

This work is on a Creative Commons Attribution 4.0 International (CC BY 4.0) license, <https://creativecommons.org/licenses/by/4.0/>. Access to this work was provided by the University of Maryland, Baltimore County (UMBC) ScholarWorks@UMBC digital repository on the Maryland Shared Open Access (MD-SOAR) platform.

Please provide feedback

Please support the ScholarWorks@UMBC repository by emailing scholarworks-group@umbc.edu and telling us what having access to this work means to you and why it's important to you. Thank you.

Merger or Not: Accounting for Human Biases in Identifying Galactic Merger Signatures

ERINI L. LAMBRIDES¹,¹ DUNCAN J. WATTS^{2,1}, MARCO CHIABERGE^{3,1}, KIRILL TCHERNYSHYOV⁴,
ALLISON KIRKPATRICK⁵, EILEEN T. MEYER⁶, TIMOTHY HECKMAN¹, RAYMOND SIMONS⁷, OZ AMRAM¹, KIRSTEN R. HALL^{8,9},
ARIANNA LONG¹⁰ AND COLIN NORMAN^{7,1}

¹*Department of Physics and Astronomy, Johns Hopkins University, Bloomberg Center, 3400 N. Charles St., Baltimore, MD 21218, USA*

²*Institute of Theoretical Astrophysics, University of Oslo, P.O. Box 1029 Blindern, N-0315 Oslo, Norway*

³*AURA for the European Space Agency (ESA), ESA Office, Space Telescope Science Institute, 3700 San Martin Drive, Baltimore, MD 21218, USA*

⁴*Department of Astronomy, University of Washington, Seattle, WA, USA*

⁵*Department of Physics and Astronomy, University of Kansas, Lawrence, KS 66045, USA*

⁶*Department of Physics, University of Maryland, Baltimore County, 1000 Hilltop Circle, Baltimore, MD 21250, USA*

⁷*Space Telescope Science Institute, 3700 San Martin Drive Baltimore, MD 21218, USA*

⁸*Schmidt Science Fellow*

⁹*Atomic and Molecular Physics Division, Harvard-Smithsonian Center for Astrophysics, 60 Garden Street, Cambridge, MA, 02138, USA*

¹⁰*Department of Physics and Astronomy, University of California, Irvine, CA 92697, USA*

(Received ****; Revised *****; Accepted *****)

Abstract

Significant galaxy mergers throughout cosmic time play a fundamental role in theories of galaxy evolution. The widespread usage of human classifiers to visually assess whether galaxies are in merging systems remains a fundamental component of many morphology studies. Studies that employ human classifiers usually construct a control sample, and rely on the assumption that the bias introduced by using humans will be evenly applied to all samples. In this work, we test this assumption and develop methods to correct for it. Using the standard binomial statistical methods employed in many morphology studies, we find that the merger fraction, error, and the significance of the difference between two samples are dependent on the intrinsic merger fraction of any given sample. We propose a method of quantifying merger biases of individual human classifiers and incorporate these biases into a full probabilistic model to determine the merger fraction and the probability of an individual galaxy being in a merger. Using 14 simulated human responses and accuracies, we are able to correctly label a galaxy as "merger" or "isolated" to within 1% of the truth. Using 14 real human responses on a set of realistic mock galaxy simulation snapshots our model is able to recover the pre-coalesced merger fraction to within 10%. Our method can not only increase the accuracy of studies probing the merger state of galaxies at cosmic noon, but also can be used to construct more accurate training sets in machine learning studies that use human classified data-sets.

Keywords: statistics – methods – galaxies – mergers

1. INTRODUCTION

There is mounting observational and theoretical evidence that significant galactic mergers, where one galaxy is at least the tenth of the mass of the other, are an important component of galaxy evolution models which aim to explain the size, shape, and mass distributions

of galaxies in the Universe (see [Conselice 2014](#), for a review). Observational estimates of the rates of significant galaxy mergers have not converged for a variety of merger types. Even studies of the same observational field, with similar wavelength coverage, can yield disparate merger rate estimates ([Mantha et al. 2018](#); [Duncan et al. 2019](#)). Thus, robustly and consistently identifying systems that are ongoing (galaxy pairs or pre-coalescence) or recently have undergone a significant merger (near- or post-coalescence) is important.

At higher redshifts, merger identification can become increasingly difficult due to the potential for faint merger signatures to be undetectable (Lotz et al. 2004).

For example, some hydrodynamic simulations of galaxy mergers predict that as the galaxies coalesce, gravitational forces funnel gas toward the center, which provides a fuel reservoir to feed the central supermassive black hole and to form large numbers of stars in a nuclear starburst (Di Matteo et al. 2005). Between redshifts 1.5 and 2.5, activity of growing central supermassive black-holes (herein referred to as active galactic nuclei – AGN) and star-formation (SF) activity appear to peak (Madau & Dickinson 2014). Galaxy mergers with comparable mass ratios (i.e., major mergers) are one of the most popular mechanisms invoked to explain the similar evolution of the AGN activity and SF rates during this cosmic epoch (Conselice 2014). Some results are in tension with this picture. For example, empirical and theoretical studies find a connection between mergers and local ultra-luminous infrared galaxies (Sanders & Mirabel 1996; Veilleux et al. 2009), local AGN (Koss et al. 2010; Ellison et al. 2013, 2019), and high-luminosity AGN (Urrutia et al. 2008; Treister et al. 2012; Glikman et al. 2015; Donley et al. 2018). In contrast, ample research finds no connection between mergers and X-ray detected AGN (Gabor et al. 2009; Georgakakis et al. 2009; Kocevski et al. 2012), high-luminosity AGN (Villforth et al. 2014, 2017; Marian et al. 2019), and low-to-intermediate luminosity AGN (Grogin et al. 2005; Schawinski et al. 2011; Rosario et al. 2015).

Selection effects introduced through the construction of the AGN sample may play a role in explaining some of the disparate conclusions between AGN morphology studies. For example, dust obscuration may play a significant role in the observed (or lack of) connection between AGN and mergers. The merger fraction is higher for samples of infra-red (IR) selected AGN versus X-ray selected AGN perhaps due to the effect of dust-attenuation (Veilleux et al. 2009; Koss et al. 2010; Kocevski et al. 2015). Though, studies of sources with similar AGN selection criteria still yield conflicting merger fractions. For example the merger enhancement of X-ray selected heavily-obscured AGN at both higher and lower redshifts yield conflicting results (i.e. Schawinski et al. 2012; Kocevski et al. 2015; Koss et al. 2016; Lanzuisi et al. 2018; Li et al. 2020). An ill-studied reason for this disagreement may be the diverse array of merger detection methods and/or statistical methods used to characterize the statistical significance of the results within each study.

The variety of merger detection methods used to assess the morphology of galaxies can be broadly placed in two regimes: qualitative and automated. Qualitative methods rely on an observer or group of observers who classify each image by eye. Automated methods employ a pixel by pixel analysis of the image to identify the morphological class of the galaxy. Some automated methods require highly spectroscopic complete observations, like the close pairs method, which uses redshift and on-sky distances to identify pairs of galaxies that are within some distance threshold. Non-parametric automated methods, such as the second-order moment of the brightest 20% of light, the *Gini* coefficient, and the CAS parameters (concentration, asymmetry, clumpiness) use pixel based algorithms to detect asymmetries, double nuclei, tidal tails and/or other disturbances (for examples see Abraham et al. 1996; Conselice et al. 2000; Lotz et al. 2004). As shown in Huertas-Company et al. (2015), some of these methods can have mis-classification rates as high as 20%, and each suffers from biases where certain merging systems are preferentially identified.

Automated methods that employ deep learning techniques, a sub-field of machine learning based on artificial neural networks with representation learning, to classify galaxy morphology are promising due to their ability to classify quickly and their model independence (for example Dieleman et al. 2015; Ackermann et al. 2018; Pearson et al. 2019; Ćiprijanović et al. 2020). In particular, a variety of deep learning merger morphology studies train their algorithms on data-sets that have been visually classified by humans or test the accuracy of their schema compared to visually classified “truth” data-sets. Many of these recent deep-learning schema are trained off of the *Galaxy Zoo* catalogue of classifications of galaxies from the Sloan Digital Sky Survey (SDSS) (Lintott et al. 2008; Willett et al. 2013). Most of these ML implementations morphologically analyse galaxy samples at moderate to low redshifts. For example, Pearson et al. (2019), employed a deep learning algorithm that was trained not only on visually classified objects via *Galaxy Zoo*, but also mock images with known truths from the Eagle Simulations. When applying a convolutional neural network on the SDSS images, an accuracy of 91.5% was achieved. When passing the simulated EAGLE images through the SDSS trained neural network, the accuracy drops to 64.6%. The Pearson et al. (2019) framework uses SDSS galaxies with redshifts less than 0.1, and simulated EAGLE galaxies with redshifts less than 1.0. As is noted in Pearson et al. (2019), due to the potential redshift evolution of general galaxy properties, such as gas and dust content, a net-

work trained on low-redshift galaxies is not expected to be reliable for higher redshift galaxies.

Furthermore, any deep learning model trained on human classifications will carry any bias that still persists in the human classified training set. Despite the great potential of these classes of algorithms for automated merger identification, there currently is not a robust enough tool to handle the diverse presentations of merging galaxies, particularly at higher redshifts. Thus, visual human classification is still a method that is commonly employed in the literature to identify moderately large samples of merging galaxies at $z > 1.0$.

Image-based morphology studies of galaxies at higher-redshifts are difficult. Beyond $z \sim 1$, optical imaging surveys begin to probe the rest-frame UV morphologies of galaxies. This is useful for probing the most active regions of un-obscured star-formation, but may miss obscured gaseous and stellar features associated with merging systems (e.g., dusty tidal tails, dusty shells, and large-scale dust and gas asymmetries). When using humans as classifiers there are a variety of assumed biases most studies try to take into account. It is inevitable that any given classifier will show a particular bias. For example, some observers may be more inclined to classify objects as mergers even if the objects display minor disturbances unrelated to galaxy encounters. The most common way of accounting for human classifier bias is to construct a control sample. The classifiers assess the morphology of the control sample, and report merger fractions of their galaxy population of interest in the context of their relative differences between the control sample. In addition to constructing a control sample, some studies try to maximize the number of individual human classifiers. For projects like *Galaxy Zoo*, there is an average of 39 classifiers per object, and they report merger classifications on a per galaxy basis.

When comparing merger fractions of a population of objects to a control sample, careful analysis of the error bars is critical in order to determine if a significant difference exists between the population of interest and the control sample. The variety of statistical treatments used in reporting merger fractions from human classified datasets makes comparisons between studies difficult. For example, some studies assume a binomial distribution to model the number of mergers from aggregate classifications given by a group of human classifiers (i.e. Ellison et al. 2013; Kocevski et al. 2015; Ellison et al. 2019). Other studies have employed rank-choice voting and model the *probability* of the number of mergers using a beta distribution (i.e. Mechtley et al. 2016; Mariani et al. 2019, 2020). All of the above studies compare the significant of their merger fraction against a similar

statistically analysed control sample, with the assumption that the human classification bias is evenly applied amongst samples.

In this work, we test the critical assumption that the bias present in human classification is evenly applied to both the population and control datasets. In [section 2](#), we find it is not, and that the effect of human bias is a function of the intrinsic merger fraction of the sample being classified. In [section 3](#), we propose a self-consistent statistical framework to use estimates of an individual human classifier’s accuracy to derive a data-driven merger fraction. In [section 4](#), we describe how we can use the data-driven merger fraction and human classifier accuracy to yield merger assessments on a per-galaxy basis. In [section 5](#), we discuss the implications and applications of our statistical framework.

2. IDEALIZED PROBLEM AND ISSUES WITH THE CONVENTIONAL APPROACH

The fundamental setup for a morphology study is as follows; given a set of n galaxies and N independent classifications of each galaxy, what is the estimated merger fraction and error on the estimate for the given population? Most studies treat this as a binomial process with two outcomes: "merger" and "not merger", where the fraction of galaxies in a merger is given by f_M .

Generally what is reported is the merger fraction of the science sample, the merger fraction of the control sample, and the difference between the two. For example, suppose three classifiers assess 50 galaxies in two sets of samples, and report $\{30, 33, 31\}$ mergers in the science sample, and $\{10, 13, 15\}$ mergers in the control sample. Conventionally, the estimate of the merger fraction for each sample would be the mean of the individual measured merger fractions, and the error would be determined using binomial statistics. The significance of the estimated merger fraction in the science sample case is determined using a differential approach. In the above example, the mean merger fraction of the science sample is ~ 2.5 times greater than the control sample, and thus some significance of the estimated merger rate of the science sample would be assumed. Part of why most merger studies report results using differential or relative treatments is because the unknown biases of a classifier’s measurements is assumed to be applied evenly to both samples and thus should cancel out.

