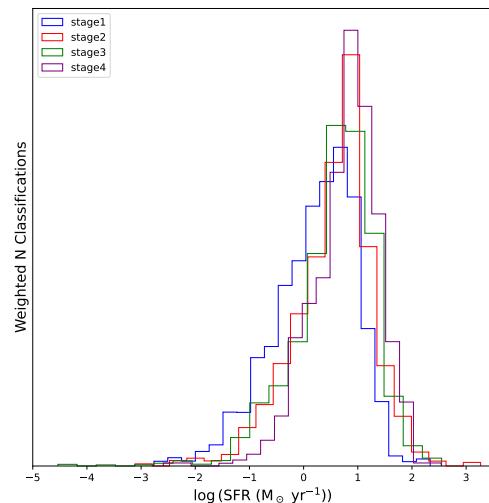


Figure 3.10: SFR distribution weighted by mass across each stage. This weighting is based on our sample of stage 4.

3.5 STAR FORMATION EVOLUTION WITH INTERACTION STAGE

and 4, 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 and are very dissimilar. However, for comparing stages 2 and 3, 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 radically changes from stage 1 to stage 2 and from stage 3 to stage 4, while remaining reasonably similar from stage 2 to stage 3.**

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 to stage 3 - where 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 radically 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 type. While the red sequence is significantly reduced 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

$$\log \text{SFR}_{\text{MS}}(z)[M_{\odot} \text{yr}^{-1}] = -7.6 + 0.76 \log \frac{M_*}{M_{\odot}} + 2.95 \log(1+z). \quad (3.1)$$

3.5 STAR FORMATION EVOLUTION WITH INTERACTION STAGE

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:

- Starburst galaxies. Here, the galaxy SFR is highly elevated compared to the SFMS. We follow Aird et al. (2019) and define a cutoff of $\log\text{SFR}/\text{SFR}_{\text{MS}} > 0.4$.
- Main sequence galaxy. The galaxy is within 0.4 dex of the SFMS and approximately has the expected SFR. Defined as $-0.4 < \log\text{SFR}/\text{SFR}_{\text{MS}} < 0.4$.
- Sub-main sequence galaxy. A galaxy whose SFR is below the majority of the SFMS, but likely not quiescent. Defined as $-1.3 < \log\text{SFR}/\text{SFR}_{\text{MS}} < -0.4$.
- Quiescent (High) galaxy. A galaxy with an SFR in the top $\approx 50\%$ of the quiescent galaxy population. Defined as $-2.3 < \log\text{SFR}/\text{SFR}_{\text{MS}} < -1.3$.
- Quiescent (Low) galaxy. A galaxy with low SFR and very likely completely quenched. Defined as $\log\text{SFR}/\text{SFR}_{\text{MS}} < -2.3$.

We split our sample into its different stages and apply these criteria. Figure 3.11 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 stages 1 to 4 and a large reduction in the number of quenched systems. Figure 3.12 shows the change in fraction of the different galaxy classifications through stage, reflecting the results found in Figure 3.8. Initially, in stage 1, 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 stage 4

3.5 STAR FORMATION EVOLUTION WITH INTERACTION STAGE

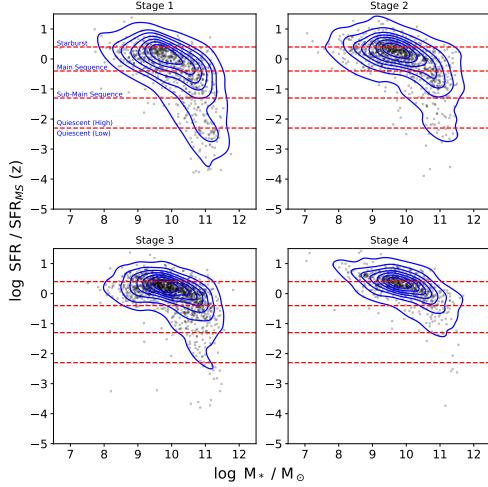


Figure 3.11: Stellar mass against the ratio of measured SFR to the expected SFR if the galaxy was on the SFMS. Black points are the individual sources, while the blue contours are as in Figure 3.8. The red dotted lines show the cutoffs for different galaxy classifications based on their SFR, with each 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.

galaxies. The complete 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 stage 2 - 3), but not hugely. 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.

Now, 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. Hence, even those mergers which are considered dry - with very little gas - would be capable of driving major rejuvenation in the star formation of these systems. However, we do not see a significant increase in the starbursting population until we the

3.5 STAR FORMATION EVOLUTION WITH INTERACTION STAGE

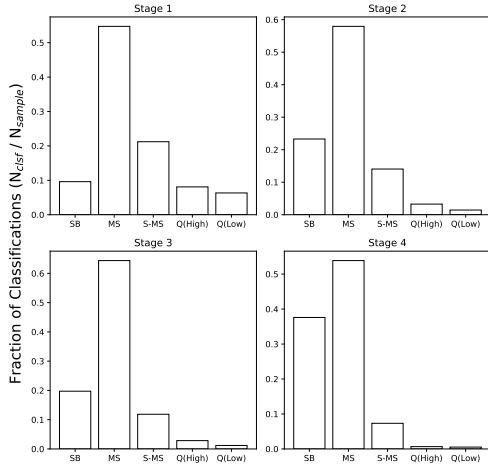


Figure 3.12: The change in fraction of different galaxy classifications from the fraction of SFR to the expected SFR on the SFMS. While galaxies on the SFMS remain dominant through each sample across interaction stage, there is significant change in the starburst and quiescent populations. The starburst galaxy population moves from being roughly half the size of the quiescent population in stage 1 to completely dominating it in stage 4. This is occurring while the quiescent population is significantly reduced to almost non-existence in stage 4.

two galaxies begin to actually coalesce. However, **we find clear, statistically robust evolution of a galaxys' SFR with interaction stage.** The SFR changes dramatically from stage 1 to 2 and stage 3 to 4 - after the initial passage of closest approach and at the point of coalescence of the interaction.

3.5.2 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 stage 1 galaxies, 146 stage 2 galaxies and 498 stage 3 galaxies. Only 3 stage

3.5 STAR FORMATION EVOLUTION WITH INTERACTION STAGE

4 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 3.3.3. 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 $\leq 50\text{kpc}$. 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 comparing 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 3.13 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.

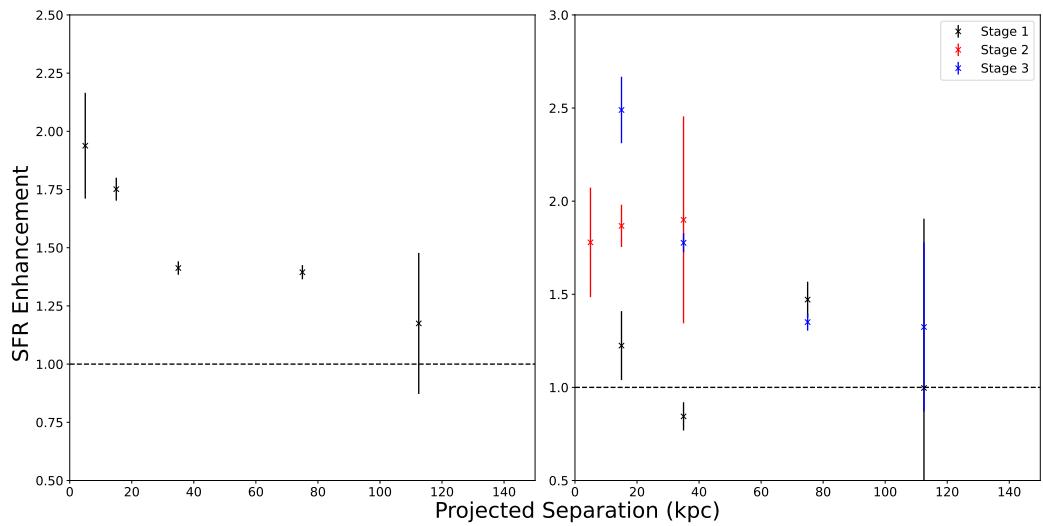


Figure 3.13: The projected separation against the star formation enhancement in average star formation at different bins of projected separation. Each bin must contain at least 10 counts to be considered. As the bulk of our galaxy pair sample is at low projected separation, we heavily sample from this 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 stage 1, red stage 2 and blue stage 3. We only investigate stages 1,2 and 3 as only three galaxy pairs were identified in stage 4. Here, we find a generally decreasing star formation enhancement with projected separation but very different individual behaviour dependent on the stage classification.

3.5 STAR FORMATION EVOLUTION WITH INTERACTION STAGE

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 stage 1, 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. Stage 1 represents when the two galaxies are close pairs, but with 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 stage 2, we see consistent enhancement with projected separation. It is also much larger than stage 1, 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 stage 3 galaxies which overlap with the stage 2 measurements. The enhancement in stage 3 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 stage 4 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.**

3.5.3 Controlling for Galactic Environment

While we have found clear evidence of evolution in the SFR of interacting galaxies with stage, it is important we remove a potential source of contamination. It is well known that the environment has a direct impact on the observed SFR of

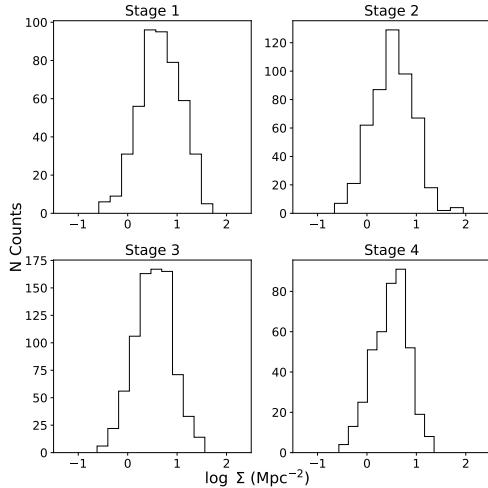


Figure 3.14: The density about each of our sources matched with the Darvish et al. (2017) catalogue. As shown, there is no existing bias in the distribution of galactic environments throughout our stages. This is confirmed by performing KS- and AD-tests. Therefore, we can conclude that the shown evolution in SFR with interaction stage is not due to environmental effects.

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. Therefore, we utilise the existing COSMOS environment catalogue described in Darvish et al. (2017) to ensure there are no environmental biases in our staging.

Upon applying the matching described in Section 3.4.2, we break our samples into different environment measures with stage. Figure 3.14 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 the evolution in the SFR with stage is due to interaction, not environment.

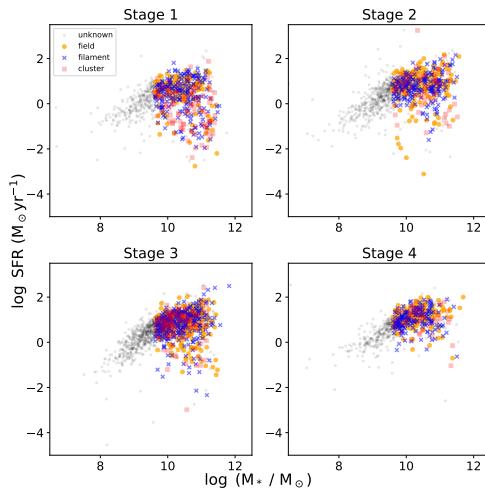


Figure 3.15: The distribution of environment classifications through our sample. This is only for sources with a $\log M_* / M_\odot \geq 9.6$. However, from this subsample, we see that there is no trend with environment and that the distribution is random throughout each stage. Therefore, the effects of increasing SFR with interaction stage is not due to environmental effects.

Figure 3.15 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 our proposed SFR evolution with stage. 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%. It is important to note, however, that the galaxies most likely to lie in a cluster in our sample would be lower mass galaxies. However, these systems are not found in the red sequence where we see the most change in SFR due to interaction. Therefore, the lack of information here does not have a large impact on this measurement.

3.6 Nuclear Activity with Interaction Stage

We now cross match our sample with the two AGN catalogues and methods described in Section 3.4.3. This provides us a sample of 1,361 AGN and star

	Stage 1	Stage 2	Stage 3	Stage 4
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.

forming galaxies (SFGs). The breakdown of number of classifications per stage is shown in Table 3.1. Figure 3.16 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 appearing throughout 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 stage 1 on the red sequence at higher masses. This shifts as we increase interaction stage, with the bulk of the AGN beginning to appear at lower masses in stages 3 and 4. Finally, in stage 4 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 these, however, reveals p-values close to 1 and therefore it is very likely that these distributions have been 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 3.17 shows the changing AGN fraction with interaction stage. We find that the AGN fraction generally increases with interaction stage. There is a large increase from stage 1 to 2, followed by the fraction mostly remaining unchanged until stage 3. Then, in stage 4 we measure another large jump in the AGN fraction. However, before we discuss this result further, we must point

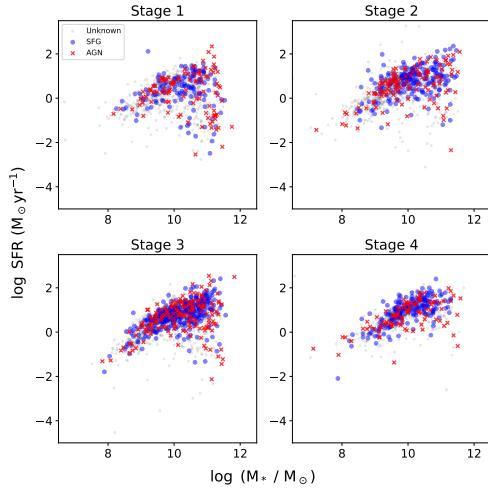


Figure 3.16: The distribution of AGN through stage with SFR and stellar mass. We do not find any major changes and evolution with AGN and SFG position through interaction stage.

