

Deep learning predictions of galaxy merger stage and the importance of observational realism

Connor Bottrell ¹★, Maan H. Hani ¹†, Hossen Teimoorinia,^{1,2} Sara L. Ellison ¹,
Jorge Moreno,^{3,4,5} Paul Torrey ⁶, Christopher C. Hayward,⁷ Mallory Thorp ¹,
Luc Simard² and Lars Hernquist⁴

¹Department of Physics and Astronomy, University of Victoria, Victoria, British Columbia V8P 1A1, Canada

²National Research Council of Canada, 5071 West Saanich Road, Victoria, British Columbia V9E 2E7, Canada

³Department of Physics and Astronomy, Pomona College, Claremont, CA 91711, USA

⁴Harvard-Smithsonian Center for Astrophysics, 60 Garden Street, Cambridge, MA 02138, USA

⁵TAPIR, Mailcode 350-17, California Institute of Technology, Pasadena, CA 91125, USA

⁶Department of Astronomy, University of Florida, 211 Bryant Space Science Center, Gainesville, FL 32611, USA

⁷Center for Computational Astrophysics, Flatiron Institute, 162 Fifth Avenue, New York, NY 10010, USA

Accepted 2019 October 15. Received 2019 October 3; in original form 2019 August 21

ABSTRACT

Machine learning is becoming a popular tool to quantify galaxy morphologies and identify mergers. However, this technique relies on using an appropriate set of training data to be successful. By combining hydrodynamical simulations, synthetic observations, and convolutional neural networks (CNNs), we quantitatively assess how realistic simulated galaxy images must be in order to reliably classify mergers. Specifically, we compare the performance of CNNs trained with two types of galaxy images, stellar maps and dust-inclusive radiatively transferred images, each with three levels of observational realism: (1) no observational effects (idealized images), (2) realistic sky and point spread function (semirealistic images), and (3) insertion into a real sky image (fully realistic images). We find that networks trained on either idealized or semireal images have poor performance when applied to survey-realistic images. In contrast, networks trained on fully realistic images achieve 87.1 per cent classification performance. Importantly, the level of realism in the training images is much more important than whether the images included radiative transfer, or simply used the stellar maps (87.1 per cent compared to 79.6 per cent accuracy, respectively). Therefore, one can avoid the large computational and storage cost of running radiative transfer with a relatively modest compromise in classification performance. Making photometry-based networks insensitive to colour incurs a very mild penalty to performance with survey-realistic data (86.0 per cent with *r*-only compared to 87.1 per cent with *gri*). This result demonstrates that while colour *can* be exploited by colour-sensitive networks, it is not necessary to achieve high accuracy and so can be avoided if desired. We provide the public release of our statistical observational realism suite, REALSIM, as a companion to this paper.

Key words: Methods: data analysis – Methods: numerical – Techniques: image processing – Galaxies: general – Galaxies: interactions – Galaxies: photometry.

1 INTRODUCTION

Theoretical predictions and observations alike show that mergers transform galaxies. Stellar bridges and tails observed in interacting

galaxy pairs are the relics of the strong gravitational and tidal forces involved in close galaxy–galaxy encounters (Toomre & Toomre 1972). But the consequences of these forces extend well beyond immediate changes to visual morphology.

Tidal torques and shocks excited by close encounters can rapidly reduce angular momentum in the dynamically cold interstellar medium (ISM) through various channels – all driving inflow of available cold gas towards the centres of interacting galaxies (e.g.

* E-mail: cbottrel@uvic.ca

† Vanier Scholar.

Hernquist 1989; Barnes & Hernquist 1992; Mihos & Hernquist 1996; Hopkins & Quataert 2010; Blumenthal & Barnes 2018). There is now a strong numerical and observational framework linking this rapid and central accumulation of gas to boosts in central star formation rates (SFRs; e.g. Ellison et al. 2008; Patton et al. 2011; Hopkins et al. 2013; Patton et al. 2013; Moreno et al. 2015; Sparre & Springel 2016; Thorp et al. 2019), dilution of central gas phase metallicity (e.g. Kewley, Geller & Barton 2006; Ellison et al. 2008; Rupke, Kewley & Barnes 2010a; Rupke, Kewley & Chien 2010b; Sol Alonso, Michel-Dansac & Lambas 2010; Perez, Michel-Dansac & Tissera 2011; Torrey et al. 2012; Moreno et al. 2015; Thorp et al. 2019), and accretion on to central black holes and triggering of active galactic nuclei (AGNs; e.g. Keel et al. 1985; Di Matteo, Springel & Hernquist 2005; Koss et al. 2010; Ellison et al. 2011; Satyapal et al. 2014; Ellison, Patton & Hickox 2015; Goulding et al. 2018; Ellison et al. 2019). Additionally, galactic outflows of gas associated with the enhancements in SFRs (e.g. Martin 2005; Rupke, Veilleux & Sanders 2005a; Strickland & Heckman 2009; Hayward & Hopkins 2017) and AGN activity (e.g. Rupke, Veilleux & Sanders 2005b; Veilleux et al. 2013; Zschaechner et al. 2016; Woo, Son & Bae 2017) can also be triggered by mergers – resulting in the growth and enrichment the circumgalactic medium (e.g. Johnson, Chen & Mulchaey 2015; Hani et al. 2018). Combined with the role of mergers in the assembly of present-day galaxies (e.g. White & Rees 1978; Blumenthal et al. 1984) and transforming their morphologies and kinematics (e.g. Toomre 1977; Negroponte & White 1983; Hernquist 1992; Naab & Burkert 2003; Hopkins et al. 2008c; Berg et al. 2014), these connections make mergers complex but unique laboratories for testing some of the most crucial aspects of galaxy formation physics.

One observationally measurable parameter that is particularly valuable for testing the statistical and cosmological role of mergers in galaxy evolution (and which is directly comparable to numerical predictions from semi-analytic models or cosmological hydrodynamical simulations) is the galaxy merger rate and its evolution with mass and redshift (e.g. Lacey & Cole 1993; López-Sanjuan et al. 2011; Lotz et al. 2011; Bluck et al. 2012; López-Sanjuan et al. 2013; Casteels et al. 2014; Rodríguez-Gomez et al. 2015; Martin et al. 2018). Estimating the merger rate requires: (1) a method with which mergers can be distinguished from non-merging galaxies and (2) an estimate of the time-scales on which the distinction can be made – which is sensitive to the method used in the former (Hopkins et al. 2008a; Lotz et al. 2008, 2010a,b). However, beyond identifying mergers, we are also particularly interested in predicting merger *stage*. Hydrodynamical simulations of galaxy mergers predict significant evolution in (among others) SFRs, ISM content, AGN accretion rates and luminosities, and subsequent stellar and AGN feedback along the merger sequence (e.g. Cox et al. 2008; Hopkins et al. 2008b; Torrey et al. 2012; Hopkins et al. 2013; Moreno et al. 2015, 2019). Consequently, in order to test the broader and detailed elements of this framework (such as feedback and outflow prescriptions), we must be able to (i) obtain large and reasonably complete observational samples of galaxy mergers and (ii) connect observed galaxy interactions to specific stages in the merger sequence.

Both of these tasks present significant challenges from an observational perspective. For example, while mergers and recent post-mergers can be selected visually on the basis of distinct (but often low-surface brightness) morphological features such as tidal tails, bridges, streams, shells, and nearby companions (Darg et al. 2010; Kartaltepe et al. 2015; Simmons et al. 2017), this process is subjective and sensitive to contrast, resolution, and surface-

brightness limits. For pair candidates, the intrinsic subjectivity of visual classification can be alleviated by obtaining relative velocities with spectroscopy. Spectroscopic pair identification is effective even at high redshifts (Lin et al. 2007; Wong et al. 2011), but is often incomplete due to the ‘fibre-collision’ problem and sparse sampling – which particularly affect close pair completeness (Patton et al. 2002; Lin et al. 2004; Patton & Atfield 2008, but see also Robotham et al. 2014). Fast and reproducible classifications can be made using automated quantitative morphologies (Conselice 2003; Lotz, Primack & Madau 2004; Pawlik et al. 2016; Rodríguez-Gomez et al. 2019). These metrics are designed to exploit the excess asymmetries, disturbed morphologies, or multiple nuclei of mergers and merger remnants relative to non-merging galaxies (e.g. Casteels et al. 2013, 2014; Patton et al. 2016). The main obstacle with quantitative morphologies is defining empirical thresholds that separate merger from non-merger classes. Like visual classification, these thresholds (particularly for asymmetries) are sensitive to resolution and surface brightness limits (e.g. Ji, Peirani & Yi 2014; Bottrell et al. 2019) but also, critically, the ‘training’ data with which these thresholds are calibrated.

To calibrate an empirical threshold for a metric that separates merger from non-merger classes, one must have a way of evaluating its performance (e.g. completeness and/or purity). The significant limitation of calibrating on observational data is that the subjective and non-subjective biases afflicting visual classifications and the incompleteness of spectroscopic samples become embedded in the calibration. In other words, observationally, one does not have access to the ground truth. This limitation can be overcome using synthetic images from hydrodynamical merger simulations as the basis for the calibration step – which simultaneously solves the problems above (Lotz et al. 2008, 2010a,b; Nevin et al. 2019). Regardless of any added ingredients to the synthetic images (sky noise, resolution degradation, additional sources, etc.), the simulations provide foreknowledge of the true properties of each target: merger stage, mass ratio, gas fractions, initial morphologies, orbital parameters, etc. Consequently, one always has unbiased target classes upon which to evaluate the performance of the method. Furthermore, one can measure the biases affecting performance from both merger and image properties. Lastly, since the morphological features of galaxy interactions are induced primarily via gravitational effects, they should be largely insensitive to the particularities of the hydrodynamic model.

Two classes of methods that have gained significant traction in general astronomy and, in particular, galaxy astronomy are machine learning and deep learning (e.g. Teimoorinia & Ellison 2014; Hezaveh, Perreault Levasseur & Marshall 2017; Bluck et al. 2019; Hausen & Robertson 2019; Jacobs et al. 2019; Ntampaka et al. 2019; Ribli, Pataki & Csabai 2019; Snyder et al. 2019). Specifically, convolutional neural networks (CNNs) have been used to improve automated image-based galaxy morphology classifications with great success (Huertas-Company et al. 2015; Domínguez Sánchez et al. 2018). The level of intricacy in the features that can be identified by CNN and other machine-learning models has made them an attractive tool for merger identification (e.g. Ackermann et al. 2018; Walmsley et al. 2019). Following the approaches adopted for quantitative morphologies by calibrating on hydrodynamical simulations, Pearson et al. (2019) train a CNN on synthetic Sloan Digital Sky Survey (SDSS) images of galaxies from the EAGLE simulation (Schaye et al. 2015) and examine biases from redshift, SFRs, and apparent brightness on merger and non-merger classifications – though with poor classification performance (65.5 per cent in a binary classification). None the

less, one of the key elements of their synthetic SDSS images is that they were inserted into a handful of SDSS survey fields in an attempt to match observational biases in real images (realistic skies, resolution, and crowding by nearby sources). Indeed, Huertas-Company et al. (2019) used a similar but more rigorous approach with CNNs trained on the Nair & Abraham (2010) SDSS visual classification sample to perform Hubble-type classifications of synthetic images of galaxies from the IllustrisTNG-100 simulations (Nelson et al. 2018; Pillepich et al. 2018; Rodriguez-Gomez et al. 2019). Huertas-Company et al. (2019) found that injecting the TNG images into real fields following the statistical observational realism approach of Bottrell et al. (2017a) was crucial to obtaining consistent classification uncertainties when testing on SDSS and TNG images.