Closer examination shows that this framework is internally inconsistent. This treatment assumes that we are showing 3 separate samples to each person, but in fact they are looking at the same galaxy and disagreeing. If for example, in the control sample each classifier identified a similar number of mergers, but they disagreed

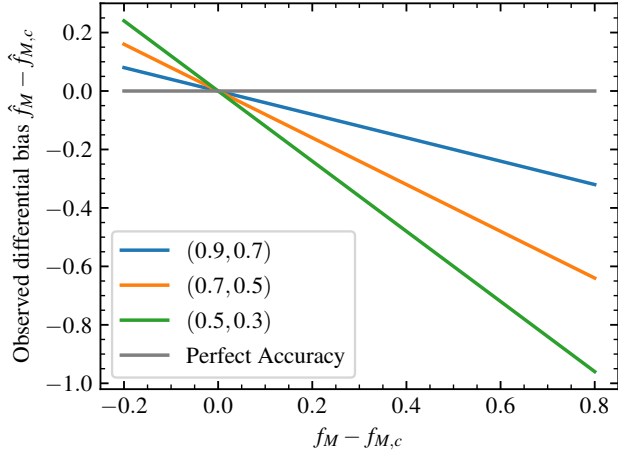


Figure 1. Observed differential merger fraction bias as a function of true differential merger fraction: Using Equation 5, the difference between a control sample and a science sample gives a result that depends both on the intrinsic differential merger fraction and the accuracy of the classifier.

with each other on the classification of individual objects, the statistical framework would not encapsulate important sources of error. At a fundamental level, if there is a disagreement amongst classifiers on a given classification: due to the binomial nature of the experimental set-up, one set of classifiers will be incorrect. To formalize this, we can say that if someone is shown a merging (isolated) galaxy, they classify it correctly with probability r_M (r_I). Therefore, if somebody is shown N_M mergers and N_I isolated galaxies, on average they will report $\hat{N}_M = r_M N_M + (1 - r_I) N_I$ mergers. The inclusion of the $(1 - r_I) N_I$ term represents the amount of galaxies that were incorrectly classified as isolated and are truly mergers.

The use of relative significance between comparing the merger fractions of the science and the control sample does not remove this issue. In the control sample, the classifier will report $\hat{N}_{M,c} = r_M N_{M,c} + (1 - r_I) N_{I,c}$ mergers on average, meaning the difference between the merger fractions depends on both the accuracy of an individual classifier and the intrinsic merger fraction of the sample. Thus by re-writing \hat{N}_M and $\hat{N}_{M,c}$ in terms of the merger fraction for each sample and taking the difference:

$$\langle \hat{f}_M \rangle = r_M f_M + (1 - r_I)(1 - f_M) \quad (1)$$

$$\langle \hat{f}_{M,c} \rangle = r_M f_{M,c} + (1 - r_I)(1 - f_{M,c}) \quad (2)$$

$$\langle \Delta \hat{f}_M \rangle = \langle \hat{f}_M \rangle - \langle \hat{f}_{M,c} \rangle \quad (3)$$

$$\begin{aligned} \langle \Delta \hat{f}_M \rangle &= r_M \Delta f_M - (1 - r_I) \Delta f_M \\ &= \Delta f_M [r_M + r_I - 1] \end{aligned} \quad (4)$$

we find the difference between the merger fractions of the two samples is still dependent on the intrinsic merger fraction of each sample. Equation (3) can then be used to quantify the systematic error due to human classification in the difference between the merger fractions as

$$b = \Delta f_M [r_M + r_I - 2]. \quad (5)$$

The only time when the bias would be equal to zero is if the intrinsic merger fraction of the two samples were identical. This is highly significant, because most morphology studies test whether there is a difference between the science sample and control sample. In Figure 1, we show three examples of this effect. Using the previous example of three classifiers assessing 50 galaxies for two sets of samples, we calculate the bias as parametrized in Equation 5 as a function of the difference of the intrinsic merger fractions for each sample. We calculate this function in four different test cases of mean observer accuracy. The blue line represents a class of observers that are very accurate in measuring merging systems and slightly less accurate at measuring isolated systems. The orange line is for a class of observers who are slightly less accurate at identifying merging galaxies and isolated galaxies. The green line represents a class of observers whose accuracy is poor for both merging and isolated systems, and the black line for classifiers with perfect accuracy. We see in all three classes of observers with non-perfect accuracy the degree of systematic bias from the truth changes as a function of the intrinsic merger fraction of each sample. Thus, if a hypothetical study finds a difference in the estimated merger fractions of their science sample and control samples, assuming the accuracy of their classifiers is not taken into account, disentangling whether the difference is due to real or simply systematic error is impossible.

We next explore how the unequal effect of this bias hinders meaningful statistical interpretation of sample difference measures between two merger fractions. Using the standard binomial statistics approach (as seen in Chiaberge et al. 2015; Villforth et al. 2017), one would use a proportion test to calculate confidence intervals of a given merger fraction and then use a hypothesis test to calculate the probability or significance of a difference between two samples given the null. In Figure 2, we show the effect of inaccurate classifiers on the recovered difference on a simulated sample of 50 galaxies. As shown in Equation 1, \hat{f}_M is a function of the true f_M and the accuracy of the classifier. Using the beta-distribution 68% confidence levels, for each r_M, r_I pair shown, there are regions in the parameter space with many standard deviations of difference between the \hat{f}_M

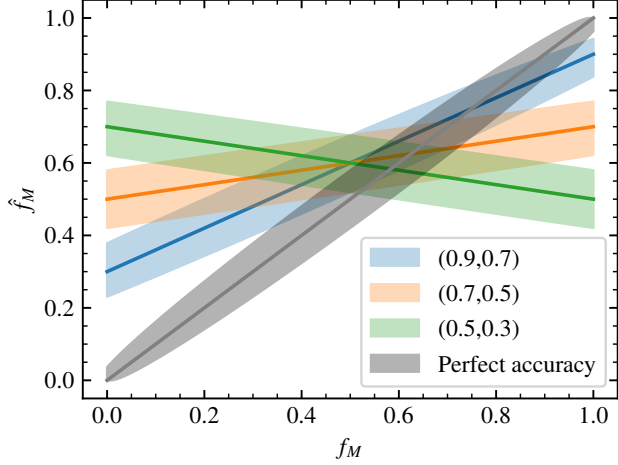


Figure 2. Intrinsic Merger Fraction vs Measured Merger Fraction: The black line corresponds to human classifiers with perfect accuracy. The blue, green, and orange lines correspond to different merger, isolated accuracy pairs. The shaded regions correspond to 68% confidence levels governed by the beta distribution.

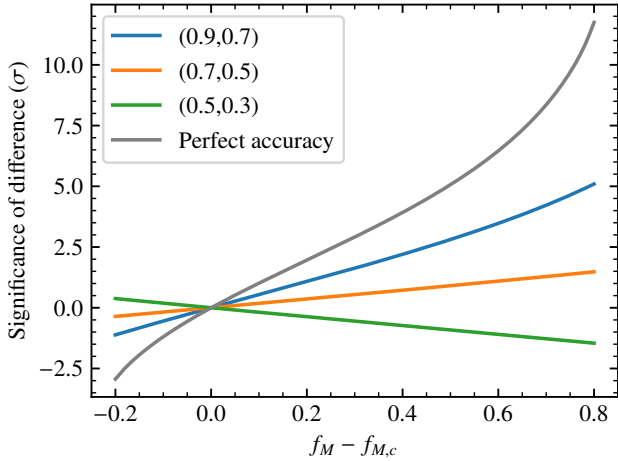


Figure 3. Intrinsic Difference of Merger Fractions vs Significance of the Measured Difference Between Merger Fractions: The colors correspond to the same definitions as of Figure 2. We test how the significance of the measured difference between a population sample and a control sample changes as a function of the intrinsic population sample merger fraction. We use the same intrinsic control sample merger fraction, $f_{M,c} = 0.2$, for each difference.

that would be reported by a perfect observer (black line) and by an inaccurate observer.

As previously mentioned, often these errors are mitigated by estimating the merger fraction of a control sample with an unknown, but likely lower, merger fraction than the science sample in question, and the relative enhancement is reported. Assuming that r_M and r_I are

independent of the class of object being classified, we can estimate the significance of the difference between a control sample's merger fraction, here $f_{M,c} = 0.2$, to see the size of the effect. We derive for each case the estimate of the control sample's merger fraction, $\hat{f}_{M,c} = r_M f_{M,c} + (1 - r_I)(1 - f_{M,c})$ and its uncertainty $\sigma_{f_{M,c}}$. To quantify the difference between a control sample and science sample, we plot $(\hat{f}_M - \hat{f}_{M,c})/\sigma$, where $\sigma^2 = \sigma_{f_{M,c}}^2 + \sigma_{f_M}^2$ using the errors derived from the proportion test. As shown in Figure 3, the general effect of this is to reduce the size of the measured difference. This does not imply that all previous merger studies have reported a lower significance than the actual truth, but rather if human classification bias is not constrained or accounted for and a null-significance is reported it is difficult to deduce whether the null result is intrinsically true or a human classifier accuracy effect. We note this reduction of significance is difficult to infer the validity of previous merger studies, due to the specific parametrization of accuracy and statistical tests used in this section. We do stress, the significance of the effect has a clear dependence on the accuracy of the classifier, and this effect is not mitigated by performing a comparison between the \hat{f}_M of the two samples using simple binomial statistical approaches.

3. A BAYESIAN UPGRADE TO THE FREQUENTIST APPROACH: THE NUMBER OF MERGERS LIKELIHOOD

As discussed in section 1, it is difficult to accurately characterize whether a galaxy is undergoing a merger or is isolated. Because of this, it is inevitable that any given classifier will obtain a merger fraction that is different than another's. Some previous works have assumed that this bias is similar for the data and the control sample, and test their results against the null hypothesis that the intrinsic fractions are identical. As shown in Figure 3, if the underlying merger fraction of the two populations are significantly different, the significance of the result will be affected by this bias. In this section, we present a method that is built upon on the standard binomial approach of determining the number of mergers in a sample while taking into account the effect of human inaccuracy.

In order to estimate the true underlying merger fraction, we can estimate the bias in an individual classifier's assessment on a sample with a known intrinsic merger fraction, and then optimally combine the individual classifier uncertainties on the sample where the intrinsic merger fraction is unknown. We perform this analysis assuming that there are two binomial processes for each classifier; (1) the probability of classifying galaxies

Table 1. Variable definitions for the Merger Fraction Likelihood

| Symbol | Definition |
|----------------------------|---|
| N_X | Number of objects of type X in the sample |
| f_M | Merger fraction of sample $f_M \equiv N_M/(N_M + N_I)$ |
| $N_{X,\text{syn}}$ | Number of mock objects of type X shown to a classifier |
| $r_{X,i}$ | Probability of classifier i identifying object type X correctly |
| $\hat{N}_{X,\text{syn},i}$ | Number of mock objects of type X correctly identified by classifier i |
| $\hat{N}_{X,1,i}$ | Number of objects <i>correctly</i> identified as type X by classifier i |
| $\hat{N}_{X,2,i}$ | Number of objects <i>incorrectly</i> identified as type X by classifier i |
| $\hat{N}_{X,i}$ | Number of objects identified as type X by classifier i , $\hat{N}_{X,i} \equiv \hat{N}_{X,1,i} + \hat{N}_{X,2,i}$ |
| \hat{f}_M | Estimated merger fraction given the set of all $\{\hat{N}_{M,i}, \hat{N}_{I,i}, \hat{N}_{M,\text{syn},i}, \hat{N}_{I,\text{syn},i}\}$. |

NOTE—We do not use the incorrectly classified mock galaxies in our likelihood, since we know the ground truth and do not need to marginalize over this parameter.

accurately as mergers, and (2) the probability of inaccurately classifying isolated galaxies as mergers.

