Galaxy Evolution Over the Last 8 Billion Years

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

Relevant Publications by the Author

Chapter 2

- Publication

Chapter 1

Introduction

1.1 GALAXY INTERACTION

Galactic interaction plays a fundemntal role in the evolution of galaxies throughout cosmic history. By galaxy interaction, what we mean is the effects and physical processes that occur when to galaxies move close enough together that they begin to have a physical affect upon on another. Most often, these affects are due to gravitational effects. These effects can range from morphological disturbance (), to

1.2 GALAXY INTERACTION IN THE COSMOLOGICAL CONTEXT

1.3 SIMULATIONS OF GALAXY INTERACTION

1.3.1 Prior Examples

1.3.2 PySPAM

1.4 STATSTISTICALLY CONSTRAINING INTERACTION

1.4.1 Bayesian Statistics & MCMC

1.4.2 Previous Examples

1.5 INTERACTING GALAXY IDENTIFICATION

1.5.1 By Citizen Scientists

1.5.2 By Machine Learning

1.5.3 Existing Spectroscopic Surveys

1.5.4 Limitations

1.6 LARGE SAMPLES OF INTERACTING GALAXIES

1.6.1 Created by Citizen Scientists

1.6.2 Found with Machine Learning

Chapter 2

Galaxy Zoo: Mergers & Constraining Interaction

2.1 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.

2.2 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 (**?**), major interactions accelerate the process of rotationally-supported systems to dispersion-dominated ones (**???**) and they likely have some link to nuclear activity (**???**). 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 **???**), 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., **????**) 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 (**???**) and are particularly pronounced in gas rich post-merger observations (e.g, **???**).

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 **????**) 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 (????). 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 (???), while others point to the kinematics of the interaction itself (??). 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 ? or the Stellar Particle Animation Module (SPAM) by ?. However, finding the best fit underlying parameters for more than a handfull 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, ? 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. ???) 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) interacting 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 ? in the Galaxy Zoo: Mergers (GZM) project (for details on Galaxy Zoo, see Lintott et al. (2008)). 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 (?). 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 2.3, 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 2.4 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 2.5 followed by our conclusions and future work in section 2.6. Where necessary, we use a Flat ΛCDM cosmology with $H_0 = 70\,\mathrm{km/s/Mpc}$ 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.

2.3 DATA

2.3.1 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 2.1 (note, this is very similar to that from **?**, with decimalised coordinates and redshift data). The SDSS cutout of each system is given in figure 2.1, with the best fit simulation cutouts being found in Figure 6 of **?**. 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 $\times$ 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 found from the NASA Extragalactic Database. The effects of incorrect/missing redshift are discussed in section 2.5.6.1, 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.

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	588011124116422756	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	758874209603223717	0.368333	31.427778	0.024

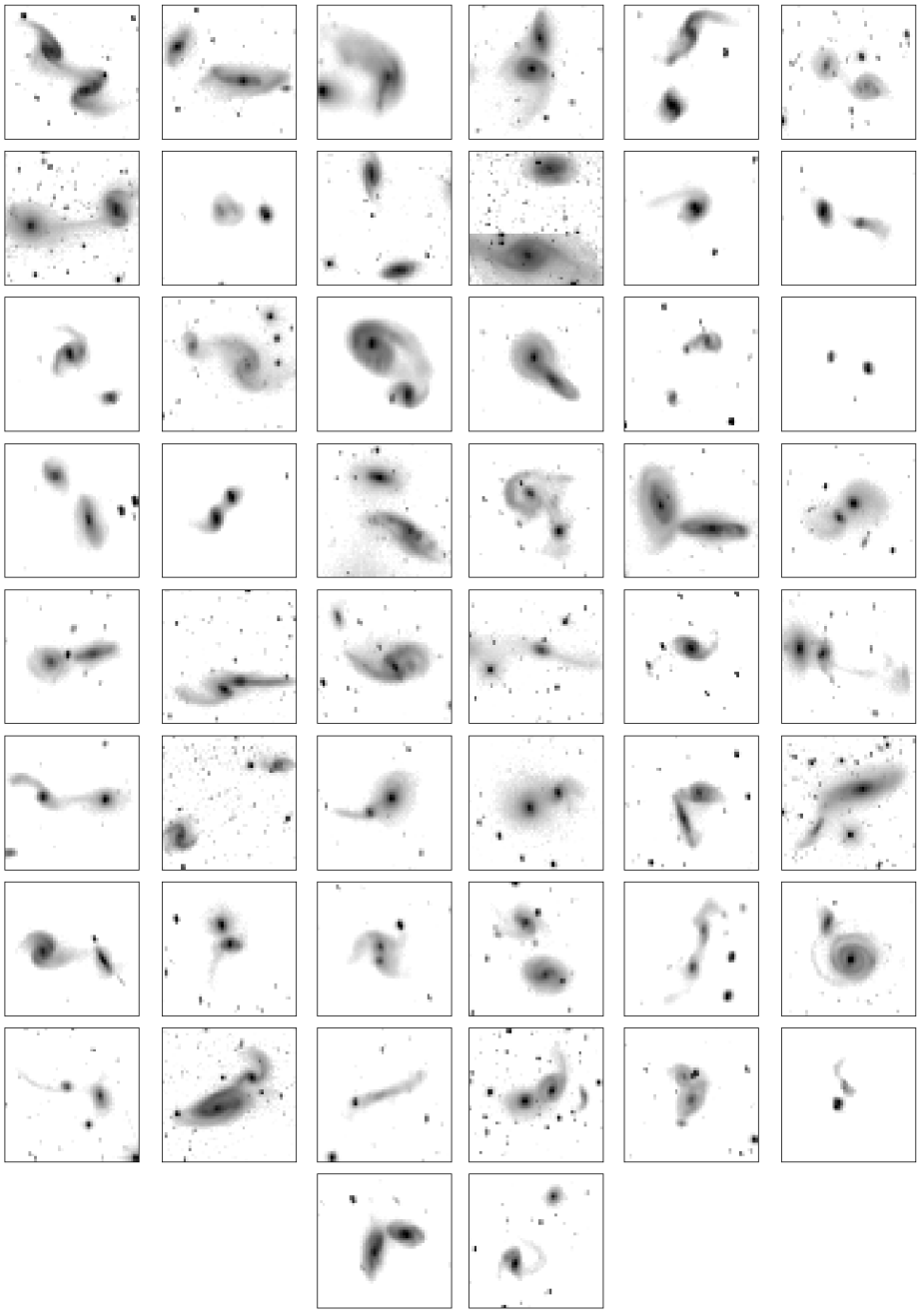

Figure 2.1: Cutouts from SDSS of the sixty two interacting systems that we test our pipeline on. For the best fit simulations see Figure 6 of **?**.

2.3.2 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 **?** 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 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

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	169.34kms[1]	-1693kms^{-1} - 1693kms^{-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 2.2: 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.

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 website[1].

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 2.4). 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 full parameter distribution. Table 2.2 shows the full parameter space we explore in this work for each system.

2.3.3 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

[1] `https://data.galaxyzoo.org/mergers.html`

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 (**??**).

2.4 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 2.2.

2.4.1 Simulations: APySPAM

The original GZM project utilised the numerical simulation code Java Stellar Particle Animation Module (JSPAM) (**?**). For an indepth description of the underling code, we direct the reader to **??** 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. **?**). 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 (**?**) 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).

2.4.1.1 Stellar Population Evolution

To model the underlying stellar populations, we utilise a **?** (BC03) simple stellar population. These contain SEDs generated from flux libraries from a **?** initial mass function. We set this stellar population model to have a delayed exponentially declining star formation rate (**????**). 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 2.4.1.2). We assume the onset of star formation is $\tau = 1.5$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 galacitc 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 **?** 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.

2.4.1.2 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 (**?**). 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 (**?**) 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

$$SFR_{\text{enhancement}} = \beta(\frac{t}{\tau^2})\exp(-\frac{t}{\tau})M_{\text{baryonic}}M_{\odot}\text{yr}^{-1}. \quad (2.1)$$

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

$$\beta = M_{\text{ratio}}D^2_{\text{ratio}}. \quad (2.2)$$

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 (**???**). We base our semi-analytic approach on the star formation histories found in a range of high resolution N-body simulations (**???**) which measure the change in star formation directly from the Kennicutt-Schmidt (**?**) relation. These simulations directly model the star form-

ing gas through the interaction, measuring the change its evolution. This is a computational expense we cannot afford.

The output of equation 2.1 is global star formation of each interacting galaxies at any given timestep. However, the aim of our models is to be able to match the flux distribution across the entire galaxy (especially the tidal features) to any observation that the code is given. Therefore, we must distribute the star formation throughout the particles. To keep computational efficiency, this is done by utilising weights which have been assigned to each particle. These weights are based on the ratio of the gas mass of the particle which has been assigned at initialisation to the total gas mass of the galaxy. So, to find the star formation rate of a single particle at any given timestep, the following equation is applied,

$$SFR_{\text{Particle}} = \frac{M_{\text{gas,Particle}}}{M_{\text{gas,Galaxy}}} SFR_{\text{Galaxy}}. \tag{2.3}$$

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 occuring 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 it's 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 2.4.1.1.

2.4.2 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 2.3. 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.

2.4.2.1 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 code which uses flux distribution rather than morphology matching is due to the prior difficulties of using such a method. Therefore, the introduction of extra algorithms which require extra computation time has been necessary. The runtimes of JSPAM and APySPAM are shown in table 2.3. 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

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 2.3: 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.

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 2.4.2, 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 2.3 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.

2.4.3 Defining the Likelihood Function

2.4.3.1 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 descibe the observed system well.

The MCMC algorithm we use is the Python package EMCEE (**?**); this is an ensemble MCMC package with numerous predefined moves and algorithms to make getting to the area of high likelihood more efficient. We construct contours of 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 2.4,

$$P(H_i|D_{obs}, C) = P(H_i|C)\frac{P(D_{obs}|H_i, C)}{P(D_{obs}|C)}). \tag{2.4}$$

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 constrains,P(D_{obs}—H_i,C), all divided by a normalisation constant, P(D_{obs}—C).

2.4.3.2 Simplifying the Prior

In order to simplify this expression, we make assumptions about the underlying parameter space to increase efficiency and simply our computations. We first assume uniform priors for each of our 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 2.2. These ranges can be tweaked, or a different prior function defined, by the user. Therefore, the priors part of equation 2.4 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 **?** (there it is called β but we call it γ here to not be confused with our star formation enhancement parameter)

$$\gamma_{min} = \frac{M_1 + M_2}{r_{min}^2 V_{r_{min}}}. \tag{2.5}$$

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 **?**, 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.

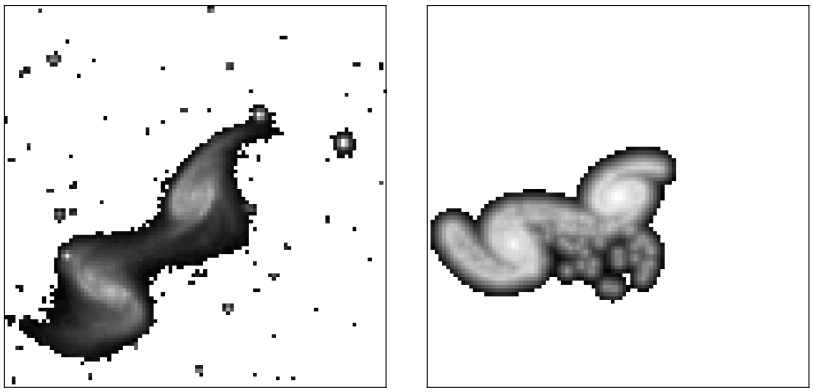

Figure 2.2: The example system used to test our pipeline: the Arp 240 interacting system. This system is considered an easy one to constrain. It is composed of two clearly distinct galaxies, with strong tidal features that our pipeline can fit. These tidal features are the two tidal tails formed in the 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 **?** and the first test image used in our pipeline. The different in scale and orientation are discussed below.

This probability, or prior, is defined as

$$C = \begin{cases} \exp(0.5\frac{\gamma}{\gamma_{min}}), & \text{if } \gamma < 0.5 \\ 0, & \text{if } \gamma \geq 0.5 \end{cases} \tag{2.6}$$

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.