These previous studies touch upon core unanswered questions for training deep neural networks based on hydrodynamical simulations (and that are particularly relevant for characterizing merger stage). Namely, what kind of synthetic images should be used when training using simulations? How realistic do the images have to be in order to achieve high performance in identifying and characterizing mergers by stage in real images? Does a network that is trained on images which include contaminating effects [such as realistic skies, resolution degradation, and additional sources in the image field of view (FOV)] perform better when handling new data which also contain these contaminants? In other words, what is gained by making synthetic images more realistic? Dust-inclusive radiative transfer can be used to generate photorealistic images (e.g. Jonsson 2006; Jonsson, Groves & Cox 2010; Baes et al. 2011; Camps & Baes 2015) but it is costly from computational and data storage perspectives. Is radiative transfer essential to merger classifications or can it be replaced with simpler images? These questions come at an important time when the state-of-the-art cosmological hydrodynamical simulations produce realistic and statistically representative populations of galaxies (e.g. IllustrisTNG; Nelson et al. 2018; Pillepich et al. 2018). Crucially, mergers identified from these simulations’ merger trees cover a range of mass ratios, orbital parameters, and initial galaxy properties that are comparable to the real Universe. Consequently, synthetic images generated along each merger sequence can be used to generate and calibrate deep network models to identify and characterize mergers in current and next-generation observational imaging surveys.

The goals of this paper are to: (1) provide the methodology with which CNNs, trained and calibrated using hydrodynamical simulations, can be used to identify mergers and predict merger stage in realistic images and (2) assess the importance of realism in the synthetic training images.¹ To realize these goals, we construct synthetic images with various levels of observational realism from a set of binary hydrodynamical merger simulations run with the FIRE-2 model (Moreno et al. 2019). Specifically, we generate images in two branches, starting with: (a) 2D projections of the stellar particles and (b) photometry from dust-inclusive SKIRT radiative transfer. In each branch, images are constructed with three levels of realism: (i) no observational effects (idealized), (ii) realistic skies and point spread functions (PSFs) (semirealistic), and (iii) statistical insertion into real survey images (fully realistic). These levels are designed to expose the roles of particular ingredients of realism in classification performance. Each image is assigned a

¹It should be noted that, based on the results of Huertas-Company et al. (2019), who use our methods for Hubble type classifications, the applications of our methods and results are not restricted to mergers.

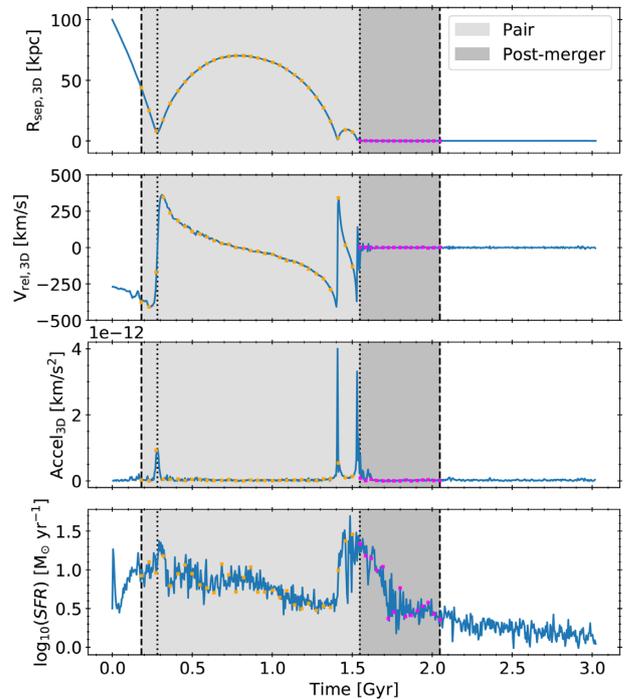


Figure 1. Radial separation, relative velocity, absolute acceleration, and total SFR sequences for the G2G3 ‘e’ orbit 1 merger (the fiducial run from Moreno et al. 2019) showing class definitions and snapshot selection. From left to right in each panel, the first thick dashed and dotted lines correspond to 100 Myr before first pericentric passage and the moment of first pericentric passage, respectively. The second dotted line corresponds to coalescence – which we take to be the last time the central black holes of each galaxy are more than 500 pc apart. The second thick dashed line corresponds to 500 Myr after coalescence. Shading between lines corresponds to the pair (light grey) and post-merger (darker grey) classes. In each panel, we show the snapshot selection for the pair (orange) and post-merger (magenta) classes. The SFRs are measured from the full simulation volume and so include contributions from both galaxies when they are separate.

target classification (isolated, pair, or post-merger) corresponding to the definitions described in Section 2.1.3 and illustrated in Fig. 1. Given that all training, validation, and test data are drawn from the same set of isolated/merger simulation runs, the training data are (by construction) highly generalizable to the test data in terms of the range of galaxy/merger properties covered. This experimental design allows us to isolate the role of realism in the performances of the networks.

This paper is laid out as follows. The simulations, construction of the synthetic images, and neural network architecture are described in Section 2. Our experiments and their results are presented in Section 3. Our results are discussed in Section 4 and summarized in Section 5. We adopt a cosmology in which ($H_0 = 70 \text{ km s}^{-1} \text{ Mpc}^{-1}$, $\Omega_m = 0.3$, $\Omega_\Lambda = 0.7$). Additionally, the Bottrell et al. (2017a) observational realism suite, REALSIM, is released publicly as a companion to this paper (see Section 2.2.5).

2 METHODS

In this section, we describe the merger simulations (Sections 2.1.1 and 2.1.2), merger stage definitions and snapshot selection (Section 2.1.3), creation of the synthetic images (Section 2.2), and CNN architecture (Section 2.4).

Table 1. Initial properties of the four galaxies in the Moreno et al. (2019) merger suite. The columns are the galaxy ID, total stellar mass, halo mass, gas fraction, stellar bulge-to-total mass fraction, gas disc scale length, and stellar disc scale length.

Galaxy ID	$M_*/10^{10} M_\odot$	M_*/M_{halo}	f_{gas}	$(B/T)_*$	$R_{\text{d, gas}}/\text{kpc}$	$R_{\text{d, stars}}/\text{kpc}$
G1	0.206	0.0157	0.681	0.0185	4.73	1.42
G2	1.24	0.0361	0.392	0.0497	6.04	1.92
G3	2.97	0.0383	0.264	0.0773	5.32	1.61
G4	5.50	0.0228	0.192	0.103	5.26	1.57

2.1 Simulations

We use the suite of galaxy interaction simulations from Moreno et al. (2019) in this study. The suite is similar to a previous merger suite from those authors (Patton et al. 2013; Moreno et al. 2015) but with much higher resolution and a new physical model and hydrodynamic solver. We describe the salient features of the suite here but refer the reader to Moreno et al. (2019) and Hopkins et al. (2018) for full details of the suite and model, respectively.² We discuss the limitations with respect to the scope of the merger suite in detail in Section 4.2. Briefly, we emphasize that the suite does not offer sufficiently representative statistics and diversity in galaxy/merger properties to train networks that will be useful in applications to a real population of galaxies. Therefore, we do not apply our trained networks to real galaxies. However, for the objectives highlighted at the end of Section 1, the suite is appropriate.

2.1.1 FIRE-2 model

The simulations were run using the FIRE-2 physics model (Hopkins et al. 2018) and the ‘meshless finite-mass’ (MFM) hydrodynamics solver, GIZMO (Hopkins 2015, 2017).³ The model includes treatment of radiative cooling and heating from free–free, photoionization and recombination, Compton, photoelectric, dust-collisional, cosmic ray, molecular, metal-line and fine-structure processes. It accounts for the UV background (Faucher-Giguère et al. 2009) and locally driven heating and self-shielding. Gas that is locally self-gravitating, self-shielding, Jeans unstable, and sufficiently dense (defined by critical gas density, $n_{\text{crit}} = 1000 \text{ cm}^{-3}$) can form stars stochastically in a sink-particle approach (see appendix C of Hopkins et al. 2018). A stellar particle is treated as a single stellar population with a known age, $t_* = t - t_{\text{form}}$, and a metallicity and mass that are inherited from its progenitor gas particle. Masses, ages, metallicities, luminosities, energies, mass-loss rates, and stellar feedback event rates are tabulated (without tuning) using the STARBURST99 stellar population synthesis model (Leitherer et al. 1999) assuming a Kroupa (2001) initial mass function (IMF). Stellar feedback includes (i) mass, metal, energy, and momentum injection from supernova type Ia & II; (ii) continuous stellar mass-loss through OB/AGB winds; (iii) photoionization and photoelectric heating; and (iv) radiation pressure. The model does not account for feedback generated via accretion of gas on to supermassive black holes (SMBHs). SMBH feedback is omitted because coupling between an AGN and the circumnuclear interstellar medium (ISM) is not yet well understood (though see Torrey et al. 2017, for an examination of the stability of feedback regulated star formation

in galactic nuclei). The MFM dark matter, gas, and stellar particle masses are $(m_{\text{dm}}, m_{\text{gas}}, m_{\text{star}}) = (19, 1.4, 0.84) \times 10^4 M_\odot$. The highest gas density and spatial resolution are $5.8 \times 10^5 \text{ cm}^{-3}$ and 1.1 pc, respectively. The typical snapshot resolution is 5 Myr. The gravitational softening lengths are 10 pc for dark matter and stellar components and 1 pc for the gaseous component.

2.1.2 Merger suite

Moreno et al. (2019) used FIRE-2 physics to generate a suite of non-cosmological binary galaxy interaction simulations covering a range of orbital parameters and mass ratios between four disc galaxies (G1, G2, G3, and G4, in order of increasing total stellar mass). The suite is complemented by secular runs (the controls used by Moreno et al. 2019) in which each galaxy is allowed to evolve in isolation. Individual galaxies are set up following the procedure described in Springel, Di Matteo & Hernquist (2005) using the analytic framework provided by Mo, Mao & White (1998). Stellar bulges and dark matter haloes are initialized analytically with Hernquist (1990) profiles. Halo masses are adopted for a given stellar mass following the abundance matching results from Moster, Naab & White (2013). Stellar bulge-to-total fractions are assigned on the basis of median trends with total stellar mass using the Mendel et al. (2014) estimates of bulge, disc and total stellar masses for galaxies in the SDSS. Similarly, gas fractions are assigned based on mean atomic and molecular gas mass fractions estimates along the main sequence (MS) of star-forming galaxies in the SFR– M_* plane from Saintonge et al. (2016). The properties of each galaxy are shown in Table 1.

