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 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.

Quote

Contents

List of Figures

Relevant Publications by the Author

Chapter 2

- Publication

Chapter 1

Introduction

Chapter 2

Chapter

2.1

Chapter 3

Chapter

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 imagery. 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[1]; 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/`

[1]`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

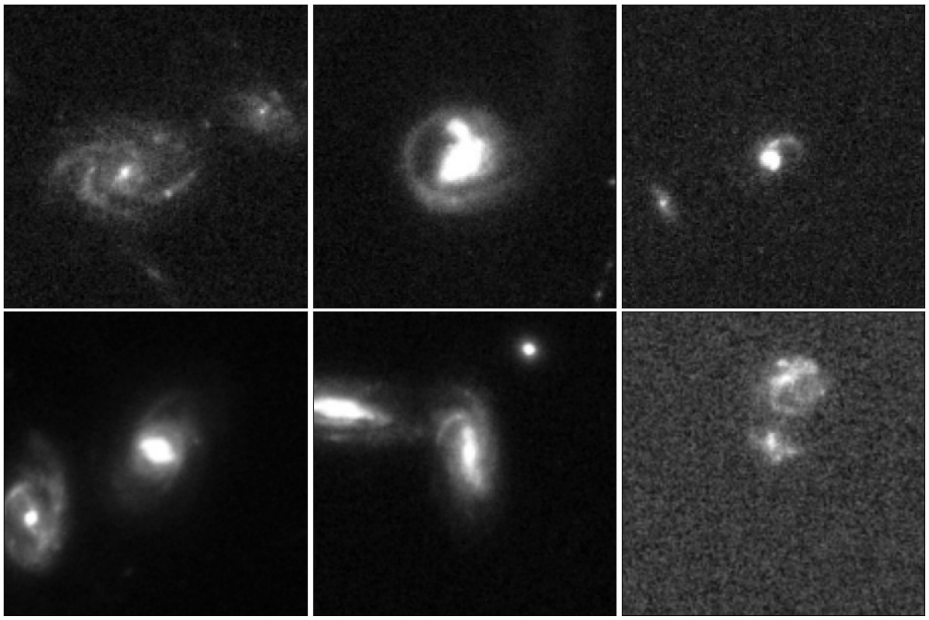

Figure 3.1: Example images of the labelled interacting galaxy systems used to train `Zoobot`. Each galaxy had a weighted vote fraction $\geq$0.75 in Galaxy Zoo. *Top Row*: Three examples from the Galaxy Zoo: *Hubble* project of the training set. *Bottom Row*: Three examples from the other Galaxy Zoo 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.

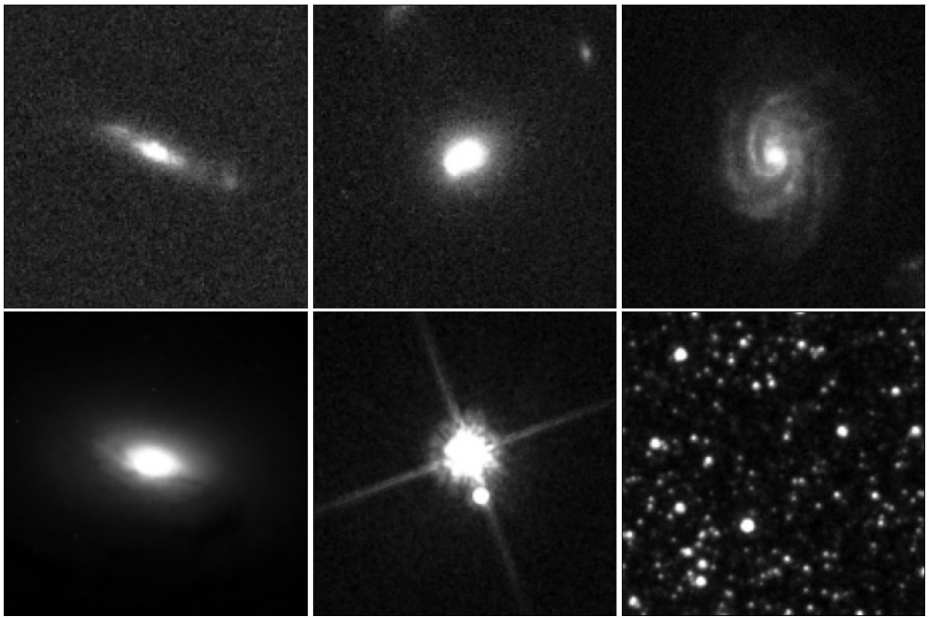

Figure 3.2: Example images of the labelled non-interacting galaxy systems used to train `Zoobot`. *Top Row*: Three examples from the Galaxy Zoo: *Hubble* project of the training set. *Bottom Row*: Three examples from the other Galaxy Zoo projects. These are, from right to left, Galaxy Zoo 2, Galaxy Zoo 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 A.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