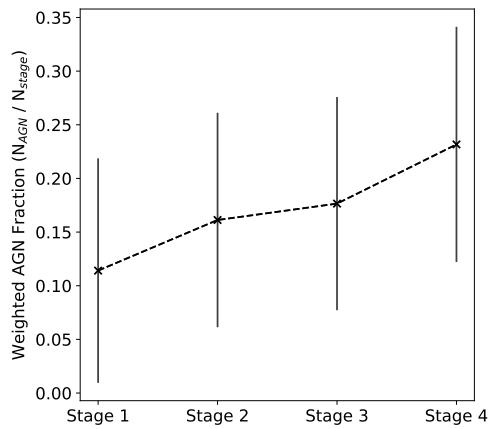


Figure 3.17: The change in AGN fraction with stage. We define the AGN fraction as the ratio of number of confirmed AGN divided by the number confirmed AGN and SFGs in the sample. This is then weighted by the relevant sizes of each subsample and the number of unclassified sources in each. Errors on each point are found by bootstrapping. We do not include any sources of which we have no information. We find a sudden drop in AGN fraction from stage 1 to stage 2, followed by a rapid increase from stage 3 to 4.

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 on our sample from bootstrapping the underlying low-count population of AGN and SFGs and comparing to the change in fraction measured. Thus, while our initial values show a general increase, because of the large error bars, there could be no increase at all with stage to even a slight decline.

It is also difficult to use interaction stage as a proxy for the projected separation in this context. As shown previously, stages 2 and 3 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 3.18 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 stage 2 galaxies, which are close to pericentre. The second peak, from a projected separation of 30kpc - 60kpc is dominated by stage 3 galaxies, although some stage 1 AGN are in this peak as well. Finally, the third peak from 85 - 125kpc is representative of a mixture of stage 1 and stage 3 galaxies. None of our stage 4 galaxies which had their secondaries identified overlapped with our AGN and SFG identified sample.

We find that the AGN fraction generally increases with interaction stage, however, it again remains constant from stage 2 to stage 3. This indicates that the mechanism driving AGN ignition in interaction is primarily at the point of the closest approach between the two systems. This is further reinforced when the system begins to merge, and we see the AGN fraction increase significantly in our sample. 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 stage 2 of the interaction where the galaxies are overlapping and secondly in the early parts of

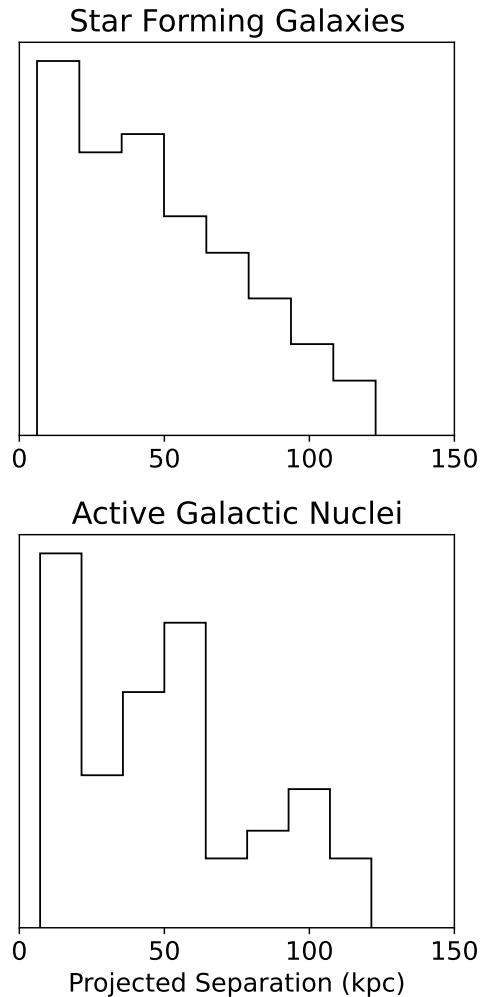


Figure 3.18: The distribution of both AGN and SFGs which have been cross matched with our confirmed galaxy pairs. While we find, as expected, the number of AGN and SFGs decreases with projected separation we find interesting structure within the AGN sample. Here, there appears to be a double peak in the number of counts of AGN. The first peak, between a projected separation of 0 and 25kpc is primarily from stage 2 galaxies while the second peak of 40 to 65kpc is of stage 3 galaxies in our sample. As shown in Figure 3.17, we expect the fraction of AGN to be similar in these two stages while they are at very different times in the dynamical timescale of the interaction. The third peak, from 75 to 125kpc is only from stage 1 and stage 3 galaxies.

stage 3. This could be evidence of a delayed AGN ignition depending on the underlying parameters of the interaction. Thus, **we find that the AGN fraction generally increases with stage, however, the mechanisms responsible for an increase in the fraction primarily occur when the two systems are at pericentre or are actually merging.** We find evidence of a delay of ignition, as there is an increase in fraction at early projected separations in stage 3.

3.7 Discussion

3.7.1 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., 2000b; Ellison et al., 2008b; 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 about to coalesce or are simply at the closest point of flying by each other.

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., 2013b; Karman et al., 2015; Moreno et al., 2021b). 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 stage 2 systems to stage 3 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 stage 3 systems (for further examples, see Li et al., 2008a; Scudder et al., 2012b). However, it is important to note that our stage 3 defined classification is a ‘wide net’ that captures many systems that may be very soon after the initial passage in the dynamical timescale. The criteria defining stage 2 and stage 3 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 stage 3 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 stage 3 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 stage 3 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., 2020b; 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 stage 2 / stage 3 to stage 4 galaxies. Stage 4 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 is completely at odds when compared to galaxies that move into stage 3 and do not merge. These stage 3 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 stage 4 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 stage 2 - where the two galaxies overlap and are at pericentre. From this point, if the galaxies then move off to stage 3 and escape, we will see a gradual decline in the SFE with the stage 3 galaxy only minorly increased in its star formation classification. Whereas, if the galaxies move into a stage 4 interaction and merge, 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.

3.7.2 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., 2007b; Ellison et al., 2013b; 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. Observations of interacting galaxies are often interesting, as they contain examples of duel 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 an increasing AGN fraction with both stage and 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., 2011b; Steffen et al., 2023). However, we specifically find that the AGN fraction increases rapidly from stage 1 to stage 2 and stage 3 to stage 4. Again, it holds quite constant between stages 2 and 3. This shows, again, shows two distinct end stages 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 interacting galaxies are actually flying past each other and being morphologically disturbed. There is evidence that the onset of nuclear ignition from interaction may be delayed (Ellison et al., 2011b), or even flicker (Schawinski et al., 2015) after ignition. Both of these possibilities are reflected distributions of AGN fraction with stage and projected separation.

Again, it is important to note the rather ‘wide net’ that our stage classification takes. There is overlap between our stage 2 and stage 3 classifications at high stage 2 projected separation and low stage 3 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 3.18 are two peaks. The second of these peaks is at the cross over point of stage 2 to stage

3; 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 3.16 where we see a slight increase in AGN fraction in stage 3. 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. Other works do also support this interpretation.

We also see a dramatic increase in AGN fraction in stage 4 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 stage 2, before the coalescence has even begun, but the actual coalescence is what sends it into overdrive. This matches observations mentioned previously of increased gas densities in nuclear regions and cores, and that stage 4 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 3.18 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.

3.8 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 stage 1 to 4 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 stage 1 to stage 2 the fraction of starbursting galaxies increases while the quiescent galaxy fraction reduces. In stage 4, 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 stage 1 to stage 2 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 10kpc. However, when broken down into their constituent stages, we find dramatically different behaviour in the star formation efficiency. This is best seen in stage 2 enhancement, which does not seem to change with projected separation, while there is a dramatic decrease in stage 3 enhancement from 2.5 at 10kpc down to 1.2 at 100kpc. 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 increases with stage. This is also true when related to the projected separation, where we find an almost bi-modal distribution. This is from a higher than average AGN fraction existing in the stage 2 and stage 3 samples, likely driven by stochastic inflows of gas during the interaction. 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 very much are. We have shown that one cannot simply use the projected separation of interacting galaxies as a proxy for the stage ongoing. We find very different behaviour in the star forming behavior of interacting galaxies based on stage they are at in the dynamical timescale of the interaction. We find the beginning of an enhancement in star formation occurs in stage 2, where the two galaxies are moving past each other at pericentre. Then, through to stage 3 the enhancement remains although with no further increase. There is then another dramatic starburst in stage 4, where the two galaxies actually begin to undergo the merging processes. However, the results are not so clear for AGN. While we find, definitively, that the AGN fraction increases with stage we are plagued by low counts and selection effects in these results. This is due to the limited classifications available of active galactic nuclei in our sample.

Our work shows the importance of considering morphological stage when considering interaction, and that there is a fine interplay between underlying processes and the point of the dynamical timescale. 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.

Acknowledgements

The Acknowledgements section is not numbered. Here you can thank helpful colleagues, acknowledge funding agencies, telescopes and facilities used etc. Try to keep it short.

This work utilised data from the Cosmic Evolution Survey (COSMOS) which has a digital object identifier of 10.26131/IRSA178.

Chapter 4

Galaxy Zoo: Mergers & Constraining Interaction

4.1 Abstract

Interacting galaxies are of fundamental importance to understanding galaxy evolution. Interaction affects multiple physical processes within galaxies but, most importantly for this work, leads to significant morphological and flux change. This morphological and flux change is driven by a host of underlying parameters, which are often used to recreate the morphological distribution of observed interacting systems. In this work, we combine a restricted N-body numerical simulation with a Markov-Chain Monte Carlo methodology to put constraints on these underlying parameters for different systems. We apply this pipeline to the best fit simulations of the previously successful interaction constraining methodology Galaxy Zoo: Mergers. In this project, citizen scientists successfully identified the underlying parameters of sixty two interacting systems. We then apply this pipeline to some test examples of observations of these systems. We find that we are able to put constraints on all underlying parameters of the best fit simulations, with the truth values all within our 2σ constraint. This pipeline runs in a fraction of the time of the Galaxy Zoo: Mergers project, and is fully automated. However, when applied to observations, we find our pipeline is more

restricted in its constraint and often fails even with those systems which were well constrained in the best fit simulations. We discuss the future of this methodology, and how this can be applied to large scale catalogues and surveys of interacting galaxies.

4.2 INTRODUCTION

Galaxy interaction and mergers are of fundamental importance to galaxy evolution. It has been shown that minor mergers likely contribute significantly to the star formation budget at the current epoch (Kaviraj et al., 2015), major interactions accelerate the process of rotationally-supported systems to dispersion-dominated ones (Hopkins et al., 2009b; Toomre & Toomre, 1972a; Yoon et al., 2022) and they likely have some link to nuclear activity (Byrne-Mamahit et al., 2023; ?; ?). However, the precise interplay between these observed processes and their underlying physical processes is difficult to quantify. To explore this interplay, we create and explore large, statistical samples of interacting galaxies. Large observational catalogues of interacting galaxies do exist (e.g Darg et al., 2010b; O’Ryan et al., 2023; Pearson et al., 2022b), however are often highly contaminated by galaxies which are close by projection effects rather than physically. It is also difficult to extract these underlying parameters from observations. Early investigations into individual systems used dynamical models (e.g., Barnes, 1988; Boylan-Kolchin et al., 2009; Hernquist, 1993; White, 1978) to recreate the morphology of interacting galaxies or merger remnants and began to explore and link underlying physical processes like increases in star formation history, or cosmic star formation rates with merger rates on a cosmological level. These effects are confirmed in observations of galaxies in the post-merger state (Di Matteo et al., 2008; Knapen & Querejeta, 2015; Mihos & Hernquist, 1994) and are particularly pronounced in gas rich post-merger observations (e.g, Hibbard & van Gorkom, 1996; Schweizer, 1982; Serra et al., 2019).