2.4.3.3 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 2.5 this is found to not always hold true; particularly for the orientations of the system. However, making this assumption allows us to utilise the following equation to

compare our mock observations to our observed data;

$$P(D_{obs}|H_i, C) = (2\pi\sigma_j^2)\exp(-\frac{1}{2\sigma^2}\sum_{j=1}^{n}(x_j - \mu)^2) \times C. \tag{2.7}$$

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 (**?**). Where χ^2 is given by;

$$\chi^2 = \frac{1}{N - n_{dof}}\sum_{0}^{N_x}\sum_{0}^{N_y}\frac{(p_{x,y} - q_{x,y})^2}{\sigma_{x,y}^2}. \tag{2.8}$$

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 (**??**). 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,

$$\log_{10}(P(H_i|D_{obs}, C)) = \log_{10}(L) = -\frac{\chi^2}{2} + \log_{10}(C). \tag{2.9}$$

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

2.4.4 Architecture for Exploring Parameter Space: `EMCEE`

As stated in the previous section, we utilise the MCMC code `EMCEE` (**?**) to explore parameter space of interaction. For full details of EMCEE and the different modes that it can use, see the extensive readthedocs[1]; but here we will briefly state the hyper parameters that we 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 (**?**) and the Snooker Differential Evolution (DES) Move (**?**). An identical version of our setup can be found on GitHub[2]. Here, a user can download our setup to reproduce our results, or to update the model for their own purposes.

2.5 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 online[3]. 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

[1] `https://emcee.readthedocs.io/en/stable/`
[2] `https://github.com/AstroORyan`
[3] All results are found here: `Link_to_results`

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.

2.5.1 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 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 2.2.

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 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 2.3. This corner plot is created using the Corner Python Package (**?**).

However, displaying our results as is shown in Figure 2.3 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 2.5.

Figure 2.5 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 2.5 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.

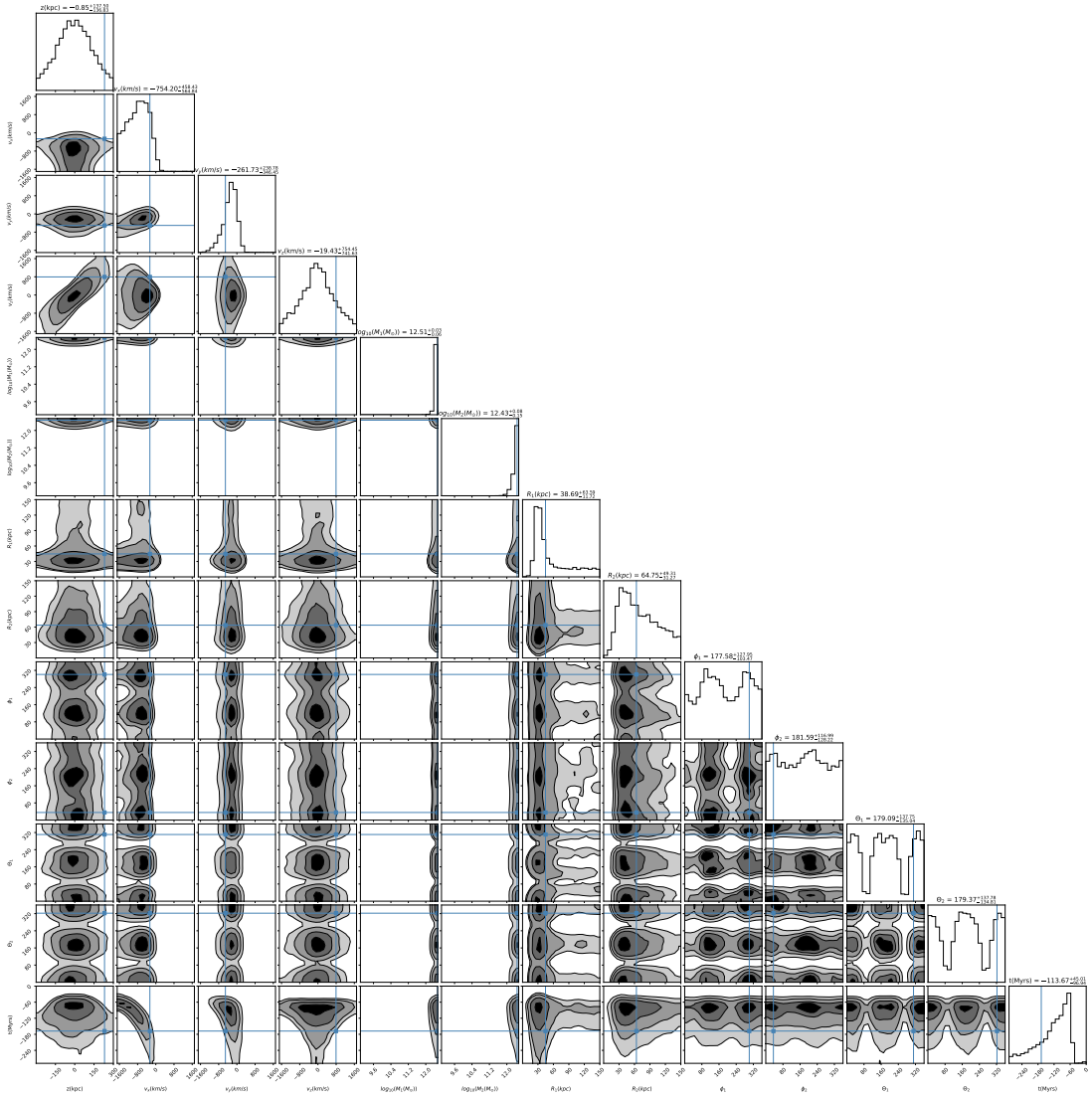

Figure 2.3: Corner plot showing the constraints made on all thirteen parameters we are exploring. Contours show the 0.5, 1, 1.5 and 2-sigma levels of the constraint. this corresponds to containing 11.8%, 39.3%, 67.5% and 86.4% of the samples across all walker chains. Displaying our results using the 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 2.5. 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.

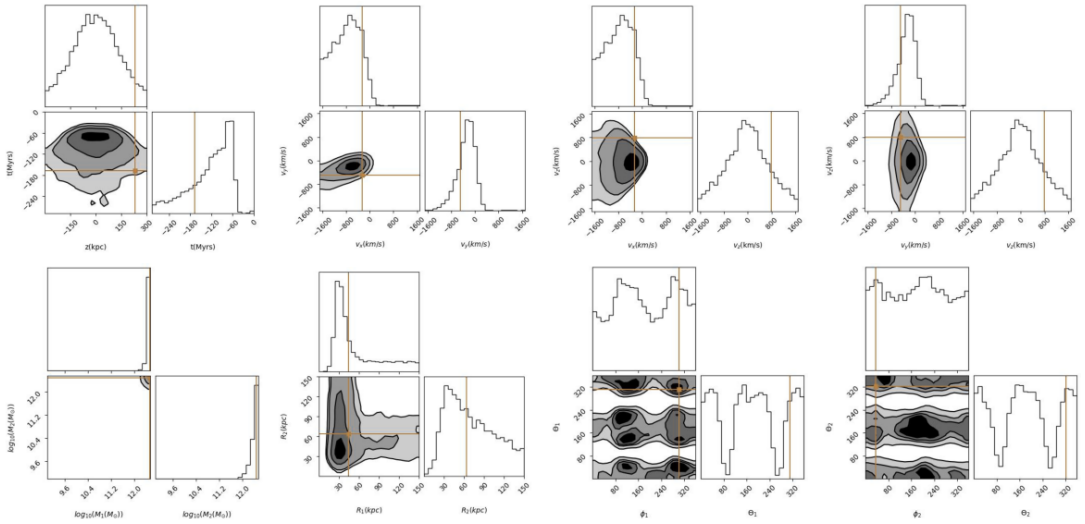

Figure 2.4: Same as Figure 2.3, but reduced to only matching parameters.

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 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 priamry, 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 2.5 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 out constraints in the z-direction: the z-position and the z-velocity. As seen in Figure 2.5, 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 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

Figure 2.5: 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.

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.

2.5.2 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 (**?**) Python package. For this case, every parameter passes this convergence test, with the exception of the orientation parameters.

Figure 2.5 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 Gewke diagnostic measure, which is looking for a single peak in parameter space. Therefore, by folding the parameter space over and only exploring over 0° - 180°, we achieve a single peak in parameter space. These 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 2.6 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

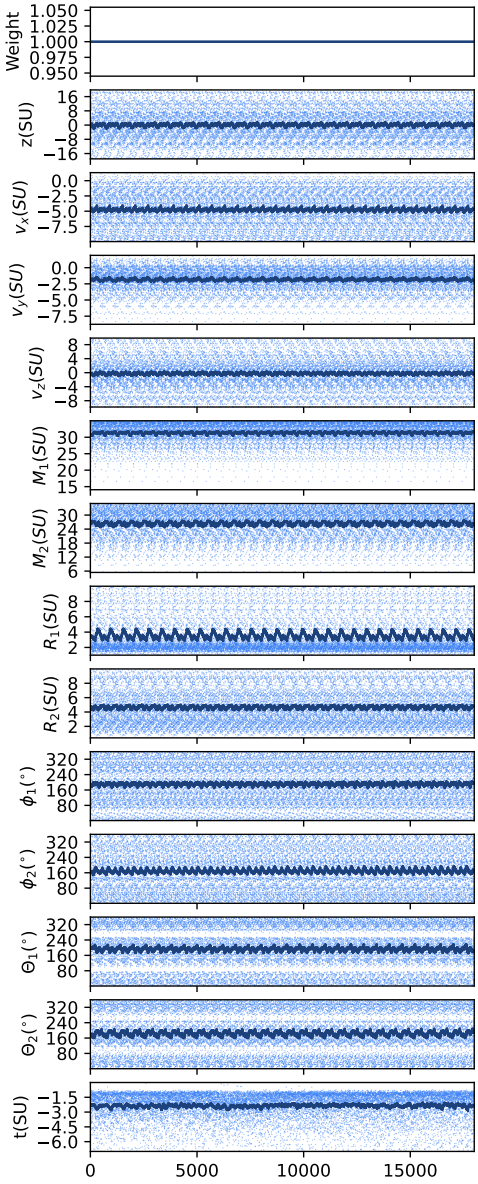

Figure 2.6: Steps taken by each walker in our MCMC chain to constrain the Arp 240 best fit simulation. Note, the y-scales here do not extend over the full parmaeter space for some galaxies, and only show where the walkers have stepped after the burn-in phase. The deeper the blue, the more walkers have stepped at that point. This figure shows that our MCMC has successfully burnt in and very quickly goes to high areas of probability for parameter space. They then oscillate around the best fit values while searching the remaining parameter space. The z, z-velocity and ϕ_2 parameters show dignificant uncertainty as the walkers move around the entire parameter space. In the $\phi_1, \theta_1, \theta_2$, we can see the two fold degeneracy form very early on and then the walkers do not explore across them at any point.

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.

2.5.3 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 2.7 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 2.7 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

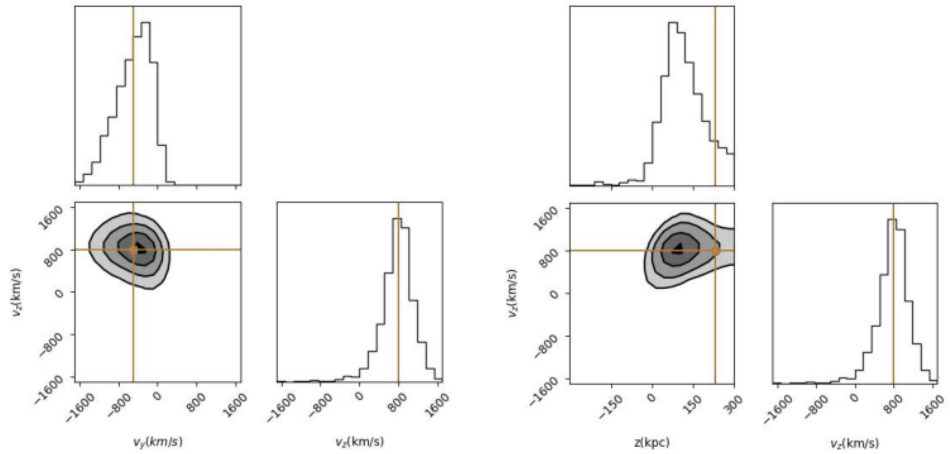

Figure 2.7: The constraints on the 3D velocity and z-position when including velocity in our MCMC pipeline. To add this extra information, we ran our best fit simulation of Arp 240 and summed the LOS velocity (z-velocity) in each pixel of our image. Comparing between here and Figure 2.5, we can see that 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 (**?**) 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 (**?**). 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.