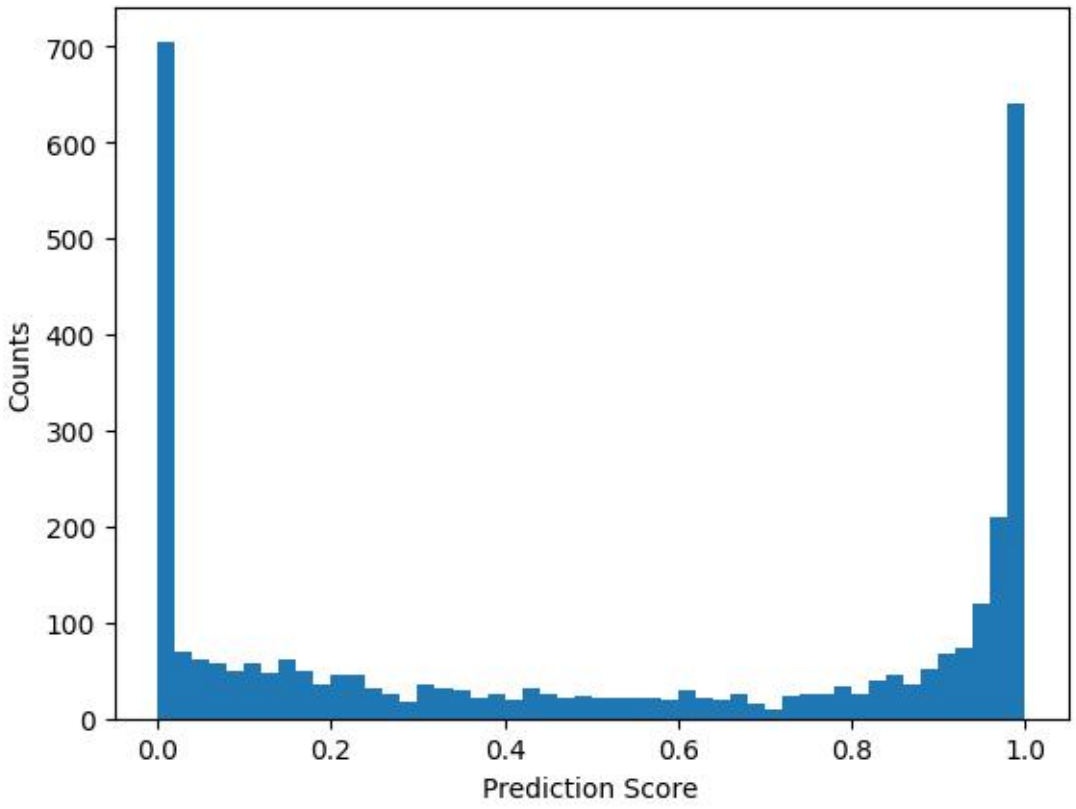

Figure 3.3: The distribution of prediction scores given to our validation set of 3,270 labelled sources set aside by `Zoobot` in training. These were split into 1,648 non-interacting sources and 1,622 interacting sources. As can be seen from the distribution, our model is often confident when a source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

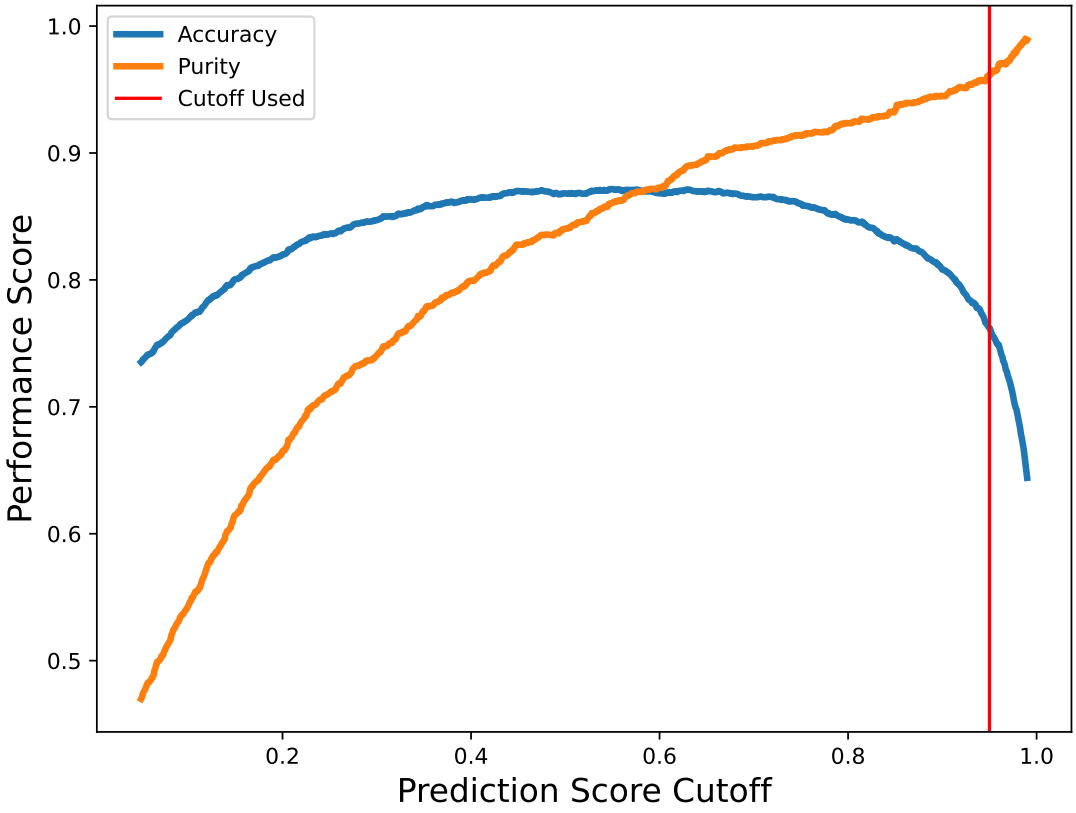

Figure 3.4: A measure of accuracy and purity against prediction score. The accuracy (in blue) is a direct measure of the number of sources `Zoobot` correctly predicted vs the total number of predictions made. The measure of purity (in orange) is the the number of predictions `Zoobot` correctly made vs 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 A.2 shows this change with prediction cutoff in the Appendix.

contamination in our final catalogue. Of the 1,622 galaxies predicted to be interacting, 218 would be non-interacting systems - approximately 13%. Our main aim in this work is to present a highly pure, large interacting galaxy catalogue that can be used for statistical exploration of interacting galaxy parameter space. Therefore, we use a very stringent cutoff of 0.95.

Using a cutoff of 0.95 reduces contamination significantly. Figure 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