3.1. The Merger Fraction Likelihood

The total number of claimed mergers is $\hat{N}_M = \hat{N}_{M,1} + \hat{N}_{M,2}$. Given N_M mergers in a sample, the probability of a classifier correctly measuring $\hat{N}_{M,1}$ mergers in a given sample is

$$P(\hat{N}_{M,1} | r_M, N_M) = \binom{N_M}{\hat{N}_{M,1}} r_M^{\hat{N}_{M,1}} (1 - r_M)^{N_M - \hat{N}_{M,1}}. \quad (6)$$

At the same time, if we have N_I isolated galaxies, the classifier will incorrectly classify an isolated galaxy as a merger with probability $1 - r_I$. We define the number of isolated galaxies incorrectly identified as mergers as $\hat{N}_{M,2}$, which follows the probability distribution

$$P(\hat{N}_{M,2} | r_I, N_I) = \binom{N_I}{N_I - \hat{N}_{M,2}} r_I^{N_I - \hat{N}_{M,2}} (1 - r_I)^{\hat{N}_{M,2}}. \quad (7)$$

Since $\hat{N}_{M,1}$ and $\hat{N}_{M,2}$ are drawn independently, we can represent the distribution of all measured galaxies using the triangular sum

$$\begin{aligned} P(\hat{N}_M | r_M, r_I, N_I, N_M) \\ = \sum_{\hat{N}_M = \hat{N}_{M,1} + \hat{N}_{M,2}} P(\hat{N}_{M,1} | r_M, N_M) P(\hat{N}_{M,2} | r_I, N_I). \end{aligned} \quad (8)$$

or equivalently

$$\begin{aligned} P(\hat{N}_M | r_M, r_I, N_I, N_M) \\ = \sum_{\hat{N}_{M,1}=0}^{\hat{N}_M} P(\hat{N}_{M,1} | r_M, N_M) P(\hat{N}_M - \hat{N}_{M,1} | r_I, N_I). \end{aligned} \quad (9)$$

Additionally, since we know the total number of galaxies N_{tot} and are interested in the true underlying number of mergers N_M , we can write the likelihood as a function of N_M , r_M , and r_I ,

$$\begin{aligned} \mathcal{L}(N_M, r_M, r_I | \hat{N}_M) \\ = P(\hat{N}_M | r_M, r_I, N_{\text{tot}} - N_M, N_M) \end{aligned} \quad (10)$$

One benefit to this formalism is that it easily generalizes to an arbitrary number of classifiers, each with their own measurements and accuracies. Assuming that each classifier is independent, the set of all observations is distributed as

$$\begin{aligned} P(\{\hat{N}_{M,i}\} | \{r_{M,i}\}, \{r_{I,i}\}, N_{\text{tot}} - N_M, N_M) \\ = \prod_i P(\hat{N}_{M,i} | r_{M,i}, r_{I,i}, N_{\text{tot}} - N_M, N_M) \end{aligned} \quad (11)$$

and we can write the likelihood as

$$\begin{aligned} \mathcal{L}(N_M, \{r_{M,i}\}, \{r_{I,i}\} | \{\hat{N}_{M,i}\}) \\ = P(\{\hat{N}_{M,i}\} | \{r_{M,i}\}, \{r_{I,i}\}, N_{\text{tot}} - N_M, N_M). \end{aligned} \quad (12)$$

In this statistical model, classifiers' accuracies are nuisance parameters that need to be marginalized over since the true merger fraction is the variable of interest. Using Bayes' theorem, we write the posterior distribution

$$\begin{aligned} P(N_M, \{r_{M,i}\}, \{r_{I,i}\} | \{\hat{N}_{M,i}\}) \\ \propto P(\{\hat{N}_{M,i}\} | \{r_{M,i}\}, \{r_{I,i}\}, N_{\text{tot}} - N_M, N_M) \\ \times P(N_M) P(\{r_{M,i}\}, \{r_{I,i}\}) \end{aligned} \quad (13)$$

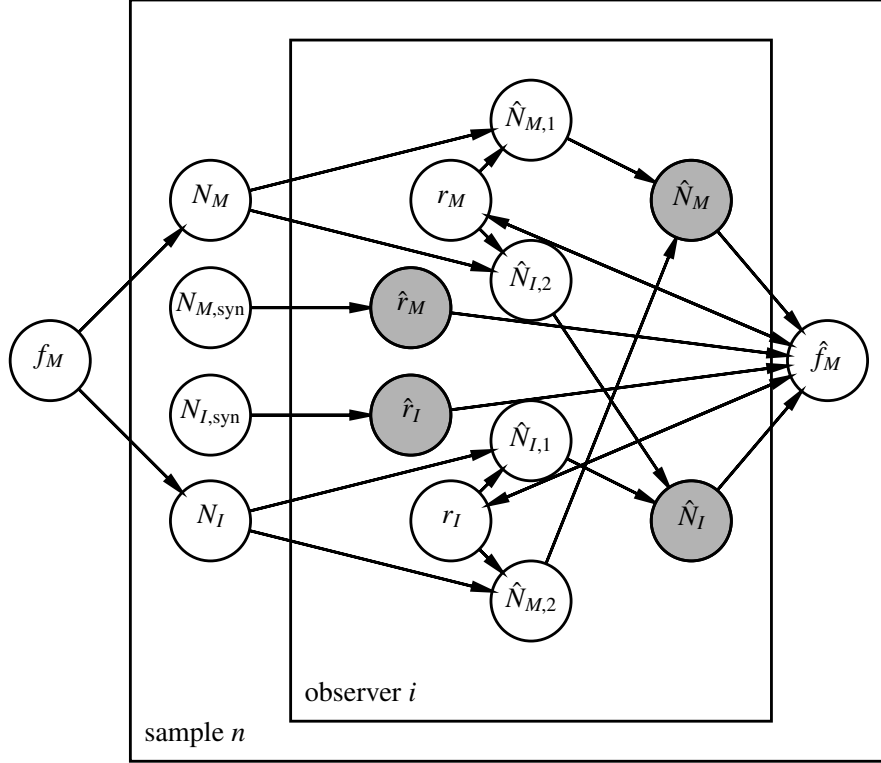


Figure 4. Graphical model of obtaining the merger fraction using the likelihood defined in Equation 13.

where $P(N_M)$ and $P(\{r_{M,i}\}, \{r_{I,i}\})$ are the prior distributions. We sample this posterior distribution using the Markov Chain Monte Carlo sampler **emcee**.¹

In principle, we can obtain the prior distribution of the classifiers' accuracies using their performances on mock galaxies whose underlying state is known a priori, although the applicability of this prior depends on (1) the number of mock galaxies and (2) the extent to which the mock galaxy sample can be treated as real data. We implement our prior using the classifiers' performance on the mock galaxies,

$$P(\{r_{M,i}, r_{I,i}\}) = \prod_i P(r_{M,i} \mid \hat{N}_{M,\text{syn},i}, N_{M,\text{syn}}) \times P(r_{I,i} \mid \hat{N}_{I,\text{syn},i}, N_{I,\text{syn}}) \quad (14)$$

such that the classifier's accuracies are beta distributed such that $r_{M/I} \sim \text{Beta}(\hat{N}_{M/I} + 1, N_{M/I} + 1)$. In principle, we can apply an additional prior on N_M , $\{r_{M,i}\}$, and $\{r_{I,i}\}$, but we find that the full likelihood results are not noticeably affected by altering the prior.

The strength of this method is its internal consistency; given a set of observed mergers, $\{\hat{N}_{M,i}\}$, the likelihood

is maximized when a value of N_M shown to all classifiers is most plausible, given a set of accuracies $\{r_{M,i}, r_{I,i}\}$. This is in contrast to the usual approach, which assumes each classifier has perfect accuracy, and can only reflect reality if each classifier was shown a different set of galaxies. In Figure 4, we show the graphical model of our likelihood analysis where all the variables are defined within this section.

3.2. Testing the Likelihood Model on a Simulated Galaxy Catalogue

To validate this model, we first simulate a data galaxy catalog with observations, best-fit values, uncertainties, and offsets from the input value;

- Choose a true underlying merger fraction f_M , with N_{tot} galaxies, $f_M N_{\text{tot}}$ mergers, and $(1 - f_M) N_{\text{tot}}$ isolated galaxies.
- Assign n accuracy pairs $(r_{M,i}, r_{I,i})$ drawn from a uniform distribution $\mathcal{U}(0.5, 0.9)$ for each classifier, and calculate the mean accuracy for merging and isolated systems. Note the exact choice of the mean accuracies is unimportant for this exercise, but rather whatever choice is made is accounted for in the statistical modelling.

¹ **emcee** is an implementation of the Goodman & Weare (2010) Affine Invariant MCMC Ensemble sampler ²

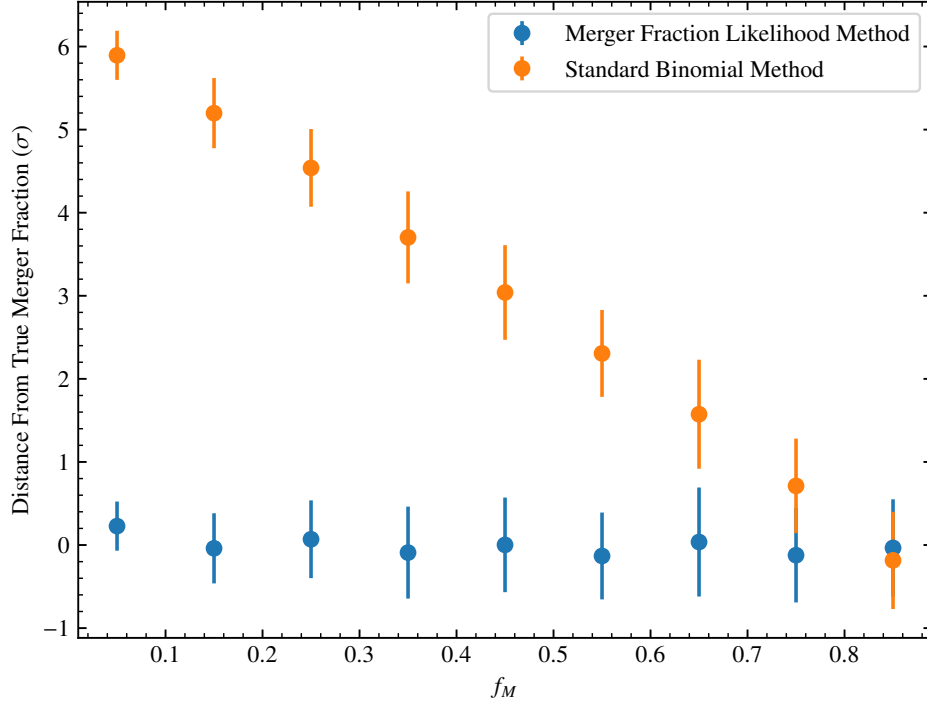


Figure 5. Recovery of Intrinsic Merger Fraction Using the Simulated Galaxy Catalogue Sample: Assuming the average accuracy of 14 simulated classifiers for merging and isolated systems is 80% and 55% respectively, we show how the sigma difference from the intrinsic fraction changes as a function of intrinsic merger fraction using the standard binomial method. The blue points are derived using our likelihood implementation. The orange points use the standard binomial method. The error bars from each use the standard deviation from each method’s own probability distribution function.

- For each classifier, draw $\hat{N}_{M1,i}$ correctly identified mergers and $\hat{N}_{M2,i}$ incorrectly identified mergers, using the accuracies from the previous step.

In the standard binomial distribution approach, nN_{tot} galaxies have been observed, $\hat{N}_M = \sum_i \hat{N}_{M,i}$ mergers have been observed, and it is assumed that this observation is drawn from a binomial distribution,

$$p(\hat{N}_M | nN_{\text{tot}}, f_M) = \binom{nN_{\text{tot}}}{\hat{N}_M} f_M^{\hat{N}_M} (1 - f_M)^{nN_{\text{tot}} - \hat{N}_M}.$$

The likelihood $p(f_M | \hat{N}_M, nN_{\text{tot}})$ is a beta distribution with parameters $\alpha = \hat{N}_M + 1$ and $\beta = nN_{\text{tot}} - \hat{N}_M + 1$, and has mean and variance

$$\frac{\hat{N}_M + 1}{nN_{\text{tot}} + 2}, \quad \frac{(\hat{N}_M + 1)(nN_{\text{tot}} - \hat{N}_M + 1)}{(nN_{\text{tot}} + 2)^2(nN_{\text{tot}} + 3)}.$$

In Figure 5, we compare the standard binomial approach against a test of the merger fraction likelihood model. We assume the average accuracies of 14 simulated classifiers for merging and isolated systems are 80% and 55% respectively. We then simulate 9 different data sets of 50 galaxies with an intrinsic merger fraction that spans from 0.05 to 0.95. We compare the distance in units of sigma from the true merger fraction for these

different intrinsic merger fractions. The orange points represent the sigma difference from the truth for the standard binomial approach, and the blue points the sigma difference from the truth for the merger fraction likelihood method presented in this work. The difference between the standard binomial method from the intrinsic merger fraction varies as a function of the intrinsic merger fraction. In the case where the intrinsic merger fraction is 0.85, the standard binomial method is able to recover the true fraction due to the particular classifier accuracy values chosen in this test. In real classification studies, where the intrinsic merger fraction is a priori unknown, it is impossible to predict the deviation from the truth without accounting for the accuracy of the classifiers. The merger fraction likelihood method presented in this work takes into account the accuracy of the human classifiers. As is seen in Figure 5, any biases inherent in our method should not significantly depend on the intrinsic merger fraction of the sample.

3.3. Testing the Likelihood Model on Mock Galaxies with Real Human Classifiers

In this section, we detail a method where one can systematically estimate a real classifier’s accuracy us-



(a) Noiseless Mock Galaxy

(b) 3DHST GOODS-S *HST* Survey
Noise Applied Mock Galaxy

Figure 6. VELA+SUNRISE Noise-Added Mock Galaxy Example: In the left most image, we show the rgb (r: *HST* WFC3/IR F160W, g: *HST* ACS F775W, b: ACS F435W) VELA+SUNRISE image of a galaxy at redshift 1.7. This image has already been convolved with the *HST* PSF in each of the wavelength bands. In the right most image, we show the same mock galaxy but with our noise model applied. The physical scale of the cutout is $7.8'' \times 7.8''$ or at $z = 1.7$, $67.2 \text{ kpc} \times 67.2 \text{ kpc}$. In the appendix we provide all merging and isolated mock galaxy noise applied images.

ing mock observations from a suite of galaxy formation simulations with known true classifications.

To do so, we use mock images created from the VELA zoom-in hydrodynamical galaxy formation simulations (Ceverino et al. 2014; Snyder et al. 2015; Simons et al. 2019). The VELA simulation suite comprises 35 galaxy halos, spanning virial masses of $\sim 1\text{--}20 \times 10^{11} M_{\odot}$ at $z = 2$. These simulations were run in a full cosmological context using the Adaptive Mesh Refinement Tree code (ART; Kravtsov et al. 1997) and the subgrid physical recipes used are described in detail in Ceverino et al. (2010, 2012, 2014).