The suite is divided into two components: (1) an orbit suite and (2) a mass ratio suite. The orbit suite comprises several interaction scenarios between the G2 and G3 galaxies (see fig. 3 of Moreno et al. 2019). It covers three unique spin-orbit orientations corresponding to the ‘e’, ‘f’, and ‘k’ orbits from Robertson et al. (2006) (see fig. 1 of Moreno et al. 2015), three impact parameters, and three impact velocities (see fig. 3 of Moreno et al. 2019). The ‘e’, ‘f’, and ‘k’ spin-orbit orientations correspond to approximately prograde, polar, and retrograde orbits, respectively. Permutation of the orbital parameters gives a total of $3 \times 3 \times 3 = 27$ unique mergers at a fixed mass ratio of $\mu \sim 2.5:1$. The mass suite adopts a single orbit (the fiducial ‘e’ orbit in Moreno et al. 2019) in which each of the four galaxies interacts with every other galaxy and itself for a total of $4 \times 4 = 16$ interactions. The range in stellar mass ratios covered by the suite is 1:1 to 1:16. Combined, the orbit and mass components of the suite cover a broad range of encounter strengths and merging time-scales.

2.1.3 Class definitions, merger selection, and snapshot sampling

For each merger simulation, we select snapshots from which to construct images based on a set of simple definitions of the pair and post-merger phases. For example, Fig. 1 shows the radial separation,

²Videos of the Moreno et al. (2019) galaxy merger simulations are available at research.pomona.edu/galaxymergers.

³For more information on the FIRE Project and FIRE-2, visit <https://fire.northwestern.edu>.

relative velocity, absolute acceleration, and total SFR sequences for the fiducial merger simulation from Moreno et al. (2019) measured from central SMBHs of each galaxy. The pair phase is defined to begin 100 Myr before first pericentric passage, $t_p - 100$ Myr, and to end just before coalescence, t_c – which, as in Moreno et al. (2019), we define as the last time the central black holes are more than 500 pc apart (light grey shading). We then define the post-merger phase as any time in $[t_c, t_c + 500$ Myr] (darker grey shading).

We use a temporal definition for the beginning of the pair phase primarily for convenience with the simulations. But additionally, a temporal definition may circumvent biases that can arise from selection based on projected separation or relative velocity. For example, a 300 km s^{-1} relative velocity cut (Ellison et al. 2008; Patton et al. 2013) may have missed snapshots around first and second pericentre for the merger in Fig. 1 depending on line of sight.

Likewise, our definition of the post-merger phase is motivated by simplicity and the availability of temporal information in the simulations. Observability time-scales for post-merger features (shells, streams, etc.), as for the mergers overall, are sensitive to surface brightnesses, gas fractions, mass ratios, and initial orbital parameters (e.g. Lotz et al. 2008, 2010a,b; Ji et al. 2014; Nevin et al. 2019). Consequently, we select a clear-cut post-merger phase in which post-merger features are still expected to be prominent. Several mergers in the suite are either fly-bys or do not evolve to 500 Myr after coalescence. In order to preserve several mergers that coalesce but do not have 500 Myr of snapshot coverage after coalescence, we set a minimum post-coalescence criterion of 250 Myr. Consequently, the post-merger stage is defined as starting at coalescence and ending at $\max(t_{\text{last}}, t_c + 500 \text{ Myr})$, where $t_{\text{last}} \geq t_c + 250$ Myr. The criteria that (a) the galaxy must coalesce and (b) the simulation has run for at least 250 Myr post-coalescence reduce the sample from 43 to 23 mergers. In particular, the interactions with the largest mass ratios and widest impact parameters are rejected on the basis of these criteria. We discuss the consequences of our class definitions on network performance in Section 4.4.

For each merger simulation, 30 (15) snapshots are selected, which uniformly sample the pair (post-merger) phase as shown with the orange (magenta) squares in Fig. 1. This approach provides a sampling cadence that is sparse enough that imaging in neighbouring snapshots are not overly correlated (see Appendix B) yet fine enough that a large number of images for each merger can be generated. 10 snapshots are selected from the isolated runs with uniform sampling cadence as with the pairs and post-mergers. Due to the significantly smaller number of snapshots corresponding to isolated galaxy runs, we bolster the isolated galaxy data set by increasing the number of orientations in which their synthetic images are generated (as described below).

2.2 Synthetic observations

Synthetic images are made for isolated, pair, and post-merger snapshots along four lines of sight corresponding to the arms of a tetrahedron whose vertex is coincident with the point of minimum potential. Consequently, images in the pair phase are always centred on the more massive galaxy.⁴ Producing only four camera angles for the isolated galaxies would make the training data highly imbalanced – with an order of magnitude more images

with pair and post-merger classes than the isolated class.⁵ As a first step in balancing the data, we increase the number of camera angle orientations for snapshots from the isolated runs by 11 inclinations and 11 position angles. Before any augmentation (see Section 2.3), there are $23 \times 30 \times 4 = 2760$ images with the pair class, $23 \times 15 \times 4 = 1380$ with the post-merger class, and $4 \times 10 \times (11 \times 11 + 4) = 5000$ with the isolated class.

Synthetic images are generated for each snapshot/orientation with various levels of realism. There are two distinct image types: (1) images originating from two-dimensional projections of stellar particles (stellar maps, SM) and (2) from photometry generated using dust-inclusive radiative transfer (PH). We produce images with three different levels of realism: (1) noiseless with high resolution; (2) include realistic (but analytically generated) noise and resolution degradation; and (3) are inserted into real SDSS survey fields that may contain additional sources. We refer to these increasing levels of realism as ‘idealized’, ‘semireal’, and ‘full real’, and they allow us to examine the importance of observational biases that are introduced level by level. The image types are described in detail in the sections that follow and are summarized in Table 2.

2.2.1 StellarMap (SM)

The zeroth-order stellar image that can be produced from a hydrodynamical simulation is a two-dimensional projection of the stellar particles along a given line of sight. This *idealized* image type has several important features: noiseless without resolution degradation; insensitivity to variations in mass-to-light ratio (M/L) from different stellar populations or dust absorption; and low computational and data management overhead. Initially, we adopt a fixed 50 kpc FOV for each image with spacial resolution of $0.097 \text{ kpc pixel}^{-1}$ (512×512 pixels).⁶ All images (including other types) are mock-observed with the SDSS camera ($0.396 \text{ arcsec pixel}^{-1}$) at a fixed redshift of $z = 0.046$ (the median redshift of galaxies in the DR14 MaNGA galaxy sample; Bundy et al. 2015) where the scale is approximately $0.9 \text{ kpc arcsec}^{-1}$. Consequently, the SM images are rebinned to a physical scale of $0.36 \text{ kpc pixel}^{-1}$ (139×139 pixels or $56 \times 56 \text{ arcsec}^2$). This still offers high resolution – particularly with respect to images that are further degraded by realistic (or real) SDSS PSFs.

2.2.2 Photometry (PH)

We generate idealized SDSS *gri* photometric images using the Monte Carlo dust radiative transfer code, SKIRT (Baes et al. 2011;

⁵Class imbalance – where the occurrence of one class in a data set significantly outnumbers other classes – is a common problem in applications of deep learning to classification tasks including medical diagnosis (Mac Namee et al. 2002; Grzymala-Busse et al. 2004), fraud detection (Chan & Stolfo 1998), and others (Cardie & Howe 1997; Radiwojac et al. 2004; Haixiang et al. 2017). An example is the natural imbalance in medical data between images of a particular diagnostic class (e.g. contains tumour), which might be 1000 times less frequent than images of another class (e.g. healthy). Unbalanced data can have significant detrimental effects on deep classifiers such as CNN, and a recent systematic study has investigated various methods that address class imbalance (Buda, Maki & Mazurowski 2017). Indeed, the method that appears to best address class imbalance is oversampling of data from the underrepresented class (e.g. through augmentation) – which is the method we adopt in this paper.

⁶The higher the initial resolution, the greater the range of instrumental angular scales (arcsec pixel^{-1}) and redshifts (kpc arcsec^{-1}) that can be explored.

⁴For our purposes, this bias towards the more massive galaxy is not of any consequence. But for a data set that must be large and general enough to handle real data, separate images should be constructed that are centred on both the primary and secondary – ideally in along different lines of sight.

Table 2. Reference summary of image types used for training and testing of networks. Intensities in the STELLARMAP images are scaled to match the total surface brightnesses in the PHOTOMETRY *i*-band images before adding realism effects.

Image Type	Shortform	Radiative transfer	Bands	Gaussian sky	Gaussian PSF	Real sky	Real PSF
STELLARMAP	SM	no	<i>i</i> *	no	no	no	no
STELLARMAP SEMIREAL	SMSR	no	<i>i</i> *	yes	yes	no	no
STELLARMAP FULLREAL	SMFR	no	<i>i</i> *	no	no	yes	yes
PHOTOMETRY	PH	yes	<i>gri</i>	no	no	no	no
PHOTOMETRY SEMIREAL	PHSR	yes	<i>gri</i>	yes	yes	no	no
PHOTOMETRY FULLREAL	PHFR	yes	<i>gri</i>	no	no	yes	yes

Table 3. SDSS sky and angular resolution measurements used to generate the background noise levels and convolution kernels for SemiReal images. Table quantities are computed from the ensemble of SDSS Field table values corresponding to the *full* Simard et al. (2011) galaxy sample. Columns: (1) SDSS bandpass; (2) mean sky noise [AB mag arcsec⁻²]; (3) standard deviation in sky noise values [AB mag arcsec⁻²]; (4) mean PSF FWHM [arcsec]; (4) standard deviation in PSF FWHM values [arcsec]. Individual SemiReal sky noise values and PSFs are drawn from normal distributions formed from these quantities.

SDSS band	$\langle \sigma_{\text{sky, Field}} \rangle$ [AB mag arcsec ⁻²]	stdev($\sigma_{\text{sky, Field}}$) [AB mag arcsec ⁻²]	$\langle \text{FWHM}_{\text{PSF}} \rangle$ [arcsec]	stdev(FWHM_{PSF}) [arcsec]
<i>u</i>	23.87	0.15	1.55	0.24
<i>g</i>	24.88	0.14	1.47	0.22
<i>r</i>	24.38	0.11	1.36	0.22
<i>i</i>	23.82	0.12	1.29	0.22
<i>z</i>	22.36	0.19	1.31	0.20

Camps & Baes 2015). SKIRT predicts the light contribution from stellar particles and star-forming regions while modelling the effects of dust on the absorption, scattering, and re-emission of stellar light (note that we ignore radiation from the central engine). We model the stellar light from old stars (older than 10 Myr) using a Kroupa (2001) IMF and the associated STARBURST99 single-age spectral energy distributions (SEDs) (Leitherer et al. 1999). Emission from star-forming regions (stellar particles younger than 10 Myr) is represented by MAPPINGS-III SEDs (Groves et al. 2008), which include contributions from young stars and H II regions. The dust contribution is modelled assuming that the dust distribution traces the metal distribution where 30 per cent of the metals are locked in dust particles. We adopt the multicomponent dust mix of Zubko, Dwek & Arendt (2004), which includes graphite grains, silicate grains, and polycyclic aromatic hydrocarbons (PAHs). We ignore dust re-emission (and the associated self-absorption) that has a negligible contribution in the wavelength regime studied in this work. The underlying radiation field is discretized by SKIRT using 10⁵ photon packages per wavelength. SKIRT’s output (spectral data cubes) are converged for >10⁵ photon packages per wavelength.