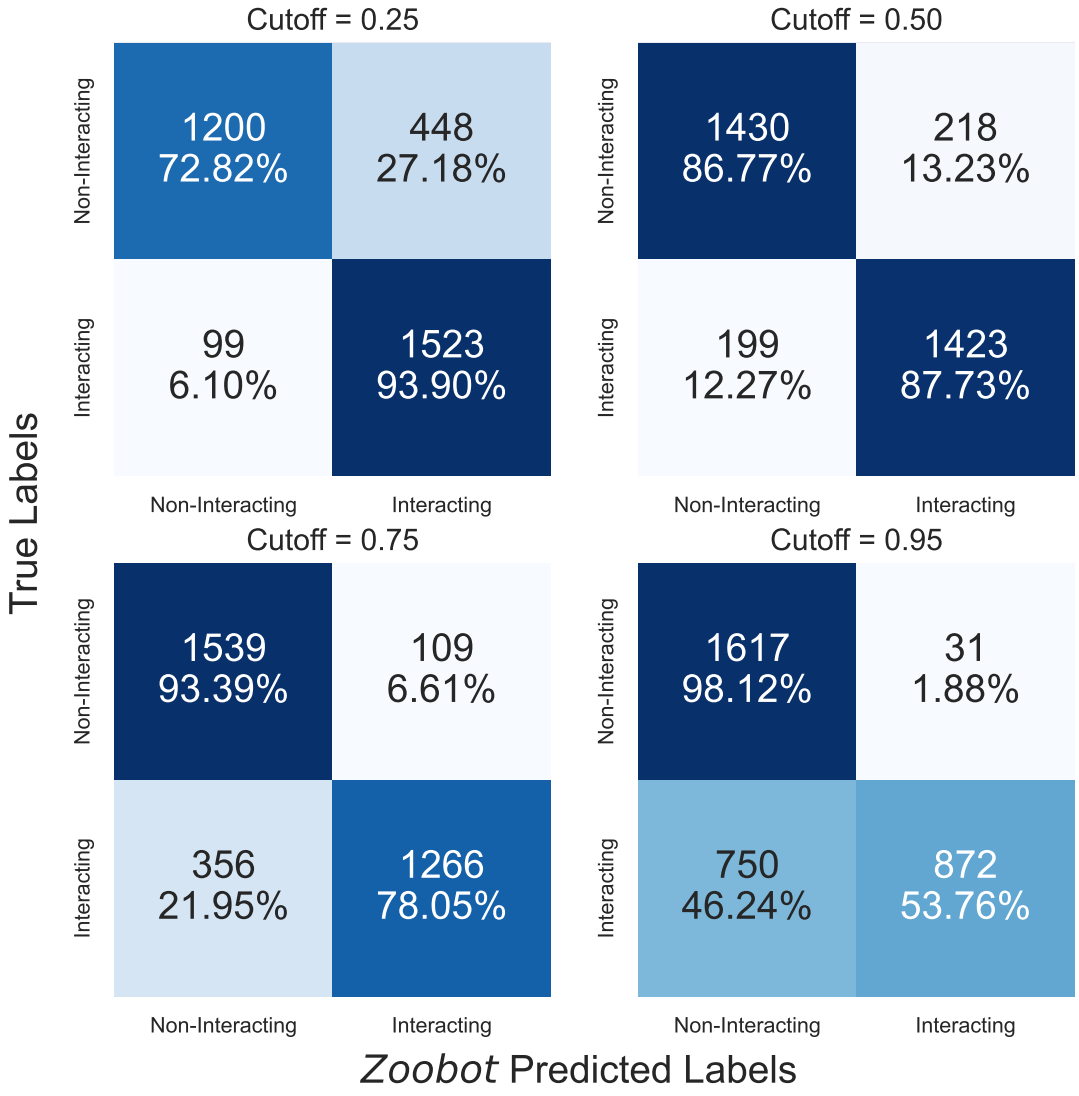

Figure 3.5: Confusion matrices of four different cutoffs of prediction score defining a binary classification of interacting galaxy or not. Confusion matrices break down our accuracy measurement into how `Zoobot` is misclassifying sources. At a cutoff of 0.50, the accuracy is highest at 88.2%. However, 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.

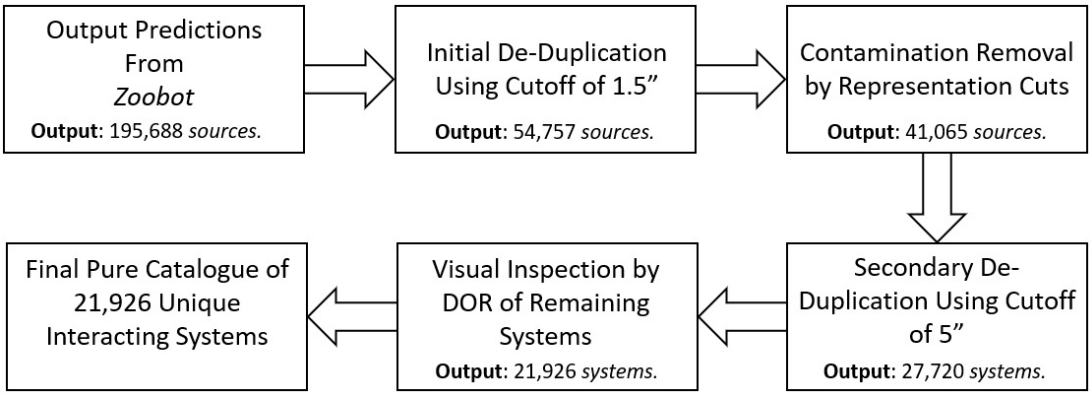

Figure 3.6: Flow diagram of our contamination and duplication removal process. De-duplication used agglomerative clustering based on sky separation. The first step of de-duplication uses a cutoff of 1.5″. This significantly reduced duplication in the catalogue, as well as the size of the catalogue to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

central source. Therefore, many interacting systems were duplicated in our final catalogue, appearing in cutouts were the central source was not interacting.

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

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

Initially, we utilise a limiting sky separation of $1.5''$ to remove the duplication. This reduced the size of our potential catalogue to 54,757 interacting galaxy candidates. We then applied contamination removal as described in Section 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