For each timestep of each simulated VELA halo, the true classification (isolated or merging) of the central mock galaxy is determined from the kinematics and spatial distribution of its stars (described in Simons et al. 2019). Galaxies are selected as mergers if they have undergone a merger within the last 100 Myr, or if they have a companion galaxy within 35 kpc. We randomly select a set of 24 simulation outputs where the central galaxy is merging, and 29 simulation outputs where the central galaxy is isolated. These simulation outputs span redshifts of 1.0 to 3.5, and the redshift distribution of the isolated and merging galaxies are similar Table 2.

Mock *Hubble* ACS/WFC3 images were created for each galaxy in the VELA suite in Snyder et al. (2015) and Simons et al. (2019), using the dust-radiative transfer code SUNRISE (Jonsson et al. 2010). The production of the mock images are described in detail in Simons

et al. (2019). The mock images are available as high level science products on a public repository.³

We downloaded noise-free versions of the mock images of our selected mock galaxies in three *Hubble* bands: ACS F435W, ACS F775W, and WFC3 F160W. The mock images include the appropriate spatial resolution and pixel scale of each band, but do not include noise. In addition to the 24 merging galaxies and 17 isolated galaxies, we create 10 images for a set of "fake" mergers. This set of "fake" mergers is used to assess how well a classifier can distinguish galaxies that are interpolating by chance alignment (i.e., not interacting) from galaxies that are merging. To do this, we superimpose the images of two mock isolated galaxies using a random separation less than $8''$.

We then add the appropriate amount of Poisson noise to simulate the well-studied real data-set of the 3DHST reduction of GOODS-South. We first calculate a normalization factor to match a background pixel in the VELA mock galaxy cutouts to the background in each *HST* band of the 3DHST GOODS-S maps. We calculate the normalization factor by first performing aperture photometry on a real galaxy where the background is sky dominated. We multiply the aperture flux of the image by the exposure time, and get the instrument counts

³ <https://archive.stsci.edu/prepds/vela/>

of the image. We then get the background counts from the *HST* Exposure Time Calculator ⁴.

Next, we matched a VELA image to the real galaxy in redshift and flux for each individual *HST* band. We again perform aperture photometry on the VELA image in all three bands. We finally apply the normalization factor of each band by multiplying the VELA image with the real image instrument counts divided by the VELA image counts. After pixel matching the VELA pixels to the 3DHST survey pixels we then apply additional sources of noise. Using IRAF's `mknoise` function we apply Gaussian read noise, gain, shot noise, and the background counts found from the ETC. In Figure 6 we show an example of a mock galaxy with and without the applied survey derived noise. In the left most image, we show the rgb (r: *HST* WFC3/IR F160W, g: *HST* ACS F775W, b: ACS F435W) VELA+SUNRISE image of a galaxy at $z = 1.7$. This image has already been convolved with the *HST* PSF in each of the wavelength bands. In the right most image, we show the same mock galaxy but with our noise model applied. The physical scale of the cutout is $7.8'' \times 7.8''$ or at $z = 1.7$, $67.2 \text{ kpc} \times 67.2 \text{ kpc}$. In the appendix we provide all merging and isolated mock galaxy noise applied images.

After creating the noise-added mock galaxy sample, we then showed fourteen different human classifiers the entire sample of mock images. The samples are intermixed, and are classified using the criteria enumerated below. The classifiers were also told there may be background or foreground galaxies in the images. The classifiers' backgrounds ranged from eight professors of astronomy, a post-doctoral fellow in astronomy, and four graduate students. The first author of this study was not included as a classifier as to minimize potential bias. We created a website where the mock images asked hosted, and asked each classifier to classify the image over the following options:

1. Merging: Major (approximately similar size)
2. Merging: Minor (approximately 1:4 size ratio)
3. Disturbance: Major
4. Disturbance: Minor
5. No Evidence of Merger/Interaction.

We provided the classifiers with the redshift of the central galaxy, and defined merging as an on-going interaction (which can include evidence of gravitational disturbances i.e., tidal tails with distinct galaxy systems,

pairs). We defined a disturbance as a post-merger in the final stages of (or post-) coalescence. A disturbance classification can include large asymmetry/gravitational disturbance and/or tidal tails. Ultimately, for our analysis we use only two morphological classes: merging and not merging. Merging includes major mergers, minor mergers, and major disturbances. The non-merging class includes minor disturbances and no-evidence of gravitational interactions. This is due to the difficulty in constraining merger stage and mass ratio from images alone. Nonetheless, when the human classifiers are presented with the images they are given multiple morphological divisions to choose from to help aid in the human classification process.

We then use the raw accuracies of the classifiers to inform a data driven model of determining the merger fraction of the sample. In the simulated galaxy case, we assume perfect knowledge of the accuracies of each classifier, or a δ -function prior for each accuracy parameter that is the same as the input value. For the real human classifications on the VELA+SUNRISE noise-added mock galaxy sample, we estimate the accuracies from the mock images, where $r_M = \hat{N}_{M,s}/N_{M,s}$ and $r_I = \hat{N}_{I,s}/N_{I,s}$. We collapsed the classification options of the mock galaxies in two options: merging and non-merging. Merging includes major mergers, minor mergers, and major disturbances. The non-merging class includes minor disturbances and no-evidence of gravitational interactions.

In Figure 7, we show the estimation of the merger and isolated classification accuracies for each individual classifier. As shown in Equation 13, the classifier accuracies are estimated using the raw accuracies from the mock galaxy classifications and the individual agreement on the number of galaxies in a merger in the mock galaxy sample. We show the likelihood distribution for assessing isolated systems (filled histogram) and merging systems (unfilled histogram). We find that some classifiers have higher accuracies assessing isolated systems, some have higher accuracies assessing merging systems, and some that are equally accurate for both. As mentioned in section 2, the effect of a classifiers bias for or against a specific morphological class depends on the intrinsic merger fraction of the population. The fourteen classifiers chosen have a diverse range of accuracies, and a standard binomial statistical approach would not capture this significant source of error. Using the likelihood model, we recover a total merger fraction of $56.8\% \pm 0.06$ (25/41) at the 95% confidence level. We are well within 1σ of the true merger fraction of the mock sample which is 54.5% (24/41). In Figure 8, we show the probability distribution of the merger fraction for the mock galaxy

⁴ <http://etc.stsci.edu/etc/input/wfc3ir/imaging/>

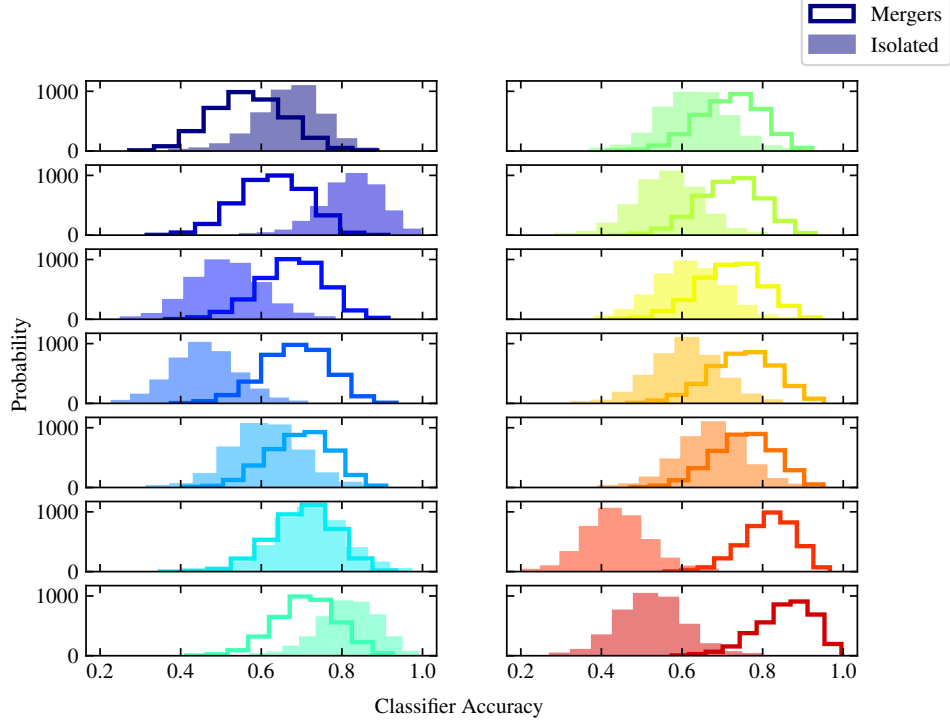


Figure 7. Estimation of the Merger and Isolated Classification Accuracies of Each Individual Human Classifier using their Classifications of the VELA+SUNRISE Noise-Added Mock Galaxy Sample: For each fourteen classifiers, we show the likelihood distribution for assessing isolated systems (filled histogram), and merging systems (unfilled histogram) using Equation 13.

sample. The dashed, orange line is the intrinsic merger fraction of the mock galaxy sample. The blue histogram is the probability distribution derived from the likelihood model or equation 11, with mean 0.57 ± 0.06 .

There are some important caveats with this approach and implementation. First, we do not know if classifiers will characterize the mock galaxies the same way as they do real images. Though, our algorithm is developed such that any metric of estimating a classifier’s accuracy can be used instead. Second, we use a point estimate, the raw merger fraction of each classifier, when it would be more appropriate to use a beta distribution prior in our fits. We test whether these affects will significantly bias our results, and we find when we run the analysis on the mock galaxies, we recover the input merger fractions correctly, and the fit has a similar likelihood surface. We also find when we run a full Monte Carlo Markov Chain with flat priors on the accuracies, the output mean and variance are consistent with the mock image estimate within a standard deviation.

Most importantly, a drawback to this method is the comparison of aggregates rather than individual galaxies. For example, two classifiers could disagree on which specific galaxies are in mergers, but find similar merger fractions in the sample. The aggregate method could

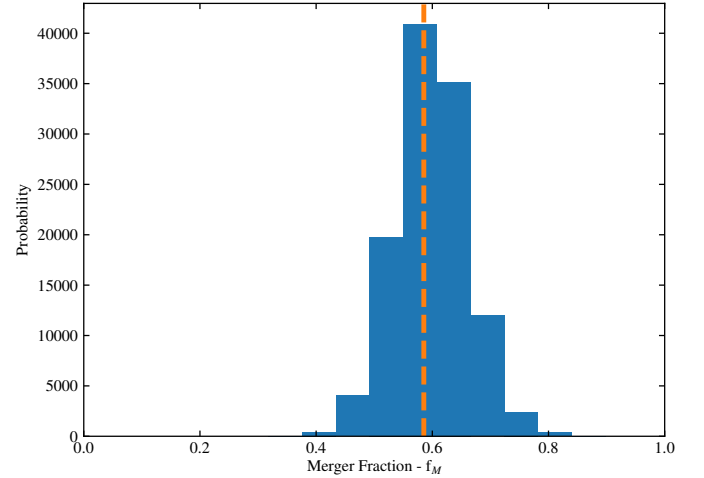


Figure 8. Measured Merger Fraction of the VELA+SUNRISE Noise-Added Mock Galaxy Derived from the Likelihood Model: The dashed, orange line is the intrinsic merger fraction of the mock galaxy sample. The blue histogram is the probability distribution derived from the likelihood model or equation 11, with mean 0.58 ± 0.06 .

erroneously imply that the two classifiers agree on their classifications, when in fact they do not.

4. A NEW APPROACH: SIMULTANEOUSLY ESTIMATING THE MERGER FRACTION AND THE PROBABILITY OF AN INDIVIDUAL GALAXY BEING IN A MERGER

Another way to statistically estimate the merger fraction of a sample is to include the level of agreement, or the amount of classification agreement between individual classifiers on a given galaxy, to estimate the merger fraction probability of a sample. In this section, we construct a method that accounts for the accuracy of a classifier using the level of individual galaxy agreement in addition to their assessment on mock images. In this new method, we are able to simultaneously estimate the merger fraction of a population and the probability of each individual galaxy being in a merger.

4.1. Statistical Framework for Per Galaxy Merger Assessments

Only a few recent works have had enough human classifiers to assume a frequentist approach and use the mean of the individual classifications of a given galaxy to estimate the individual galaxy morphology (i.e., *Galaxy Zoo* 1 and 2, Lintott et al. 2008; Willett et al. 2013). In *Galaxy Zoo* 2, their main sample of 283,971 galaxies had a median of 44 classifications; the minimum was 16, and > 99.9% of the sample had at least 28 classifications. Even in the case of many individual classifications of a given galaxy, it is unclear what minimum number of classifications is needed in order to ignore intrinsic merger fraction dependent biases.

Since a large fraction of merger studies have smaller samples and consequently less human classifiers, often there are not enough individual classifications on a given galaxy to robustly report a classification and error of the classification for that galaxy in the manner that *Galaxy Zoo* studies can. In this new approach, we constrain an individual's accuracy (similar to the method presented in section 3), and using this information we show we can estimate the merger fraction of a sample and the probability of an individual galaxy being in a merger.