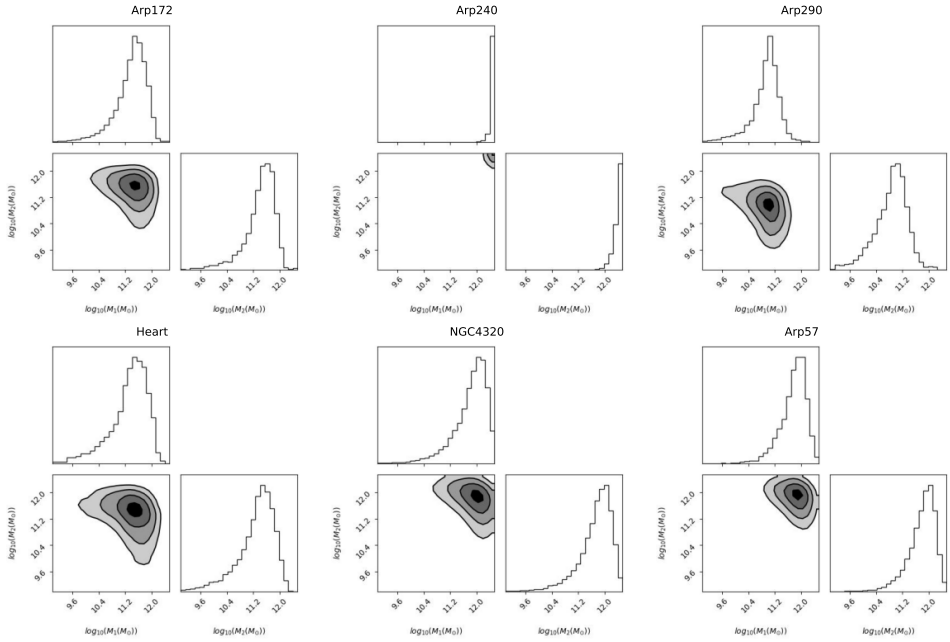

Figure 2.8: Mass corner plots for our best and worst fits from the pipeline. This was judged by the FWHM of the marginalised posteriors. *Top*: Our top three best fits. *Bottom*: Our worst fits. Even with our worst fits here, the masses are very well constrained.

2.5.4 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 2.8.

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 2.4: 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 2.4 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.

2.5.5 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 2.4. We reiterate here that the cutout resolution is reduced from its native resolution to images of 100 $\times$ 100 pixels. Before we input the images into the pipeline, we find the central pixel of the secondary galaxy and convert this into a physical x- and y-position. We also find the total size physical size of the cutout in kpc, convert it to simulation units and provide this to the pipeline. We find that the physical size of the cutouts are significantly different from the cutout sizes used by GZM in their work. As an example, we create the cutout of Arp 240 at 600 $\times$ 600 pixels at native resolution. With SDSS data, this corresponds to a physical size of 111.50kpc $\times$ 111.50kpc at the redshift of Arp 240. The best fit simulation from GZM for Arp 240 is 785.35kpc $\times$ 785.35kpc. Figure 2.2 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 2.9 shows the reduced corner plot for the observed Arp 240 system. The constraints on the velocity parameter have significantly worsened when compared to the constrains on the best fit simulation. They are also in a different area of parameter space when compared. The velocity and spatial parameters are the most likely to be affected by the change in scale between the two input images. As the secondary galaxy position has been completely altered the best fit trajectory of the interaction has also completely changed. The secondary is significantly closer to the primary, therefore meaning the secondary velocity must be significantly slower than previously. Due to the increase in noise in the observational image, and the extent of the tidal features of the primary and secondary galaxies, the constraint has also significantly weakened.

However, for the remainder of the parameters the level of constraint remains similar even with the change in the best fit values. We retain the two- and four-fold degeneracy in the orientation parameters. This follows on that the pipeline outright rejects disks which are edge on and loses accuracy when attempting to pin point the precise inclination of each galaxy. We are unable to make any constraint on the θ parameters, while the degeneracy of the ϕ parameters is still visible. The radius parameter is well constrained at 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

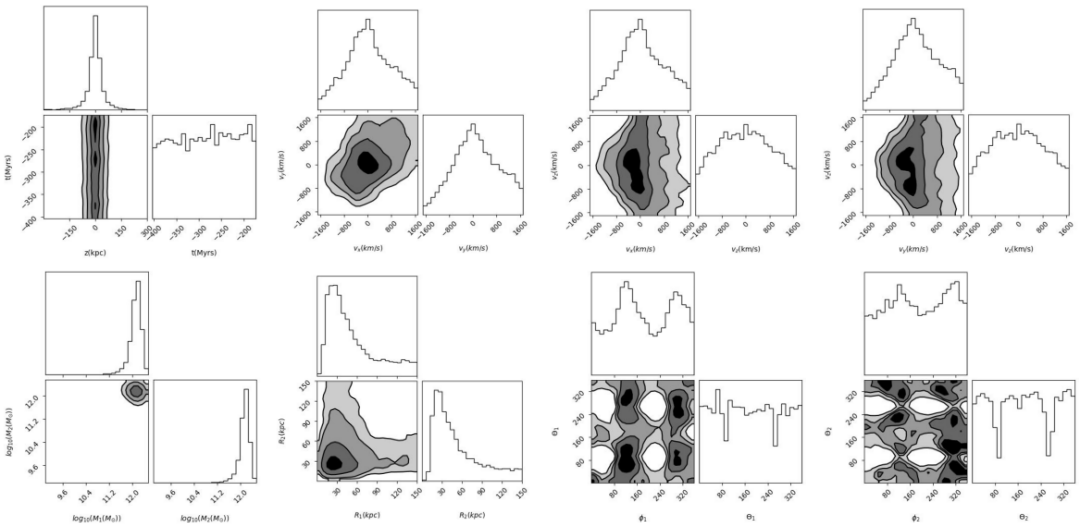

Figure 2.9: Reduced corner plot of the constraints made on the observational image of Arp 240. As shown, there is significantly more uncertainty in these measurements than those of the best fit simulation. However, we are able to make constraints on almost all parameters in the sample. The degeneracy 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, are 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.

2.5.6 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.

2.5.6.1 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 a 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/arcsec^2. 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 preform 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 $\times$ 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 2.4.2, which is an imperfect solution. It is recommended for any user to make a balanced trade off between resolution and computational expense.

2.5.6.2 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.

2.6 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 **?** 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

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 3

Creating a Large Sample of Interacting Galaxies

3.1 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.

3.2 INTRODUCTION

Interacting and merging galaxies are important to our current theory of ΛCDM cosmology, in which structure typically assembles hierarchically (Abadi et al., 2003; De Lucia & Blaizot, 2007; Guo & White, 2008; Springel et al., 2005). Galaxy interaction leads to highly disturbed morphologies (Hernández-Toledo et al., 2005; Toomre & Toomre, 1972; Wallin et al., 2016), intense starbursts (Mihos & Hernquist, 1996; Moreno et al., 2021; Saitoh et al., 2009; Springel, 2000) and, potentially, quenching of some systems (Das et al., 2022; Hani et al., 2020; Hopkins et al., 2013; 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., 2008; Pearson et al., 2019; Scudder et al., 2012). Interaction also has a direct impact on the gas angular momentum within each galaxy, causing it to decrease. This, potentially, leads to funnelling of gas into into their nuclear regions and igniting activity. This could be a connection with active galactic nuclei (Comerford et al., 2015; Ellison et al., 2008, 2011; Li et al., 2008). However, such a connection remains debated (Alonso et al., 2007; 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., 2008), 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., 2010b). These studies (and many others; for further examples, see Alonso et al., 2004; Barton et al., 2000; Ellison et al., 2013; Holincheck et al., 2016; Silva et al., 2021) illustrate the complex parameter space involved in understanding the role of interaction in

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

The first large-scale catalogues of interacting galaxies are from the mid 20th century (Arp, 1966; Vorontsov-Velyaminov, 1959, 1977, hereafter VV). These catalogues primarily used visual inspection to identify mergers (e.g., 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. (2022) demonstrated the power of CNNs for finding interacting and merging galaxies specifically, finding 2,109 in 5.4 deg^2 of Hyper Suprime-Cam imagery - a large sample for the small area covered.

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

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

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

This paper is laid out as follows: Section 3.3 describes the HSC and all the criteria we applied to create the images we predict over. This Section also introduces ESA Datalabs[2]; a new platform which allows the user to directly access the *Hubble* Science Archive. Section 3.4 gives an in depth description of the `Zoobot` CNN we utilise for our predictions, and how it differs from a commonly used CNN. Section 3.5 explains the process of creating the training set for our CNN to find interacting galaxies, with Section 3.6 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 3.7 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 ΛCDM cosmology with $H_0 = 70\,\mathrm{km/s/Mpc}$ 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

[1] See `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.

3.3 DATA

3.3.1 The *Hubble* Archives & ESA Datalabs

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

The criteria we apply are: the observational data must be from the Advanced Camera for Surveys (ACS), it must be final product data of *HST* (i.e. within a .drc file, where the data has been drizzle (Avila et al., 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 3.3.2) 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., 2013, 2018) 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.

3.3.2 The Shapely Python Package

A large computational expense in our pipeline was matching FITS files to sources. Conventionally, the `Astropy` CONTAINS function would be used to match source

coordinates to the FITS file WCS. We instead use the `Shapely`[1] Python package. `Shapely` is a geometry orientated package primarily focused on geospatial data. We found converting the FITS image footprints into `Shapely` Polygons and the source coordinates to `Shapely` Points and then checking if they overlapped had significant speed up. Per iteration, Astropy's CONTAINED_BY function matches a source to a FITS file on the order of 500ms. Using `Shapely's` CONTAINS function, the same process is on the order of 6μs.

3.4 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.

3.4.1 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

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

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 3.4.2 we describe using transfer learning to use Zoobot effectively outside of this magnitude and redshift range.

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

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

`Zoobot` was trained to give a prediction score to an image of a galaxy based on the question it is answering. The type of prediction score is set by the users choice of the model final layer in `Zoobot`. We elect to use a SoftMax output,

which returns an output score as a float between 0 and 1. This prediction score is not a probability score, although it may seem analogous. A well behaved prediction score will map to probability, though not necessarily linearly. The mapping between prediction score and probability is not considered in this work, and we use the prediction score as an indicator of `Zoobot`'s confidence a source is an interacting galaxy.

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

3.4.2 Transfer Learning

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

We construct a classification section that maximises accuracy and only allow the weights of this section to change. As the classification section has fewer parameters than the feature extraction section (the classification section contains 86,209 parameters compared to the feature extraction sections' 4,048,989 parameters) we need significantly less data to completely retrain it (in our case, a factor

of 15 less). Once this retraining is complete, the weights of the feature extraction sections of the model can be unfrozen and tweaked using our smaller dataset with a very low learning rate to further boost overall model accuracy.

An example of taking an existing model and applying it to a new problem with transfer learning is shown in Walmsley et al. (2022b). Here, they take the trained model and finetune it to finding ring galaxies. They retain an accuracy of 89% while only needing to train the model on 10^3 ring galaxies. This significantly reduces computational expense and training time of the model, while keeping the required training set very small. Interacting galaxies are rare, and interacting galaxy catalogues not expansive. So retraining the full network on hundreds of thousands of interacting galaxies is not feasible. Using transfer learning, and following the example from Walmsley et al. (2022b), we only need to create a training set of 10^3 - 10^4 interacting galaxies to achieve an accuracy of $\approx$90%.

3.5 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. (2022b), and aim to create a balanced training set. This has the advantage that it significantly improves the performance and accuracy of machine learning classifiers, but the disadvantage that it can bias our final model if few interacting galaxies exist compared to the general population. However, such a bias will be mitigated by using a high prediction cutoff to define an interacting galaxy. This is discussed in in Section 3.6.1. To create this large training set we use the Galaxy Zoo collaboration (initial data release described in Lintott et al., 2008).

3.5.1 Interacting Galaxies and Galaxy Zoo

The data in Galaxy Zoo is volunteer classifications on galaxy images spanning multiple projects. We incorporate classifications from all major Galaxy Zoo

projects; Galaxy Zoo 1 (Lintott et al., 2008), Galaxy Zoo 2 (Willett et al., 2013), Galaxy Zoo: *Hubble* (Willett et al., 2017), Galaxy Zoo: CANDELS (Simmons et al., 2017) and Galaxy Zoo: DECaLS (Walmsley et al., 2022a). These projects contain a total of 1,367,760 labelled galaxy images that we must extract the interacting galaxies from. We only use labels that are from citizen scientists, and no labels generated by previous versions of `Zoobot`. We apply three criteria to each interacting or non-interacting label. Firstly, it must have greater than 20 volunteer votes on it. Applying this allows us to use a statistically robust weighted vote from a crowd answer rather than trusting any volunteers individually. Secondly, the calculated weighted vote (i.e. the combination of the 20 or greater votes) must then be greater than 75% in favour of being an interacting galaxy or less than or equal to 25% for it not to be; this ensured purity in our training set. If the question given to volunteers was more specific (such as 'Is this a minor disturbance?' and 'Is this a major disturbance?') then if either answer was the majority vote we classified it as an interacting galaxy. Thirdly, the object must exist in the *Hubble* footprint so that we could make a cutout of it.