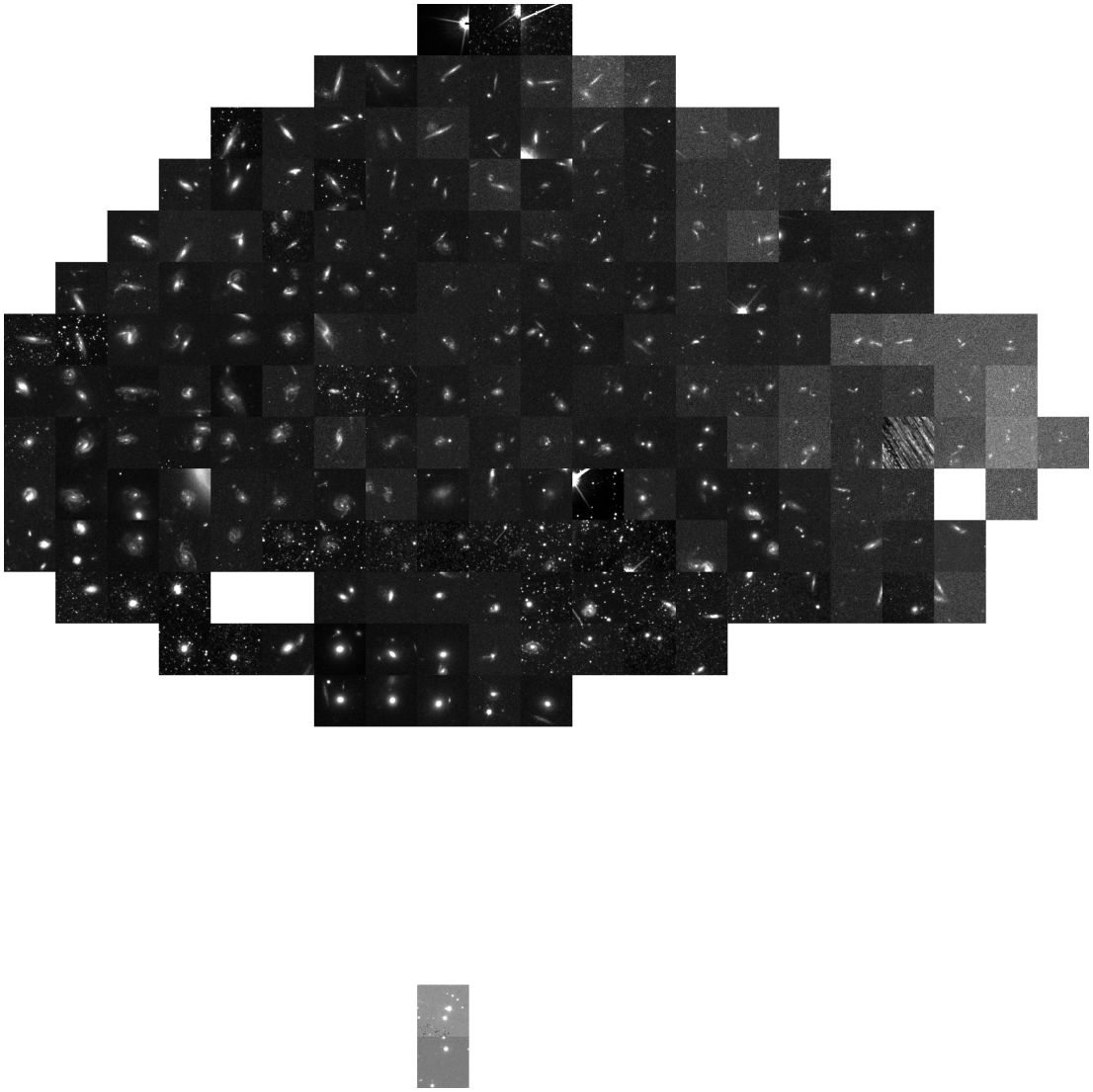

Figure 3.7: The representation distribution of 54,757 candidate interacting galaxies. This distribution is the compressed 2D representation of the 1,280 dimensional representation that `Zoobot` has learned of each image. Each image is a randomly selected one from sources within each bin in the distribution.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

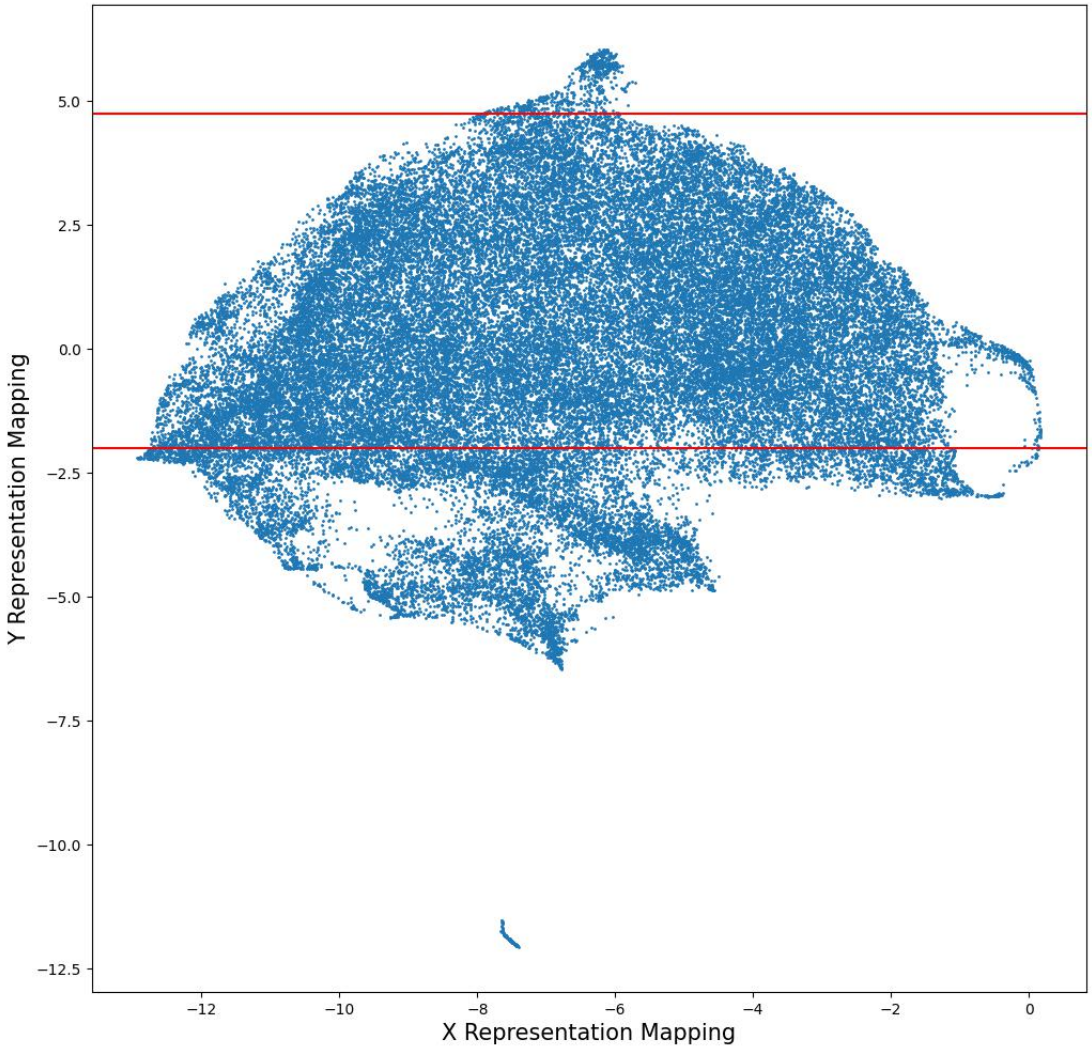

Figure 3.8: Scatter plot showing the precise distribution of each representation of sources in the remaining 54,757 sources. This is the unbinned version of Figure 3.7. The two red lines show the cutoffs utilised to remove the majority of close pairs by projection as well as the very obvious contamination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

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

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

As described in Section 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.