If a respondent is shown a merger, they will say it is a merger with probability r_M , or say it is isolated with probability $1 - r_M$. Conversely, if it is isolated, they will say it is a merger with probability $1 - r_I$ or say it is isolated with probability r_I . Thus respondent i classifies j th galaxy G with classification m as

$$p(m_i | G_j) = \begin{cases} r_M & m_i = G_j = \text{merger} \\ 1 - r_M & m_i \neq G_j = \text{merger} \\ r_I & m_i = G_j = \text{isolated} \\ 1 - r_I & m_i \neq G_j = \text{isolated} \end{cases} \quad (15)$$

With more sub-categories, this can be generalized to $p(m_i | G_j) = r_{ij}$, where $\sum_i r_{ij} = 1$. The mock galaxy sample presented in section 3 is a Bernoulli trial, although technically the respondents were asked to choose one option out of five. The generalization is described by a multinomial distribution, and its conjugate distribution is Dirichlet.

The likelihood of the classifications of a single galaxy by multiple classifiers given a merger fraction and classifier accuracies can be written

$$p(\{m_i\} | \{r_i\}, f_m) = f_m \prod_i p(m_i | G = M) + (1 - f_m) \prod_i p(m_i | G = I). \quad (16)$$

In this expression, the true nature of the galaxy in question is marginalized out. Expanding to multiple galaxies, we get the likelihood for the classifications of a collection of galaxies:

$$p(\{m_{ij}\} | \{r_i\}, f_m) = \prod_j p(\{m_{ij}\} | \{r_i\}, f_m). \quad (17)$$

Multiplying this likelihood by a prior on the merger fraction and, if the classifier accuracies are not held fixed, by a prior on accuracies gives the unnormalized posterior probability distribution function for this model.

If we wish to recover the probability that a particular galaxy is a merger, we can use the expression

$$p(G = M | \{m_i\}, \{r_i\}, f_m) = \frac{f_m \prod_i p(m_i | G = M)}{p(\{m_i\} | \{r_i\}, f_m)}. \quad (18)$$

The probability that this galaxy is isolated is the complement of this expression. The classifier's observations of simulated galaxies can be used as a prior on the observer's accuracies $r_{M/I}$, depending on how they classify the known synthetic population. This gives an informative prior, which inherently assumes that the synthetic catalog is statistically similar to the real catalog.

The strength of this method is its internal consistency; given a set of observed mergers, the likelihood is maximized when a value of f_M shown to all classifiers is most plausible given a set of individual classifications for each galaxy. We evaluate Equation 17 using the Markov chain Monte Carlo No-U-Turn Sampler algorithm (details within Hoffman & Gelman 2011) using the open source probabilistic programming framework PyMC3 (Salvatier et al. 2016).

The likelihood function of a given galaxy having a specific morphological classification requires a robust statistical description of a human classifiers accuracy in

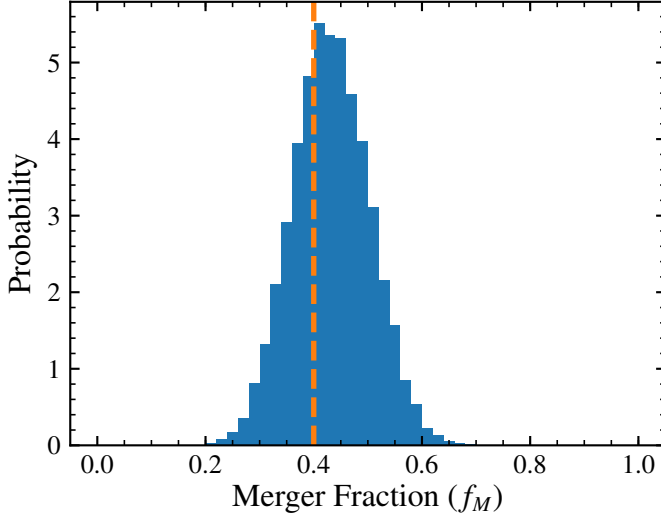


Figure 9. Merger Fraction Probability for a Simulated Galaxy Catalogue Using Level of Classifier Agreement per Galaxy: The orange dashed lines corresponds to the merger fraction truth value of 0.4. The blue histogram is the merger probability of a simulated galaxy catalogue (50 objects) using simulated classifications (14 classifiers) with mean accuracies $r_M=0.75$ and $r_I=0.65$ for identifying mergers and isolated galaxies respectively.

assessing both merging and isolated systems. In the previous step, where we maximize the likelihood of a population’s merger fraction, our algorithm also maximizes the likelihood of an individual galaxy’s classification. This allows for deeper data exploration on galaxy samples that are normally too small to do anything but population averages.

We also provide to the community the full code repository to calculate the merger fraction probability and probability of an individual galaxy being in a merger given a set of individual galaxy classifications and an estimate of the classifier accuracy.⁵

4.2. Testing Per Galaxy Model on Simulated Data

Similarly as in subsection 3.2, we simulate a galaxy catalog with imaginary classifications for each individual galaxy to obtain the probability that each galaxy is in a merger. We test with a true underlying merger fraction $f_M = 0.4$. We randomly assign 14 accuracy pairs from a uniform distribution $r_M \sim \mathcal{U}(0.5, 0.9)$ and $r_I \sim \mathcal{U}(0.5, 0.9)$ for each classifier. For each classifier, we assign 50 observations, and use mean accuracies $r_M=0.75$ and $r_I=0.65$ for identifying mergers and isolated galaxies respectively. In Figure 9, we show

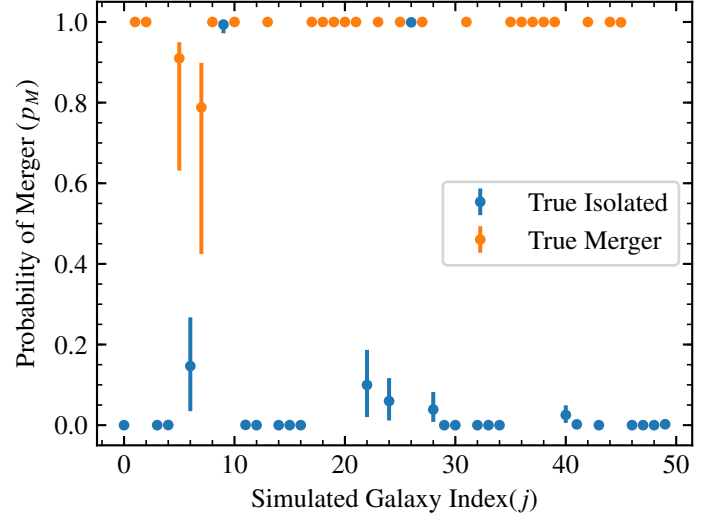


Figure 10. Probability of Individual Galaxy Being in a Merger for a Simulated Galaxy Catalogue: The simulated galaxies are labeled by simulated classifiers (14) with the same perfectly known r_M and r_I values used in Figure 9.

the merger fraction probability for the simulated galaxy catalogue using the statistical framework presented in section 4. The mean estimated merger fraction probability is within a σ of the true merger fraction of 0.4: $f_M = 0.43 \pm 0.07$. As a consistency test, we evaluated the likelihood function with 50 different randomly generated galaxy catalogues and classifier accuracies, and we find, for every test, the mean merger fraction is within 1σ of the input true merger fraction.

As shown in Equation 18, we can also estimate the probability of an individual galaxy being in a merger. Using the same example parameters as in Figure 9, we show the probability of 50 galaxies being in a merger given the above simulated set-up in Figure 10. We find two galaxies that are mis-classified, which yields an overall accuracy of 96%.

4.3. Testing on Mock Galaxies with Real Human Classifiers

We now test how well our method recovers the properties of mock galaxies as observed by real classifiers. This is an important test, as the mock images are constructed to be realistic, and the ability of classifiers to identify them correctly should be closely related to classifiers’ ability to identify real galaxies’ properties. In addition, this test allows us to look at the failures of the model on a per-image basis and determine whether the issue comes from the algorithm or the data.

Using the formalism of this section, we can estimate the probability that each mock galaxy is a merger given the accuracies of our classifiers and their agreement on

⁵ https://github.com/elambrid/merger_or_not

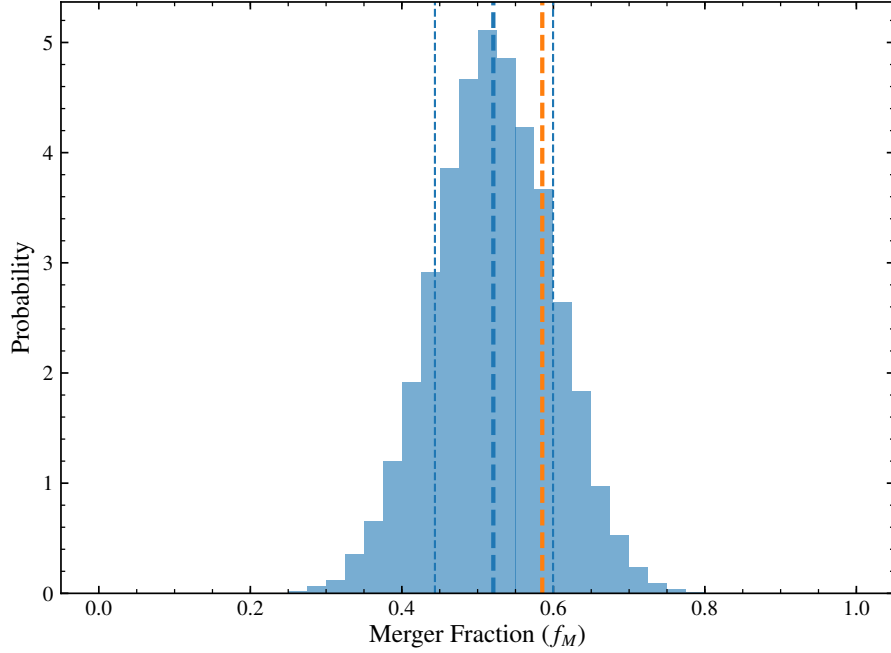


Figure 11. Merger Fraction Probability for a sample of Mock Galaxies Classified by Real Humans: The orange dashed line corresponds to the merger fraction truth value of 0.59. The blue histogram is the merger probability of the mock galaxy catalogue (41 objects) using real human classifications (14 classifiers) with the accuracies at identifying mergers and isolated galaxies estimated using Equation 17.

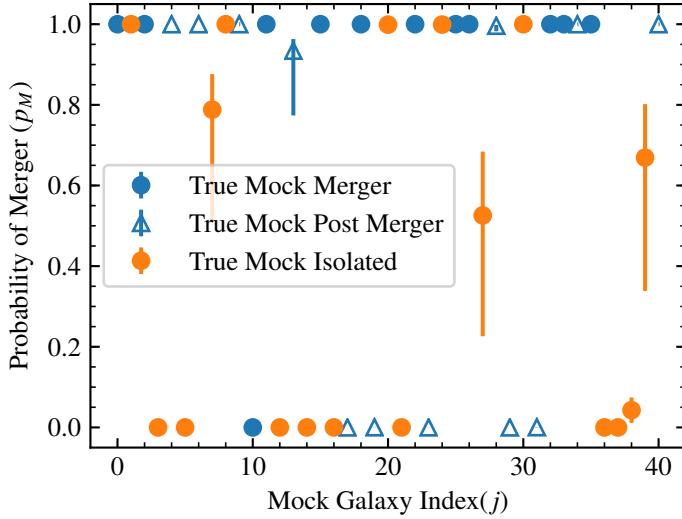


Figure 12. Probability of Individual Mock Galaxies Being in a Merger: The mock galaxies are classified by 14 real human classifiers with a range of accuracies estimated using Equation 17. The solid blue circles correspond to the mock galaxies undergoing a merger, the empty blue triangles are mock galaxies that have coalesced within 100 Myrs, and the orange filled circles are mock galaxies in isolated systems.)

classifications of individual galaxies. We use the full Monte Carlo Markov Chain samples to estimate the scatter in this value. In other words, for each step

in the chain i , there is a vector of parameters $N_{M,i}$, $\{r_{M,i,i}\}$, and $\{r_{I,i,i}\}$ that can be used to estimate the probability of a galaxy being in a merger for that step, $p_i(\text{merger} | \{m_{ij}, r_{M,ij,i}, r_{I,ij,i}\})$. We can then report the probability of this galaxy being in a merger by finding the 5%, 50%, and 95% percentiles, corresponding to 2σ Gaussian errors.

Using the above methodology, we calculate the per-galaxy merger probability for the mock galaxy sample. We simultaneously estimate the probability of the merger fraction, the accuracies of the classifiers, and the probability that each mock galaxy is in a merger given every classifier's label. We first separate the "fake mergers" from the merging and isolated mock galaxy sample. We do this to strictly test the ability of human classifiers to identify objects that are not intrinsically merging but have small on-sky separations. We found 50% of the "fake mergers" were erroneously classified as mergers, and 100% of sources that were within $2''$ of each other were erroneously classified as mergers. Thus, even with the usage of mock galaxy samples aimed to mimic the color differences due to objects at different redshifts, in this test, humans struggled to interpret their merger status correctly. We do not include the "fake mergers" in the merger fraction analysis presented in the rest of this section. To truly construct a test in the context of the effect of "fake close pairs" on the error of the merger

fraction the entire mock galaxy sample would need to incorporate fake galaxies at various redshifts $>8''$ of the central source in such a fashion to simulate real galaxy cut-outs. While this is an important exploration of a potential source of human bias on merger fraction measurements, it is outside the scope of this study and will be incorporated in future work.