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

Upon applying these criteria we find 3,167 labelled interacting galaxies in Galaxy Zoo: *Hubble* project, the largest contribution to our training set. These were paired with 3,167 labelled non-interacting systems (following the previous criteria) to balance the training set. From all other projects, we find 869 labelled

interacting systems which fitting the creation criteria. The primary limiting factor for Galaxy Zoo's 1 and 2 was that many found interacting galaxies did not exist in the Hubble footprint. For Galaxy Zoo: CANDELS and Galaxy Zoo: DECaLS the limiting factor was the required calculated weighted vote. These labelled interacting systems were then paired with 869 labelled non-interacting systems, 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 3.1 and 3.2 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. (2022), 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.

3.5.2 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

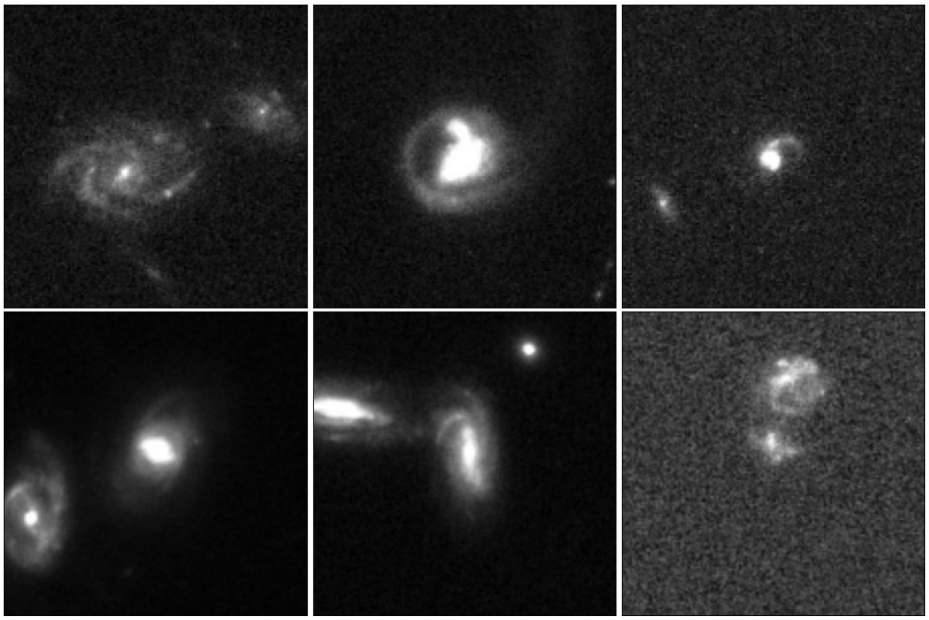

Figure 3.1: Example images of the labelled interacting galaxy systems used to train `Zoobot`. Each galaxy had a weighted vote fraction $\geq$0.75 in Galaxy Zoo. *Top Row*: Three examples from the Galaxy Zoo: *Hubble* project of the training set. *Bottom Row*: Three examples from the other Galaxy Zoo projects. These are, from right to left, Galaxy Zoo 2, Galaxy Zoo CANDELS and Galaxy Zoo DECaLS. The priority with this training set was that the interactors had clear tidal features and disruption so `Zoobot` would learn to highly weight them and not misclassify close pairs.

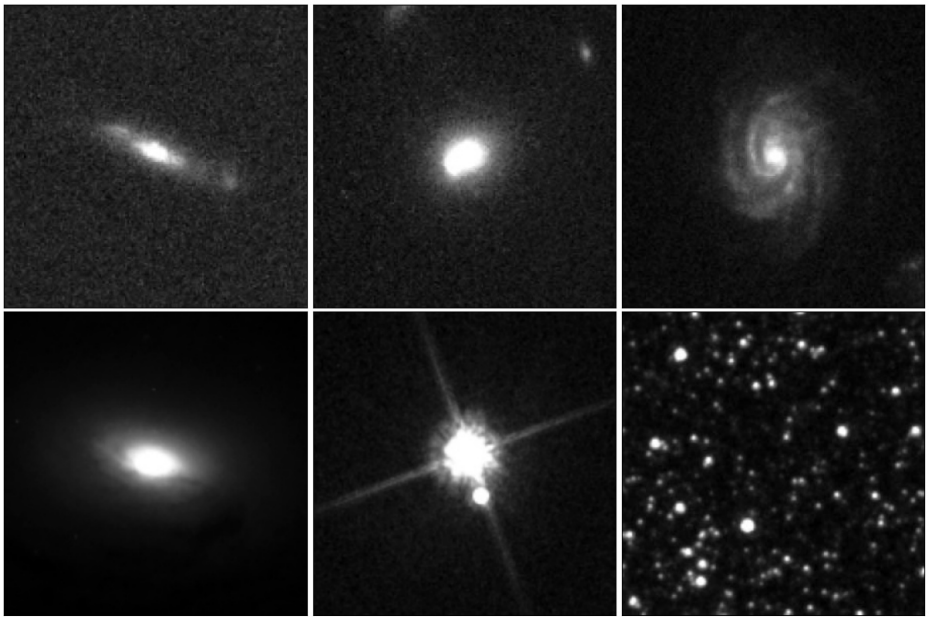

Figure 3.2: Example images of the labelled non-interacting galaxy systems used to train `Zoobot`. *Top Row*: Three examples from the Galaxy Zoo: *Hubble* project of the training set. *Bottom Row*: Three examples from the other Galaxy Zoo projects. These are, from right to left, Galaxy Zoo 2, Galaxy Zoo 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 $\geq$0.75. We correct the predictions `Zoobot` made and balance this set to 5,698. During this cycle, a large number of globular clusters/starfields/open clusters were given a very high prediction score. Figure 3.2 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.

3.6 DIAGNOSTICS

3.6.1 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 3.3 shows the resulting distribution. This distribution shows that our model has high confidence in what is or isn't an interacting system due to the high counts at very low and very high probability scores. It is likely the use of a balanced training set, and the very low volunteer score needed to define a source as non-interacting that leads to a strongly bi-model prediction score distribution. Using a balanced training set is an intrinsic trade off between ease of training, and potential biases introduced. Having a balanced dataset does not reflect reality, and leads `Zoobot` to over-predict interacting galaxies. Using very stringent volunteer classification cutoffs also leaves few ambiguous systems in the validation set, further enhancing this bi-modality.

The prediction score must be reduced to a binary classification for our problem. We use Figure 3.3 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 3.4 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 3.4 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 3.5 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 3.5 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

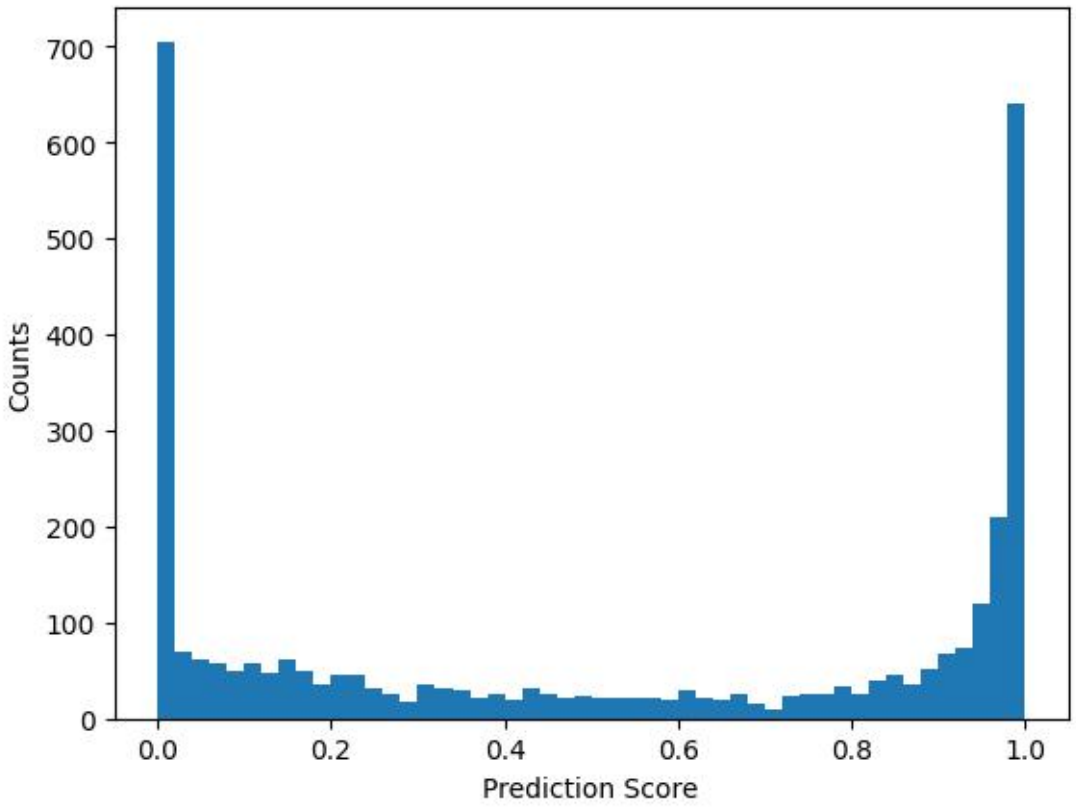

Figure 3.3: The distribution of prediction scores given to our validation set of 3,270 labelled sources set aside by `Zoobot` in training. These were split into 1,648 non-interacting sources and 1,622 interacting sources. As can be seen from the distribution, our model is often confident when a source 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.

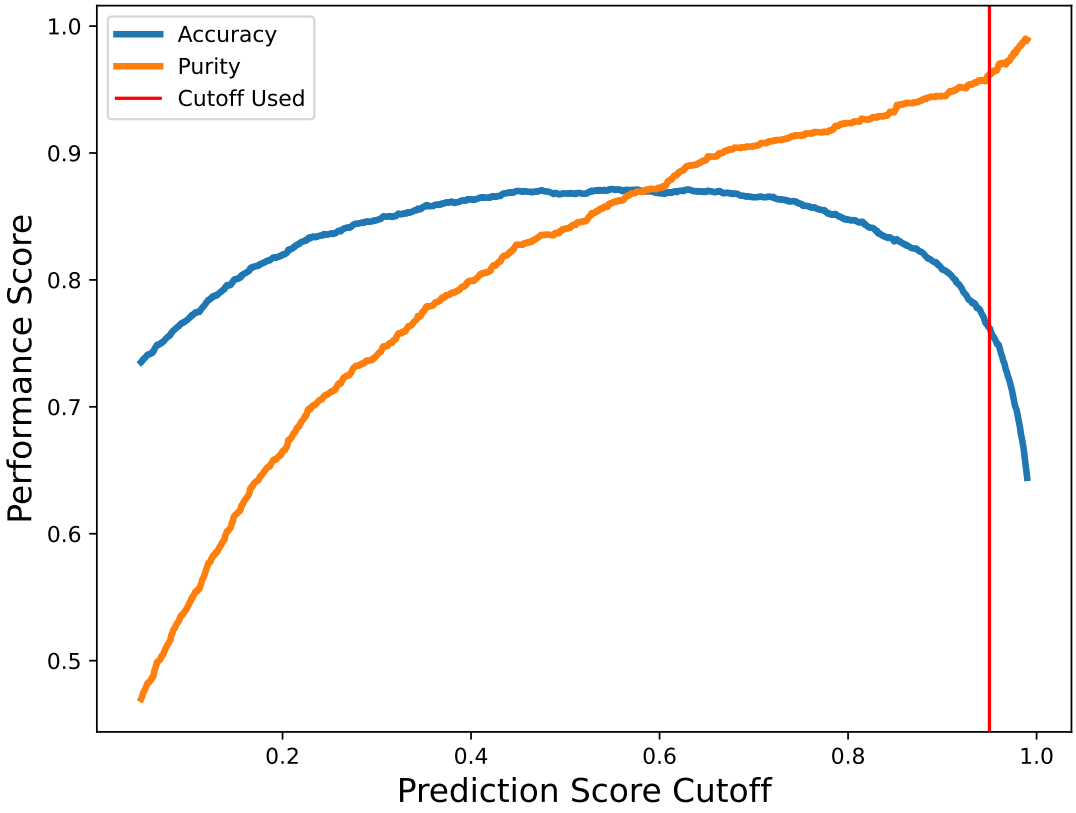

Figure 3.4: A measure of accuracy and purity against prediction score. The accuracy (in blue) is a direct measure of the number of sources `Zoobot` correctly predicted vs the total number of predictions made. The measure of purity (in orange) is the the number of predictions `Zoobot` correctly made vs 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 3.5 shows the final contamination in our validation catalogue would be ≈2%, where Figure 3.4 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 ≈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.