The clearest indication that two galaxies are interacting is from their disturbed morphologies. These can range from tidal tails, to stellar streams, to tidal bridges, to the warping of each galactic disk, to the disks complete destruction. From

numerical simulations, it has been found that the variability in these features relies heavily on the interactions underlying parameters. These parameters range from the mass ratios between the interacting galaxies, the orientation of the interaction, the galaxies relative sizes and the three dimensional velocity of the interaction. Early numerical simulations (e.g Hernquist, 1993; Toomre & Toomre, 1972a; Wallin, 1990; White, 1978) showed that the mass ratio has a major impact on the final morphology of different. This has been further explored in more modern works, and found to be consistent (Bournaud et al., 2005; Lotz et al., 2008a; Tanaka et al., 2023; Xu et al., 2012). Which parameters have an effect on increasing the star formation in interacting systems remains debated. Many works point to the mass ratio of the two systems being the primary driver of changes in star formation rates (Knapen et al., 2015; Li et al., 2023; Renaud et al., 2022), while others point to the kinematics of the interaction itself (Morales-Vargas et al., 2020; Xu et al., 2021). To resolve such debates, we must be able to link different system morphologies, flux distributions and star formation rates and with the responsible underlying parameters.

Many numerical algorithms have been built for this purpose, with examples being 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. Parameter space can often be greater than 10 individual parameters, meaning tens of thousands of models would have to be run to fully explore the parameter space. The parameter space has also been found to be degenerate, often with multiple parameter values able to describe the same system. For example, Smith et al. (2010) found that four of their best-fit models were able to accurately match the pair of interacting galaxies NGC 7715/5. Therefore, it is often preferable to use large cosmological simulations (e.g. Hani et al., 2020c; Hopkins et al., 2018; Schaye et al., 2015) to create large samples of synthetic interacting galaxies and analogues to the observed system found within it. 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) inter-

acting systems will remain unexplored without significant further computational expense.

The problem of direct comparison between simulations and observations was solved in a novel way by Holincheck et al. (2016a) in the Galaxy Zoo: Mergers (GZM) project (for details on Galaxy Zoo, see Lintott et al. (2008a)). GZM worked with citizen scientists to run the simulations and directly compare the outputs of them to observations visually. GZM studied a small sample of sixty two interacting galaxies from the Arp catalogue of interacting galaxies (Arp, 1966c). Citizen scientists' were given a selection of simulation outputs around an observation of one of the Arp galaxies. They would then choose the best fit one. The simulations would then be again with tweaked parameters and continue to select the best fit output. With enough citizen scientists on the project, enough new simulations run and enough time, GZM were able to fully constrain their interacting galaxy sample of sixty two; 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 a pipeline for automating the GZM process. We combine their fast restricted numerical simulation code with a Markov-Chain Monte Carlo (MCMC) framework and Bayesian statistics. While GZM was found the set of best fit parameters for their observed interacting galaxies, we constrain the probability distribution of each parameter and form a posterior of the likely parameter values. This approach allows us to marginalise over each parameter, provide their best fit as well as the error on each measurement. First, we apply this to the best fit simulations of GZM (where the true values of the ‘observation’ is known) followed by applying this to observations of a subset of the sixty two interacting galaxies, and compare the differences in results. We also discuss the limitations of this approach, with a particular emphasis on the computational expense required for our pipeline, and the potential future solutions.

The layout of this paper is as follows. In section 4.3, we describe the sample of sixty two galaxies we will study as well as give a brief description of the results of Galaxy Zoo: Mergers. Section 4.4 will breakdown our methodology

as well as describe our statistical approach, our simulation code and how it has been updated from previous iterations. A full discussion of our results will be given in section 4.5 followed by our conclusions and future work in section 4.6. Where necessary, we use a Flat Λ CDM cosmology with $H_0 = 70 \text{ km/s/Mpc}$ and $\Omega_M = 0.3$. Hereafter in this paper, 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 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.

4.3 DATA

4.3.1 Galaxy Zoo: Mergers sample

In this work, we use the same dataset as was defined in the Galaxy Zoo: Mergers project. This was a catalogue of sixty two major interacting systems whose tidal features are in the high surface brightness regime. Fifty five of the images were from the SLOAN Digital Sky Survey (SDSS) with the remaining seven coming from the Hubble Space Telescope (HST). A full description of each object is given in table 4.1 (note, this is very similar to that from Holincheck et al. (2016a), with decimalised coordinates and redshift data). The SDSS cutout of each system is given in figure 4.1, with the best fit simulation cutouts being found in Figure 6 of Holincheck et al. (2016a). Those images that were originally from HST are indicated by having a * against their name.

The best fit simulation images remain centered on the primary galaxy, with the coordinates of the secondary galaxy being used to calculate the size of the cutout. This could change largely between different cutouts, and therefore the resolutions of each image are not identical. These images were then reduced from their native resolution to 100×100 cutouts for use in our pipeline. This image size was found to be the best compromise between detail in the tidal features, and minimising the effects of using a numerical simulation with limited particle number. As our new pipeline also matches the flux distribution of each system, it requires the redshift of the interacting system. The redshifts for our sample were

4.3 DATA

Image Order	Name	SDSS ID	RA	Dec	z
1	Arp 240	587722984435351614	204.980417	0.835278	0.02250
2	Arp 290	587724234257137777	30.946317	14.72365	0.01171
3	Arp 142	587726033843585146	144.429583	2.763056	0.02329
4	Arp 318	587727177926508595	32.380516	-10.158508	0.0132
5	Arp 256	587727178988388373	4.710417	-10.369167	0.02730
6	UGC 11751	587727222471131318	322.247796	11.382539	0.02909
7	Arp 104	587728676861051075	203.037083	62.733889	0.01082
8	Double Ring, Heart	587729227151704160	238.287292	54.147861	0.040
9	Arp 285	587731913110650988	141.040000	49.226111	0.00967
10	Arp 214	587732136993882121	173.145221	53.067922	0.00331
11	NGC 4320	587732772130652231	185.740516	10.548328	0.02668
12	UGC 7905	587733080814583863	190.952917	54.900278	0.01648
13	Arp 255	587734862680752822	148.290000	7.870000	0.04106
14	Arp 82	587735043609329845	122.811250	25.193056	0.01368
15	Arp 239	587735665840881790	205.423852	55.672324	0.02489
16	Arp 199	587736941981466667	214.265833	36.573333	0.01024
17	Arp 57	587738569246376675	199.198750	14.424444	0.048
18	(HWB2016)Pair 18	587738569249390718	206.209583	13.921361	0.089
19	Arp 247	587739153356095531	125.889478	21.342976	0.01108
20	Arp 241	587739407868690486	219.461958	30.481222	0.03472
21	Arp 313	587739505541578866	179.418333	32.285556	0.01045
22	Arp 107	587739646743412797	163.069583	30.065278	0.03318
23	Arp 294	587739647284805725	174.931624	31.920108	0.00892
24	Arp 172	587739707420967061	241.389583	17.597222	0.029
25	Arp 302	587739721376202860	224.251667	24.612222	0.03286
26	Arp 242	587739721900163101	191.544583	30.727222	0.02205
27	Arp 72	587739810496708646	236.733750	17.878333	0.01100
28	Arp 101	587739845393580192	241.124946	14.800192	0.026
29	Arp 58	587741391565422775	127.990209	19.211523	0.03722
30	Arp 105	587741532784361481	167.804167	28.724722	0.021
31	Arp 97	587741534400217110	181.439583	31.068889	0.02305
32	Arp 305	587741602030026825	179.655833	27.490833	0.004
33	Arp 106	587741722819493915	183.902522	28.173576	0.02199
34	NGC 2802/3	587741817851674654	139.172619	18.963463	0.02914
35	Arp 301	587741829658181698	167.470000	24.259722	0.02059
36	Arp 89	587742010583941189	130.665852	14.285624	0.00687
37	Arp 87	587742014353702970	175.185000	22.437778	0.02373
38	Arp 191	587742571610243080	166.834167	18.431111	0.02739
39	Arp 237	587745402001817662	141.933458	12.286750	0.02899
40	Arp 181	587746029596311590	157.118946	79.818129	0.026
41	Arp 238	58801124116422756	198.886667	62.126944	0.03106
42	MCG +09-20-082	588013383816904792	181.161667	52.956111	0.113
43	Arp 297	588017604696408086	221.330417	38.761389	0.0298
44	NGC 5753/5	588017604696408195	221.328663	38.805889	0.01374
45	Arp 173	588017702948962343	222.869434	9.328297	0.028
46	Arp 84	588017978901528612	209.649167	37.438889	0.01158
47	UGC 10650	588018055130710322	255.060770	23.106346	0.00986
48	Arp 112	588018055130710322	255.060770	23.106346	0.00986

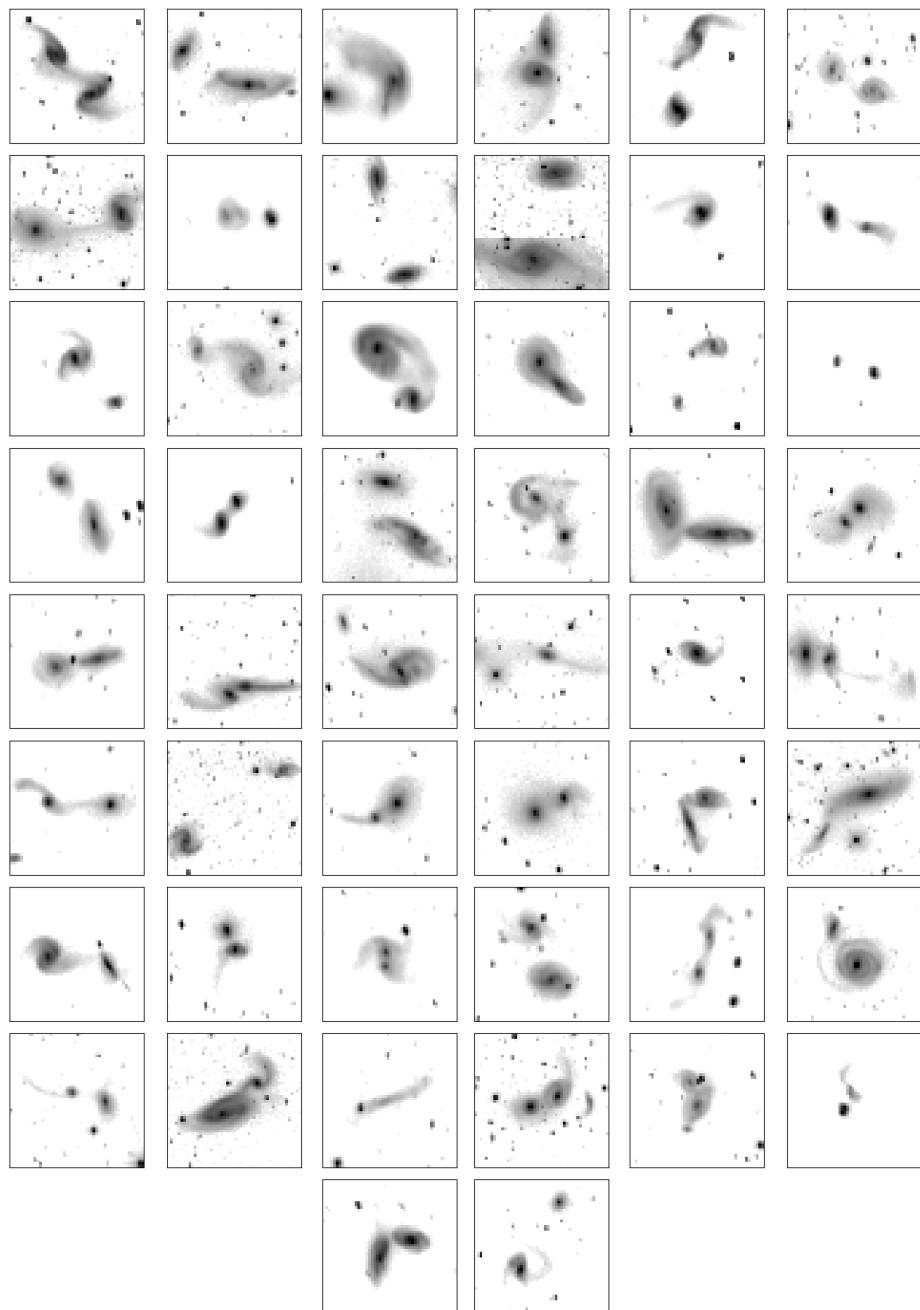


Figure 4.1: Cutouts from SDSS of the sixty two interacting systems that we test our pipeline on. For the best fit simulations see Figure 6 of Holincheck et al. (2016a).

found from the NASA Extragalactic Database. The effects of incorrect/missing redshift are discussed in section 4.5.6.1, and the effect of incorporating it into a population parameter space explored.

The full range redshift range of our sample is $0.003 < z < 0.113$. To follow GZM, we also assume that the secondary galaxy in each system is on its first pericentre passage and before the interaction began each galaxy in the interaction was disc galaxies. Therefore, we are assuming that the merger history of each galaxy is merger-free. We also only look at systems where two galaxies are involved.