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

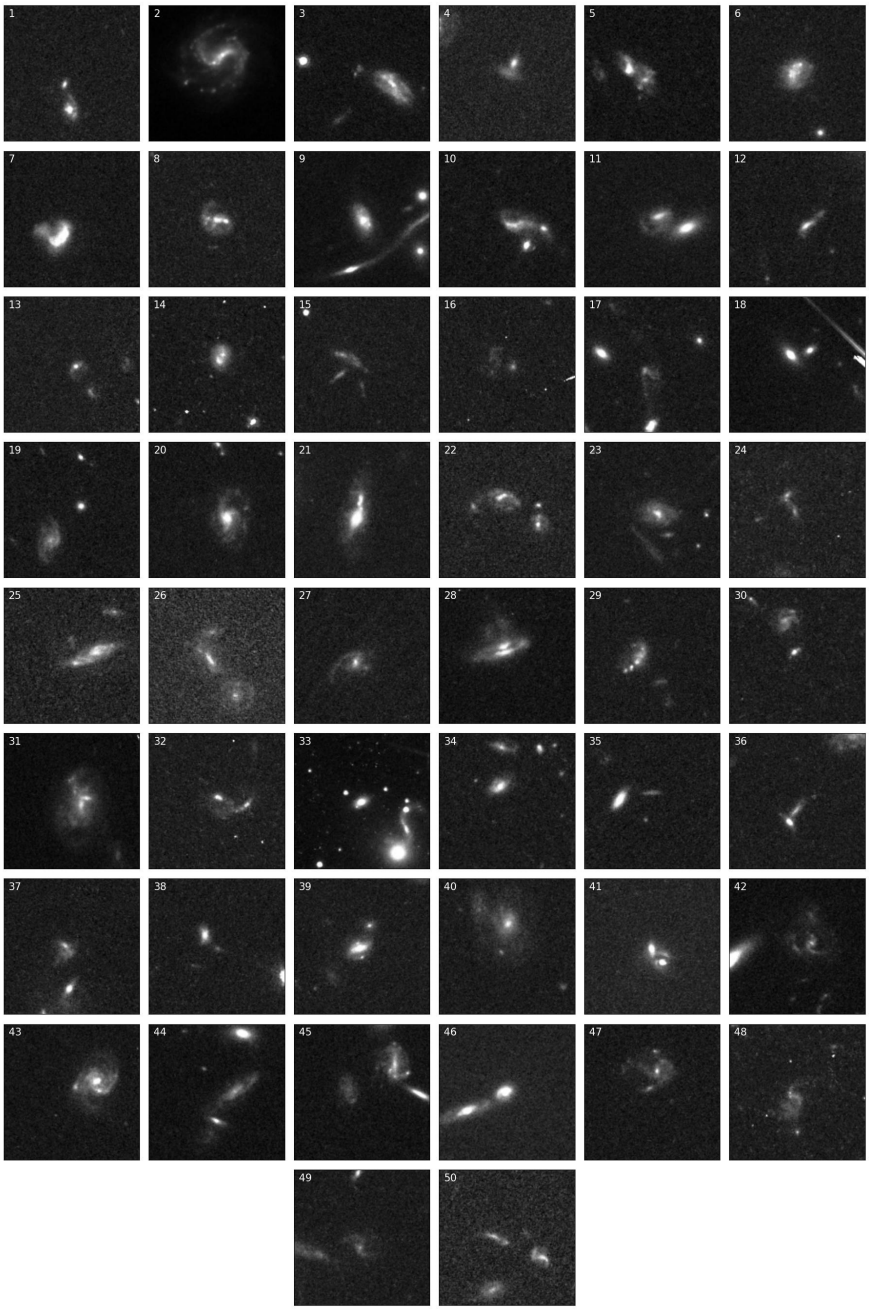

Figure 3.9: An example of 50 of the final interacting systems found with `Zoobot`. These were selected randomly from the de-duplicated and de-contaminated 21,926 sources. Each of these examples have extended tidal features and distortion. Not all of the final interacting systems have two galaxies within them (for example, image 2), but are clearly very disturbed by a tidal event. These were kept in as they would form a large part of the interacting galaxy population and would be flagged as disturbed or interacting in Galaxy Zoo. Each of these images is a 1-colour image using the $F814W$ *HST* filter.

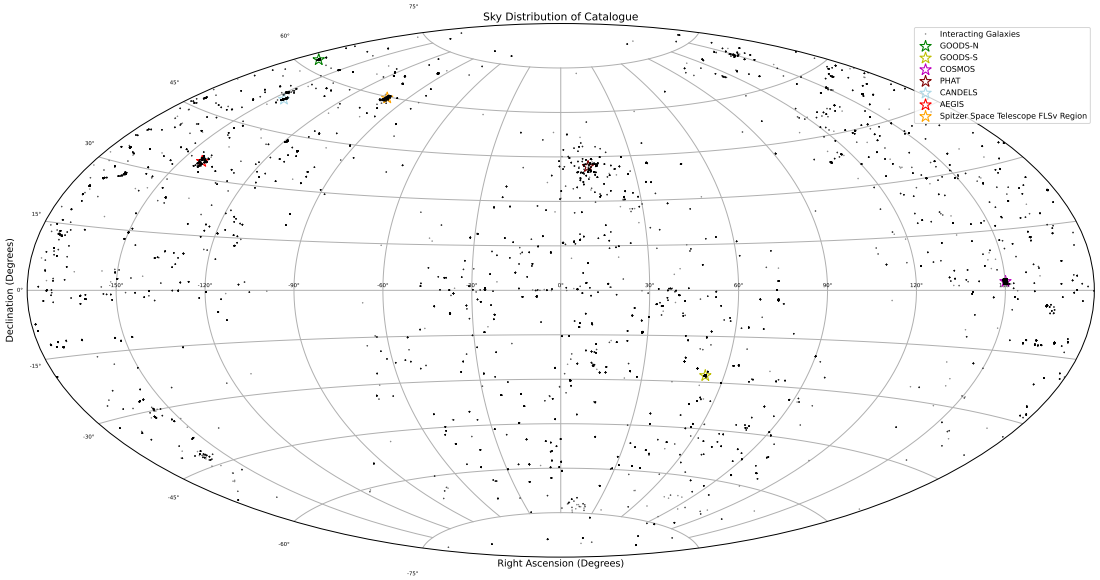

Figure 3.10: Sky Distribution of our catalogue, with marked positions of well known deep surveys conducted by *HST*. *HST* is able to observe almost the entire sky and therefore the interacting galaxies are scattered throughout. Large clusters of sources are found in the locations of surveys. This shows that often our sources are in the background of larger surveys and observations.

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 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,

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	

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

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

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

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

Table 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.

[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.

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`. 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

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.

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 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 A.2.