3.6.2 Duplication Removal

The fully trained `Zoobot` made predictions on ≈126 million extended sources from the HSC that had passed our creation criteria. Of these, 195,688 sources were given a score of 0.95 or greater, ≈0.2% of the total number of sources. Upon visually inspecting a subset of sources, it is clear that our `Zoobot` model had predicted for an interacting galaxy even if it was not the central (and, therefore, target) source in the image. This is due to the misalignment of sources from the centre in the training set as described in Section 3.5. `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.

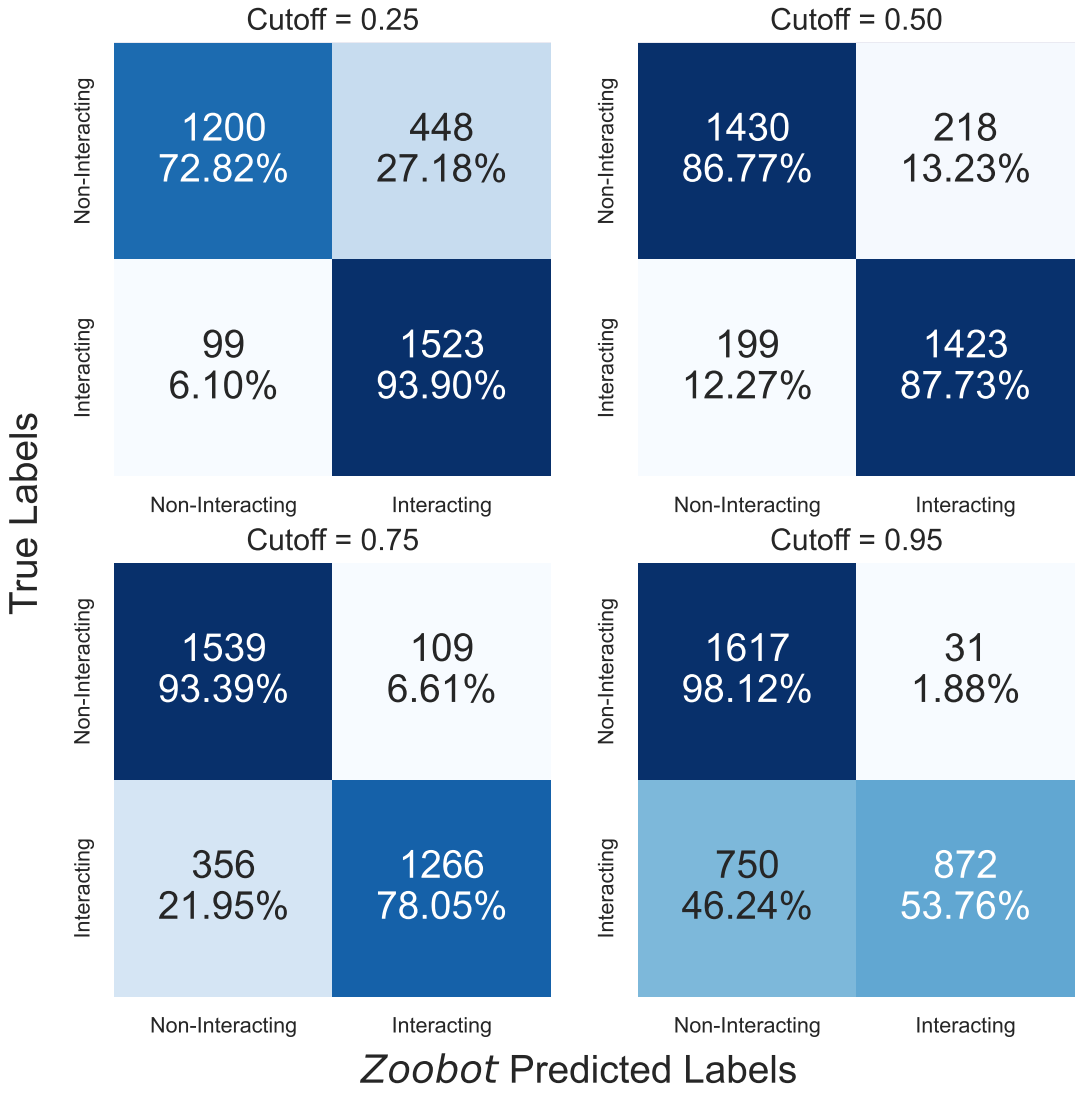

Figure 3.5: Confusion matrices of four different cutoffs of prediction score defining a binary classification of interacting galaxy or not. Confusion matrices break down our accuracy measurement into how `Zoobot` is misclassifying sources. At a cutoff of 0.50, the accuracy is highest at 88.2%. However, at this cutoff, ≈10% of our final catalogue would contain contamination. We elect to use the very stringent prediction cutoff of 0.95 for the rest of this work as it will return the lowest contamination.

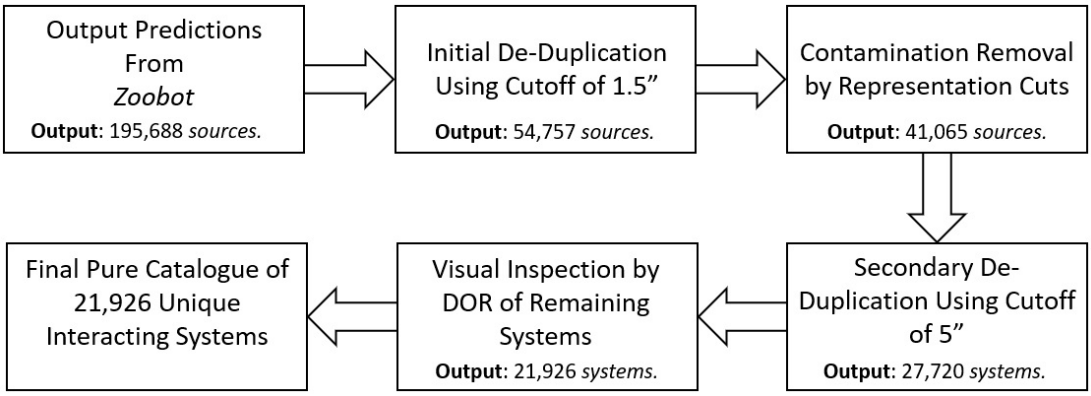

Figure 3.6: Flow diagram of our contamination and duplication removal process. De-duplication used agglomerative clustering based on sky separation. The first step of de-duplication uses a cutoff of 1.5″. This significantly reduced duplication in the catalogue, as well as the size of the catalogue to 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 3.6.3. 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 3.6 shows a full breakdown of the steps in our de-duplication and contamination removal process.

3.6.3 Bad Predictions & Removal

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

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

We remove the classification head of `Zoobot` and directly output the final layer of the feature learning section of the model. This gives 1,280 features (the representations) for each of our 27,720 candidate systems. However, there will be much redundant information in this very high dimensional feature space. We compress this using incremental principal component analysis (PCA) (Ross et al., 2008). An excellent demonstration of using this approach can be found in Walmsley et al. (2022b). We reduce the dimensionality from 1,280 to 40(as in Walmsley et al. (2022b)), and input the resultant components into the Auto-Encoder `UMAP`

(McInnes et al., 2018). `UMAP` projects the 40 dimensional components of each candidate system onto a 2 dimensional manifold. The position of each galaxy on this manifold is directly linked to its visual morphology. Close pairs have similar visual features which will then appear as a cluster in our representation space.

Figure 3.7 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 3.7 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 3.8 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

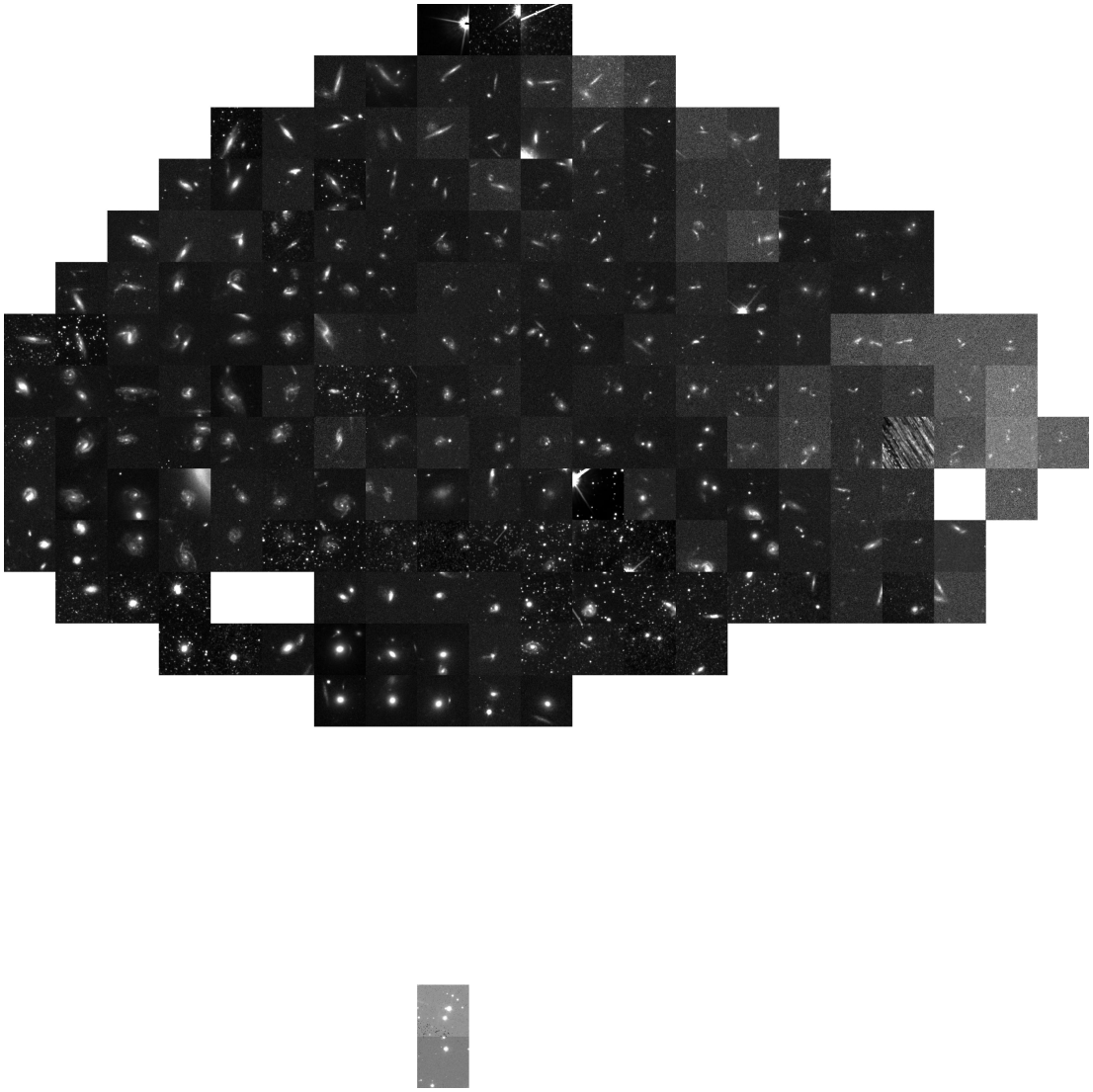

Figure 3.7: The representation distribution of 54,757 candidate interacting galaxies. This distribution is the compressed 2D representation of the 1,280 dimensional representation that `Zoobot` has learned of each image. Each image is a randomly selected one from sources within each bin in the distribution. 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 3.8.