4.3.2 Galaxy Zoo Mergers: Constraining Interaction

The aim of the GZM project was to find the best fit parameters and recreate the morphology of each of the sixty two interacting systems. This was achieved using an efficient restricted numerical simulation algorithm and the help of citizen scientists. Here, we will briefly summarise their methodology.

They created each of their observational images of interacting galaxy systems using the SDSS ImgCut service, based on the Lupton et al. (2004b) 3-colour image creator. Each image was centered on one of the two galaxies involved in the interaction, the primary galaxy in the interaction. This image was then converted into a grayscale image, though any citizen scientist involved in the project would have access to the 3-colour image as well. By converting the SDSS pixel scale to physical distance with the measured galactic redshift, the position of the secondary galaxy was then found. The approximate z-position of the secondary relative to the primary could also be estimated from the redshift measurements. By measuring the luminosity of each system, an approximate mass could be calculated for each. Finally, the inclination of each system was estimated. Each of these approximations were used as a starting location for how they would explore the underlying parameter space. By reducing their overall parameter space around these values, they could sufficiently reduce the size of the parameter space to explore to put more concrete constraints on their sample.

To achieve estimates on the best fit values for each observational image, this project involved the help of thousands of citizen scientist. A citizen scientist

Parameter	Description	Conversion(Spec. Units)	Parameter Range(Spec. Unit)
r_z	Secondary z-position	15kpc	-300kpc - 300kpc
v_x, v_y, v_z	Secondary velocities	169.34km s ⁻¹	-1693km s ⁻¹ - 1693km s ⁻¹
M_1, M_2	Total Masses of Galaxies	$10^{11} M_{\odot}$	$1 \times 10^9 M_{\odot}$ - $3.5 \times 10^{12} M_{\odot}$
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 Closest Distance	57.7Myr	173.1Myr - 1.2Gyr

Table 4.2: 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.

would be given one of the sixty two observational images (in both grayscale and 3-colour), with eight simulation outputs around it. These eight simulations would have different underlying parameters which varied about the approximated parameter values. These simulations would, therefore, have differing morphologies and may not match the observational image at all. The citizen scientist would then compare each of the simulation outputs to the observation, and pick the one which they judged to be the most similar. The parameters of all eight surrounding would be tweaked again and the simulation outputs recreated (one recreation would have similar parameters to the output previously selected). This would be done multiple times, with the idea that each selected simulation was gradually becoming more similar to the observational image. Thus, over time with enough volunteers and enough selections the underlying parameters of the observed image were constrained and the best fit values found. The best fit parameters for each interacting system can be found in the GZM website¹.

In this work, we attempt to automate this workflow. We directly compare simulation outputs to the observed image using a χ^2 measurement in a MCMC sampling algorithm (see Section 4.4). As this is fully automated, and therefore much quicker than visually checking each image, we differ from GZMs methodology by exploring the entire parameter space. We don't attempt to measure any approximations of the galactic parameters, but constrain them directly from the

¹<https://data.galaxyzoo.org/mergers.html>

full parameter distribution. Table 4.2 shows the full parameter space we explore in this work for each system.

4.3.3 Observation Preparation

Our method of observational cutout preparation differs slightly from that of GZM and therefore we detail it here. As stated, the majority of the GZM systems are found in SDSS. We downloaded the FITS files from SDSS which contained the full system of interacting galaxies. These were then used to create smaller cutouts of the systems. The central coordinates of primary galaxy as started in GZM were used as the centre of the cutout. 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 filter of SDSS (ugriz), keeping the image dimensions equal through all filters.

As our pipeline constrains the images in terms of counts, we convert each cutout into counts. Using the conversion value in each FITS header, we converted each image from nanomaggies to counts. 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 that would be used in our pipeline. Each cutout was visually inspected to ensure that, even with the reduced resolution, the tidal features were still clear and prominent in the image.

Now following GZM, we took the 0.396" pixel scale of SDSS on the sky and used the measured redshifts of each system to convert it into physical sizes. As we know the pixel-to-distance conversion for each image, the position of each secondary galaxy was then calculated from the central pixel. This was judged by eye. No attempt was made to approximate the z-position of the secondary galaxy, as this is a parameter to constrain in the pipeline. We made these preparations using the Astropy Python package (Astropy Collaboration et al., 2013b, 2018b).

4.4 METHODS

The cornerstone of our methodology is to utilise Bayesian statistics. We will not only find a best fit value for each simulation parameter, but will quantify the posterior distribution around it and provide a diagnostic on the goodness of the fit. In this way, we do not have to conduct any post-processing to find the uncertainties in our parameter values. The uncertainty will be found simultaneously with the parameter value. To explore parameter space, we utilise the MCMC Python package EMCEE combined with a modified version of the underlying code of GZM. We take the input parameters to the simulation as those we are trying to constrain. This translates to a 14 dimension parameter space to explore. These are listed in Table 4.2.

4.4.1 Simulations: APySPAM

The original GZM project utilised the numerical simulation code Java Stellar Particle Animation Module (JSPAM) (Wallin et al., 2016b). For an indepth description of the underling code, we direct the reader to Wallin (1990); Wallin et al. (2016b) but give a very brief description of the base code here. JSPAM is a restricted N-body code focused on recreating the morphology of interacting systems. It approximates the interaction as a restricted three-body problem by approximating the interacting system as a set of massless test particles with two massive bodies moving through a galactic potential. The simulation is computationally efficient, being able to run thousands of particles in a matter of seconds on a regular work PC. The primary particle integrator is a fourth-order Runge-Kutta which updates the velocities and positions of particles over a fixed timestep.

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, and updates their position and velocity. Therefore, dependent on which model used, will slightly change the output morphologies. We elect to utilise the softened point mass approximation, as this has improved computational efficiency over the higher accuracy of N-body approximation. Upon testing, we do note that the morphologies differ in our results,

but they do not differ enough to make recreating our results dependent on the selected force model.

The base code of JSPAM was purely a morphology matching code, which has been shown in prior works to be very difficult to use in an automated fashion to constrain interaction (e.g. Barnes & Hibbard, 2009). JSPAM itself has been used in multiple genetic algorithms to find the best fit parameters of different systems. However, this lacks the exploration of parameter space and the quantification of uncertainties we aim to achieve in our approach. Utilising both morphology and flux matching between mock and observations has been shown improve accuracy of constraint (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. To preserve computational efficiency, this is applied in a semi-analytic fashion. We have also created this enhanced version in Python 3.7.4 (hence, we shall refer to this algorithm as Advanced Python Stellar Particle Animation Module, APySPAM).

4.4.1.1 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 (Boquien et al., 2019; Johansson et al., 2009; Pacifici et al., 2013; Simha et al., 2014). However, to capture any star formation due to the galaxy starbursting from the interaction, we add further star formation based on the conditions in the interaction (see section 4.4.1.2). We assume the onset of star formation is $\tau = 1.5\text{Gyrs}$. Our simulation outputs a spectra normalised to $1M_{\odot}$ with each particle assumed to be the same age as the galaxy. We then scale each spectra to the assumed mass at the position of the particle

Upon initialisation, we assign each massless test particle total baryonic mass. This does not impact the calculation of the force upon it during the interaction, but is only used in the calculation of what star formation may occur at its position.

The mass is assigned by calculating the total baryonic mass in the galaxy and distributing it equally for each particle in the galaxy. These particles are then distributed as an exponential disk around the galactic position. We therefore create an exponentially thin disk of particles with equal mass about the galactic centre, where the expected declining baryonic mass of the galactic disk is captured by the particle spatial distribution.

In the underlying simulation, there are three assumed components of the user inputted mass: the bulge mass, the disk mass and the halo mass. In order to assign a correctly scaled spectra to each particle, we follow the prescription as stated in Wallin et al. (2016b) to distribute the mass between these components. They state that the mass distribution is $M_{\text{bulge}} = 0.05M_{\text{galaxy}}$, $M_{\text{disk}} = 0.14M_{\text{galaxy}}$ and $M_{\text{halo}} = 0.81M_{\text{galaxy}}$. We assume that the bulge and disk masses are fully baryonic, with the remaining mass being the dark matter halo. We only utilise the baryonic mass for the star formation at each particle. We further divide the 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, but can be updated by the user.

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 over, 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.

4.4.1.2 Star Formation

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., 2009b). However, in our simulations we must approximate the behaviour of the starburst 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 CIGALE (Boquien et al., 2019) algorithm, and where we base our process from. Here, we detail this enhancement parameter.

At any given timestep the total stellar mass is given by

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

Here, τ is the e-folding time of star formation (set at $\tau = 1.5 \text{Gyrs}$), 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

$$\beta = M_{\text{ratio}} D_{\text{ratio}}^2. \quad (4.2)$$

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. If our interaction has one galaxy being significantly more massive than the other, then when we calculate the enhancement for the massive galaxy the mass ratio parameter will rapidly go to zero. If vice versa, it will become much larger. Therefore, 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 galaxy's

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. However, as the point of closest approach gets further away, the star formation will approach that of a non-interacting galaxy. It is important to note, however, that this has significantly less impact on strengthening star formation than that of the mass ratios.

This parameter successfully reflects 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; Li et al., 2008b; Lotz 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, 1996b; Rodríguez Montero et al., 2019b; Springel, 2000b) 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. This is a computational expense we cannot afford.

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,

$$SFR_{\text{Particle}} = \frac{M_{\text{gas,Particle}}}{M_{\text{gas,Galaxy}}} SFR_{\text{Galaxy}}. \quad (4.3)$$

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 disc galaxy prior to the interaction when we assign gas masses to each particle. This may not be true if either galaxy has an intense merger history before the interaction we are attempting to model. We also assume that all of the gas mass assigned can be used in star formation, whereas only cold molecular gas can be used in forming stars. We make no account of gas ionisation or the intense turbulence in the ISM that likely occurs during these interactions. We also assume that the disruption occurring to the massless 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 using the methodology defined in Section 4.4.1.1.

4.4.2 Flux Distribution

There remains a problem with our simulations, however. 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. This is a result of the limited number of particles not covering each pixel which within the galaxy. Therefore, to mitigate this effect, we calculate the flux at each particle position and then 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 users choice. These convolved SEDs are then integrated to give an absolute value of the flux in counts at each particle. We then create a grid of pixels where the distance between the centre of each is the physical distance in the image. This is calculated from the pixel physical scale and image scale as described in Section 4.3. We then calculate the flux contribution at each pixel from every particle in the simulation. We then stack these images together into one final image. Using the known redshift, we then calculate the flux we would observe from each pixel. After this step, any pixels where the total counts within them are found to be less than one are set to zero.

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 probabilities being calculated, even when the simulation itself has reproduced the tidal features very well. As a result, we set the value of any particle with no neighbour within 10×10 pixels to zero.

4.4.2.1 Impact on Computation Time

These new algorithms have been added to the original JSPAM code while preserving computational efficiency. The choice to create an interaction constraining

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 4.3: 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.

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 run-times of JSPAM and APySPAM are shown in table 4.3. As shown here, even with our extra flux calcualtions, our simulation code outperforms the original JSPAM code by several times. 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.

For our purposes in an MCMC, we must run the simulation 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 4.4.2, we can gain better resolution while not sacrificing computational efficiency. By distributing the particle flux rather than binning it, we do find we introduce the most computational overhead. We have found that this part of the algorithm is not highly effected by particle size, but mainly by image size. The timing calculations shown in Table 4.3 used an image size of 100×100 . When increased to 500×500 , the computation time for 2,500 particles increased dramatically to over 30s. Therefore, it is imperative that the user keeps this in mind when selecting cutout size.

4.4.3 Defining the Likelihood Function

4.4.3.1 Bayes Theorem

We combine APySPAM with a Markov-Chain Monte Carlo (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, calculating the likelihood at each walker position. They then compare the likelihood between the old and new position, and if it the likelihood is higher they move. If not, they remain in place and attempt to go to a place of higher likelihood. In this way, 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 underlying 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.

The MCMC algorithm we use is the Python package EMCEE (Foreman-Mackey et al., 2013); this is an ensemble MCMC package with numerous pre-defined moves and algorithms to make getting to the area of high likelihood more efficient. We construct contours of increasing likelihood and calculate the errors and probability distribution of our best fit measurement; i.e. we can construct a posterior for each of our parameters.

We define a likelihood function to compare the observation images to simulated mock observations. By Bayes Theorem, the probability that a set of underlying parameters which produced a mock observation also describe the observed image follows equation 4.4,

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

$P(H_i|D_{obs}, C)$ 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 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 constraints $P(D_{obs}|H_i, C)$, all divided by a normalisation constant, $P(D_{obs}|C)$.

4.4.3.2 Simplifying the Prior

In order to simplify this expression, we make assumptions about the underlying parameter space to increase efficiency and simplify our computations. We first assume uniform priors for each of our thirteen 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 4.2. These ranges can be tweaked, or a different prior function defined, by the user. Therefore, the priors part of equation 4.4 can simply be made equal 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. (2016a) (there it is called β but we call it γ here to not be confused with our star formation enhancement parameter)

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

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 Holincheck et al. (2016a), this parameter is used to directly filter out simulations where tidal distortion is unlikely and to not show the volunteers lots of featureless simulations. In our case, we use it to inform our prior as the MCMC continues. This significantly enhances the efficiency of the pipeline. 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

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

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.