In Figure 11, we show the merger fraction probability for a sample of merging and isolated mock galaxies classified by real humans. The orange dashed line corresponds to the merger fraction truth value of 0.59. The blue histogram is the merger probability of the mock galaxy catalogue (41 objects) using real human classifications (14 classifiers) with the accuracies at identifying mergers and isolated galaxies estimated using Equation 17. The mean estimated merger fraction probability is within a σ of the true merger fraction of 0.59: $f_M = 0.52 \pm 0.08$. In Figure 12, we show the probability of 41 galaxies being in a merger given the above simulated set-up. We find 6 galaxies that are mis-classified as mergers ($p_M \geq 0.8$, which yields a merger accuracy of 85%.

We can understand the implications of Figure 12 by determining the completeness, and overall accuracy of our model at inferring true merger classifications in the context of their merger state. We define completeness as the intrinsic merging mock galaxies with $p_M \geq 0.8$ compared to the total amount of intrinsic mock merging galaxies, and accuracy as the comparison between the estimated merger fraction and the true merger fraction of the mock galaxy sample. Using the classified mock galaxy intrinsically merging sample, we split the classifications of intrinsically pre-coalesced vs intrinsically post-coalesced (post-merger) systems. Note that, a mock galaxy is defined as a post-coalesced system if in the previous time-stamp, or 100 Myrs prior, the system was undergoing a merger with at least a mass ratio of 0.25. By our framework, pre-coalesced and post-coalesced systems are both defined as "mergers", and we aim to test whether intrinsically post-coalesced systems have similar classification accuracies as pre-coalesced systems.

We find 15% of the intrinsically merging systems are mis-classified as isolated galaxies, thus a completeness of 85%. We note all but one of the mis-classified mock galaxies are post-mergers. Upon deeper inspection, all of the falsely categorized mock post-mergers are in the top 50% of the mock merger redshift distribution with mean and median $z = 1.98$, $z = 2.03$ respectively. When removing post-merging galaxies with $z > 2.0$, we find a completeness of 90% and total accuracy of 92%. Our results show that human visual classification can identify

post-merging systems within 100 Myrs of coalescence up to $z = 2.0$ with 92% accuracy. For on-going merging galaxies, using our model, humans are able to robustly classify on-going significant mergers up to $z = 3.0$ with 92% accuracy. The decrease in accuracy of human classified post-merging systems is not surprising because merger features become more faint as time from coalescence increases. Previous studies that have combined pre- and post-coalesced merging galaxies across a large redshift range may be particularly susceptible to underestimating the overall merger fraction of their sample. Due to the relatively coarse time resolution of VELA-Sunrise snapshots, we are unable to further test the effect of merger feature dimming post-coalescence.

We also note five isolated galaxies are erroneously measured as mergers. When removing higher redshift post-merging galaxies from the sample, these misclassified isolated systems drive the inaccuracy of our results. These five galaxies in particular have 95% agreement of a "merging" classification from the 14 human classifiers. Future work will consist of understanding how the accuracy of human classification varies as a function of additional galaxy properties (i.e stellar mass, minor mergers) to understand why there can be such high classifier agreement on mis-classified sources.

5. SUMMARY AND CONCLUSIONS

In this work we propose a method of quantifying and accounting for merger biases of individual human classifiers and incorporate these biases into a full probabilistic model to determine the merger fraction of a population, and the probability of an individual galaxy being in a merger. We find the bias introduced from human classification is dependent on the intrinsic merger fraction of the population, and thus in order to report robust results from human visually classified data-sets, the bias from humans must be quantified.

We then construct a likelihood model to determine the merger fraction of a sample given a set of human classifications. We apply this model using two different data-sets: (1) A simulated galaxy catalogue with simulated classifications (2) Real Human Classifications on a sample of mock galaxies derived from the VELA-SUNRISE sample, a catalogue of zoom-in hydro-dynamical galaxy simulations with synthetic Hubble ACS/WFC3 images Simons et al. (2019). We recover the merger fractions to within 1% of the truth for the simulated galaxy catalogue with simulated classifiers. For the real human classifications on a sample of mock galaxy images, we recover the merger fraction to within 1% of the true merger fraction.

We then create a model to simultaneously determine the merger fraction, human accuracies and probability of each individual galaxy being in a merger. Using simulated human responses and accuracies, we are able to correctly label a galaxy as a "merger" or "isolated" to within 3% of the truth. Using the mock galaxies with real human classifications, our model is able to recover the pre-coalescing merger fraction to within 10%. For galaxies that have coalesced within 100 Myrs, our model recovers the intrinsic merger fraction to within 10% for the sources that occupy the lowest 50% of the redshift distribution. For the post-coalesced sources in the top 50% of the redshift distribution (i.e $z \sim 2.0$), the accuracy of human classifiers significantly drops, and our model infers a merger fraction within 15% of the truth. Note, this specific bound is observed at this redshift due to the mock galaxy images incorporating a noise model that will reflect the sensitivity of GOODS-S Hubble Observations. Thus, this important estimate on human classifier accuracy must be incorporated in merger studies that contain high redshift post-merger sources in the GOODS-S field.

The implementation of our Bayesian model in studies that assess the merger state of $0.5 < z < 2$ galaxies using

human classifiers yields better understood errors on the merger fraction. In addition, this statistical framework is able to more robustly constrain the probability of individual galaxies being in mergers with a smaller number of human classifiers than was previously possible.

ACKNOWLEDGMENTS

We thank the anonymous referee for their thoughtful insight and important contributions to this work. In addition, we thank [insert] for useful discussions and insight. ELL is supported by [].

Software: `astropy` (Astropy Collaboration et al. 2013; Price-Whelan et al. 2018), `corner` (Foreman-Mackey 2016), `emcee` (Foreman-Mackey et al. 2013), `IPython` (Pérez & Granger 2007), `matplotlib` (Hunter 2007), `numpy` (van der Walt et al. 2011), `pandas` (McKinney 2010; Pandas Development Team 2020), `scipy` (Virtanen et al. 2020), `statsmodels` (Seabold & Perktold 2010)

APPENDIX

A. MOCK GALAXY IMAGES AND CLASSIFICATIONS

As described in subsection 3.3, Figure 13, Figure 14, Figure 16 comprise the VELA+SUNRISE noise-added merging and isolated mock galaxy sample. The images are identified by their ID number. We also label each image as "correct" or "incorrect". Using Equation 18, we label a galaxy as "correctly" measured if the probability of being in a merger, p_M , is $> 99\%$ and the error on the probability is less than 10%. In Table 2, we list the z , intrinsic type, and measured probability of each VELA+SUNRISE noise-added mock galaxy.

| ID | VELA ID | z | Type | p_M | err |
|----|-----------|------|------|-------|-----|
| 0 | 32,320,10 | 2.13 | i | 0.00 | n |
| 1 | 7,370,10 | 1.70 | i | 1.00 | n |
| 2 | 27,330,10 | 2.03 | i | 0.53 | y |
| 3 | 20,290,5 | 2.45 | f | 1.00 | n |
| 4 | 4,460,10 | 1.17 | i | 1.00 | n |
| 5 | 2,290,10 | 2.45 | m | 1.00 | n |
| 6 | 29,480,10 | 1.08 | m | 1.00 | n |
| 7 | 6,310,10 | 2.23 | m | 1.00 | n |
| 8 | 33,330,5 | 2.03 | m | 0.00 | y |
| 9 | 29,500,5 | 1.00 | f | 0.00 | n |
| 10 | 30,300,10 | 2.33 | i | 0.04 | n |
| 11 | 33,380,5 | 1.63 | m | 1.00 | n |
| 12 | 10,420,5 | 1.38 | m | 0.00 | n |
| 13 | 25,270,5 | 2.70 | f | 0.00 | n |
| 14 | 12,340,10 | 1.94 | i | 0.00 | n |
| 15 | 25,300,5 | 2.33 | m | 1.00 | n |
| 16 | 33,380,10 | 1.63 | m | 1.00 | n |
| 18 | 21,400,10 | 1.50 | i | 0.79 | y |
| 20 | 19,220,10 | 3.55 | i | 0.00 | n |
| 21 | 21,400,5 | 1.50 | f | 0.00 | n |
| 22 | 1,470,10 | 1.13 | i | 1.00 | n |
| 24 | 33,250,5 | 3.00 | m | 0.00 | n |
| 25 | 23,450,0 | 1.22 | m | 1.00 | n |
| 26 | 9,300,10 | 2.33 | i | 0.00 | n |

| ID | VELA ID | z | Type | p_M | err |
|----|-----------|------|------|-------|-----|
| 27 | 12,380,10 | 1.63 | i | 0.00 | n |
| 28 | 6,290,5 | 2.45 | m | 1.00 | n |
| 30 | 23,320,10 | 2.13 | i | 0.67 | y |
| 31 | 28,330,10 | 2.03 | m | 0.00 | y |
| 32 | 1,410,0 | 1.44 | m | 0.93 | y |
| 33 | 5,390,10 | 1.56 | m | 0.00 | n |
| 34 | 27,330,5 | 2.03 | f | 1.00 | n |
| 35 | 25,270,10 | 2.70 | i | 0.00 | n |
| 36 | 33,250,10 | 3.00 | m | 1.00 | n |
| 37 | 29,280,10 | 2.57 | m | 1.00 | n |
| 38 | 8,360,5 | 1.78 | m | 1.00 | n |
| 39 | 22,460,5 | 1.17 | f | 1.00 | n |
| 40 | 9,390,10 | 1.56 | m | 1.00 | n |
| 41 | 29,480,5 | 1.08 | m | 1.00 | n |
| 42 | 25,480,5 | 1.08 | f | 0.00 | n |
| 43 | 3,380,10 | 1.63 | i | 1.00 | n |
| 44 | 33,330,10 | 2.03 | m | 0.00 | y |
| 46 | 17,310,5 | 2.23 | m | 1.00 | n |
| 47 | 20,290,10 | 2.45 | i | 1.00 | n |
| 48 | 32,320,5 | 2.12 | f | 1.00 | n |
| 49 | 20,390,5 | 1.56 | m | 1.00 | n |
| 50 | 24,370,5 | 1.70 | m | 1.00 | n |
| 51 | 25,480,10 | 1.08 | i | 0.00 | n |
| 52 | 22,460,10 | 1.17 | i | 0.00 | n |
| 53 | 28,270,10 | 2.70 | m | 1.00 | n |

Table 2. Probabilities of an Individual Mock Galaxy Being in a Merger: The column ID refers to our internal mock galaxy catalogue identification number. The VELA ID column is the combination of the simulation number (sim), the time snapshot (snap), and camera id (id) as defined in [Simons et al. \(2019\)](#). The z column corresponds to the redshift. The true type column refers to the intrinsic morphological type of the galaxy as determined by [Simons et al. \(2019\)](#), where "i" refers to isolated, "m" refers to merger and "f" refers to a fake merger. A fake merger designation is for cutouts where two mock isolated galaxies were superimposed using a random separation $< 8''$ to represent real world observations of apparent galaxy pairs that have small on-sky separations but exist at different redshifts. The merger designation includes minor-, major-, pre- post- merging/coalesced systems. The p_M column refers to the probability of the object being in a merger given the classifications of 14 human classifiers and evaluated using [Equation 18](#). The final column, err, indicates whether the probability of an object being in a merger is unconstrained where an unconstrained probability is defined as when the standard deviation of the probability distribution, p_M , is greater than 10%.