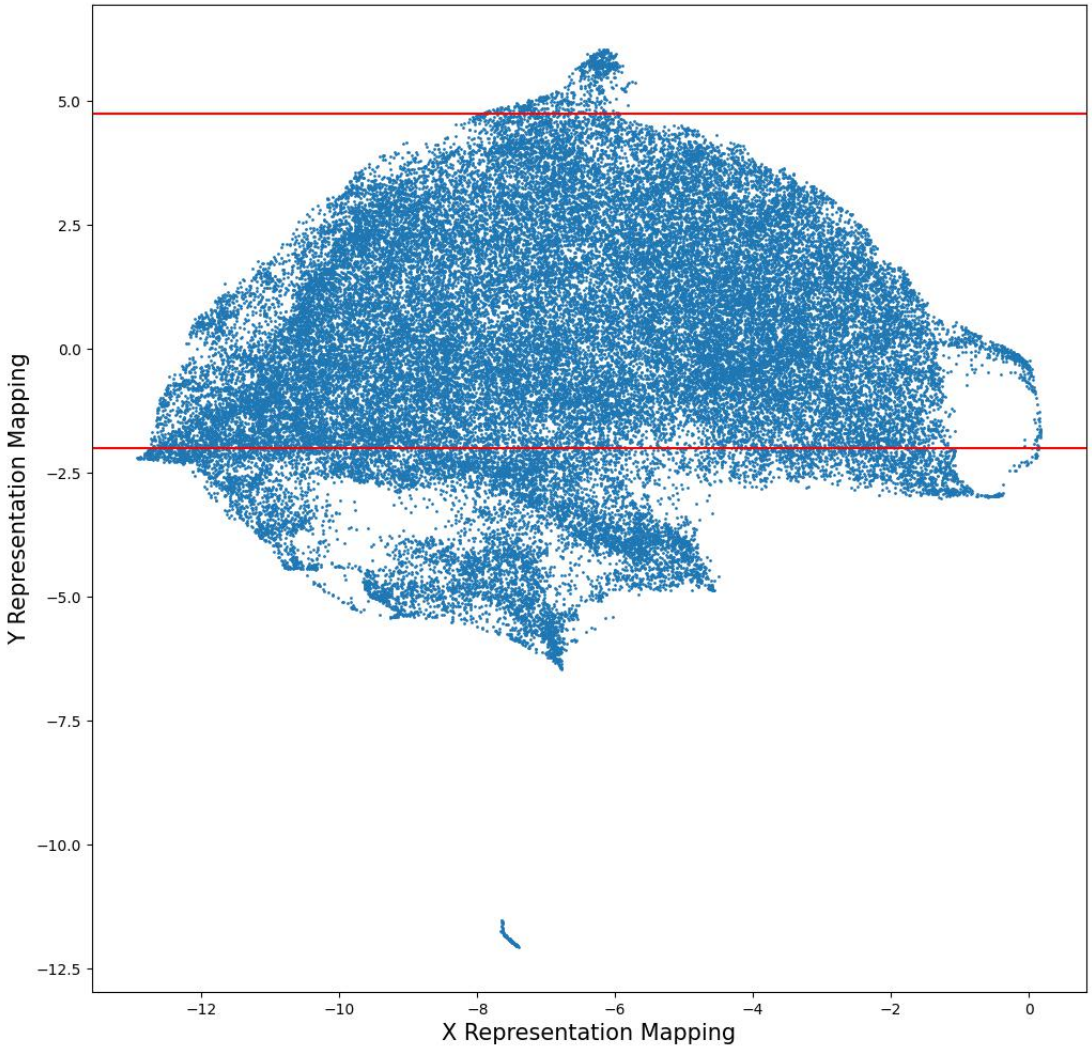

Figure 3.8: Scatter plot showing the precise distribution of each representation of sources in the remaining 54,757 sources. This is the unbinned version of Figure 3.7. The two red lines show the cutoffs utilised to remove the majority of close pairs by projection as well as the very obvious contamination 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 3.6.2, 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 3.7.2.

3.7 RESULTS & DISCUSSION

3.7.1 An Interacting Galaxy Catalogue

Upon de-duplication and contamination removal described in Sections 3.6.2 and 3.6.3, our final catalogue contains 21,926 interacting systems. Figure 3.9 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.

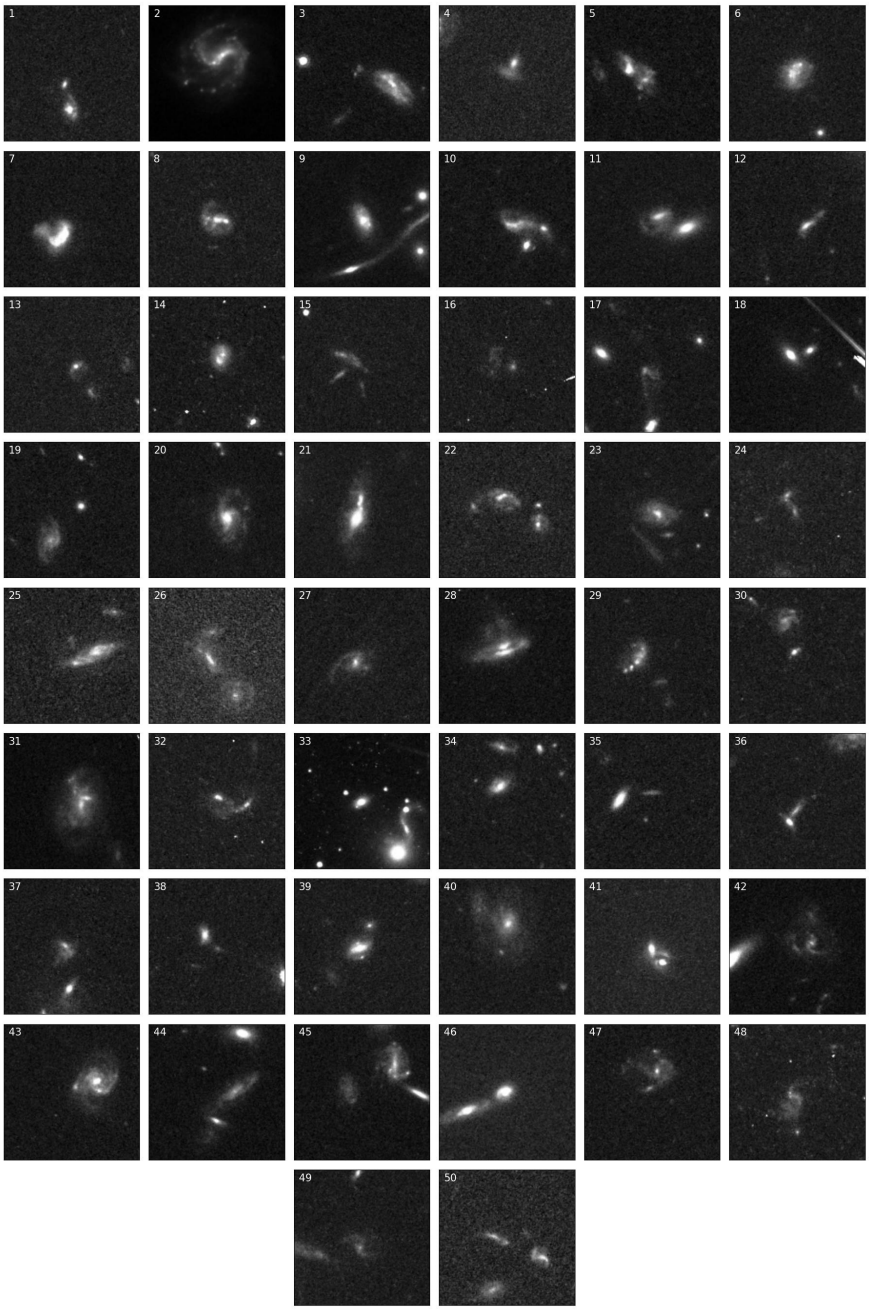

Figure 3.9: An example of 50 of the final interacting systems found with `Zoobot`. These were selected randomly from the de-duplicated and de-contaminated 21,926 sources. Each of these examples have extended tidal features and distortion. Not all of the final interacting systems have two galaxies within 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 and 14,404 do not. A flag exists in the catalogue data release which shows whether a system has references associated with it or it could be considered a 'new' system. We, however, do not claim that these systems are discovered by ourselves. These systems have always existed in the backgrounds of large surveys or observations and been discovered by others, it is only with ESA Datalabs that we can apply a methodology such as in this work to extract those systems from these observations. We also do not claim that these unreferenced systems are particularly interesting or phenomenal. It is most likely that these systems are the very faint background galaxies in surveys or observations whose main objective was something other than finding interacting galaxies. This will be further discussed in Section 3.7.3.

Figure 3.10 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 3.1 shows an example of the data and format of the 50 sources shown in Figure 3.9. We also bootstrap the final catalogue as an estimate of contamination remaining. As described in Section 3.6.3, the final step of contamination removal was visual inspection by DOR of the 27,720 candidate interacting systems

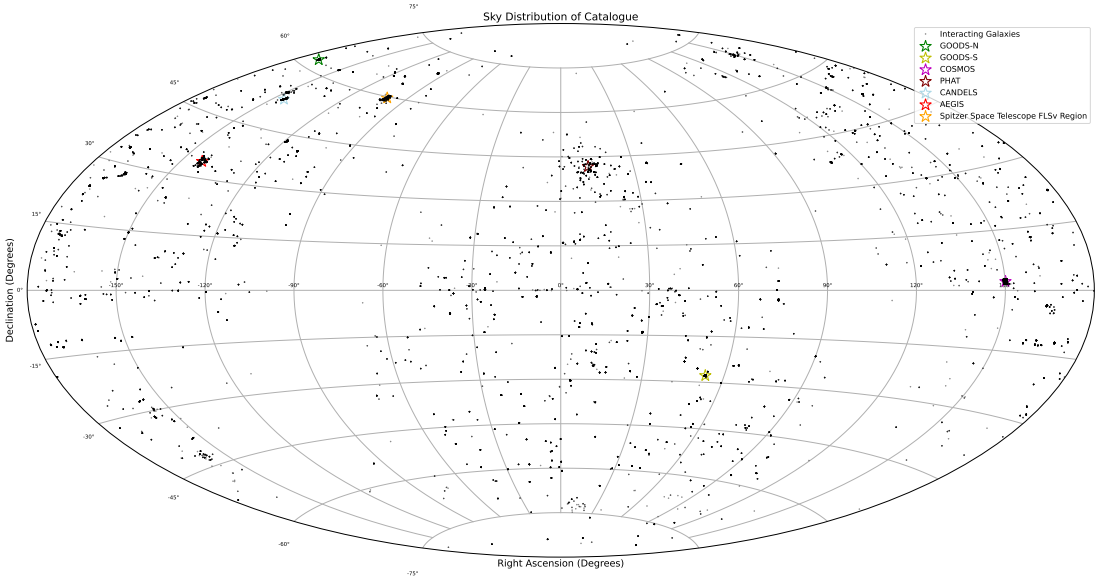

Figure 3.10: Sky Distribution of our catalogue, with marked positions of well known deep surveys conducted by *HST*. *HST* is able to observe almost the entire sky and therefore the interacting galaxies are scattered throughout. Large clusters of sources are found in the locations of surveys. This shows 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 ≈3% of our interacting system in the final catalogue remains contamination.

3.7.2 The Gems

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

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	[201
4	4553390202	73.581297	2.903528	0.968280	
5	4000907600174	259.037474	59.657617	0.999978	
6	4575187799	150.001883	2.731942	0.974649	[200
7	4000717342023	149.527791	2.126945	0.993912	[200
8	4001174802281	28.593114	-59.643515	0.982890	
9	4182689774	186.709991	21.835419	0.973232	[2016ApJS..224
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
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	[200
18	4000877021787	116.211231	39.462563	0.979178	
19	4000878525229	149.834893	2.516816	0.963694	[2007ApJS..172
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	[200
24	4001200639012	54.037618	-45.170026	0.991404	
25	4000921402261	150.417634	2.313781	0.990775	[2018ApJ...858
26	4001224732336	337.217339	-58.444885	0.955972	
27	4000781402752	216.968619	34.575819	0.974076	
28	4001283017901	120.202582	36.058927	0.994169	[201
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	[201
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	[201
45	4001282607544	333.765783	-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 ESASky[1](Merín et al., 2017), NASA Extragalactic Database (NED) and the Sloan Digital Sky Survey were also used to investigate any unreferenced objects. ESASky was of paramount importance as we could investigate many objects across a range of wavelengths with many instruments.

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

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

Table 3.2 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.

3.7.3 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`.

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

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

Table 3.2: 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 3.11 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 3.4), given that finetuning is based primarily on the $z \lesssim 1$ imagery of Galaxy Zoo: *Hubble*. Therefore, the currently measured redshift distribution in Figure 3.11 is likely due to some combination of selection bias and training bias.