4.4.3.3 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

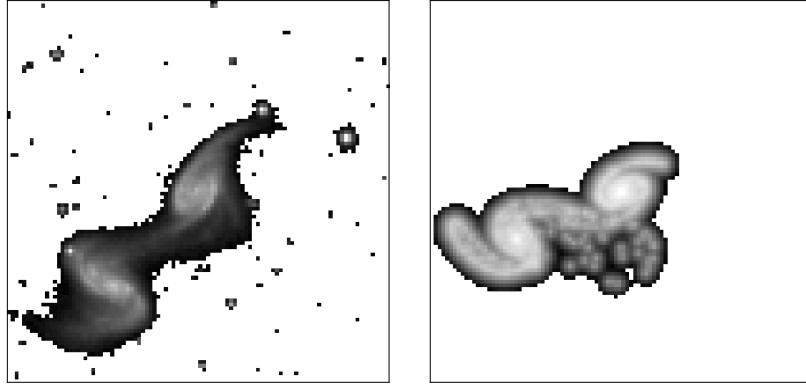


Figure 4.2: The example system used to test our pipeline: the Arp 240 interacting system. This system is considered an easy one to constrain. It is composed of two clearly distinct galaxies, with strong tidal features that our pipeline can fit. These tidal features are the two tidal tails formed in the 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. (2016a) and the first test image used in our pipeline. The different in scale and orientation are discussed below.

compare our mock observations to our observed data;

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

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;

$$\chi^2 = \frac{1}{N - n_{dof}} \sum_0^{N_x} \sum_0^{N_y} \frac{(p_{x,y} - q_{x,y})^2}{\sigma_{x,y}^2}. \quad (4.8)$$

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 sigma value of the (xth,yth) pixel; this is the uncertainty in the observed images

pixel value and is the same as that which is defined in GALFIT (Peng et al., 2002; Peng et al., 2010). This is then summed over all pixels in the entire image, giving us a single χ^2 value between each simulated image and observed image. For now, we define the σ image in the same way as described in GALFIT.

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 mock observation image also describe the observed image input into the algorithm,

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

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

4.4.4 Architecture for Exploring Parameter Space: EMCEE

As stated in the previous section, we utilise the MCMC code EMCEE (Foreman-Mackey et al., 2013) to explore parameter space of interaction. For full details of EMCEE and the different modes that it can use, see the extensive readthedocs¹; but here we will briefly state the hyper parameters that we used.

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 dealing with 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 GitHub². Here, a user can download our setup to reproduce our results, or to update the model for their own purposes.

¹<https://emcee.readthedocs.io/en/stable/>

²<https://github.com/AstroORyan>

4.5 RESULTS & DISCUSSION

We now apply our described pipeline to the best fit simulations from the Galaxy Zoo: Mergers project. While showing results for every system cannot be done in this paper, they are presented online¹. Here, we discuss the constraints on a specific system: the Arp 240 system. We discuss how the constraints could be improved, and what extra information we required for this. We explore the results we find when applying to all simulations, only presenting the results of our three best and our three worst fits, while discussing the trends we see in the remaining results. We briefly present our pipeline applied to observational images. We discuss the limitations of our approach, and how they could be improved upon.

When discussing our pipeline applied to observations, we only look at our best fit subset from applying the pipeline to the best fit simulations of GZM. The computational expense must be significantly increased to make constraints on the observational systems and for our MCMC to reach convergence. We compare our found best fit values and uncertainties to any measured values in the literature, 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.

4.5.1 Testing on a Single System: Arp 240

We apply our automated pipeline to the best fit output simulations of GZM. Specifically, to estimate our performance, we test on the best fit simulation of the Arp 240 system. We elect to use this system as it 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. Our prepared observation and the best fit simulation from GZM are shown in Figure 4.2.

We first apply it to the best fit simulation as this is an excellent noiseless example. This synthetic observation is created with 10,000 particles, and a high time resolution of 0.57Myrs per timestep. We then attempt to constrain this

¹All results are found here: [Link_to_results](#)

synthetic observation by running the 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 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 mock observation. The contours then correspond to 0.5, 1.0, 1.5 and 2-sigma levels (these are default values). Figure 4.5 shows that each of the truth values is within 1-sigma of the peak of the probability distribution with the exception of the z position and time. We also see multiple maxima of probability in the orientation space of the galactic disks.

We get excellent constraints on the masses of the two galaxies. The Arp 240 interacting system is the most massive system in our sample, and therefore we expect the mass at the very limits of the range we are exploring. Our pipeline the two masses the easiest to constrain out of all the parameters, and this is the first parameter which converges in the MCMC. We provide the algorithm with the secondary two dimensional position, and therefore it has to only fit the flux distribution of the inner disk correctly to get good constraints on the galactic masses. It knows where to place the secondary and then just has to match the pixels on the inner parts of the galaxy to quickly increase the likelihood. While the formation of the tidal features is dependent on the mass ratio, we have found that they are also very dependent upon the orientation of the interaction as well as the relative sizes.

We can see that the relative sizes of the disks is important for the constrainingt he tidal features as there is a very large spread in the constrains. 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

4.5 RESULTS & DISCUSSION

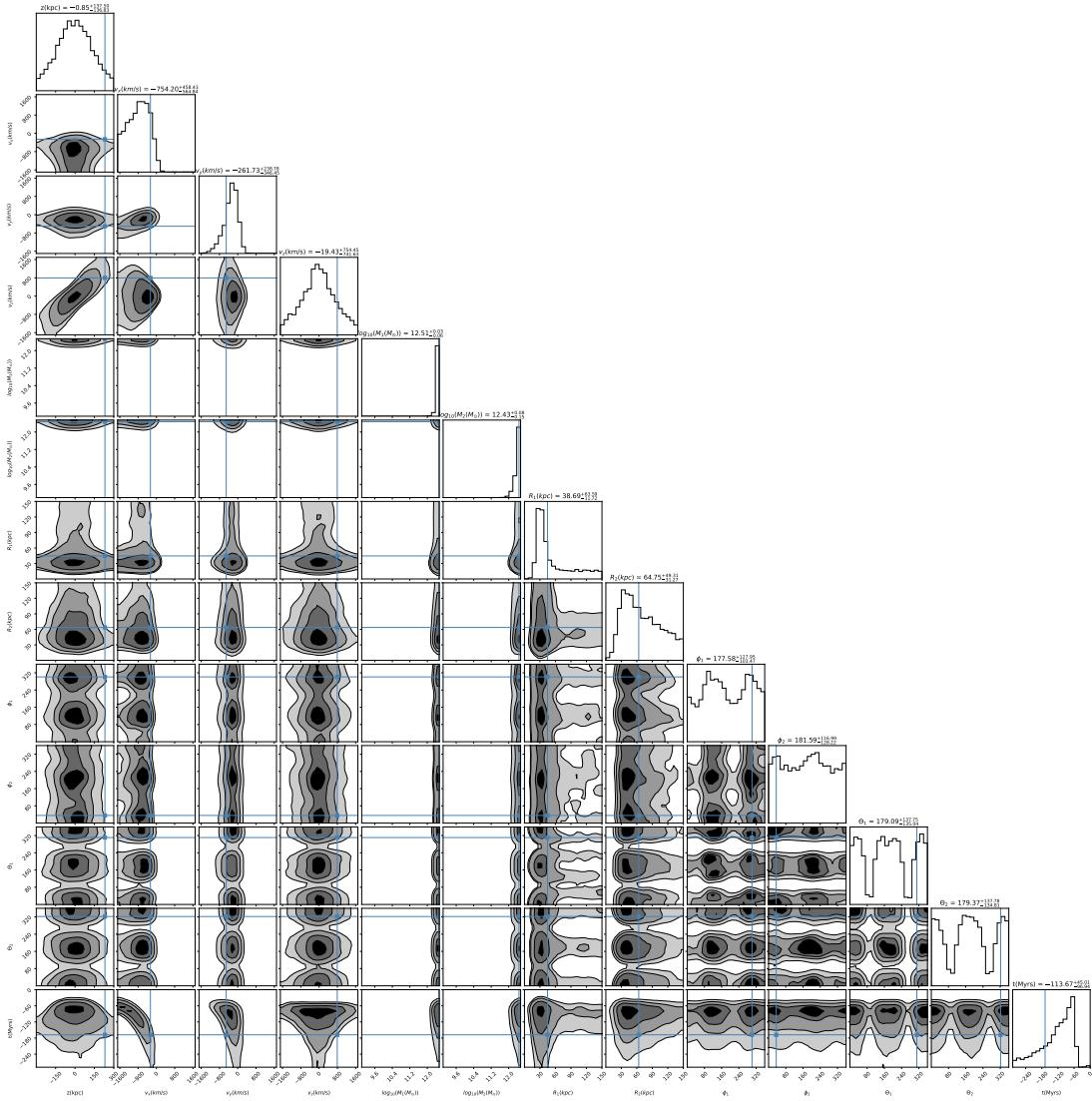


Figure 4.3: Corner plot showing the constraints made on all thirteen parameters we are exploring. Contours show the 0.5, 1, 1.5 and 2-sigma levels of the constraint. this corresponds to containing 11.8%, 39.3%, 67.5% and 86.4% of the samples across all walker chains. Displaying our results using the 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 find the full corner plots of each system, find them at the results website for this paper.

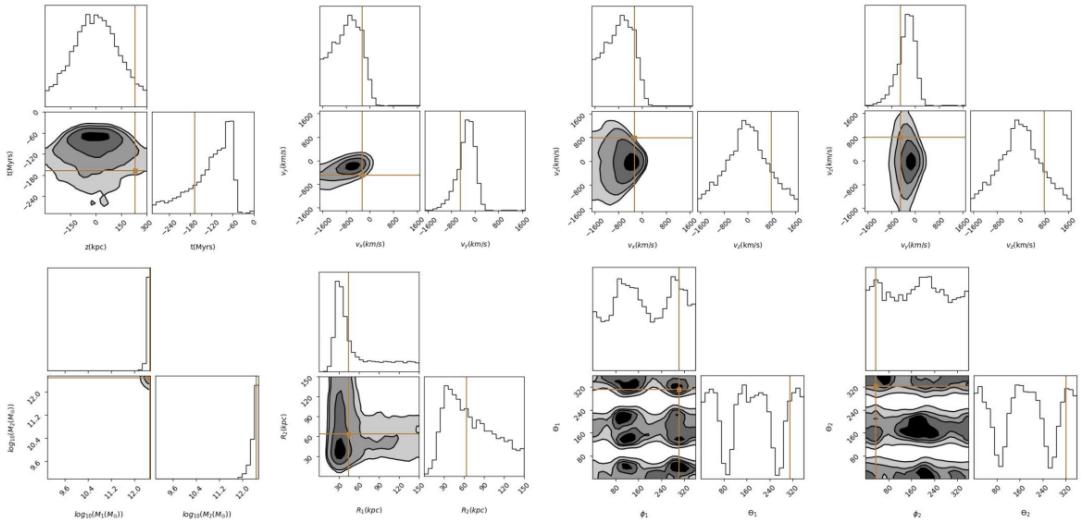


Figure 4.4: Same as Figure 4.3, but reduced to only matching parameters.

pixels within the galaxy being missed than a pixel containing galaxy in it when it shouldn't. We therefore have a bias effect where the peak of probability drifts towards just below the true value of the radius. However, the true value of the radius remains within 0.5-sigma of the found contours, and therefore this acts as a good approximation.

The secondary galaxy has significantly larger uncertainty of its radius. The large tail in the 1σ in the marginalisation shows this. 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 (by ± 1 bin in the x or y direction). This change transfers further uncertainty in the secondary disk size. However, once again, the best fit values found are within 1σ 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 .

It is important to note here that ϕ , in our simulation, 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 three 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 actually 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, a more subtle approach will be needed to solve the degeneracy in θ .

The lack of three dimensional information also affects our constraints in the z-direction: the z-position and the z-velocity. As seen in Figure 4.5, the algorithm fails to constrain the z-position, while there is significant uncertainty in the z-velocity. First, the z-position is difficult to constrain as we lack three dimensional information. The simulation is run in the reference plane of the primary galaxy, therefore the secondary can only be behind or in front of the primary galaxy. 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 z-velocity remains largely unconstrained due to similar reasons as above. The true value lies at the very edge of the probability distribution found, at

approximately 2σ . This would, again, be rectified readily by introducing velocity information into our constraints. The simulation works so that the secondary galaxy always is at the same x and y position that is defined by the user (or calculated from the observation). Therefore, an output simulation with a positive or negative LOS will have the same flux distribution. Hence, this constraint simply peaks about zero for the z-velocity.