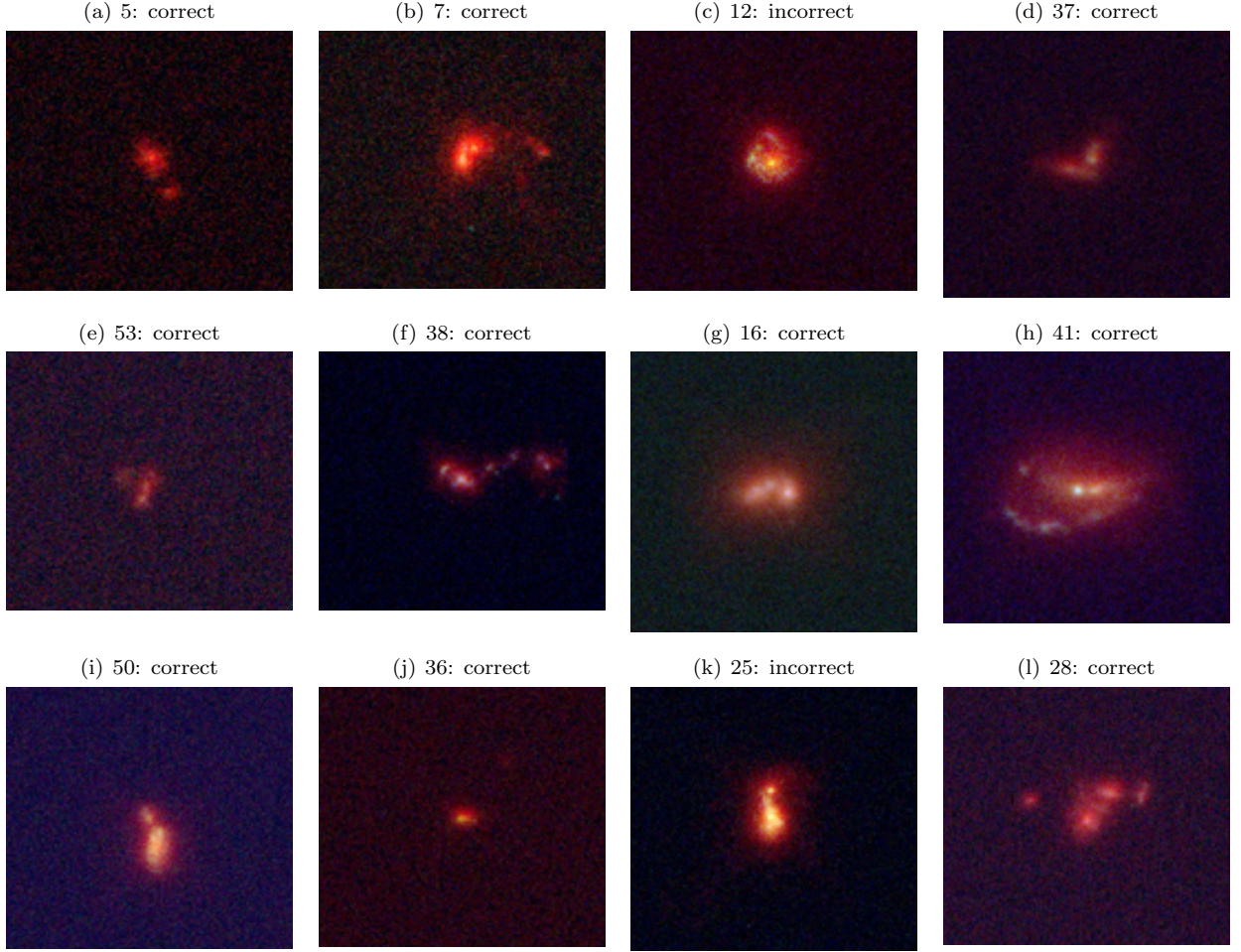


Figure 13. Pre-coalescence VELA+SUNRISE Noise Added Mock Merging Galaxies. The ID and measured classification are provided for each image (r:HSTWFC3/IR F160W, g:HSTACS F775W, b: ACS F435W). If a galaxy is classified correctly, $p_M > 99\%$, we identify it as "correct". All cutouts are $8'' \times 8''$.

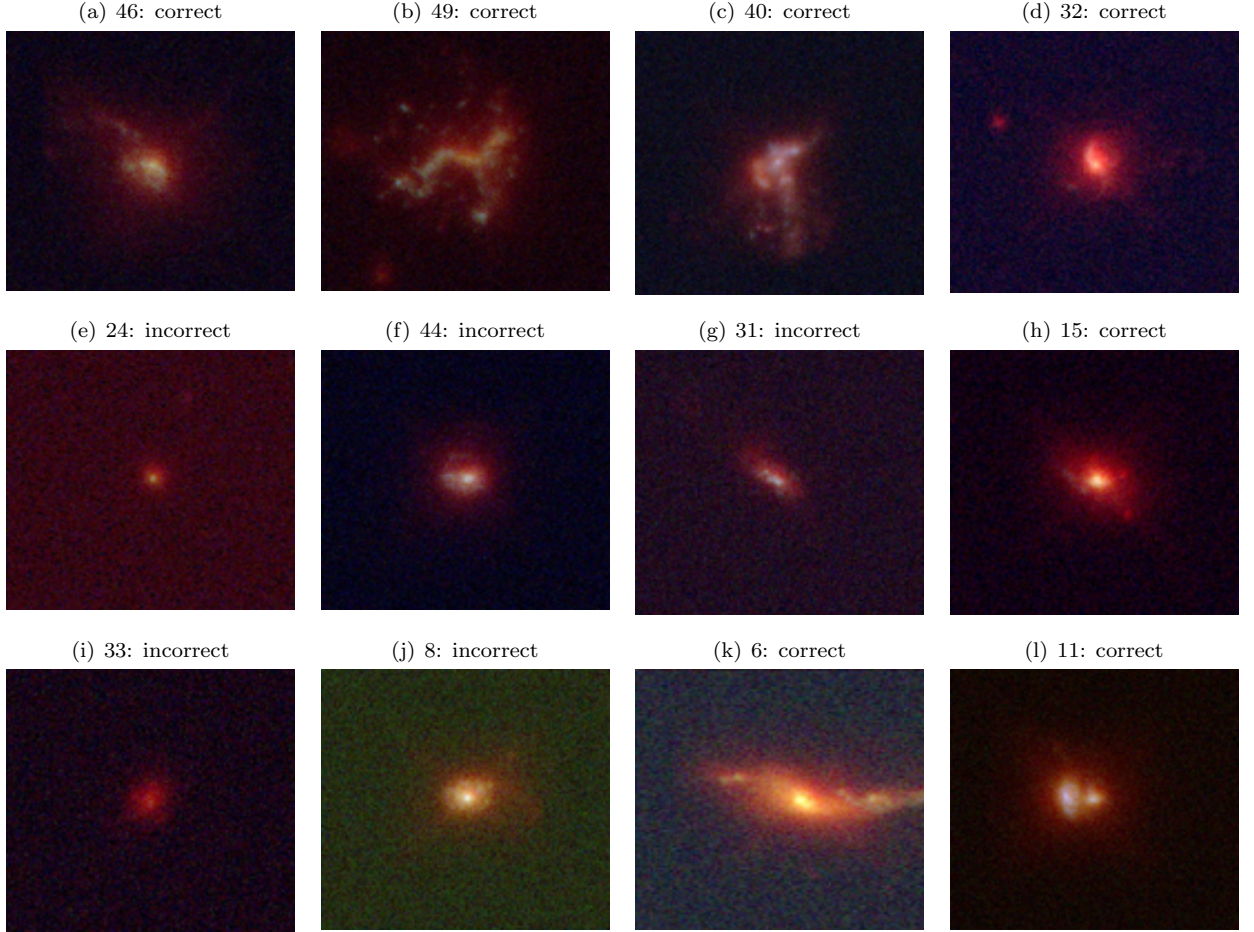


Figure 14. Post-coalescence VELA+SUNRISE Noise Added Mock Merging Galaxies. The ID and measured classification are provided for each image (r:HSTWFC3/IR F160W, g:HSTACS F775W, b: ACS F435W). If a galaxy is classified correctly, $p_M > 99\%$, we identify it as "correct". All cutouts are $8'' \times 8''$.

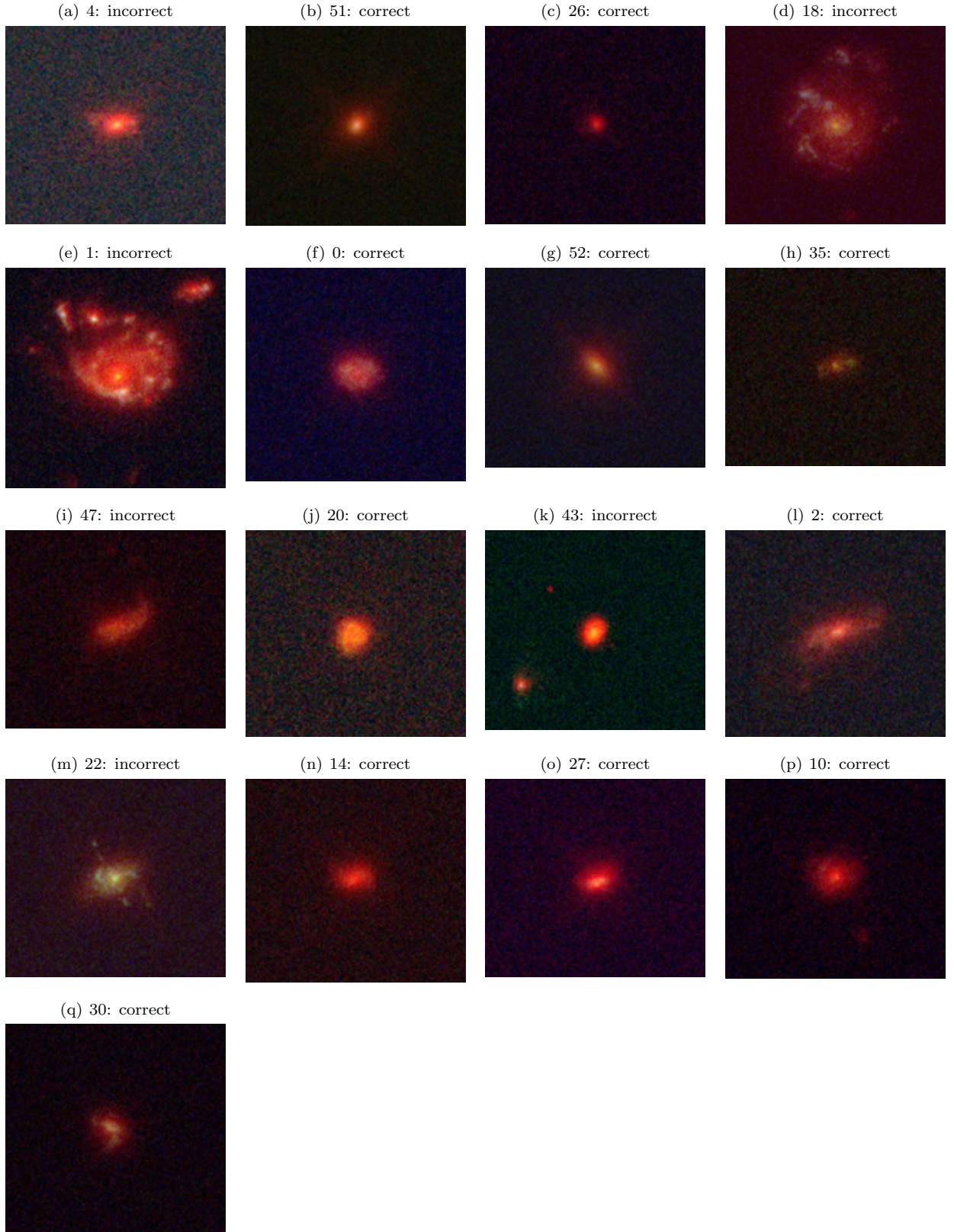


Figure 15. Isolated VELA+SUNRISE Noise Added Galaxies. The ID and measured classification are provided for each image (r:HSTWFC3/IR F160W, g:HSTACS F775W, b: ACS F435W). If a galaxy is classified correctly, $p_M < 1\%$, we identify it as "correct". All cutouts are $8'' \times 8''$.

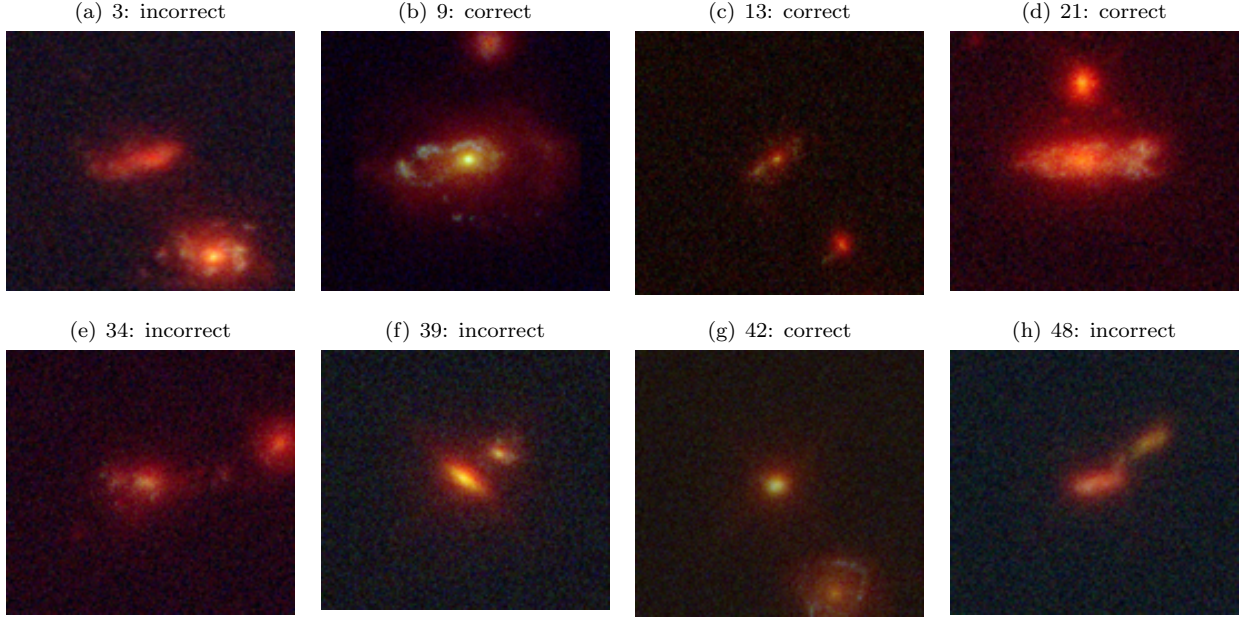


Figure 16. Fake Merging VELA+SUNRISE Noise Added Galaxies. The ID and measured classification are provided for each image (r:HSTWFC3/IR F160W, g:HSTACS F775W, b: ACS F435W). If a galaxy is classified correctly, $p_M > 99\%$, we identify it as "correct". All cutouts are $8'' \times 8''$.