We extract the measured magnitudes of the $F606W$ and $F814W$ filters, giving us two-band photometry for 9,876 sources. Cross referencing with each source that had a redshift yields 2,993 sources from our catalogue. We calculate the colour of each source and plot it against the absolute magnitude in the $F814W$

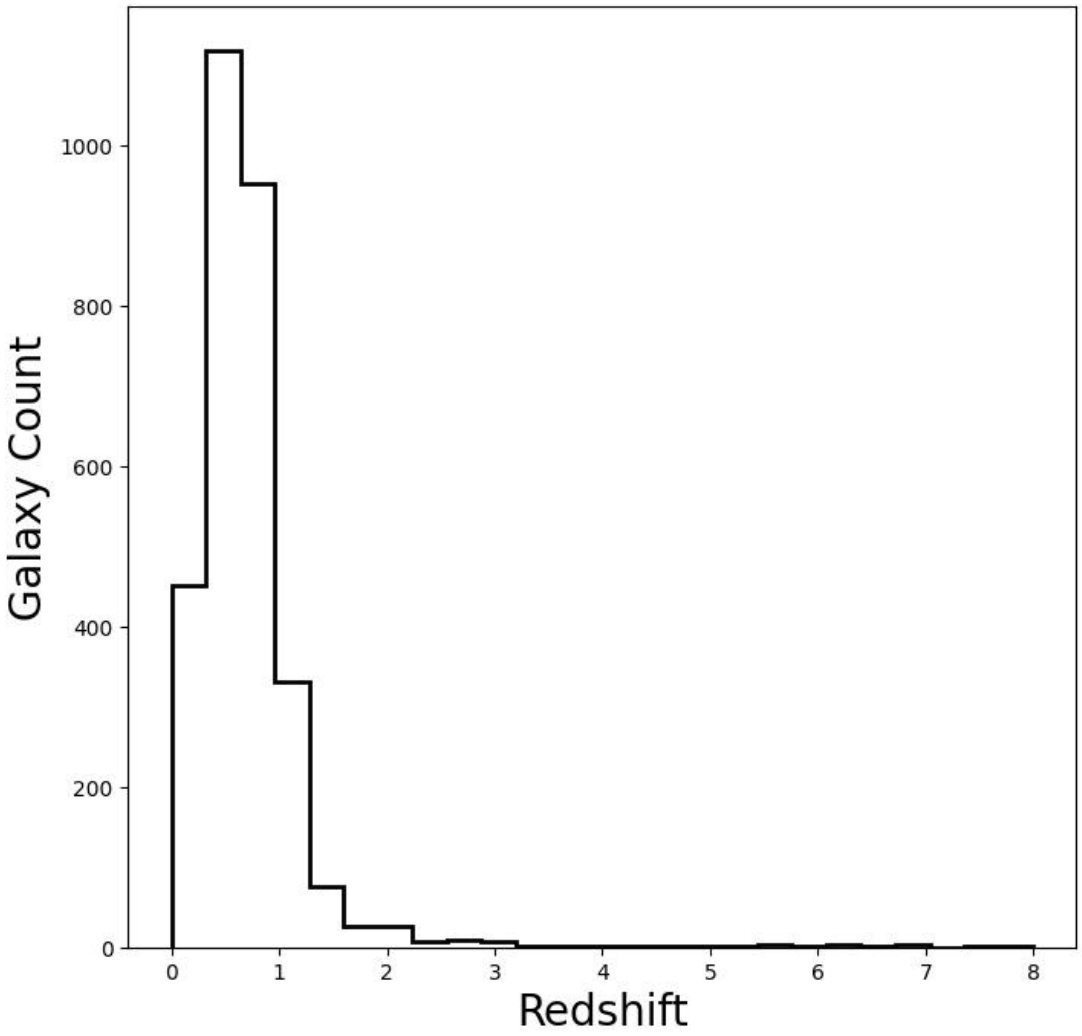

Figure 3.11: The redshift distribution of a subsample of our catalogue. Of the 7,583 referenced systems, 3,037 of them had redshift measurements in the NED, MAST or Simbad. This redshift distribution shows that our model confidently predicted interacting systems primarily for $z < 1$ systems. This was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