We adopt the same initial FOV and spatial resolution as for the SM images in construction of the SKIRT data cubes. Since we are generating broad-band photometry, a relatively coarse spectral resolution is adopted with 241 spectral elements that linearly and uniformly sample the rest-frame optical spectrum from the near-UV to near-infrared (250–850 Å). These data cubes are redshifted to $z = 0.046$ and convolved with the SDSS *gri* response functions to produce idealized photometry – accounting for stretch in the spectrum and $(1 + z)^{-5}$ reduction in specific intensities in each spectral element.⁷ As with the SM images, the PH images are rebinned to the SDSS camera pixel scale. The PH images are:

noiseless with high resolution; light weighted and sensitive to variations in *M/L* for different stellar populations and dust; and very expensive from a computational and data management perspective when compared to SM (see Section 3.1.2). Furthermore, training networks with all three *gri* bands as input allows networks to develop sensitivity to colour.

2.2.3 StellarMap SemiReal (SMSR)

Ground-based imaging surveys are affected by sky surface-brightness limitations and blurring from the atmospheric PSF. These biases can be emulated using the statistics of sky brightnesses and PSF sizes measured in SDSS fields. Crucially, we match the *statistics* of sky noise levels and PSF resolution to the field properties for 1.12 million galaxies in the SDSS Legacy images (Abazajian et al. 2009) using the Simard et al. (2011) quantitative morphology catalogue and ancillary data measured by the PHOTO pipeline (Lupton et al. 2001, 2002, 2012). We compute the means and standard deviations in the resulting sky noise and PSF resolution distribution functions. The results are tabulated in Table 3. We use these results to generate analytic Gaussian profiles from which sky noise and PSF resolution levels are sampled independently for each synthetic image.

The idealized SM images are not light weighted and therefore do not offer straightforward conversion to calibrated AB flux units. To approximate the intensities of the stellar maps in realistic images, we scale each normalized stellar map by the total intensity in its corresponding idealized *i*-band photometry image. We choose to scale by the *i*-band light because it is less sensitive to variations in *M/L* from young stellar populations or starbursts compared to *g* or *r*. With the idealized SM images effectively ‘light weighted’, we sample the distribution function for the PSF and convolve. Before adding sky noise, we use the average SDSS photometric zero-point magnitude, airmass, extinction, and gain over all SDSS fields to convert the PSF-convolved images to electron counts from which source Poisson noise can be added. We then convert back to calibrated flux units, sample our sky noise distribution, and add Gaussian sky noise to the image.

⁷We provide a code that performs all of these tasks (for a specified redshift) from the default rest-frame specific intensity data cubes from SKIRT ($\text{W m}^{-2} \mu\text{m}^{-1} \text{arcsec}^{-2}$). The code produces output photometry in convenient AB mag arcsec⁻² units for each filter (Oke & Gunn 1983) and can be found at the following url: https://github.com/cbottrell/RealSim/blob/master/SpecToSDSS_gri.py.

2.2.4 *Photometry SemiReal* (P_{HSR})

The procedure for creating PHSR images in each band is the same as for creating SMSR images but without the normalization and ‘light-weighting’ step. One feature of the current SEMIREAL procedure that can be remedied in the future is that the sky noise and PSF estimates are drawn independently in each band and so are not correlated as they should be. However, our results do not give us reason to suspect that this is a significant limitation of our methods.

2.2.5 *StellarMap FullReal* ($SMFR$)

Synthetic images with extensive observational realism are generated following the methods presented in Bottrell et al. (2017a) and Bottrell et al. (2017b). Similar to SEMIREAL, the FULLREAL procedure is designed to incorporate *statistical* observational realism into the synthetic images so that real survey field statistics are matched between the simulated and observed galaxies. The main difference from the SEMIREAL procedure is that the synthetic images are added quasi-randomly to real survey fields in the FULLREAL procedure. In this approach, the insertion statistics are guided by a basis catalogue of real galaxies (Simard et al. 2011). As such, the statistics of sky brightness, PSF resolution and crowding by nearby sources for real galaxies are reproduced in the synthetic images (along with any other field-dependent characteristics). The FULLREAL procedure is described in detail in Bottrell et al. (2017a). We provide a summary here of the procedure that is followed for every synthetic image to complement the public release of the realism suite:⁸

(i) A galaxy is randomly selected from the Simard et al. (2011) basis catalogue. The SDSS *gri*-band fields in which that galaxy resides are extracted and converted to calibrated flux units using ancillary data queried from the SDSS Data Archive Server. A source mask is generated for the *r*-band field using SEXTRACTOR (Bertin & Arnouts 1996) and deblending parameters optimized for SDSS in Simard et al. (2011) (specifically for the Patton et al. 2011 pair sample). A common injection site for each band (where applicable) is selected randomly with the restriction that the *centre* of the injected image cannot land on another object in the source mask.

(ii) The PSFs for each band corresponding to the injection site are reconstructed using the PSFIELD files and the standalone READ.PSF software. Each band of the synthetic image (in the SMFR case, the single ‘light-weighted’ idealized stellar map) is converted to electrons using the ancillary data from Step (i) and convolved with the local SDSS PSF for that band. Source Poisson noise is then added.

(iii) A PSF-convolved and Poisson noise-added synthetic image (in each desired band) is finally converted back to calibrated flux units and inserted into the SDSS field at the injection site selected in Step (i). A cut-out corresponding to the desired FOV (in our case, 50 kpc or approximately 56 arcsec at $z = 0.046$) is extracted

⁸A public version of the REALSIM observational realism suite is available at the following url: <https://github.com/cbottrell/RealSim> for Python 3. It includes the Simard et al. (2011) bulge + disc decomposition catalogue from which it draws galaxy and field statistics; a Python 3 version of the SDSS `sq1c1.py` code that queries field information directly from the SDSS Data Archive Server; *ugriz* filter response functions from Doi et al. (2010); the Simard et al. (2011) SEXTRACTOR configuration files required for deblending of images when inserting into real SDSS fields; a Python notebook of example executions; a code for converting SKIRT output to SDSS images; and a sample of input images.

around this location. This cut-out now includes real sky, real PSF degradation, and real additional sources in the FOV that track the statistics for galaxy image properties in the basis catalogue.

Some particularities of this procedure for generating SMFR images are that (1) images are only generated in the *i* band and (2) the *r*-band image is still used to generate the source mask as described in Step (i).

2.2.6 *Photometry FullReal* ($PHFR$)

Construction of PHFR images follows the same procedure as for the SMFR but for each of the SDSS *gri* bands. Because the PHFR images incorporate light weighting from radiative transfer and the full rigour of statistical observational realism, the PHFR data set is our benchmark for how a given network would be expected to perform on realistic data and is often discussed as such in Sections 3 and 4 (i.e. the PHFR images are the closest representation of observable galaxies in our suite and are hence used as the ultimate training set for likely ‘real-life’ performance). As with the PH and PHSR data sets, the PHFR has three channels of input corresponding to the three bands in which we produce photometry.

Fig. 2 shows a recent post-merger for each of the six image types and demonstrates the impact of each level of realism. The upper panels show images originating from stellar maps and lower panels show images originating from radiative transfer. In the idealized SM and PH images (left-hand column), morphological features post-merger are visually prominent including shells, streams, and tidal tails that have not yet decayed from the pair phase. In the PH image, a dust lane obscures light emanating from the nucleus – giving it an asymmetric appearance with respect to the SM image. Additionally, the PH image has bright knots associated with the low *M/L* of young stellar populations, whereas the SM image is insensitive to these features. The SEMIREAL images in the middle panels show the results of adding realistic SDSS noise and resolution degradation to the images. Many of the features that made this object easily identifiable as a post-merger in the idealized images are drowned by the sky noise and PSF blurring. Features of the post-merger remain in the SEMIREAL images but are more subtle than in idealized images. The right-hand panels show FULLREAL images for the post-merger. In addition to real skies and degradation by real PSFs, these images incorporate contamination by nearby sources. The lower right-hand panel shows particularly striking chance projection with an interloping disc galaxy. Taken together, these images nicely encapsulate the rationale of this work.

2.3 Image normalization and augmentation

Normalizations and augmentation (oversampling) are applied to images in each data base. We generate augmented images by applying zoom, rotation, and small translational transformations to the set of original images in order to (1) reduce class imbalance and (2) achieve rotational invariance in the models. The augmentations are performed with using the `IMAGEDATAGENERATOR` class from the KERAS Python API (Chollet et al. 2015). All images of a particular class are augmented N times until the total number of images exceeds 10 000. Consequently, the final numbers of images in each class are $(N_{ISO}, N_{Pair}, N_{Post}) = (10\,000, 11\,040, 11\,040)$ after augmentation for each image type.

The augmented images (starting in linear intensities) are normalized following a standardized algorithm that is applied to all image types:

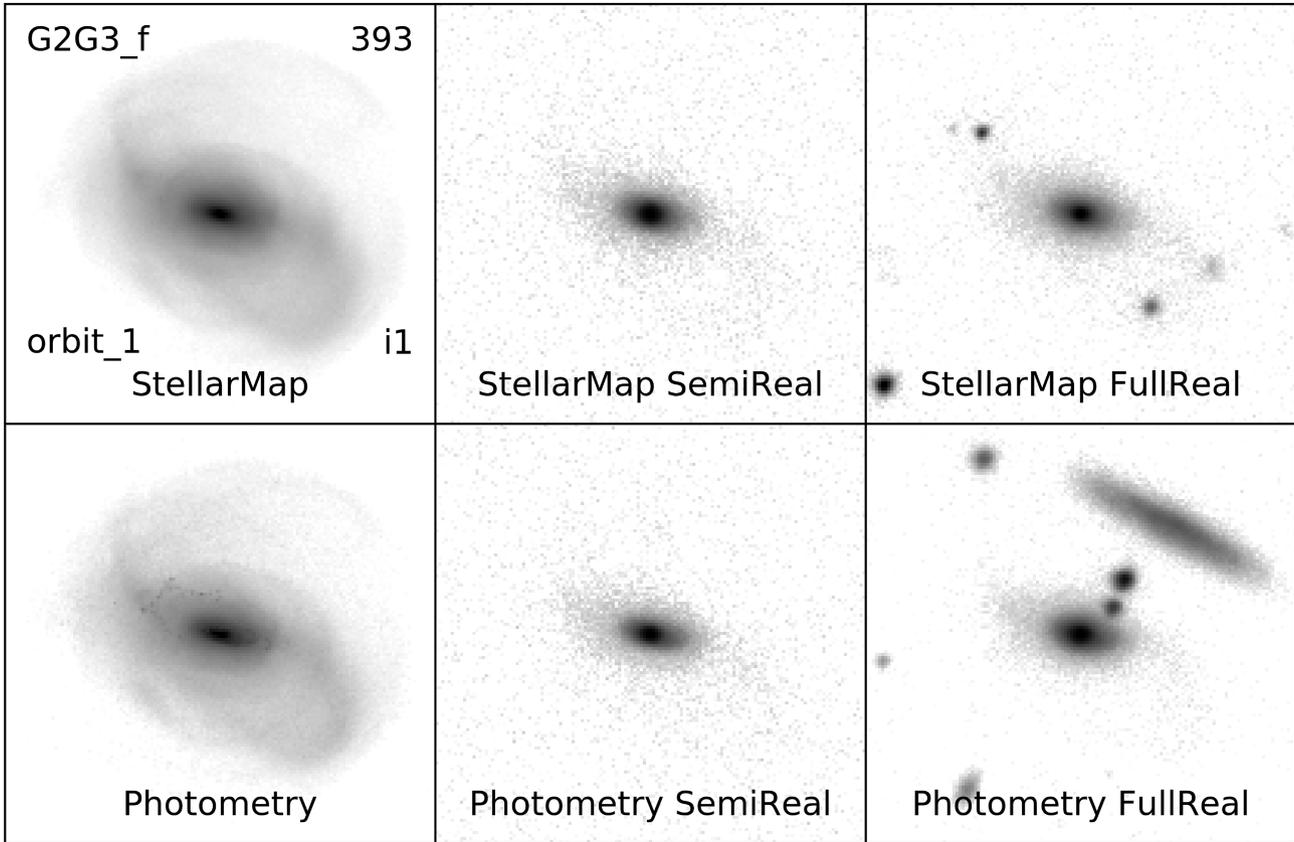


Figure 2. Visualization of a single post-merger galaxy realized with every image type. The post-merger image taken from orbit 1 of the G2G3 ‘f’ orbit suite at snapshot 393 (post-coalescence). All images show i -band intensities. Note the potential for misclassification in the PHFR realization of this post-merger due to a chance projection with a field galaxy.