REFERENCES

- Abraham, R. G., van den Bergh, S., Glazebrook, K., et al. 1996, *ApJS*, 107, 1, doi: [10.1086/192352](https://doi.org/10.1086/192352)
- Ackermann, S., Schawinski, K., Zhang, C., Weigel, A. K., & Turp, M. D. 2018, *MNRAS*, 479, 415, doi: [10.1093/mnras/sty1398](https://doi.org/10.1093/mnras/sty1398)
- Astropy Collaboration, Robitaille, T. P., Tollerud, E. J., et al. 2013, *A&A*, 558, A33, doi: [10.1051/0004-6361/201322068](https://doi.org/10.1051/0004-6361/201322068)
- Ceverino, D., Dekel, A., & Bournaud, F. 2010, *MNRAS*, 404, 2151, doi: [10.1111/j.1365-2966.2010.16433.x](https://doi.org/10.1111/j.1365-2966.2010.16433.x)
- Ceverino, D., Dekel, A., Mandelker, N., et al. 2012, *MNRAS*, 420, 3490, doi: [10.1111/j.1365-2966.2011.20296.x](https://doi.org/10.1111/j.1365-2966.2011.20296.x)
- Ceverino, D., Klypin, A., Klimek, E. S., et al. 2014, *MNRAS*, 442, 1545, doi: [10.1093/mnras/stu956](https://doi.org/10.1093/mnras/stu956)
- Chiaberge, M., Gilli, R., Lotz, J. M., & Norman, C. 2015, *ApJ*, 806, 147, doi: [10.1088/0004-637X/806/2/147](https://doi.org/10.1088/0004-637X/806/2/147)
- Ćiprijanović, A., Snyder, G. F., Nord, B., & Peek, J. E. G. 2020, *Astronomy and Computing*, 32, 100390, doi: [10.1016/j.ascom.2020.100390](https://doi.org/10.1016/j.ascom.2020.100390)
- Conselice, C. J. 2014, *ARA&A*, 52, 291, doi: [10.1146/annurev-astro-081913-040037](https://doi.org/10.1146/annurev-astro-081913-040037)
- Conselice, C. J., Bershad, M. A., & Jangren, A. 2000, *ApJ*, 529, 886, doi: [10.1086/308300](https://doi.org/10.1086/308300)
- Di Matteo, T., Springel, V., & Hernquist, L. 2005, *Nature*, 433, 604, doi: [10.1038/nature03335](https://doi.org/10.1038/nature03335)
- Dieleman, S., Willett, K. W., & Dambre, J. 2015, *MNRAS*, 450, 1441, doi: [10.1093/mnras/stv632](https://doi.org/10.1093/mnras/stv632)
- Donley, J. L., Kartaltepe, J., Kocevski, D., et al. 2018, *ApJ*, 853, 63, doi: [10.3847/1538-4357/aa9ffa](https://doi.org/10.3847/1538-4357/aa9ffa)
- Duncan, K., Conselice, C. J., Mundy, C., et al. 2019, *ApJ*, 876, 110, doi: [10.3847/1538-4357/ab148a](https://doi.org/10.3847/1538-4357/ab148a)
- Ellison, S. L., Mendel, J. T., Scudder, J. M., Patton, D. R., & Palmer, M. J. D. 2013, *MNRAS*, 430, 3128, doi: [10.1093/mnras/sts546](https://doi.org/10.1093/mnras/sts546)
- Ellison, S. L., Viswanathan, A., Patton, D. R., et al. 2019, *MNRAS*, 487, 2491, doi: [10.1093/mnras/stz1431](https://doi.org/10.1093/mnras/stz1431)
- Foreman-Mackey, D. 2016, *The Journal of Open Source Software*, 1, 24, doi: [10.21105/joss.00024](https://doi.org/10.21105/joss.00024)
- Foreman-Mackey, D., Hogg, D. W., Lang, D., & Goodman, J. 2013, *PASP*, 125, 306, doi: [10.1086/670067](https://doi.org/10.1086/670067)
- Gabor, J. M., Impey, C. D., Jahnke, K., et al. 2009, *ApJ*, 691, 705, doi: [10.1088/0004-637X/691/1/705](https://doi.org/10.1088/0004-637X/691/1/705)
- Georgakakis, A., Coil, A. L., Laird, E. S., et al. 2009, *MNRAS*, 397, 623, doi: [10.1111/j.1365-2966.2009.14951.x](https://doi.org/10.1111/j.1365-2966.2009.14951.x)
- Glikman, E., Simmons, B., Mailly, M., et al. 2015, *ApJ*, 806, 218, doi: [10.1088/0004-637X/806/2/218](https://doi.org/10.1088/0004-637X/806/2/218)
- Goodman, J., & Weare, J. 2010, *Communications in Applied Mathematics and Computational Science*, 5, 65, doi: [10.2140/camcos.2010.5.65](https://doi.org/10.2140/camcos.2010.5.65)
- Grogin, N. A., Conselice, C. J., Chatzichristou, E., et al. 2005, *ApJL*, 627, L97, doi: [10.1086/432256](https://doi.org/10.1086/432256)
- Hoffman, M. D., & Gelman, A. 2011, arXiv e-prints, arXiv:1111.4246. <https://arxiv.org/abs/1111.4246>
- Huertas-Company, M., Gravet, R., Cabrera-Vives, G., et al. 2015, *ApJS*, 221, 8, doi: [10.1088/0067-0049/221/1/8](https://doi.org/10.1088/0067-0049/221/1/8)
- Hunter, J. D. 2007, *Computing In Science & Engineering*, 9, 90, doi: [10.1109/MCSE.2007.55](https://doi.org/10.1109/MCSE.2007.55)
- Jonsson, P., Groves, B. A., & Cox, T. J. 2010, *MNRAS*, 403, 17, doi: [10.1111/j.1365-2966.2009.16087.x](https://doi.org/10.1111/j.1365-2966.2009.16087.x)
- Kocevski, D. D., Faber, S. M., Mozena, M., et al. 2012, *ApJ*, 744, 148, doi: [10.1088/0004-637X/744/2/148](https://doi.org/10.1088/0004-637X/744/2/148)
- Kocevski, D. D., Brightman, M., Nandra, K., et al. 2015, *ApJ*, 814, 104, doi: [10.1088/0004-637X/814/2/104](https://doi.org/10.1088/0004-637X/814/2/104)
- Koss, M., Mushotzky, R., Veilleux, S., & Winter, L. 2010, *ApJL*, 716, L125, doi: [10.1088/2041-8205/716/2/L125](https://doi.org/10.1088/2041-8205/716/2/L125)
- Koss, M. J., Assef, R., Baloković, M., et al. 2016, *ApJ*, 825, 85, doi: [10.3847/0004-637X/825/2/85](https://doi.org/10.3847/0004-637X/825/2/85)
- Kravtsov, A. V., Klypin, A. A., & Khokhlov, A. M. 1997, *ApJS*, 111, 73, doi: [10.1086/313015](https://doi.org/10.1086/313015)
- Lanzuisi, G., Civano, F., Marchesi, S., et al. 2018, *MNRAS*, 480, 2578, doi: [10.1093/mnras/sty2025](https://doi.org/10.1093/mnras/sty2025)
- Li, J., Xue, Y., Sun, M., et al. 2020, *ApJ*, 903, 49, doi: [10.3847/1538-4357/abb6e7](https://doi.org/10.3847/1538-4357/abb6e7)
- Lintott, C. J., Schawinski, K., Slosar, A., et al. 2008, *MNRAS*, 389, 1179, doi: [10.1111/j.1365-2966.2008.13689.x](https://doi.org/10.1111/j.1365-2966.2008.13689.x)
- Lotz, J. M., Primack, J., & Madau, P. 2004, *AJ*, 128, 163, doi: [10.1086/421849](https://doi.org/10.1086/421849)
- Madau, P., & Dickinson, M. 2014, *ARA&A*, 52, 415, doi: [10.1146/annurev-astro-081811-125615](https://doi.org/10.1146/annurev-astro-081811-125615)
- Mantha, K. B., McIntosh, D. H., Brennan, R., et al. 2018, *MNRAS*, 475, 1549, doi: [10.1093/mnras/stx3260](https://doi.org/10.1093/mnras/stx3260)
- Marian, V., Jahnke, K., Mechtley, M., et al. 2019, *ApJ*, 882, 141, doi: [10.3847/1538-4357/ab385b](https://doi.org/10.3847/1538-4357/ab385b)
- Marian, V., Jahnke, K., Andika, I., et al. 2020, *ApJ*, 904, 79, doi: [10.3847/1538-4357/abbd3e](https://doi.org/10.3847/1538-4357/abbd3e)
- McKinney, W. 2010, in *Proceedings of the 9th Python in Science Conference*, ed. S. van der Walt & J. Millman, 51 – 56
- Mechtley, M., Jahnke, K., Windhorst, R. A., et al. 2016, *ApJ*, 830, 156, doi: [10.3847/0004-637X/830/2/156](https://doi.org/10.3847/0004-637X/830/2/156)
- Pandas Development Team. 2020, *pandas-dev/pandas: Pandas 1.1.0, v1.1.0*, Zenodo, doi: [10.5281/zenodo.3964380](https://doi.org/10.5281/zenodo.3964380)

- Pearson, W. J., Wang, L., Trayford, J. W., Pettillo, C. E., & van der Tak, F. F. S. 2019, *A&A*, 626, A49, doi: [10.1051/0004-6361/201935355](https://doi.org/10.1051/0004-6361/201935355)
- Pérez, F., & Granger, B. E. 2007, *Computing in Science and Engineering*, 9, 21, doi: [10.1109/MCSE.2007.53](https://doi.org/10.1109/MCSE.2007.53)
- Price-Whelan, A. M., Sipőcz, B. M., Günther, H. M., et al. 2018, *AJ*, 156, 123, doi: [10.3847/1538-3881/aabc4f](https://doi.org/10.3847/1538-3881/aabc4f)
- Rosario, D. J., McIntosh, D. H., van der Wel, A., et al. 2015, *A&A*, 573, A85, doi: [10.1051/0004-6361/201423782](https://doi.org/10.1051/0004-6361/201423782)
- Salvatier, J., Wieckia, T. V., & Fonnesbeck, C. 2016, *PyMC3: Python probabilistic programming framework*. <http://ascl.net/1610.016>
- Sanders, D. B., & Mirabel, I. F. 1996, *ARA&A*, 34, 749, doi: [10.1146/annurev.astro.34.1.749](https://doi.org/10.1146/annurev.astro.34.1.749)
- Schawinski, K., Simmons, B. D., Urry, C. M., Treister, E., & Glikman, E. 2012, *MNRAS*, 425, L61, doi: [10.1111/j.1745-3933.2012.01302.x](https://doi.org/10.1111/j.1745-3933.2012.01302.x)
- Schawinski, K., Treister, E., Urry, C. M., et al. 2011, *ApJL*, 727, L31, doi: [10.1088/2041-8205/727/2/L31](https://doi.org/10.1088/2041-8205/727/2/L31)
- Seabold, S., & Perktold, J. 2010, in 9th Python in Science Conference
- Simons, R. C., Kassin, S. A., Snyder, G. F., et al. 2019, *ApJ*, 874, 59, doi: [10.3847/1538-4357/ab07c9](https://doi.org/10.3847/1538-4357/ab07c9)
- Snyder, G. F., Lotz, J., Moody, C., et al. 2015, *MNRAS*, 451, 4290, doi: [10.1093/mnras/stv1231](https://doi.org/10.1093/mnras/stv1231)
- Treister, E., Schawinski, K., Urry, C. M., & Simmons, B. D. 2012, *ApJL*, 758, L39, doi: [10.1088/2041-8205/758/2/L39](https://doi.org/10.1088/2041-8205/758/2/L39)
- Urrutia, T., Lacy, M., & Becker, R. H. 2008, *ApJ*, 674, 80, doi: [10.1086/523959](https://doi.org/10.1086/523959)
- van der Walt, S., Colbert, S. C., & Varoquaux, G. 2011, *Computing in Science & Engineering*, 13, 22, doi: [10.1109/mcse.2011.37](https://doi.org/10.1109/mcse.2011.37)
- Veilleux, S., Rupke, D. S. N., Kim, D. C., et al. 2009, *ApJS*, 182, 628, doi: [10.1088/0067-0049/182/2/628](https://doi.org/10.1088/0067-0049/182/2/628)
- Villforth, C., Hamann, F., Rosario, D. J., et al. 2014, *MNRAS*, 439, 3342, doi: [10.1093/mnras/stu173](https://doi.org/10.1093/mnras/stu173)
- Villforth, C., Hamilton, T., Pawlik, M. M., et al. 2017, *MNRAS*, 466, 812, doi: [10.1093/mnras/stw3037](https://doi.org/10.1093/mnras/stw3037)
- Virtanen, P., Gommers, R., Oliphant, T. E., et al. 2020, *Nature Methods*, 17, 261, doi: <https://doi.org/10.1038/s41592-019-0686-2>
- Willett, K. W., Lintott, C. J., Bamford, S. P., et al. 2013, *MNRAS*, 435, 2835, doi: [10.1093/mnras/stt1458](https://doi.org/10.1093/mnras/stt1458)