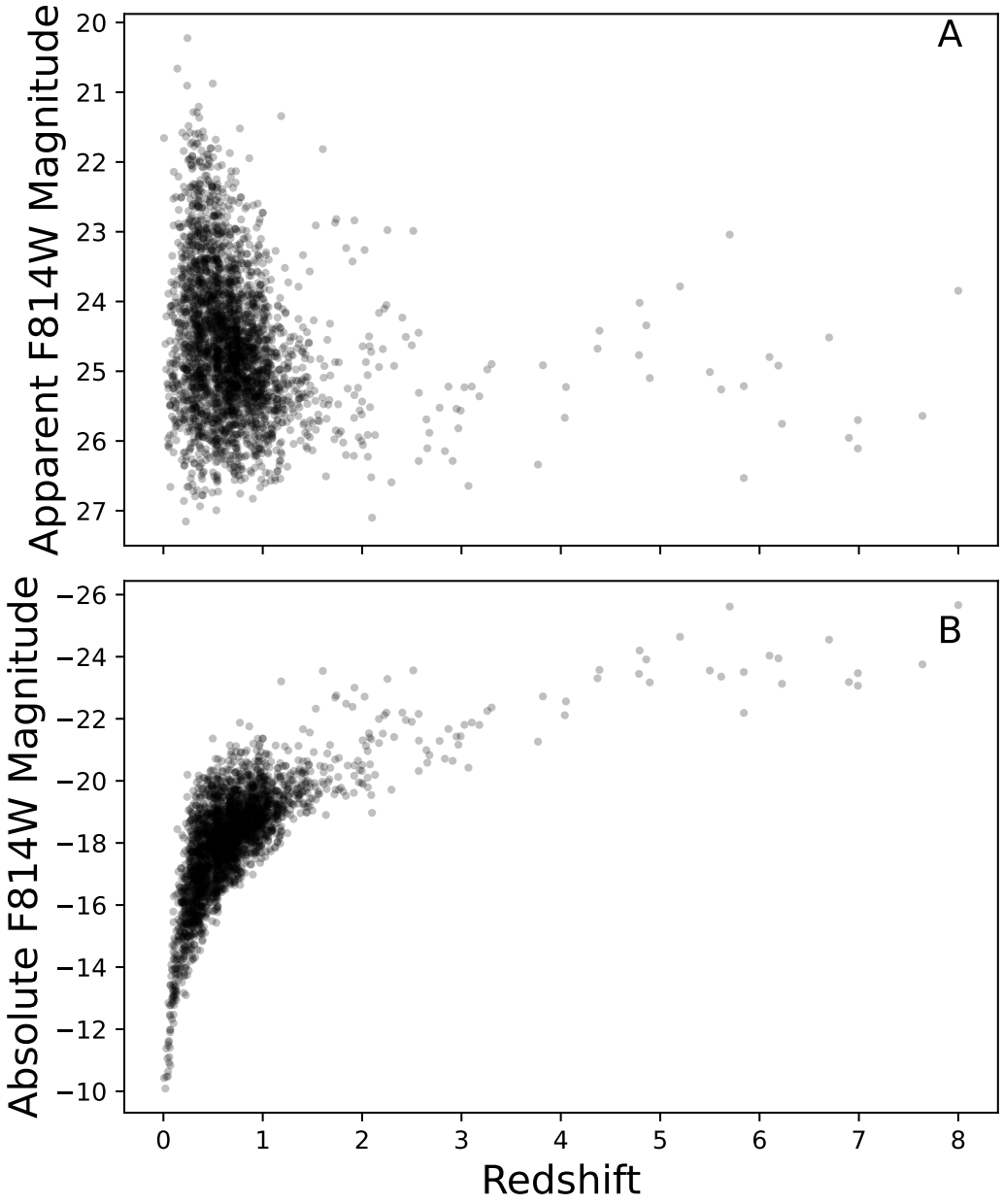

Figure 3.12: The distribution of redshift with magnitude for all sources with available data. This shows the parameter space we are sampling in this catalogue. Panel A shows that the majority of our sources are dim, background sources at low redshift. Panel B shows the faintest objects we find are at the limiting magnitudes of the different surveys this data is from.

filter. 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.,

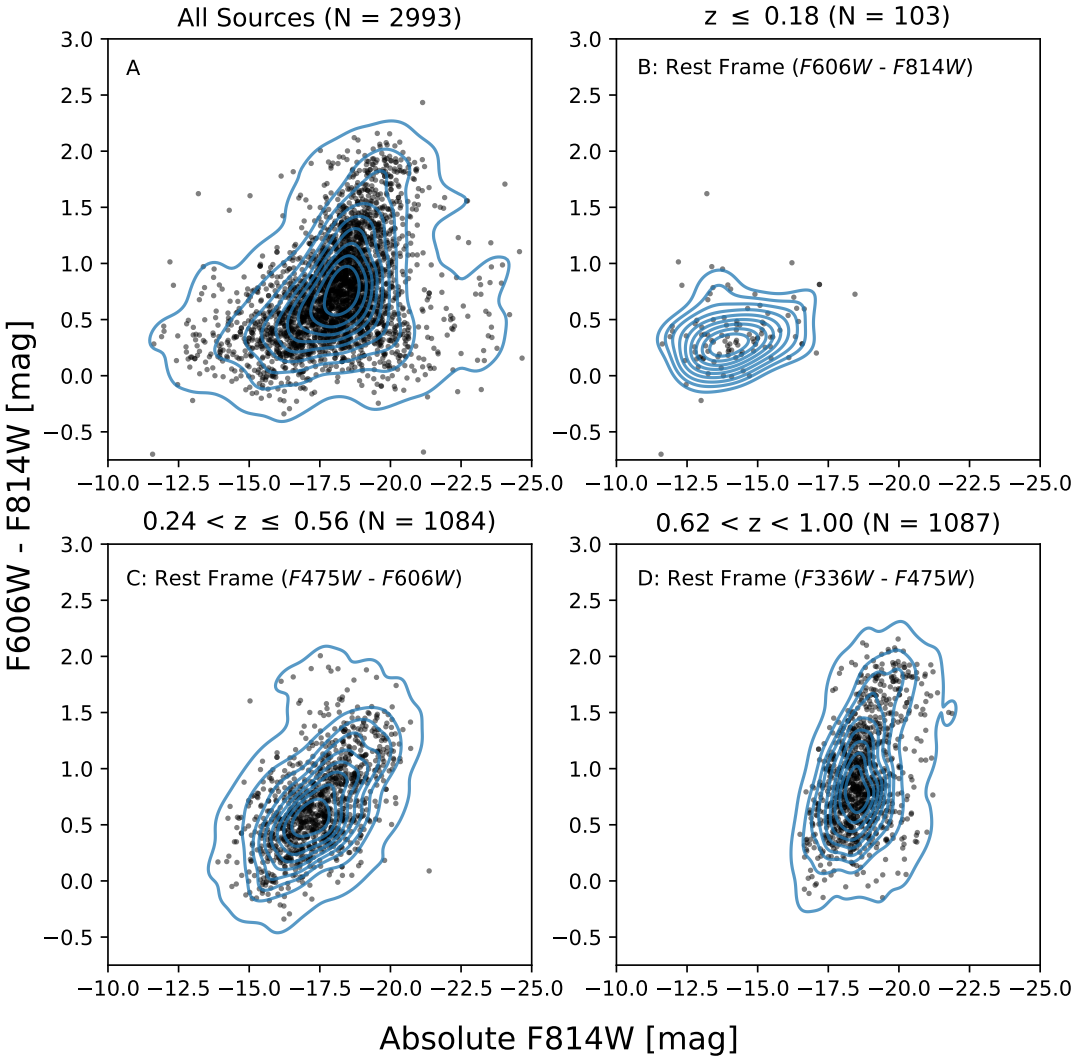

Figure 3.13: The colour-magnitude distribution of sources with a redshift measurement associated. Panel A shows the distribution of all galaxies, without controlling for redshift or dust extinction. The remaining panels then split these sources into distinct redshift bins where the $F606W$ and $F814W$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

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, 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

Chapter

4.1

Chapter 5

Conclusion

Appendix A

Appendices

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

These are both as expected. Both curves show that the model behaves well, and are much better than a random classifier (which would have a 1:1 relation). The ROC plot shows that we are minimising our false positive rate when using a prediction score cutoff of 0.95. However, we are misclassifying approximately