Our pipeline works significantly better, however, with the x- and y-velocity of the secondary galaxy. We find the the truth value of this is within 1σ 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 pipeline is informed by the flux distribution of the resultant and gives an excellent constraint on the parameters.

Finally, we discuss the attempting to constrain the time of the interaction. Shown here is not the total time of the interaction, but how long ago the time of closest approach was. 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 calculating 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. I.e., 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 2σ . 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 had the highest log probabilities throughout



Figure 4.5: Simulations from the areas of parameter space that lay within the 0.5σ of our constraints. *Top:* The best fit simulations from this parameter space. *Bottom:* The worst fit simulations from this parameter space. Our pipeline has 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 wildly different tidal features in the areas of high probability. Therefore, identifying specific systems with the sought after tidal features remains needing to be a manual process.

each walker chain and take the top 5 as an illustration here. What we find is that our pipeline is not able to precisely reproduce the best fit simulation from GZM. 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 pipeline 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 constraint pipeline. Without the ability to filter the simulations quickly, the parameter space is simply too large to fully explore in a computationally reasonable time. 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 some amount 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 2σ of the true parameters. The only missing parameter is in the secondary ϕ parameter. There is significant degeneracy in constraining the orientations of this interaction, but this is not unexpected. 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 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.

Figure 4.5 clearly shows that ϕ_2 has not converged, given the marginalised posterior is has very limited structure with two insignificant peaks. With the remaining three orientation measures, this is more complicated. Here, we have a two or even four fold degeneracy in the output models due to three dimensional information not being available. This disrupts our Geweke diagnostic measure, which is looking for a single peak in parameter space. Therefore, by folding the parameter space over and only exploring over $0^\circ - 180^\circ$, we achieve a single peak in parameter space. These remaining three results then pass the Geweke diagnostic

test. The failure of ϕ_2 to pass the Geweke test means that more steps in each chain must be run to reach true convergence.

The second diagnostic we check is that the walkers have fully explore 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, 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. Simply 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, as stated previously, the orientations do not have a massive impact on the flux distribution of the disks of the output interacting system. The Arp 240 interacting system is face on, and therefore the degeneracies form where the disks are face on and the empty areas are where they would be edge on. 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 the cost of little improvement for the remaining parameters. This is also at the cost of much larger computational expense. 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

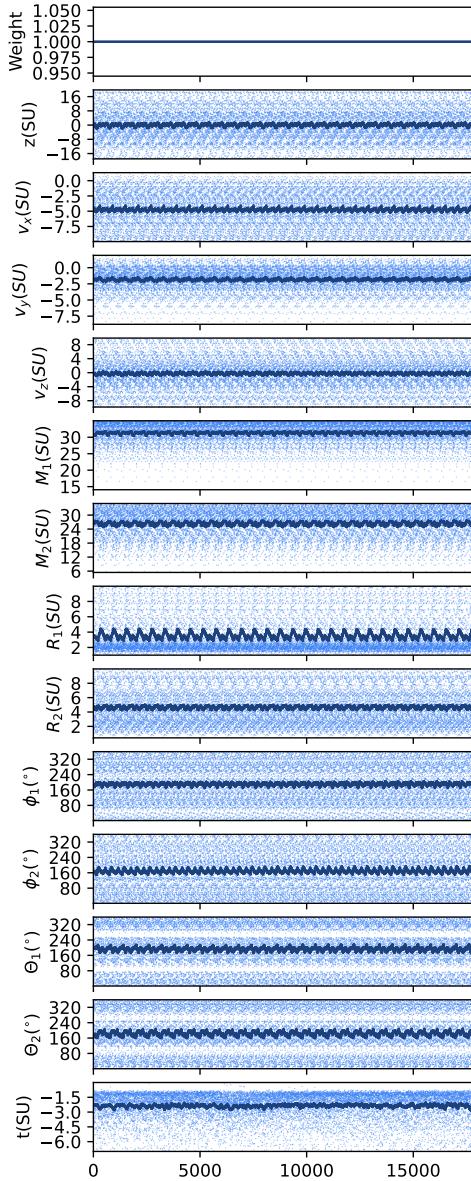


Figure 4.6: Steps taken by each walker in our MCMC chain to constrain the Arp 240 best fit simulation. Note, the y-scales here do not extend over the full parameter space for some galaxies, and only show where the walkers have stepped after the burn-in phase. The deeper the blue, the more walkers have 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 , z -velocity and ϕ_2 parameters show significant 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.

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 the GZM sample galaxies. Therefore, we run this test using the same best-fit simulation but taking 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 in the GZM sample have associated IFU data in order to get LOS velocities to incorporate 3D information into our fitting pipeline. Therefore, we once again test incorporating these by using them with a best fit simulation of Arp 240. From these, we simply sum 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 allowed us to completely constrain the z-velocity.

Figure 4.7 shows the new measurements of the constraint on the z-position, y- and z- velocities. We can immediately see that the constraints on each of these parameters is significantly improved. This is with the same number of MCMC walkers and steps remaining the same from running without. For the y- and z- velocities, the constraints are improved to the point where we completely recover the true underlying parameter values within 0.5σ . With the z-position, we also gain significant constraint. The pipeline is 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.

Other parameters are not shown in Figure 4.7 as there was no change in their constraint after adding in velocity information. This was unexpected with the orientation constraints of the galaxies. However, it is important to note 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 would need higher resolution on our velocity map than the simple 100×100 binning we used. We would also need further information about the rotation, and get accurate measures of the bulk motion of the tidal features in the galaxy. Currently, the simulation is

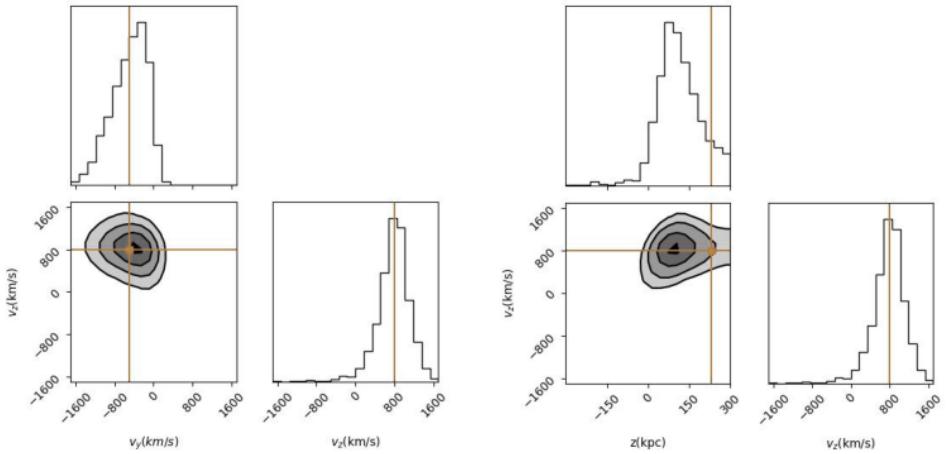


Figure 4.7: The constraints on the 3D velocity and z-position when including velocity in our MCMC pipeline. To add this extra information, we ran our best fit simulation of Arp 240 and summed the LOS velocity (z-velocity) in each pixel of our image. Comparing between here and Figure 4.5, we can see that we achieve complete constraint of the z-velocity of the interacting galaxy. We also significantly improve the constraint on the z-position parameter as the pipeline is able to distinguish which side the secondary galaxy is of the primary.

not able to accurately reproduce this outwith simply assuming circular velocities at different radii from each galactic centre.

This shows the improvement that adding velocity information to our methodology could bring, and how far interacting galaxy simulation and constraint can go once we incorporate integral field unit (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. To significantly improve our constraints, we will need this information of the tidal features of these systems. IFUs with larger fields of view are soon to come online, such as WEAVE (Dalton et al., 2014). When this does, we will be able to fully apply the velocity and rotation information of these systems to this methodology, and reach a true conclusion of the improvement we could reach.

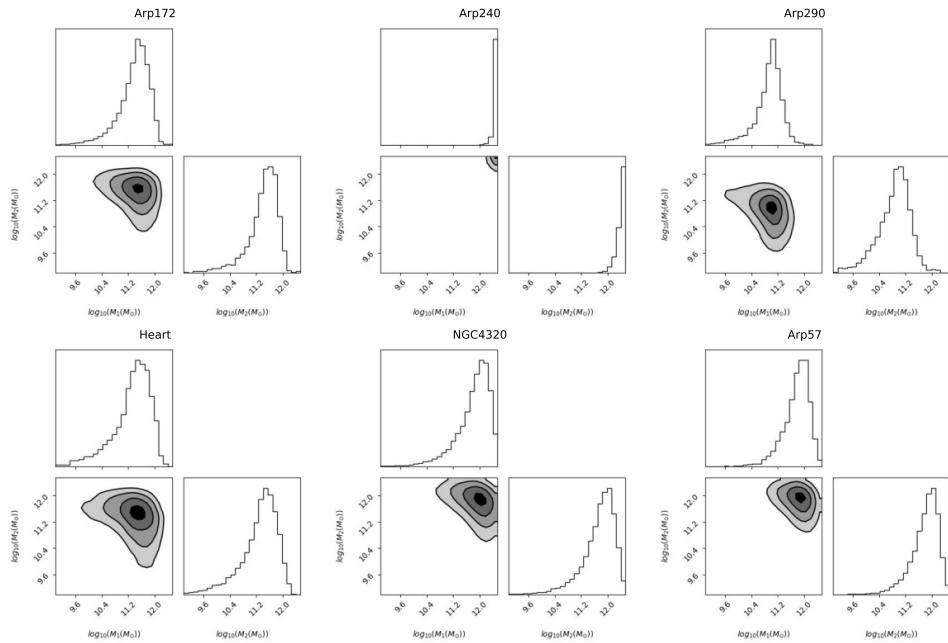


Figure 4.8: Mass corner plots for our best and worst fits from the pipeline. This was judged by the FWHM of the marginalised posteriors. *Top*: Our top three best fits. *Bottom*: Our worst fits. Even with our worst fits here, the masses are very well constrained.

4.5.4 Recovering the Galaxy Zoo: Mergers Results

With constraints being ascertained on nearly all parameters of Arp 240, we then applied our pipeline to the remaining 61 interacting galaxy systems of the Galaxy Zoo: Mergers sample. The reader is invited to see the resultant corner and reduced corner plots of each system on the website. Here, we will detail observations and trends from applying our pipeline to multiple systems. First, our ability to make constraint 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 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. This constraint on our three best fit and three worst fit systems are shown in Figure 4.8.

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. Each of these systems represent these three limitations, respectively. It is relatively easy to see why these limitations lead to difficulty in our pipeline. If two cores are close to coalescence, the flux distribution will appear similar to two overlapping disks. Therefore, our pipeline will calculate equivalent likelihoods between systems with little to no interaction and a merger with multiple flybys undergoing coalescence. Our pipeline is also not yet designed to account for multiple flybys, a key assumption being that these systems are only on their first passage. Our pipeline requires tidal features to make fits on the flux distribution. NGC 4320 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 pipeline 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. While our pipeline is able to reproduce the tidal feature, it is unable to reproduce the flux distribution as found in the original Galaxy Zoo: Mergers project.

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. For every single system, the worst fit parameters were the four orientation parameters for the two galaxies. Aside from the points that have already been raised regarding the problems of our pipeline constraining these, a further problem was that the likelihood calculation is dominated by contributions from the mass and velocity parameters. As long as the shape of the final system is correct, the orientations only alter the likelihood by between 0 - 20 points. Whereas the mass can be the difference in hundreds. There are solutions to this, however. Further walkers and a longer chain can be run to strengthen these constraints, but this comes at the cost of far more computational expense. Second, only a subset of the orientation space could be

Table 4.4: NOTE: To be added as the MCMC runs finish. Or, may go online!
Actually, likely to be a machine-readable table.

explored. Due to the 2- to 4- fold degeneracy in the orientations, only a small section of the orientation parameter space needs to be explored. For the purposes of this work, we show the full degeneracy and what is to be expected throughout all of the parameter space.

The remaining parameters, and their level of constraint, lie between our best case scenario of the mass constraint and the worst case scenario of the orientation constraint. However, the importance here is that we are able to reproduce the Galaxy Zoo: Merger best fit values and uncertainty measures without the need for human interaction. Thus, this pipeline and methodology can be applied to larger samples of galaxies than we have presented here. Table 4.4 gives a full breakdown of our results for each system. This includes the best fit value for each parameter from the MCMC runs as well as the errors upon them. Comparing with the GZM best fit results (as an accompanying machine readable table), shows where improvement is needed but also that for nearly all parameters are pipeline is successful.

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. There thus comes 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. However, this will come at an increased computational cost.

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 pipelines such as these is never an easy one. We, therefore, use the 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 4.4. We reiterate here that the cutout resolution is reduced from its native resolution to images of 100×100 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 the cutout sizes used by GZM in their work. As an example, we create the cutout of Arp 240 at 600×600 pixels at native resolution. With SDSS data, this corresponds to a physical size of $111.50\text{kpc} \times 111.50\text{kpc}$ at the redshift of Arp 240. The best fit simulation from GZM for Arp 240 is $785.35\text{kpc} \times 785.35\text{kpc}$. 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 in GZM. As a result, the parameters we will find for constraining the observation will be very different from the values found in GZM.