Figure 3.12 shows the basic parameter space sampled by the sub-sample of the catalogue with existing photometry and redshifts. We show the distributions

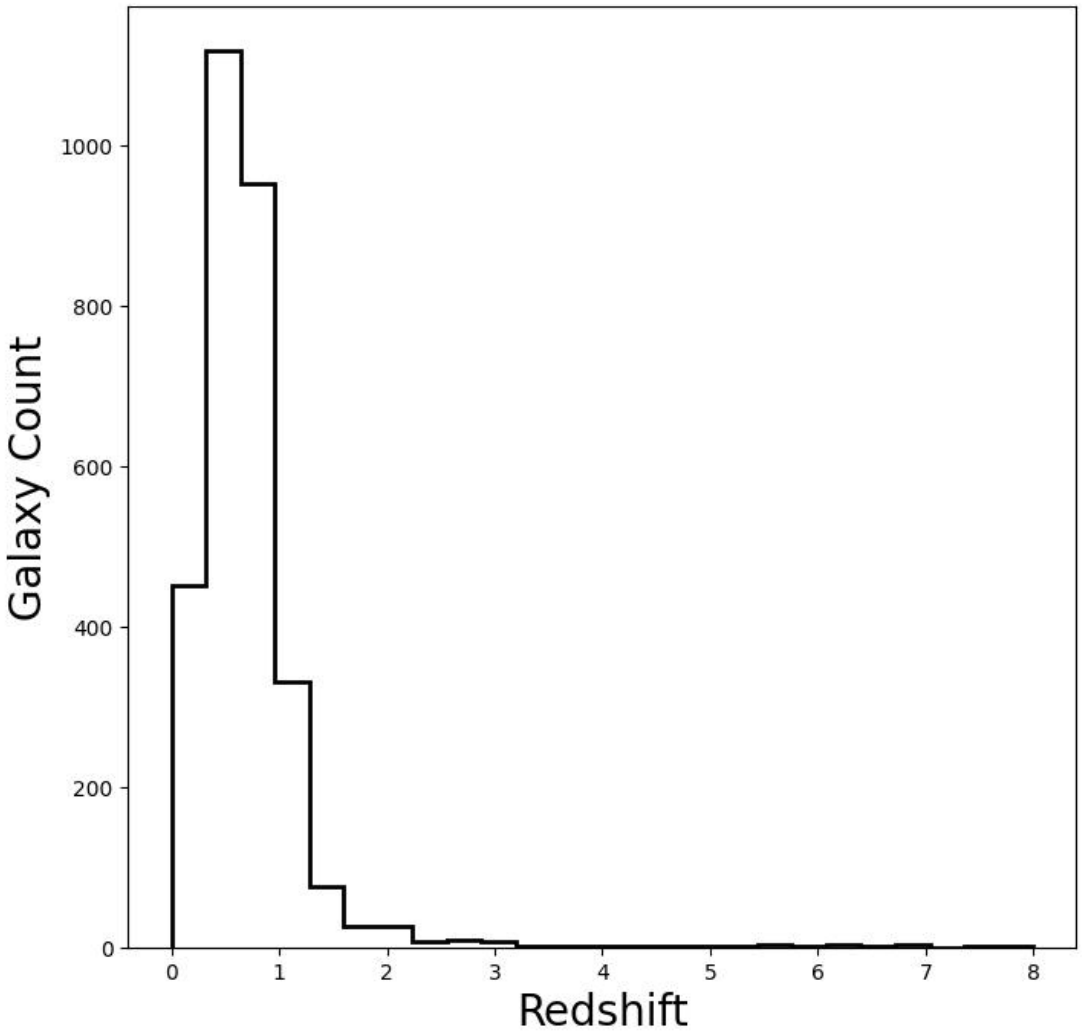

Figure 3.11: The redshift distribution of a subsample of our catalogue. Of the 7,583 referenced systems, 3,037 of them had redshift measurements in the NED, MAST or Simbad. This redshift distribution shows that our model confidently predicted interacting systems primarily for $z < 1$ systems. This was anticipated, as the model was primarily trained on systems at these redshifts. There are fifteen sources with a reported $z > 5$.

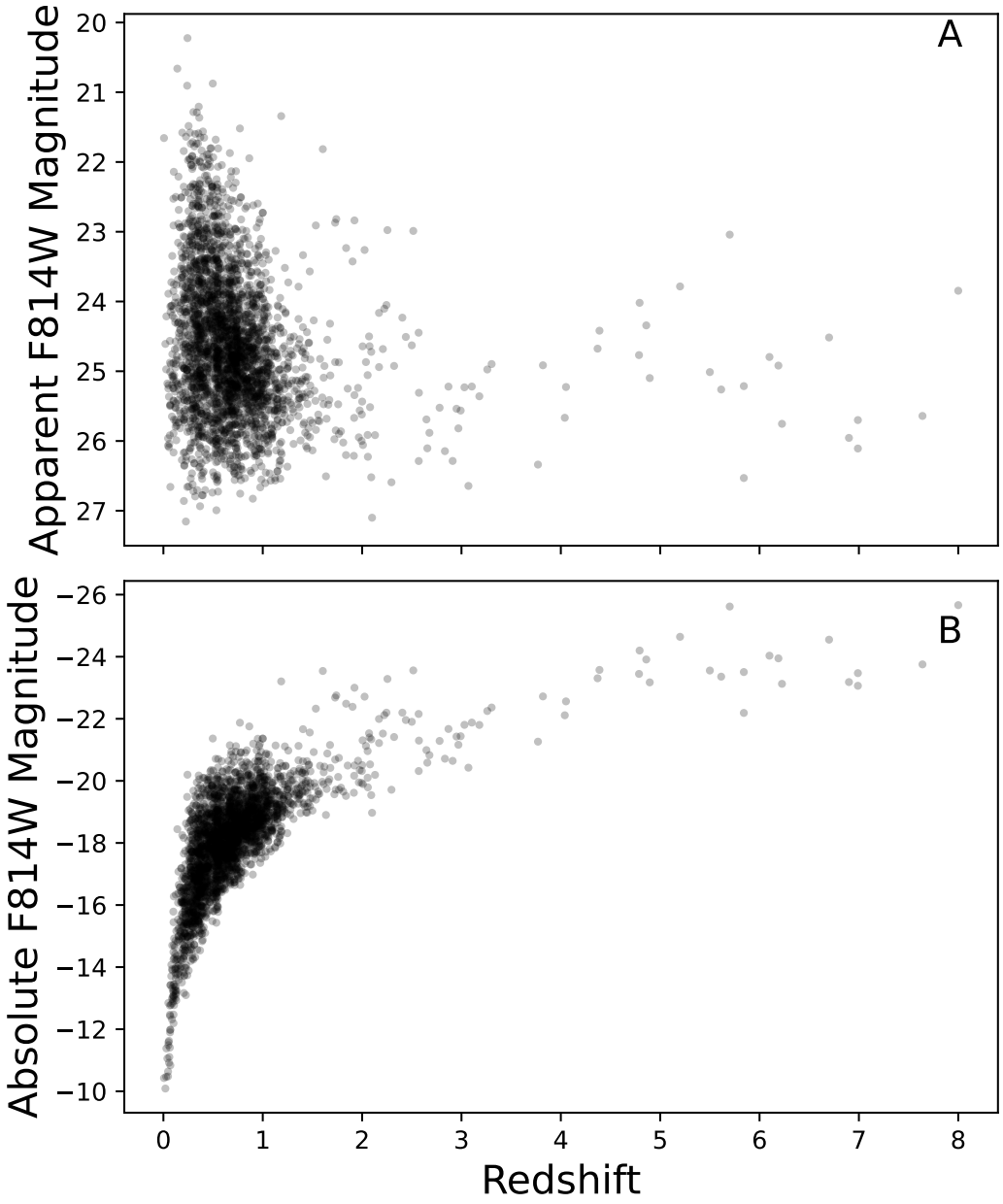

Figure 3.12: The distribution of redshift with magnitude for all sources with available data. This shows the parameter space we are sampling in this catalogue. Panel A shows that the majority of our sources are dim, background sources at low redshift. Panel B shows the faintest objects we find are at 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 3.3: 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 3.3 summarizes the filter coverage of our catalogue. 6.1% (1336 sources) have complete 3-band photometric information in the HSC. We use these to create examples of colour images from the catalogue (using the algorithm of Lupton et al., 2004). We used a scaling factor $Q = 2$ and $\alpha = 0.75$, with ($F814W$, $F606W$, $F475W$) as RGB channels and multiplicative factors of (1.25, 0.95, 2). The resultant images are shown in Appendix 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 3.13 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 3.13 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 3.13B 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 3.13C is approximately a rest-frame $F475W - F606W$ vs $F606W$ plot. At $0.62 < z < 1$, Figure 3.13D is approximately a rest-frame NUV-Blue plot ($F336W - F475W$ vs $F475W$).

The galaxies in Panel B are observed in approximately the rest frame $F606W$ and $F814W$ filters. Nearly all are blue systems (by general definitions at various redshifts, *e.g.*, Kauffmann et al., 2003; 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 3.13C, 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., 2008; 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,

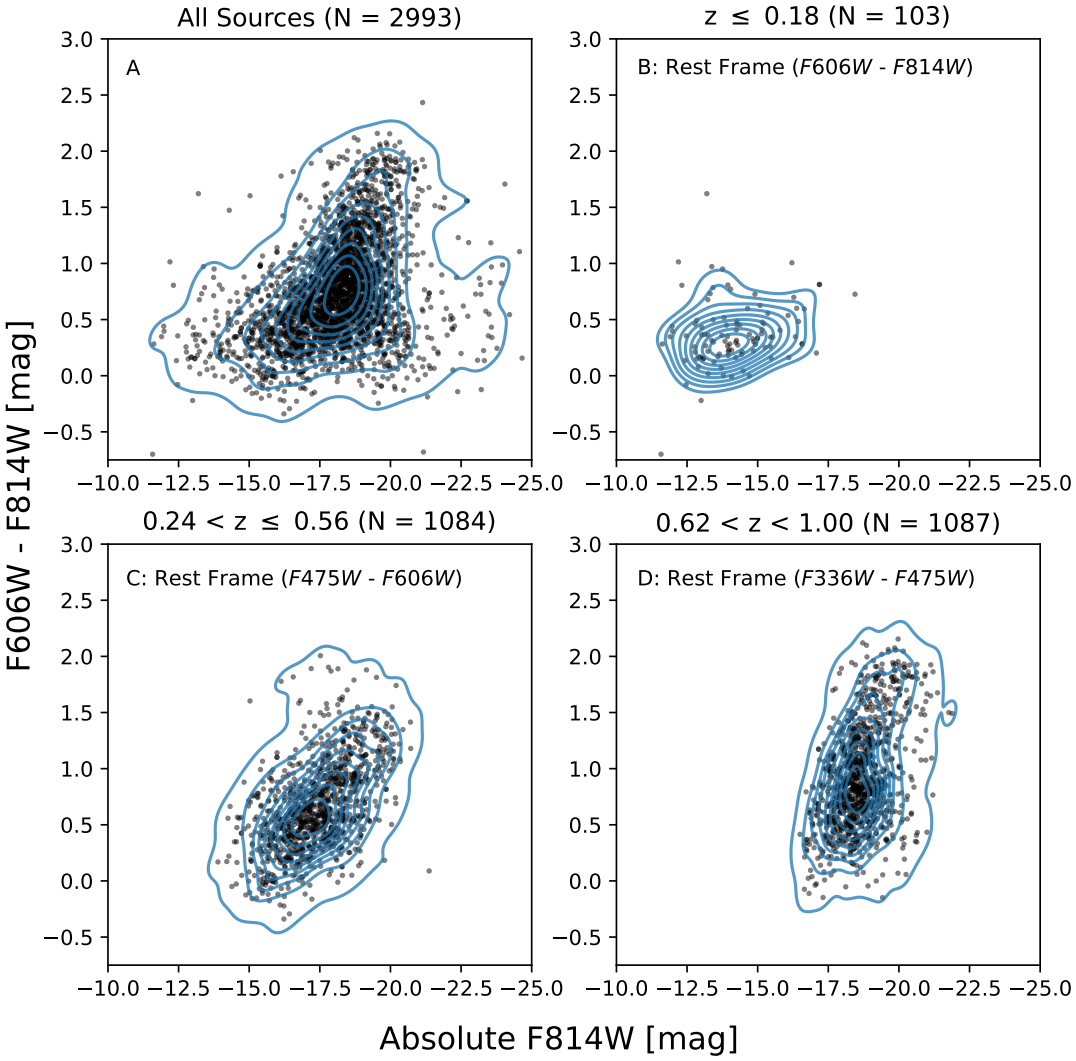

Figure 3.13: The colour-magnitude distribution of sources with a redshift measurement associated. Panel A shows the distribution of all galaxies, without controlling for redshift or dust extinction. The remaining panels then split these sources into distinct redshift bins where the $F606W$ and $F814W$ 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.