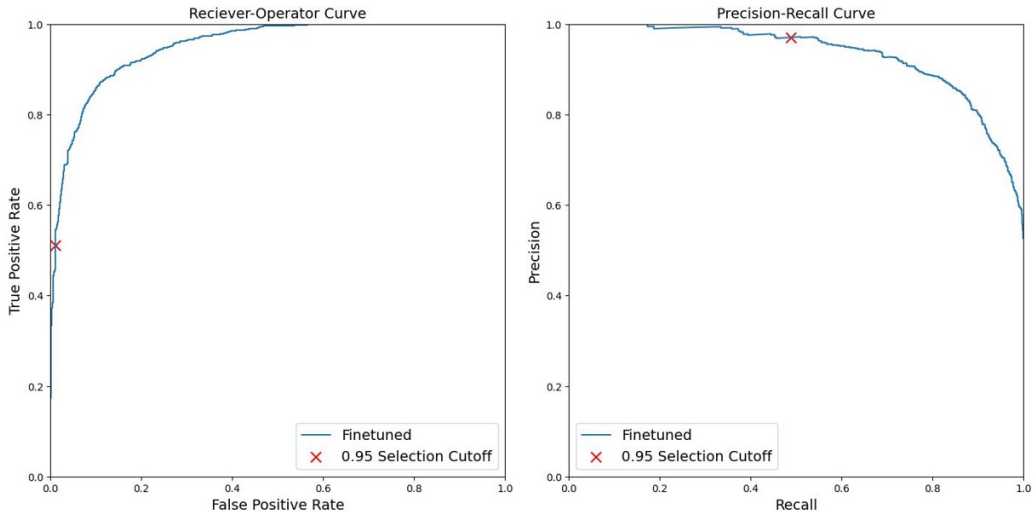

Figure A.1: The Receiver-Operator and Precision-Recall Curve for the `Zoobot` model that was used to explore the Hubble archives. The blue curves are the measured curves. These curves measure the relevant rates or characteristics based on the changing cutoff applied to how `Zoobot` defines an interacting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

50% of interacting galaxies as non-interacting galaxies. The contamination rate in our final catalogue (False Positives rate) will be very low (close to zero in this ideal validation set). The PR curve shows a similar result. Here, we are operating with a high precision (finding a pure catalogue) while keeping our recall minimal.

We also present the changing F1 score for the model used in this work, shown in Figure A.2. The F1 score is twice the ratio of precision multiplied by recall upon precision summed to recall. This combines our measure of accuracy and purity into a single metric. The cutoff we use in this work is at the point where the F1 score has began to decline. This is because we are beginning to lose recall rapidly, but gaining significantly in precision. As discussed in Section 3.6, this was an acceptable trade off in this work for a very large, pure interacting galaxy catalogue.

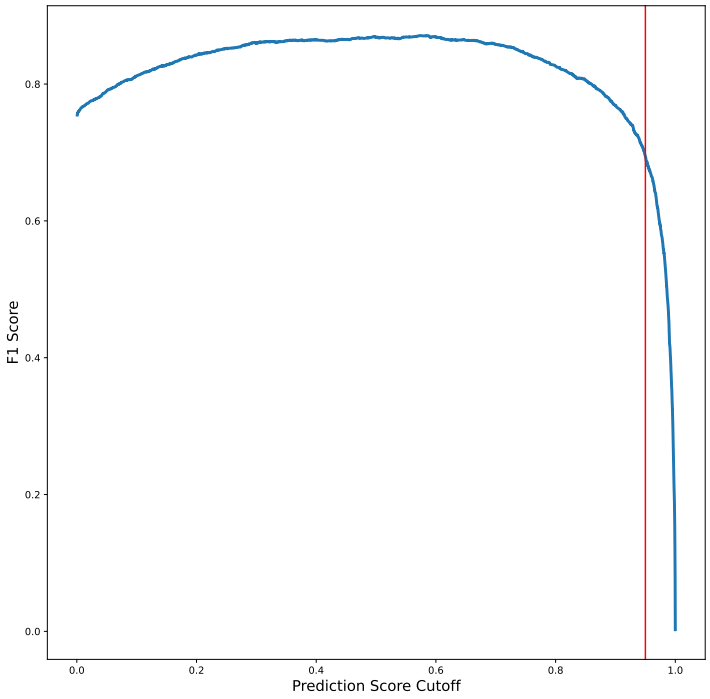

Figure A.2: The F1 score found during the diagnostics of the model used in this work. The F1 score is a measure combing the measure of accuracy and purity into one metric. The cutoff we use is at the point where the F1 score begins to rapidly decline. This point is shown by the red vertical line.

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

A.2 Examples of Sources with 3-Band Information

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

A.3 Unknown Objects

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

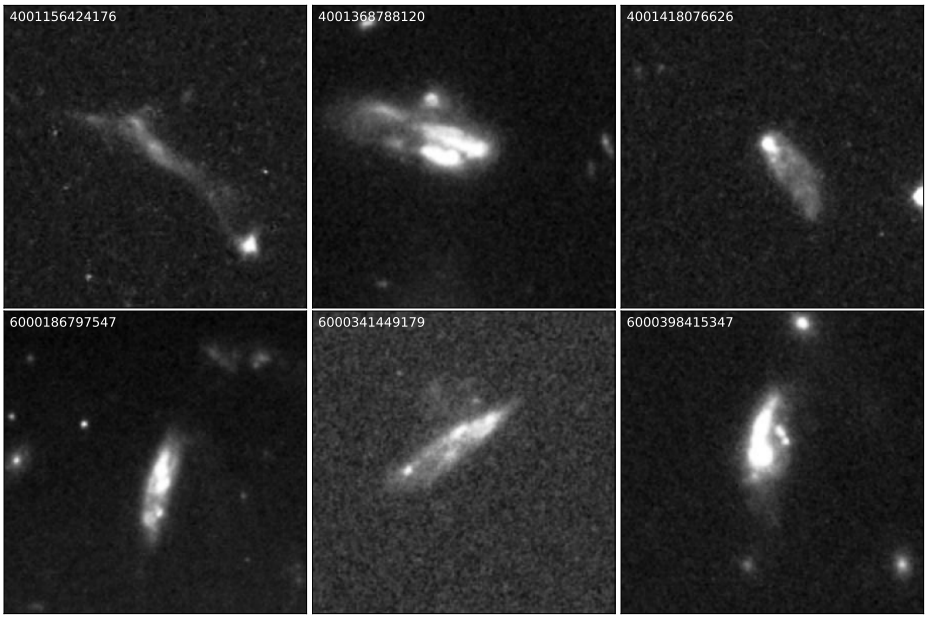

Figure A.4: The six unknown systems found in this work. These have no reference in Simbad or in NED, and their morphology could not be classified by the authors. Investigation into these six objects are presented to the community, with the authors hoping that future work and investigation of them can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

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

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

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

Proposal ID	Observation ID	Observation Date	DOI
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 A.1: Twenty example of the accompanying data table of observations used.

A.4 Acknowledging PIs

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

References

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

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

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

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

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