We apply our pipeline 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 best fit simulation to reach convergence. Second, due to the smaller scale size of our images, the defined γ

parameter is much stronger than with the GZM sample. Both of these reasons for higher computational expense are to be expected. We are now constraining over images with noise, rather than the noiseless images of the best fit simulations. The increase in the γ parameter also means that we are filtering out less candidate systems and running the base simulation code more often. This directly translates into a higher runtime for our pipeline.

Figure 4.9 shows the reduced corner plot for the observed Arp 240 system. The constraints on the velocity parameter have significantly worsened when compared to the constraints 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 0.5σ , 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 retain our precise constraint on each galaxy's mass. 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

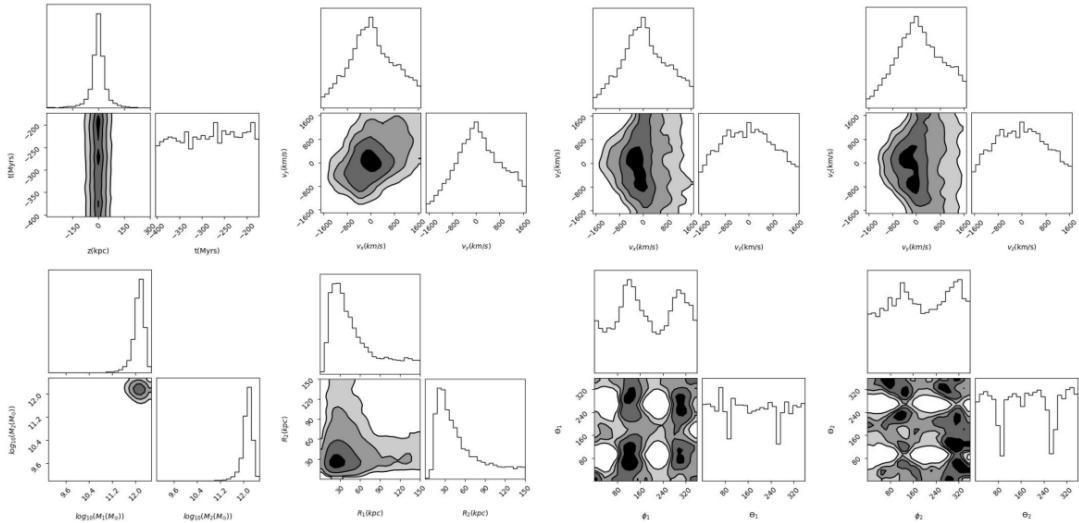


Figure 4.9: Reduced corner plot of the constraints made on the observational image of Arp 240. As shown, there is significantly more uncertainty in these measurements than those of the best fit simulation. However, we are able to make constraints on almost all parameters in the sample. The degeneracy in the orientations remains, although is less obvious in the two θ parameters. We are also unable to make conclusive constraints on the velocity parameters with the observations.

the best fit simulation, although slightly lower. This is likely, again, due to the change in scale of observation. Our model uses the mass to assign flux to each particle which is then projected across the simulation. Therefore, because of the zoom in, we see that our projection of flux is significantly larger than in the best fit simulation. This means that for the same mass, a galaxy will appear larger in size. So, to maintain the correct galactic size the mass must be reduced.

We see the same general trends when looking at the Arp 172 and Arp 290 systems. In each system, the best fit simulation scale used in GZM is much larger than the measured scale from 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.

For the all three systems, the size of the observational image was at the limits of the simulation resolution. One simulation unit is equivalent to 15kpc. Therefore, having the image itself only be seven simulation units across would give non-physical results of the final system. Therefore, when making constraints, the observational image was artificially scaled up. We elected to scale it up by four times. This allowed us to recreate the tidal features and match the flux distribution, however, the effect this has on the mass measurements could be severe if made too large. When looking at the constraints on position here, we accounted for the scaling up by dividing the results back down.

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 across populations of 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 N-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 pipeline.

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 in kpc and then converted to an assumed resolution from the known pixel scale on the sky. 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 of the 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 observations from the SDSS which is a reasonably shallow survey, being able to observe down to the 22 mag/arcsec². Therefore, for most of our major interacting systems, we retained the full extent of the tidal features which were created in the interaction and could provide better fitting. Our algorithm will perform significantly better when being run on systems where the clarity of the tidal features has not been lost, such as Arp 240. 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 binned all of the observations down from their native pixel scale to 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 4.4.2, 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 40 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 125MB per core, therefore being very cheap in the memory.

This methodology was successful in reproducing the results of GZMs methodology - a project which took months to complete. We have automated this, and 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 an updated version of the restricted numerical simulation JSPAM and modified it to calculate and match the flux distribution of interacting galaxies based on fourteen 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 simulation, we utilised the archived Galaxy Zoo: Mergers project, which had constrained a sample of sixty two interacting systems. In Galaxy Zoo: Mergers, volunteers used visual inspection to find the best fit simulation for observed interacting systems using the mass distribution. We have introduced a Markov-Chain Monte Carlo pipeline to directly compare simulations to observations. By relating the likelihood that a set of underlying parameters describes an observed system to the χ^2 between that image and the simulated flux distribution we have provided a methodology to automatically put constraints on these parameters.

Using this methodology, we have found that we can constraint the best fit simulation parameters from Galaxy Zoo: Mergers to within 2σ . Every parameter of our fourteen 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 find it is that, 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 our best fit observed systems, were able to retain the excellent constraints on eight of the fourteen parameters, losing any constraint on the orientations, the z-position and the velocity parameters. However, we were able to maintain a tight constraint on the masses of the two systems. This loss of constraint was due to a loss in the signal in the tidal features/edges of the two interacting galaxies; particularly in the secondary galaxy. Therefore, deep observations of interacting systems are crucial for our methodology to work.

There are numerous limitations to this method that any users must be aware of. First, the core simulation of this process is a restricted N-body code with

4.6 CONCLUSIONS & FUTURE WORK

a highly optimised flux distribution calculation and approximation. It is not impossible that there are systems that this pipeline 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 how it could be used 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 of it. To run this on sixty two systems, the total run time was thirty days; an unusable timescale when we are talking about interacting galaxy samples in the thousands. Therefore, future works with this algorithm and methodology will 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, one solution 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

Conclusion

- 5.1 SOFTWARE FOR STATISTICAL CONSTRAINT**
- 5.2 CREATING LARGE SAMPLES WITH MACHINE LEARNING**
- 5.3 APPLYING LARGE CATALOGUES: PROSPECTS**
- 5.4 APPLYING LARGE CATALOGUES: CURRENT LIMITATIONS**
- 5.5 FUTURE WORK**

Appendix A

MCMC Best Fit Values

A.1 Galaxy Zoo: Mergers Best Fit Parameters

Appendix B

Model Diagnostics & Further Identified Objects

B.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 B.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 B.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.

B.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 B.3. These were created using the Lupton et al. (2004a) 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).

B.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 B.4. Their Source IDs are shown in the upper left of each image, and a separate catalogue

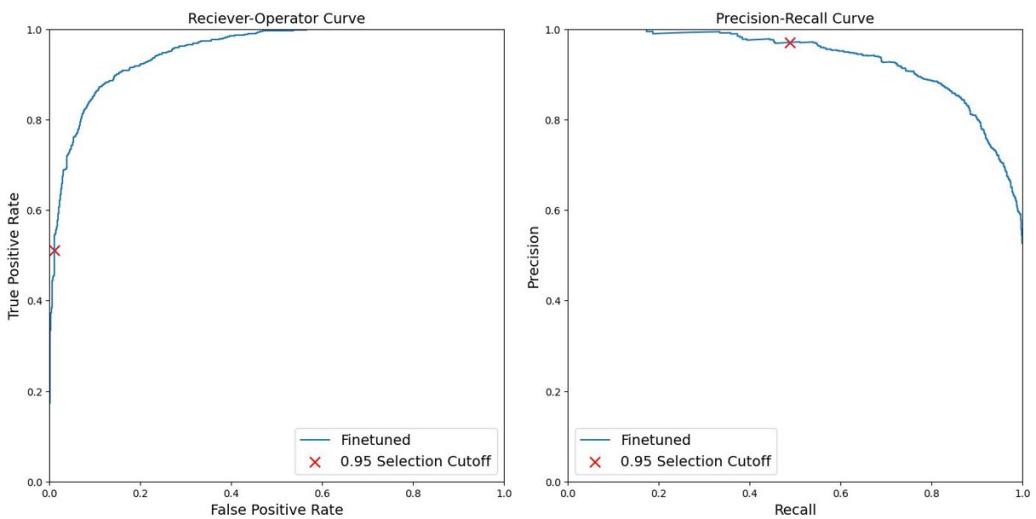


Figure B.1: The Receiver-Operator and Precision-Recall Curve for the Zoobot model that was used to explore the Hubble archives. The blue curves are the measured curves. These curves measure the relevant rates or characteristics based on the changing cutoff applied to how Zoobot defines an interacting 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 $\approx 50\%$ of interacting galaxies. This also shown in the precision recall curve where our recall is $\approx 50\%$.

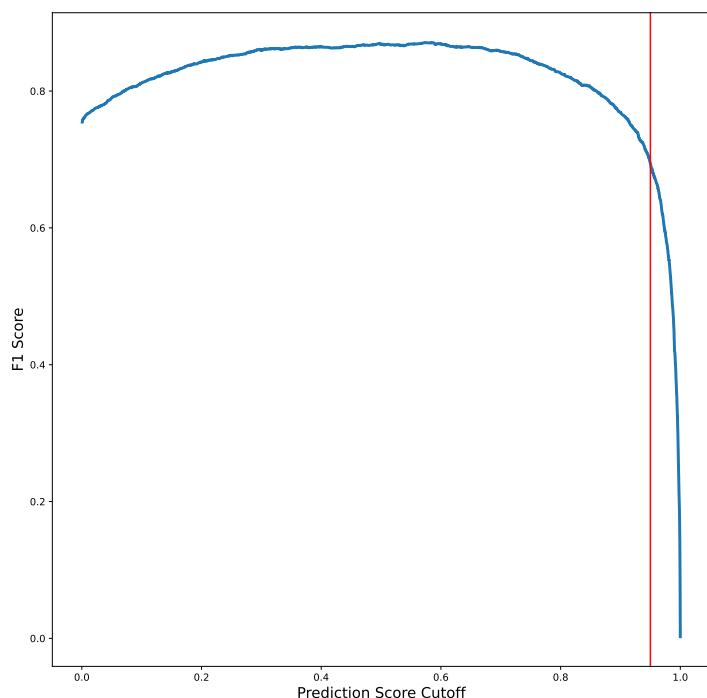


Figure B.2: The F1 score found during the diagnostics of the model used in this work. The F1 score is a measure combining the measure of accuracy and purity into one metric. The cutoff we use is at the point where the F1 score begins to rapidly decline. This point is shown by the red vertical line.



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

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 disrupted 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.