3.8 CONCLUSION

We present a large, pure catalogue of 21,926 interacting galaxy systems found from the *Hubble* Source Catalogue. This catalogue is a factor of six larger than previous works. Each interacting system was found using the European Space Agency's new platform ESA Datalabs, which allowed us to directly apply an advanced CNN - `Zoobot` - to the entire *Hubble* science archive. This corresponds to predicting over 126 million sources. The compiled catalogue has a contamination rate of ≈3% as found by bootstrapping. Table 3.1 shows an example of 50 entries in our new catalogue, Figure 3.9 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 3.8) 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 3.2 shows a breakdown of the objects found. These sub-catalogues have been released alongside our interacting galaxy catalogue. With the priority of purity in this catalogue creation, we will aim in future work to use it in the statistical analysis of interacting galaxies and begin linking the underlying parameters of interaction to the complex physical processes that occur in them. A secondary purpose of this

catalogue is to serve as a training set for future models which may wish to search for interacting or merging galaxies.

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

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

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

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

ACKNOWLEDGEMENTS

DOR gratefully acknowledges the support from European Space Agencies Visitor Archival Research program, and hosting at the European Space Astronomy Centre. DOR thanks Bruno Merín for supervising this project and Sarah Kendrew for aiding its creation. This project was conducted as part of DORs PhD program supported by the UK Science and Technology Facilities Council (STFC) under grant reference ST/T506205/1. BDS acknowledges support through a UK Research and Innovation Future Leaders Fellowship [grant number MR/T044136/1]. ILG acknowledges support from an STFC PhD studentship [grant number ST/T506205/1] and from the Faculty of Science and Technology at Lancaster University. MW gratefully acknowledges support from the UK Alan Turing Institute under grant reference EP/V030302/1. MRT acknowledges the support from an STFC PhD studentship [grant number ST/V506795/1] and from the Faculty of Science and Technology at Lancaster University.

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

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

This project used data from the *Hubble* Space Telescope and stored in the archives at the European Space Astronomy Centre. These observations are obtained from the Space Telescope Science Institute, which is operated by the Association of Universities for Research in Astronomy, Inc, under NASA contract NAS

5-26555. All sources were found using v3.1 of the *Hubble* source catalogue (Whitmore et al., 2016) and accessed using the ESA Datalabs science platform. ESA Datalabs is directly connected to the ESA *Hubble* Science Archive. This study makes use of data from AEGIS, a multiwavelength sky survey conducted with the Chandra, GALEX, Hubble, Keck, CFHT, MMT, Subaru, Palomar, Spitzer, VLA, and other telescopes and supported in part by the NSF, NASA, and the STFC.

DOR would like to thank those in the ESA Traineeship program cohort of 2022. They created a wholly welcoming environment and space of support. A special thanks must go to Karolin Frohnapfel and Emma Vellard for much technical discussion. Finally, DOR would like to acknowledge Aurélien Verdier.

Chapter 4

Galaxy Interaction in COSMOS

4.1 INTRODUCTION

4.2 DATA

4.2.1 The O'Ryan+23 Catalogue

4.2.2 The COSMOS2020 Catalogue

4.3 Galaxy Classification

4.3.1 Into Stage of Interaction

4.3.2 AGN Identification

4.4 RESULTS & DISCUSSION

4.4.1 Star Formation with Stage

4.4.2 AGN Activity with Stage

4.4.3 Controlling for Environment

4.4.4 Limitations of Approach

4.5 CONCLUSION

Chapter 5

Conclusion

5.1 SOFTWARE FOR STATISTICAL CONSTRAINT

5.2 CREATING LARGE SAMPLES WITH MACHINE LEARNING

5.3 APPYING 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 3.6 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 3.6, 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. (2004) algorithm, with a scaling factor $Q = 2$ and $\alpha = 0.75$, with ($F814W$, $F606W$, $F475W$) as RGB channels and multiplicative factors of (1.25, 0.95, 2).

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

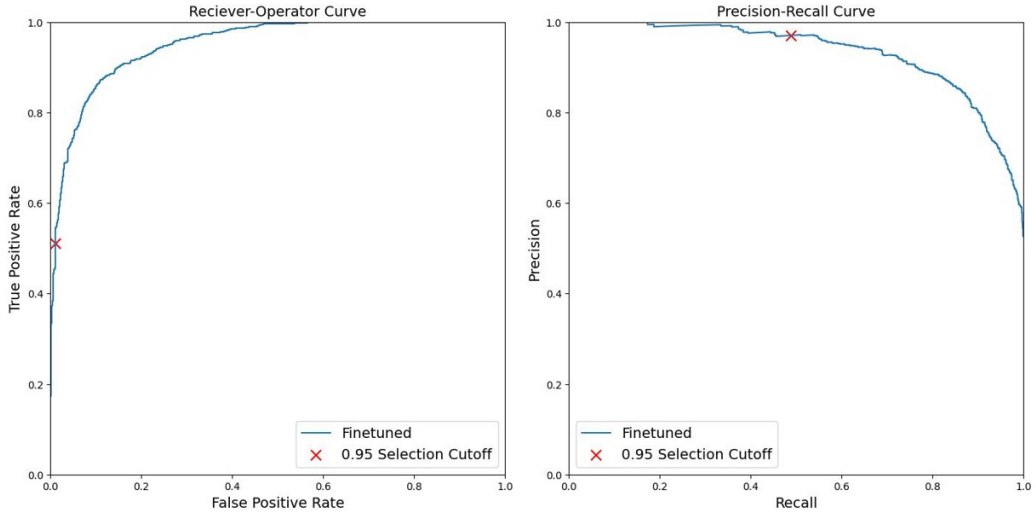

Figure B.1: The Receiver-Operator and Precision-Recall Curve for the `Zoobot` model that was used to explore the Hubble archives. The blue curves are the measured curves. These curves measure the relevant rates or characteristics based on the changing cutoff applied to how `Zoobot` defines an interacting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

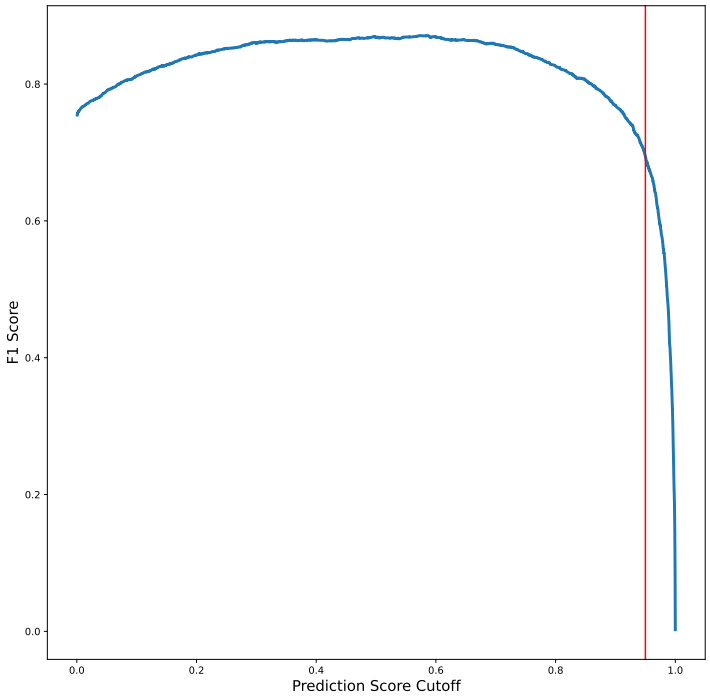

Figure B.2: The F1 score found during the diagnostics of the model used in this work. The F1 score is a measure combing the measure of accuracy and purity into one metric. The cutoff we use is at the point where the F1 score begins to rapidly decline. This point is shown by the red vertical line.

Figure 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 disruped system with a significantly elongated tidal feature. The final two unknown sources (6000186797547 and 6000341449179) have no clear central source, though there is extended structure to them. These are likely to be highly irregular galaxies, but no confirmation could be found.

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

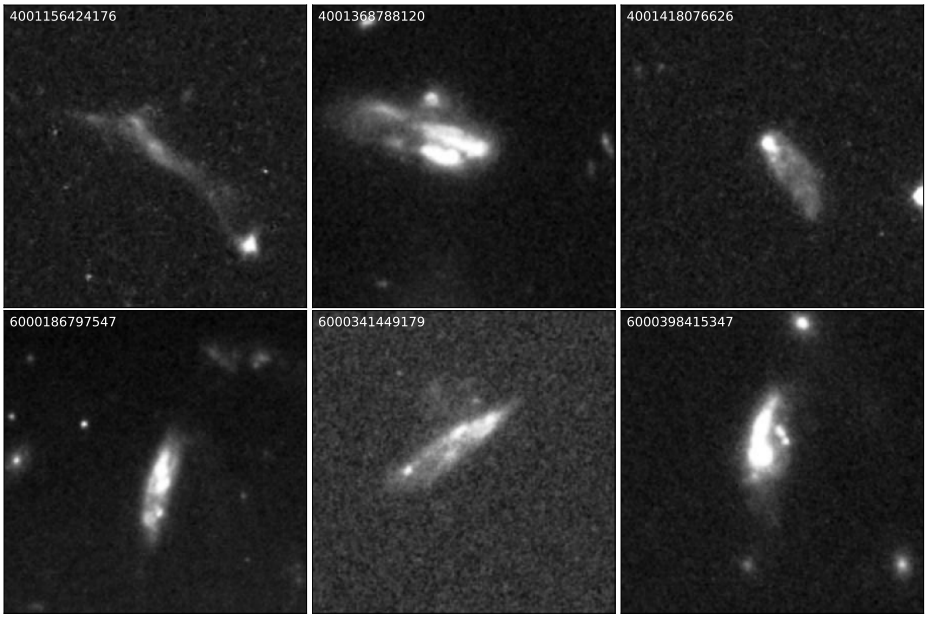

Figure B.4: The six unknown systems found in this work. These have no reference in Simbad or in NED, and their morphology could not be classified by the authors. Investigation into these six objects are presented to the community, with the authors hoping that future work and investigation of them can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

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

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

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

Arp H., 1966, ApJS, 14, 1

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

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

Astropy Collaboration et al., 2018, 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`

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

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

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

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

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

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

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

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

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

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

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

Das A., Pandey B., Sarkar S., 2022, arXiv e-prints, p. arXiv:2207.03968

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

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

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

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

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

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

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

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

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

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

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., 2020, 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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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., 1996, ApJ, 464, 641

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

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

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

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

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

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

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

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

Rejkuba M., Greggio L., Harris W. E., Harris G. L. H., Peng E. W., 2005, ApJ, 631, 262

Ross D., Lim J., Lin R., et al. 2008, Int J Comput Vis, p. 125–141

Saitoh T. R., Daisaka H., Kokubo E., Makino J., Okamoto T., Tomisaka K., Wada K., Yoshida N., 2009, PASJ, 61, 481

Schawinski K., et al., 2014, MNRAS, 440, 889

Scoville N., et al., 2007, ApJS, 172, 1

Scudder J. M., Ellison S. L., Torrey P., Patton D. R., Mendel J. T., 2012, MNRAS, 426, 549

Silva A., Marchesini D., Silverman J. D., Martis N., Iono D., Espada D., Skelton R., 2021, ApJ, 909, 124

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

Springel V., 2000, MNRAS, 312, 859

Springel V., et al., 2005, Nature, 435, 629

Toomre A., Toomre J., 1972, ApJ, 178, 623

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., Holincheck A. J., Harvey A., 2016, Astronomy and Computing, 16, 26

Walmsley M., et al., 2022a, MNRAS, 509, 3966

Walmsley M., et al., 2022b, MNRAS, 513, 1581

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

Whitmore B. C., et al., 2016, AJ, 151, 134

Willett K. W., et al., 2013, MNRAS, 435, 2835

Willett K. W., et al., 2017, MNRAS, 464, 4176

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

van Dokkum P. G., 2005, AJ, 130, 2647

van der Walt S., et al., 2014, PeerJ, 2, e453