(i) If the image contains sky noise (SEMIREAL and FULLREAL types), then the image is subtracted by its median intensity.

(ii) Take the logarithm of the sky-subtracted image. All values less than -7 are converted to NaNs.

(iii) The median of the full image, a_{\min} , and 99th percentile for the central 20×20 pixels, a_{\max} , are computed.

(iv) All values below a_{\min} (and NaNs) are set to a_{\min} and values greater than a_{\max} are set to a_{\max} .

(v) The clipped logarithmic image is subtracted by a_{\min} and normalized by $a_{\max} - a_{\min}$.

The results are images with logarithmic-scale intensities in the range of 0–1 in which each image is scaled to maximize contrast for the central target in the image. This normalization procedure avoids the problem of reduced contrast in FULLREAL images that contain bright stars or other contaminating effects (such as would be produced by a more conventional standard scalar).

2.4 Neural network architecture

We use our synthetic images to train CNNs that classify galaxies as isolated, pairs, or post-mergers. CNNs are a class of deep learning model that are particularly useful for data that exhibit topological structure such as images (Fukushima 1980; LeCun et al. 1989; LeCun, Bengio et al. 1995; LeCun et al. 1998; Krizhevsky, Sutskever & Hinton 2012; Lecun, Bengio & Hinton 2015). There is enormous flexibility in CNN architectures in terms of depth (number

of layers), layer properties (kernel sizes, etc.), and layer structures (e.g. residual blocks; He et al. 2015). Given the successes of previous works using a particular (and relatively simple) CNN architecture for galaxy morphology classifications (Dieleman, Willett & Dambre 2015; Huertas-Company et al. 2015; Domínguez Sánchez et al. 2018, 2019; Huertas-Company et al. 2019), we adopt a similar (but not identical) CNN architecture for predicting galaxy merger stage. This architecture is summarized in Table 4. The output of the network for a given image is a class probability distribution function, $(P_{\text{Iso}}, P_{\text{Pair}}, P_{\text{Post}})$. For our analyses and comparison with the known target classes, we adopt the class with the highest probability density.

A (70, 15, 15) per cent split is used for *training*, *validation*, and *test* images, respectively. The networks are optimized on the training images and corresponding known target classes. The overall performance of a network on the validation images, $N(\text{correct})/N_{\text{tot}}$, is evaluated after each training epoch. While this step does not, strictly speaking, affect optimization, it is used to determine an appropriate time to stop training and consequently prevent overfitting to the training images. In contrast, networks are never exposed to test images during training. For tests in which networks trained on a particular image type (e.g. PHFR) are tested on images of a different type (e.g. SMSR – which may *all* technically be considered distinct data), we find that there is no difference in network performance whether we test only on the corresponding test images or all images (including training and validation images). Our results show the latter for such tests.

Table 4. CNN architecture for the full-colour model that accepts three channels of input comprising *gri* images. Convolution kernel sizes (Conv2D), max-pooling windows (MaxPool), dropout rates (Dropout), output shapes, and the total number of trainable parameters for each of the layers are indicated. The convolution layers use Rectified Linear Unit (ReLU) non-linear activation functions and have a (1,1) stride. The output of the 4th convolution layer is flattened to a one-dimensional feature array and passed to the fully connected (dense) component. Dense layers also use ReLU activation functions. The output layer uses a softmax activation function.

Layer (type)	Output shape	# Parameters
Input Layer	(139,139,3)	0
Conv2D-1 (6 × 6)	(139,139,32)	3488
MaxPool-1 (2 × 2)	(69,69,32)	0
Dropout-1 (0.25)	(69,69,32)	0
Conv2D-2 (5 × 5)	(69,69,64)	51 264
MaxPool-2 (2 × 2)	(34,34,64)	0
Dropout-2 (0.25)	(34,34,64)	0
Conv2D-3 (2 × 2)	(34,34,128)	32 896
MaxPool-3 (2 × 2)	(17,17,128)	0
Dropout-3 (0.25)	(17,17,17)	0
Conv2D-4 (3 × 3)	(17,17,128)	147 584
Dropout-4 (0.25)	(17,17,17)	0
Flatten	36 992	0
Dense-1	(512)	18 940 416
DropFC-1 (0.25)	(512)	0
Dense-2	(128)	65 664
DropFC-2 (0.25)	(128)	0
Output Layer	(3)	387
Total # parameters	–	19 241 699

3 EXPERIMENTS

Each synthetic image data set is used to train neural networks that are then applied to test images of every type. This ‘handshake’ of training/testing experiments includes cases where training and test data are of the same type. We generate 10 networks for each image type by splitting the data into training, validation, and test images using 10 unique random states. This ‘bootstrapping’ of the data allows us to characterize the random error associated with the selection of a particular training set and to statistically merge our test results.

3.1 Model and image handshake

Our main experiment comprises a handshake between networks trained using the six unique image types (see Table 2) that are each tested on the six image types – resulting in 6×6 tests. Fig. 3 shows a qualitative schematic of this experiment. The results for a single test can be characterized using a confusion matrix (e.g. Fig. 4, description in Section 3.1.1). The full 6×6 matrix of confusion matrices can be found in Appendix C where the reader can zoom in on every individual test. In this section, we focus on the matrices corresponding to specific tests from the main handshake, which address our core questions.

3.1.1 Case study: training and testing with ideal PHOTOMETRY

Fig. 4 shows a single confusion matrix corresponding to the models trained on PH and tested on PH (as described in Section 2.2.2 and

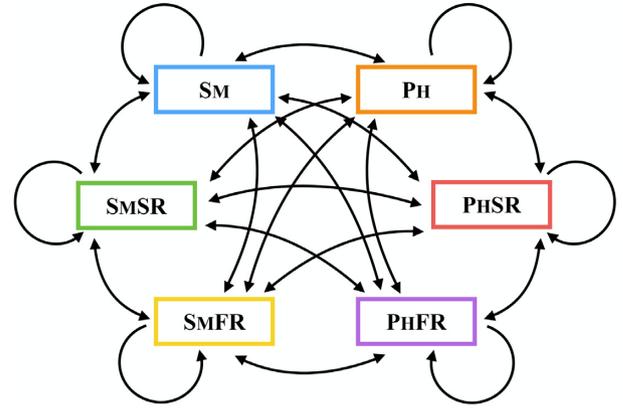


Figure 3. Schematic of the main handshake experiment. Networks are trained with images of each type. Each trained network is tested to images of every other type. Additionally, each network is tested on images of the same type but that the networks never see during training (outer looping arrows).

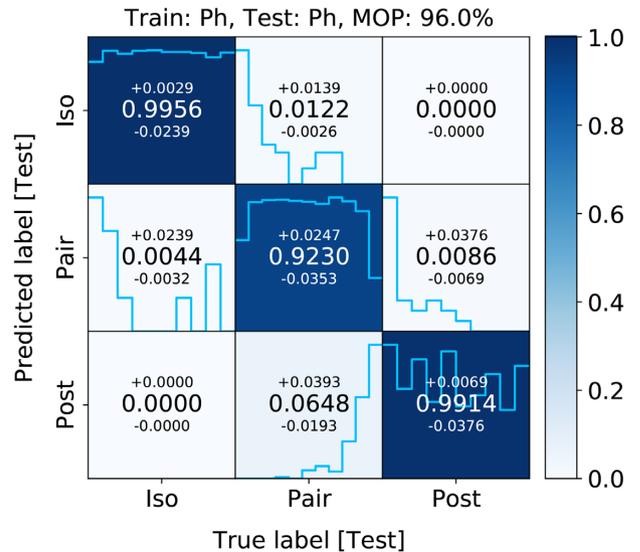


Figure 4. Confusion matrix for merger-stage classifications using the PH networks and corresponding test images. The number in each matrix element quantifies the median predicted fraction of images and 16th and 84th percentile offsets computed by bootstrapping the results of 10 unique realizations of training, validation, and test data. The inset in each matrix element is a light blue bar plot showing the median empirical distribution function of relative merger class time-scales (see equation 1) for the images that fall in that matrix element in each of the 10 bootstraps. These timing histograms in each element are normalized by their respective maximum values for visibility in cases where the total number of images in a particular matrix element is small. Networks trained on PH training data have an MOP of 96.0 per cent when applied to PH test data.

summarized in row 4 of Table 2). We use this as a simple case study to orient the reader to the general features of our analysis for subsequent experiments. Each column of the confusion matrix shows the normalized distribution of predicted labels (y-axis) for all test images with a particular truth label (x-axis). A perfect classification network would therefore produce an identity matrix where all of the power is in the diagonal elements. Off-diagonal elements correspond to misclassifications denoted by a combination of predicted and true labels. Values shown in each matrix element

are the median and 16th and 84th percentile range computed from the 10 bootstrap realizations of training, validation, and test data.

Fig. 4 demonstrates that exceptional performance is achievable under ideal but generally unrealistic conditions: (1) noiseless images; (2) high spatial resolution; (3) no contamination by projection effects or other objects in the FOV; and lastly (4) a training set that is (by construction) generalizable to the test images. The ideal PH networks have a median overall performance (MOP) of 96.0 per cent. None the less, 7.7 per cent of pairs are misclassified by the networks – with 6.5 per cent being misclassified as post-mergers and 1.2 per cent misclassified as isolated galaxies. Similarly, a small number of isolated and post-merger images are misclassified as pairs.⁹

Our confusion matrix includes an additional dimension, which allows us to more deeply investigate network systematics and misclassifications. Histograms are inset in each matrix element, which show the distributions of *relative merger class time-scales*:

$$t_{\text{rel,iso}} = \frac{t-t_1}{t_{10}-t_1}, \quad t_{\text{rel,pair}} = \frac{t-t_p}{t_c-t_p}, \quad t_{\text{rel,post}} = \frac{t-t_c}{t_{\text{last}}-t_c}. \quad (1)$$

Here, t is the simulation time-stamp associated with a particular snapshot in any simulation run. For isolated runs, t_1 and t_{10} are the timestamps for first and tenth of the 10 snapshots selected. For merger simulations, t_p is the time of first pericentric passage, t_c is the coalescence time, and t_{last} is the timestamp of the last snapshot selected for a given run. t_{last} is therefore a number between $t_c + [250, 500]$ Myr as per our class definitions and merger selection criteria. These normalizations allow us to place the timestamp of each snapshot (from each simulation) on a *relative* timeline corresponding to its target class. An image from the isolated or post-merger classes has a $t_{\text{rel,iso}}$ or $t_{\text{rel,post}}$, respectively, that is between 0 and 1 by definition – which we divide into 10 bins each. An image from the pair phase has $t_{\text{rel,pair}}$ between -0.1 and 1 because (a) we start the pair phase 100 Myr before first pericentre and (b) the shortest $t_c - t_p$ is roughly 1 Gyr. Accordingly, each pair timing histogram has 11 bins and starts at 100 Myr before first pericentre and ends at coalescence. With these definitions, a uniform timing distribution in any matrix element would indicate that there is no temporal preference for the images assigned to that element. For visibility, the timing histograms in each element are normalized by their maximum values rather than the total number of images with the corresponding truth label.