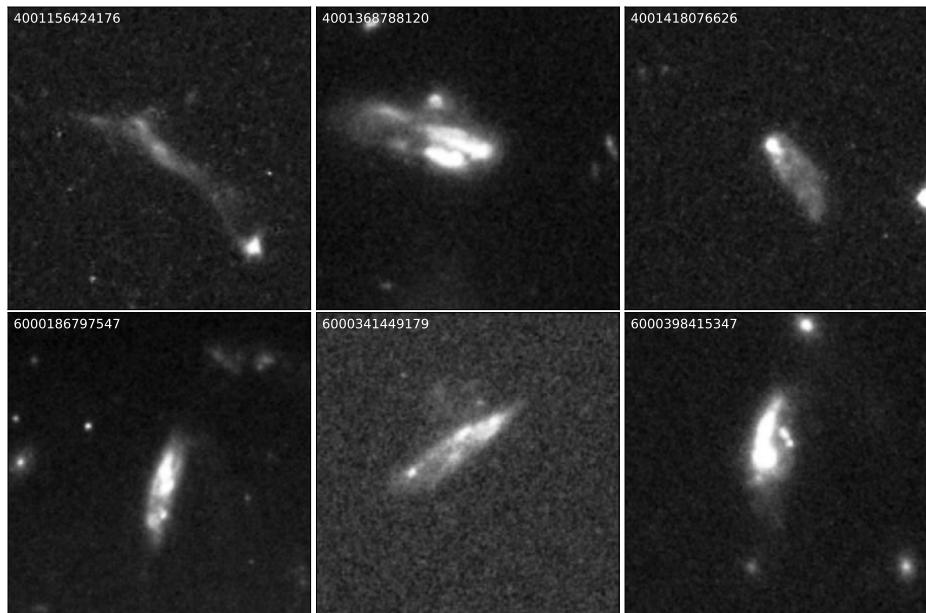


Figure B.4: The six unknown systems found in this work. These have no reference in Simbad or in NED, and their morphology could not be classified by the authors. Investigation into these six objects are presented to the community, with the authors hoping that future work and investigation of them can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

Proposal ID	Observation ID	Observation Date	DOI
8183	hst_8183_54.acs_wfc_f814w_j59l54	18/07/2002	https://doi.org/10.5270/
9075	hst_9075_2a.acs_wfc_f814w_j6fl2a	24/07/2002	https://doi.org/10.5270/
9351	hst_9351_11.acs_wfc_f814w_j8d211	31/03/2003	https://doi.org/10.5270/
9361	hst_9361_03.acs_wfc_f814w_j8d503	22/07/2003	https://doi.org/10.5270/
9363	hst_9363_09.acs_wfc_f814w_j8d809	02/07/2002	https://doi.org/10.5270/
9367	hst_9367_02.acs_wfc_f814w_j8ds02	10/06/2003	https://doi.org/10.5270/
9373	hst_9373_02.acs_wfc_f814w_j6la02	05/07/2002	https://doi.org/10.5270/
9376	hst_9376_02.acs_wfc_f814w_j8e302	13/07/2002	https://doi.org/10.5270/
9381	hst_9381_02.acs_wfc_f814w_j8fu02	13/03/2003	https://doi.org/10.5270/
9400	hst_9400_04.acs_wfc_f814w_j6kx04	29/05/2003	https://doi.org/10.5270/
9403	hst_9403_02.acs_wfc_f814w_j8fp02	09/07/2002	https://doi.org/10.5270/
9405	hst_9405_6k.acs_wfc_f814w_j8iy6k	22/05/2003	https://doi.org/10.5270/
9409	hst_9409_03.acs_wfc_f814w_j6n203	29/06/2003	https://doi.org/10.5270/
9411	hst_9411_09.acs_wfc_f814w_j8dl09	11/02/2003	https://doi.org/10.5270/
9427	hst_9427_13.acs_wfc_f814w_j6m613	21/10/2002	https://doi.org/10.5270/
9438	hst_9438_01.acs_wfc_f814w_j6me01	16/01/2003	https://doi.org/10.5270/
9450	hst_9450_02.acs_wfc_f814w_j8d402	25/08/2002	https://doi.org/10.5270/
9453	hst_9453_02.acs_wfc_f814w_j8f802	03/12/2002	https://doi.org/10.5270/
9454	hst_9454_11.acs_wfc_f814w_j8ff11	23/03/2003	https://doi.org/10.5270/

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

B.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 B.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
- 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., 2007a, MNRAS, 375, 1017
- Alonso M. S., Lambas D. G., Tissera P., Coldwell G., 2007b, 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., 1966a, ApJS, 14, 1
- Arp H., 1966b, ApJS, 14, 1
- Arp H., 1966c, ApJS, 14, 1

REFERENCES

- 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
- Avila R. J., Hack W., Cara M., Borncamp D., Mack J., Smith L., Ubeda L., 2014, DrizzlePac 2.0 - Introducing New Features, doi:10.48550/ARXIV.1411.5605, <https://arxiv.org/abs/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
- Barchi P. H., et al., 2020, Astronomy and Computing, 30, 100334
- Barnes J. E., 1988, ApJ, 331, 699
- Barnes J. E., Hernquist L. E., 1991, ApJL, 370, L65
- Barnes J. E., Hibbard J. E., 2009, AJ, 137, 3071
- 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., 2000a, ApJ, 530, 660
- Barton E. J., Geller M. J., Kenyon S. J., 2000b, ApJ, 530, 660
- Bell E. F., et al., 2004, ApJ, 608, 752
- 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

REFERENCES

- 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., 2019a, *MNRAS*, 490, 5390
- Bottrell C., et al., 2019b, *MNRAS*, 490, 5390
- Bournaud F., Jog C. J., Combes F., 2005, *A&A*, 437, 69
- Boylan-Kolchin M., Springel V., White S. D. M., Jenkins A., Lemson G., 2009, *MNRAS*, 398, 1150
- 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
- Byrne-Mamahit S., Hani M. H., Ellison S. L., Quai S., Patton D. R., 2023, *MNRAS*, 519, 4966
- 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

REFERENCES

- Comerford J. M., Pooley D., Barrows R. S., Greene J. E., Zakamska N. L., Madejski G. M., Cooper M. C., 2015a, ApJ, 806, 219
- Comerford J. M., Pooley D., Barrows R. S., Greene J. E., Zakamska N. L., Madejski G. M., Cooper M. C., 2015b, ApJ, 806, 219
- 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](https://arxiv.org/abs/1412.0843)), doi:10.1117/12.2055132
- Darg D. W., et al., 2010a, MNRAS, 401, 1043
- Darg D. W., et al., 2010b, MNRAS, 401, 1043
- Darg D. W., et al., 2010c, 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., 2022, arXiv e-prints, p. arXiv:2207.03968
- Das A., Pandey B., Sarkar S., 2023, Research in Astronomy and Astrophysics, 23, 095026
- De Lucia G., Blaizot J., 2007, MNRAS, 375, 2
- Delvecchio I., et al., 2017, A&A, 602, A3
- Dey A., et al., 2019, AJ, 157, 168

REFERENCES

- Di Matteo P., Combes F., Melchior A. L., Semelin B., 2007, *A&A*, 468, 61
- Di Matteo P., Bournaud F., Martig M., Combes F., Melchior A. L., Semelin B., 2008, *A&A*, 492, 31
- Ellison S. L., Patton D. R., Simard L., McConnachie A. W., 2008a, *AJ*, 135, 1877
- Ellison S. L., Patton D. R., Simard L., McConnachie A. W., 2008b, *AJ*, 135, 1877
- Ellison S. L., Patton D. R., Mendel J. T., Scudder J. M., 2011a, *MNRAS*, 418, 2043
- Ellison S. L., Patton D. R., Mendel J. T., Scudder J. M., 2011b, *MNRAS*, 418, 2043
- Ellison S. L., Mendel J. T., Patton D. R., Scudder J. M., 2013a, *MNRAS*, 435, 3627
- Ellison S. L., Mendel J. T., Patton D. R., Scudder J. M., 2013b, *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
- 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
- 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
- Ghosh A., Urry C. M., Wang Z., Schawinski K., Turp D., Powell M. C., 2020, *ApJ*, 895, 112

REFERENCES

- 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>
- 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
- Hani M. H., Gosain H., Ellison S. L., Patton D. R., Torrey P., 2020a, MNRAS, 493, 3716
- Hani M. H., Gosain H., Ellison S. L., Patton D. R., Torrey P., 2020b, MNRAS, 493, 3716
- Hani M. H., Gosain H., Ellison S. L., Patton D. R., Torrey P., 2020c, MNRAS, 493, 3716
- Harris C. R., et al., 2020, Nature, 585, 357
- 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., 1993, ApJS, 86, 389
- Hibbard J. E., van Gorkom J. H., 1996, AJ, 111, 655
- Hinton S. R., 2016, The Journal of Open Source Software, 1, 00045

REFERENCES

- Hogg D. W., et al., 2002, AJ, 124, 646
- Holincheck A. J., et al., 2016a, MNRAS, 459, 720
- Holincheck A. J., et al., 2016b, MNRAS, 459, 720
- 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., 2013a, MNRAS, 430, 1901
- Hopkins P. F., Cox T. J., Hernquist L., Narayanan D., Hayward C. C., Murray N., 2013b, 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
- Jacobs C., et al., 2019, ApJS, 243, 17
- 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., 2003a, MNRAS, 341, 33
- Kauffmann G., et al., 2003b, MNRAS, 341, 33

REFERENCES

- 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
- 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., Querejeta M., 2015, Galaxies, 3, 220
- Knapen J. H., Cisternas M., Querejeta M., 2015, MNRAS, 454, 1742
- Kolmogorov A., 1933, Giorn Dell'inst Ital Degli Att, 4, 89
- Lacey C., Cole S., 1993, MNRAS, 262, 627
- 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., Jing Y. P., White S. D. M., 2008b, MNRAS, 385, 1903
- Li C., Kauffmann G., Heckman T. M., White S. D. M., Jing Y. P., 2008c, MNRAS, 385, 1915
- Li Y. A., Ho L. C., Shangguan J., 2023, ApJ, 953, 91
- Lintott C. J., et al., 2008a, MNRAS, 389, 1179
- Lintott C. J., et al., 2008b, 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

REFERENCES

- 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., 2004a, PASP, 116, 133
- Lupton R., Blanton M. R., Fekete G., Hogg D. W., O'Mullane W., Szalay A., Wherry N., 2004b, PASP, 116, 133
- Marchesi S., et al., 2016, ApJ, 817, 34
- Marian V., et al., 2020, ApJ, 904, 79
- 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., 1996a, ApJ, 464, 641
- Mihos J. C., Hernquist L., 1996b, ApJ, 464, 641
- 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

REFERENCES

- 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., 2021a, MNRAS, 503, 3113
- Moreno J., et al., 2021b, MNRAS, 503, 3113
- Morganti R., Garrett M. A., Chapman S., Baan W., Helou G., Soifer T., 2004, A&A, 424, 371
- Nair P. B., Abraham R. G., 2010, ApJS, 186, 427
- Navarro J. F., Frenk C. S., White S. D. M., 1997, ApJ, 490, 493
- Nelson B., Ford E. B., Payne M. J., 2014, ApJS, 210, 11
- 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
- Pacifci C., Kassin S. A., Weiner B., Charlot S., Gardner J. P., 2013, ApJL, 762, L15
- Patton D. R., Torrey P., Ellison S. L., Mendel J. T., Scudder J. M., 2013, MNRAS, 433, L59
- 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., 2022a, A&A, 661, A52
- Pearson W. J., et al., 2022b, A&A, 661, A52

REFERENCES

- 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., 2010, AJ, 139, 2097
- Plummer H. C., 1911, MNRAS, 71, 460
- Quai S., Hani M. H., Ellison S. L., Patton D. R., Woo J., 2021, MNRAS, 504, 1888
- 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
- Rodríguez Montero F., Davé R., Wild V., Anglés-Alcázar D., Narayanan D., 2019a, MNRAS, 490, 2139
- Rodríguez Montero F., Davé R., Wild V., Anglés-Alcázar D., Narayanan D., 2019b, 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., 2009a, PASJ, 61, 481
- Saitoh T. R., Daisaka H., Kokubo E., Makino J., Okamoto T., Tomisaka K., Wada K., Yoshida N., 2009b, PASJ, 61, 481
- Salim S., et al., 2007, ApJS, 173, 267
- Schawinski K., et al., 2014, MNRAS, 440, 889
- Schawinski K., Koss M., Berney S., Sartori L. F., 2015, MNRAS, 451, 2517

REFERENCES

- Schaye J., et al., 2015, MNRAS, 446, 521
- Schweizer F., 1982, ApJ, 252, 455
- 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., 2012a, MNRAS, 426, 549
- Scudder J. M., Ellison S. L., Torrey P., Patton D. R., Mendel J. T., 2012b, MNRAS, 426, 549
- Serra P., et al., 2019, A&A, 628, A122
- 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., 2021a, ApJ, 909, 124
- Silva A., Marchesini D., Silverman J. D., Martis N., Iono D., Espada D., Skelton R., 2021b, 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
- 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](https://arxiv.org/abs/0908.3478))

REFERENCES

- Smolčić V., et al., 2017, A&A, 602, A6
- Springel V., 2000a, MNRAS, 312, 859
- Springel V., 2000b, MNRAS, 312, 859
- Springel V., White S. D. M., Tormen G., Kauffmann G., 2001, MNRAS, 328, 726
- Springel V., et al., 2005a, Nature, 435, 629
- Springel V., et al., 2005b, 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
- 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., Toomre J., 1972a, The Astrophysical Journal, 178, 623
- Toomre A., Toomre J., 1972b, ApJ, 178, 623
- Toomre A., Toomre J., 1972c, 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
- 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

REFERENCES

- Wallin J. F., Holincheck A. J., Harvey A., 2016a, *Astronomy and Computing*, 16, 26
- Wallin J. F., Holincheck A. J., Harvey A., 2016b, *Astronomy and Computing*, 16, 26
- Walmsley M., et al., 2022a, *MNRAS*, 509, 3966
- Walmsley M., et al., 2022b, *MNRAS*, 509, 3966
- Walmsley M., et al., 2022c, *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
- 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., 2016a, *AJ*, 151, 134
- Whitmore B. C., et al., 2016b, *AJ*, 151, 134
- Willett K. W., et al., 2013, *MNRAS*, 435, 2835
- Willett K. W., et al., 2017, *MNRAS*, 464, 4176
- 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

REFERENCES

- York T., Jackson N., Browne I. W. A., Wucknitz O., Skelton J. E., 2005, MNRAS, 357, 124
- 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 der Walt S., et al., 2014, PeerJ, 2, e453