Despite the good overall performance of this network and small fraction of misclassifications, the timing histograms reveal temporal preferences for misclassifying certain classes. True pairs that are misclassified as isolated are predominantly in the very early pair phase – with the largest fraction in the pre-first pericentre bin (middle-top panel of Fig. 4). In contrast, true pairs that are misclassified as post-mergers are preferentially near coalescence (middle-bottom panel of Fig. 4). Consequently, the distribution of $t_{\text{rel,pair}}$ for the correctly classified pairs is truncated in the first and final bins. The choppy $t_{\text{rel,post}}$ distribution for the correctly classified post-mergers is due to the chance temporal resonance of snapshots selected for each merger. Coarser binning reveals an essentially uniform timing distribution. Lastly, post-mergers that are misclassified as pairs show a strong preference towards snapshots shortly after coalescence (right-middle panel of Fig. 4).

⁹However, a subtle but notable feature of the networks is that the 55 arcsec fields of view (50 kpc at $z = 0.046$) of the images often do not contain both galaxies for pairs. None the less, networks distinguish pairs from isolated and post-merger images with great accuracy in these cases.

The timing histograms in Fig. 4 show that images that correspond to snapshots at the temporal interface between two neighbouring classes are the most challenging for the network to accurately classify. The subtlety is that these misclassifications are not completely spurious but rather follow intuitive temporal distribution functions – which is actually a validation that the network is behaving as it should. For example, the most common misclassification for pairs early in their interaction is ‘isolated’. Likewise, the most common misclassification of late-stage pairs, shortly before coalescence, is ‘post-merger’. Conversely, it is rare for early pairs to be misclassified as post-mergers, or for late-stage pairs to be classified as isolated. The misclassifications arise because the features of images on either side of a particular class boundary are genuinely similar. Indeed, the timing distributions for correctly and incorrectly classified pair and post-merger targets are qualitatively similar in our other tests (except where additional systematics due to strongly contrasting network/data types dominate).

However, the timing distribution of isolated galaxies that are misclassified as pairs is less intuitive. First, the timing distribution of *correctly* classified isolated galaxies is largely uniform. This result is important. Temporarily discounting secular changes to morphology (such as the emergence of bars and spiral arms), the main changes to these galaxies are their SFRs – which decay exponentially with time. The uniform timing distribution for correctly classified isolated galaxies indicates that *most* isolated galaxies are being correctly classified despite significant changes in SFR. However, the timing distribution of isolated galaxies that are misclassified as pairs suggests that early (and incidentally high-SFR) snapshots from the isolated runs are favoured.

Fig. 5 shows 16 randomly selected images of isolated galaxies that are correctly classified as isolated (left-hand panel) and incorrectly classified as pairs (right-hand panel). The comparison reveals that, occasionally, misclassified isolated galaxies are not easily visually distinguished from the correctly classified isolated targets (e.g. first row, second column of the right-hand panel). However, the right-hand panel of Fig. 5 shows that the majority of misclassified isolated galaxies have not yet dynamically relaxed from the initial conditions of the simulations and consequently have non-steady-state morphologies. Many have unusually bright spiral arms or rings of star formation, which may confuse a network that would desirably exploit morphological features such as tidal tails and shells to identify pairs. In addition, galaxies in the early stages of the merger simulations are similarly unrelaxed and should increase confusion between the pair and isolated classes. Again, there is a subtle importance to these results. The mischaracterization of these dynamically unrelaxed isolated galaxies as pairs *confirms* that the network is exploiting desirable morphological features to make pair classifications. One may also visually note that (unlike morphology) high central surface brightnesses (such as those induced by a starburst) do not necessarily translate to a high pair probability. In Section 3.2, we perform additional tests outside the main handshake, which allow us to examine the temporal misclassification of isolated galaxies more deeply.

3.1.2 Is radiative transfer necessary?

We are particularly interested in knowing whether SM images are an adequate replacement for PH. Radiative transfer makes photometric images computationally expensive to produce in large quantities, and the data products from radiative transfer can be very large depending on the desired spectral resolution. For reference, a single

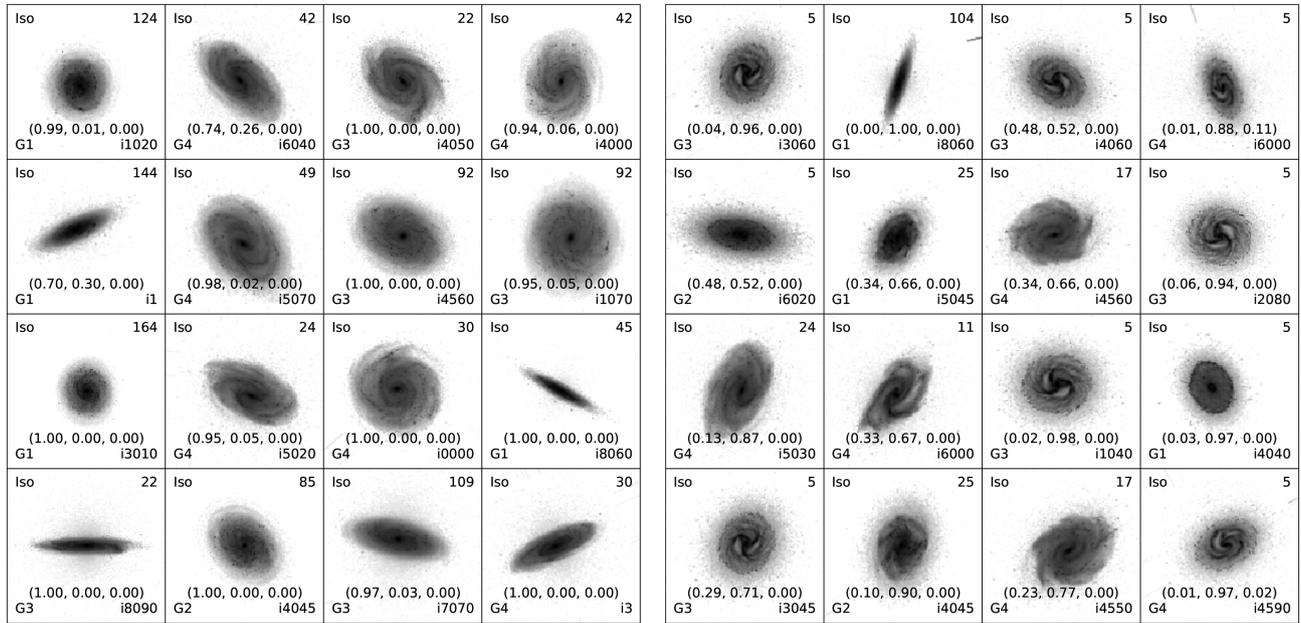


Figure 5. Randomly selected r -band idealized PH images of isolated galaxies: (left) correctly classified as isolated; (right) incorrectly classified as pairs by the networks from Fig. 4. The unique identifier for each image is composed of the four labels in the corners of each image. The upper left shows the simulation type (all Iso in this case), the upper right is the snapshot number, the lower left is the galaxy ID, and the bottom right is the camera angle ID. The normalized probability distribution for each image is given as the tuple of $(P_{\text{Iso}}, P_{\text{Pair}}, P_{\text{Post}})$. These single-band images show that there are notable differences in the surface-brightness distributions between correctly classified isolated galaxies and those misclassified as pairs. Many of the misclassified isolated galaxies have not yet relaxed from the initial conditions of the Iso simulation runs and exhibit non-steady-state morphologies. Streaks at the edges of some images are artefacts of border-handling when images are rotated/shifted during augmentation.

data cube from our data set with (512×512) spatial elements, a modest 241 spectral elements, and 32-bit floating precision is 252.7 MB. Constructing a sufficiently large training set for application to cosmological volumes ($\sim 10^6$ images) with these standards is a data management expense of the order of hundreds of terabytes for the raw data products alone. Therefore, if a neural network performs equally well when trained on SM images (1.05 MB/image for the same spatial resolution) as with images produced using radiative transfer (i.e. generation of PH), it would not only save a significant step in the image generation pipeline, but also be a massive computational saving.

Fig. 6 shows the confusion matrices corresponding to our test of the importance (or not) of including radiative transfer and the generation of multiband photometric images. The lower right panel shows the same high-performance result as in Fig. 4 where the idealized PH-trained networks are applied to PH test data. The upper left panel shows that nearly equally high performance is achieved by the SM networks on SM test data. The off-diagonal panels show the results of testing of these networks on data from the other type. The lower left panel shows that networks trained on PH and tested on SM images have significantly lower performance (median performance of 80.2 per cent) than when either SM or PH networks are tested on data of their own respective type. In contrast, the upper right panel shows that networks that are trained on SM images and tested on PH still have excellent performance (median performance of 90.2 per cent). These results are intuitive when we reflect on the differences between a single-band photometric image and a map of stellar mass. With respect to the SM images, PH images include higher order information from which the network can draw (such as locally varying mass-to-light ratios due to the ages and metallicities of stellar populations and dust). If these higher order features correlate with the target classifications, then the PH network may suffer from an unconventional form of overfitting with

respect to the corresponding SM images – because these higher order features are absent in the SM images.

In contrast, the morphological disturbances that are exploited by the SM network when training on SM images will always be present in the PH images. They will simply underlay any higher order PH features. Consequently, the SM model tests with higher performance on PH images than vice versa because the SM network is *guaranteed* to focus on lower order features and thus is generalizable to PH. Later, in Section 4.2, we argue that the disparity between PH and SM networks/data arises due to the limitations of our training data and predict it would disappear with a galaxy population that is more diverse in stellar populations, colours, and gas fractions.

The results of this section demonstrate that radiative transfer provides a network with more exploitable features than are available in SM images. These include higher order features of the surface brightness profiles and the colour information that can be made accessible by producing images in multiple bandpasses. However, we also show that idealized SM networks are, none the less, highly effective at handling idealized PH images. This means that, at least in the idealized case, one can avoid the (potentially enormous) computational and data management expenses of radiative transfer for large data sets by using SM-based images for a modest trade-off in performance.

3.1.3 Is observational realism necessary?

We are ultimately only interested in networks that will perform well on realistic images. In the last two sections, we have shown that the networks trained on idealized SM and PH images perform very well on their respective selves and reasonably well on each other. In this section, we address the question of whether networks trained with idealized images can accurately classify images with realistic

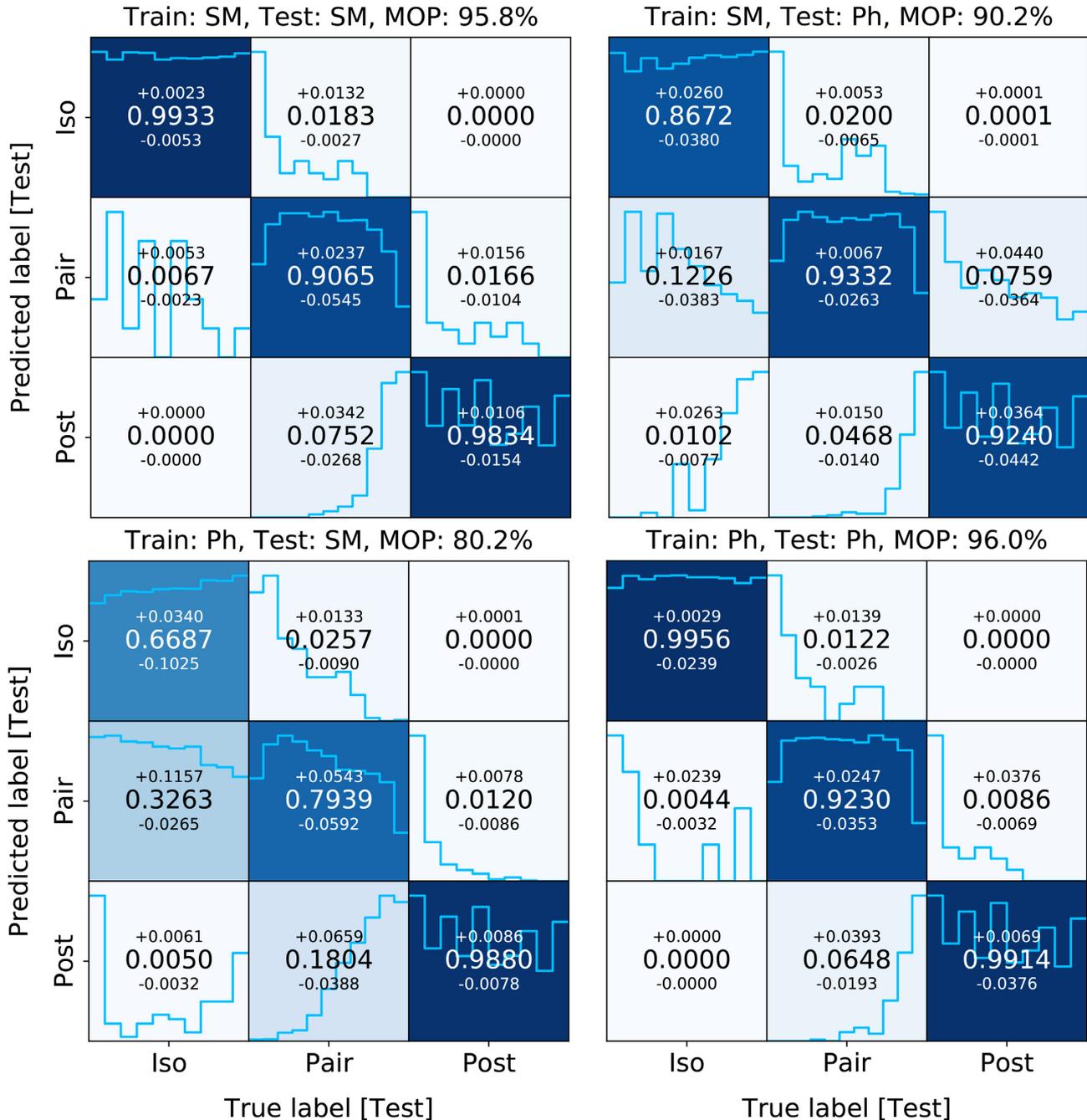


Figure 6. The importance of radiative transfer. Confusion matrices for merger-stage classifications with SM and PH images/models. Models for matrices in the top row are trained on SM images and matrices in the bottom row are trained on PH. Models for matrices in the left column are tested on SM images. Models for matrices in the right column are tested on PH images. The columns of each matrix show the normalized distribution of predicted labels for each true label. The median and 16th and 84th percentile offsets are computed by bootstrapping the results of 10 unique realizations of training, validation, and test data.

noise, resolution degradation, and contamination by nearby objects in the images’ fields of view. Our benchmark for assessing how well any of our networks will perform on real data is the Photometry FullReal data set – which is our best representation of real data. The tests in this section are designed to tell us whether it is sufficient to construct idealized STELLARMAP or PHOTOMETRY synthetic images as training data for networks that can be applied to real data.

Fig. 7 shows the results of applying the PH (left), PHSR (centre), and PHFR (right) trained networks to PHFR test data. In this section, we will focus on the left- and right-hand panels – which

demonstrate the importance of realism, returning to the central panel of Fig. 7 in the following section. The left-hand panel shows that PH networks have very poor performance when tested on realistic images. Similarly, poor results are obtained using the idealized SM networks (see the corresponding panel in Fig. C1 in Appendix C). Both idealized PH and SM networks systematically classify targets in the FULLREAL images as pairs. The histogram insets in the elements of each matrix show that there are no particular temporal preferences for post-merger or isolated targets that are misclassified as pairs by these networks.

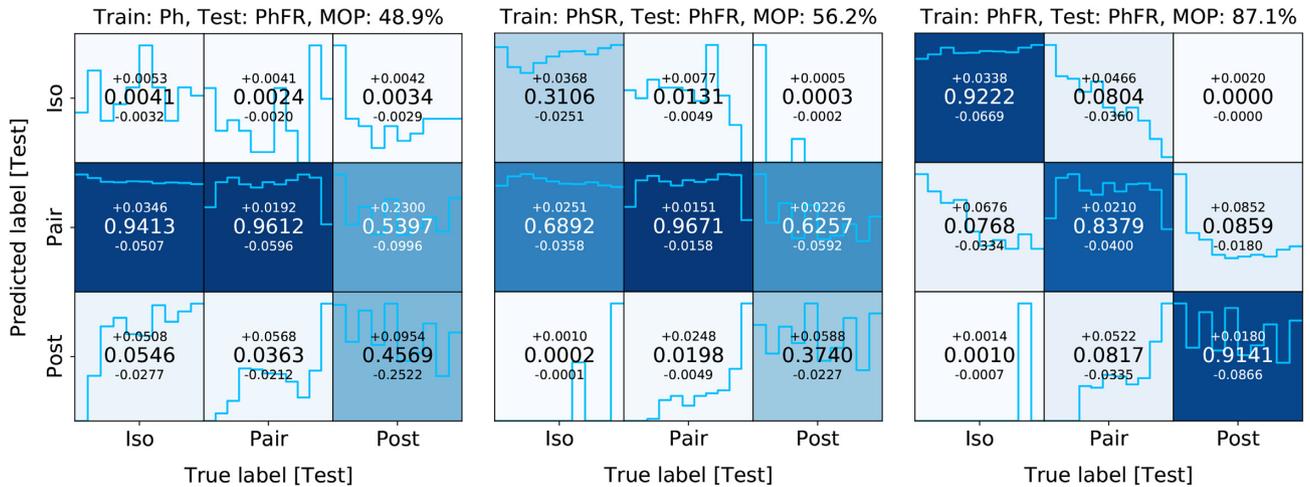


Figure 7. The importance of realism. PH (left), PHSR (centre), and PHFR (right) networks are applied to the PHFR test data. Both the idealized PH and PHSR networks systematically misclassify non-pair targets as pairs. In contrast, networks that are exposed to training data with full realism (real noise, resolution, and crowding) do not appear to be affected by this systematic, and an overall median performance of 87.1 per cent is achieved with the PHFR networks on the PHFR test data.

On the other hand, the right-hand panel of Fig. 7 shows that networks that are trained using PHFR images perform very well on other PHFR survey-realistic images – despite the full statistical rigour of noise, resolution, and contamination effects. Indeed, these results suggest that it is *only* because this network was exposed to these biases in training that it is capable of handling other realistic images. We investigate this hypothesis more closely in the next section.

3.1.4 Is the level of realism important?

In the previous section, we showed that networks that are trained on either PH or SM idealized images perform poorly when tested on realistic images, whereas the PHFR network performed very well. To interpret these results, we now focus on the following question: What ingredients of the FullReal images are key to the success of the FullReal network? Are realistic skies and resolution sufficient criteria? Or are realistic additional sources necessary too? We answer these questions using the networks trained on SemiReal images that are tested on FullReal images. Recall that the sky noise and spatial resolution in the SemiReal images are statistically the same as for the FullReal images by construction (see Section 2.2.4). Consequently, the only difference between these two data sets is that the FullReal images can contain other objects in the FOV, which may confuse a network that is not used to seeing additional sources.

The central panel of Fig. 7 shows the results of applying PHSR networks to PHFR test sets. Again, the results are qualitatively similar when we apply the SMSR networks to the PHFR or SMFR images (see the corresponding panels in Fig. C1 of Appendix C). These tests reveal that the SEMIREAL networks (whether originally deriving from SM or PH images) systematically classify targets as pairs in the FULLREAL images – as was the case for the idealized PH and SM networks. The consequently poor overall performance, particularly when compared to the successes we see when training and testing using FullReal images (right-hand panel of Fig. 7), demonstrates that the *level* of realism is crucial to network performance with realistic images. Specifically, without exposure to projection effects and additional sources of contamination

during training, the SEMIREAL networks preferentially associate secondary sources in the images as companions to the target. In contrast, networks trained on images that include contaminating effects beyond sky and resolution degradation must learn ways to separate false positives and true positives with respect to the pair class.

Fig. 8 examines some important details in the relationship between the level of realism in training data and network performance on realistic test images. The upper panel of Fig. 8 shows that SEMIREAL networks perform very well on SEMIREAL test data – while we know from the middle panel of Fig. 7 that this performance in uncontaminated FOVs does not translate to the FULLREAL images. However, the reverse of that test (training on FULLREAL images and testing on SEMIREAL images) would show whether the contaminants in the FULLREAL training data negatively affect network performance on images that do not contain contaminants. The lower panel of Fig. 8 shows the results of applying FULLREAL networks to the PHSR test images. This test confirms that the networks trained on FullReal images have no trouble testing on images which have similar skies and resolution but which do not contain contaminants. Indeed, statistically fewer isolated (−4 per cent) and post-merger (−2 per cent) targets in the SEMIREAL test images are misclassified as pairs by the PHFR network compared to the right-hand panel of Fig. 7. It is important to note that, until now, we have not seen a network that is trained on one image type and performs equally well (or better) on another image type. Given that our network architecture is the same for every model (with the exception of the number of input channels in particular cases), Fig. 8 is also a crucial validation that the networks are not predestined to overfit to their own training data due to some property of the model architecture.

The results of this section show that training images with realistic noise and resolution are important but (on their own) insufficient criteria for achieving high merger classification performance in realistic images. Networks will only learn to discriminate between true and false pairs if they have been exposed to realistic fields of view in their training data (FULLREAL images). As such, training data must include realistic noise and resolution *as well as* contamination by additional sources.

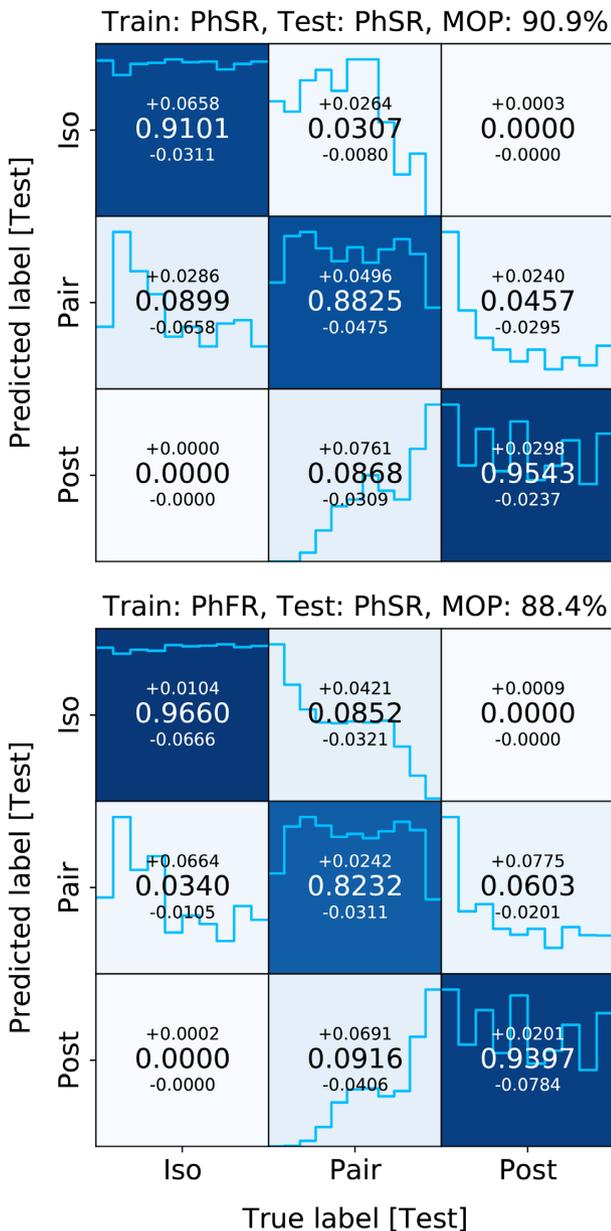


Figure 8. The importance of the *level* of realism. The upper panel shows that PHSR networks (trained on images with realistic skies and resolution but no field objects) handle other PHSR images with very good performance. In contrast, the middle panel from Fig. 7 showed that SEMIREAL networks perform very poorly when applied to fully realistic images and systematically classify targets as pairs – regardless of their true class. The lower panel here shows that the reverse of this test (training on FULLREAL images and testing on SEMIREAL images) produces even better results than when the PHFR networks to PHFR test data. This test shows that networks that are trained on images that include contaminated FOVs have no trouble handling images that have similar noise and resolution but do not contain contaminants.

3.1.5 Is realism more important than radiative transfer?

In the previous section, we showed that the highest performance on realistic test data is obtained when training with PHFR images (see Fig. 7). In addition, in Section 3.1.2 and Fig. 6 we showed that networks trained on SM images performed well when tested on PH data. So we return to the guiding question of Section 3.1.2: *Can we get away with using STELLARMAP-based images in lieu of radiative*

transfer? In other words, is the realism more important than whether the realistic images are originally derived from a STELLARMAP or from PHOTOMETRY?

Fig. 9 offers a compressed view of every test in the main handshake shown in Fig. C1 of Appendix C. Fig. 9 shows the MOPs of each network applied to each set of test data. The overall performance is computed as the number of images in the diagonal elements of a confusion matrix relative to the total number of images. Each panel shows the results of networks trained using each type of training data (labels along *x*-axis) and tested on a particular type of test data (indicated in the tan box). Coloured bars show the median and sample standard deviation test performance of each network type for the 10 networks trained using different random samplings of the training images. The dashed black line denotes a random performance of 1/3 for a three-class model. Since PHFR data are closest to what would be observed with a real instrument, the lower right panel is the focus of this section. As we showed in Fig. 7, the SM and PH networks do only slightly better than random when applied to the PHFR test images because of the lack of realism. Similarly, we showed in Section 3.1.4 that the SEMIREAL networks do only marginally better than models with no realism because they are not exposed to projection effects in training. In contrast, *both* FULLREAL networks (SMFR and PHFR) perform well on PHFR images. The SMFR and PHFR networks have MOPs of 79.6 and 87.1 per cent, respectively, when applied to PHFR images.

In contrast, networks trained on either idealized or SEMIREAL images never exceed 60 per cent performance when handling the more realistic PHFR images. This is true whether the training images are derived from photometry or stellar maps. These results show that the level of realism is more important than radiative transfer and that one can achieve strong performance with SM-based images as long as they are fully realistic. Although there is a big difference between a network that can achieve ~ 90 per cent performance compared to one that achieves ~ 80 per cent, the difference in performance may be an acceptable trade-off for being able to side-step radiative transfer and its associated computational and data management expenses – particularly on the scale of the current state-of-the-art cosmological simulations.

3.2 Single-channel experiments

We supplement the main handshake experiment with a series of additional single-band tests with PH-based networks. These tests are designed to determine the importance of colour and bandpass to network performance. In particular, we are interested in whether the timing preference for misclassified isolated galaxies seen in several PH-based tests (e.g. Fig. 4 and the right-hand panel of Fig. 7) and discussed in Section 3.1.1 persists for networks that are colourblind. If the timing preference persists and similar performance is achieved when a network is trained using a single band, then colour can be ruled out as a major factor in distinguishing pairs from isolated galaxies by the network.

More generally, we are also interested in knowing the degree to which overall network performance is sensitive to colour. There are important advantages of a network that can achieve high performance without exploiting colour and focuses primarily on morphological features. For example, star formation correlates strongly with colour. Since interactions between gas-rich galaxies are proven triggers of central star formation (Bushouse 1987; Noguchi 1991; Carlberg, Pritchet & Infante 1994; Mihos & Hernquist 1994, 1996; Barton, Geller & Kenyon 2000; Springel 2000; Smith et al. 2007; Cox et al. 2008; Ellison et al. 2008; Patton et al. 2011; Hopkins

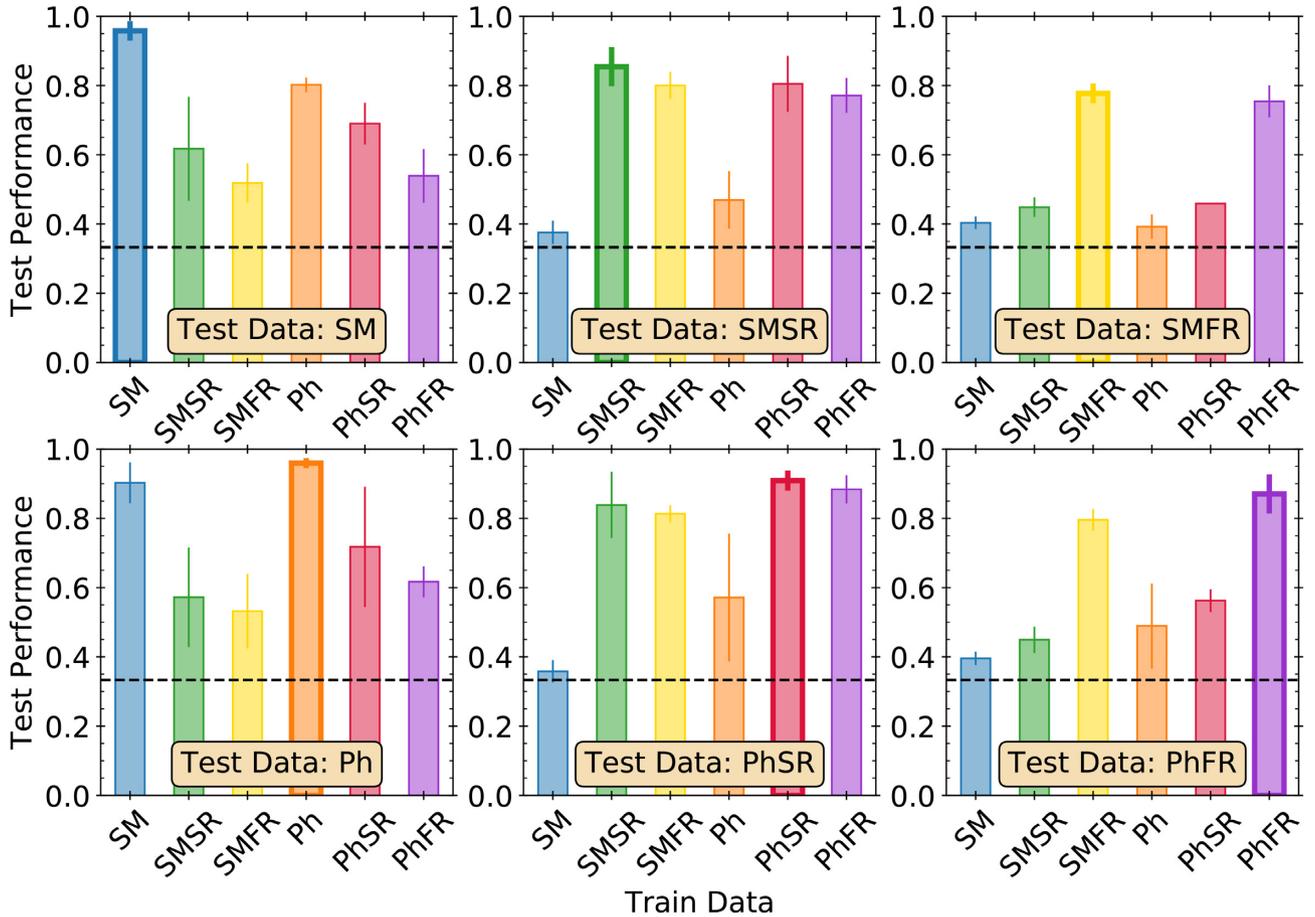


Figure 9. Overall median test performances in our network/data handshake. Each panel shows the median test results of applying a specified network to test images of every type. The ordering of test data types is indicated along the x -axes. For reference, the overall performance of a single test is computed from the number of test images in the diagonal elements of a confusion matrix as a fraction of the total number of images. MOP is computed over the 10 bootstraps of training/validation/test sets. Similarly, errorbars show the 5th to 95th percentile range in overall performance. Bolded borders are placed on bars corresponding to cases in which the test data are of the same image type as the training data. The dashed line denotes a uniformly random classification performance – which in the case of three possible classes is $1/3$. The only model that performs comparably with the PHFR model on the PHFR test data is the SMFR. This result demonstrates that matching the realism is more important than whether or not the training data derive from images generated with proper radiative transfer.

et al. 2013; Patton et al. 2013; Moreno et al. 2015; Sparre & Springel 2016; Thorp et al. 2019), a colour-sensitive network may learn to exploit central star formation to characterize interaction stage through its correlation with colour. However, the negative consequence of identifying/characterizing interactions based on triggered star formation is that any study which then examines the relationship between interaction stage and star formation is automatically biased. Therefore, it is of great value to know that our networks are able to make merger-stage classifications without exploiting colour information.

The single-channel experiments are divided into two handshakes. New PH, PHSR, and PHFR networks are each trained and tested using only the corresponding r -band and i -band images to produce a single 3×3 handshake for each band. As with the main handshake, we statistically combine the individual test results of 10 bootstraps of training/validation/test images for our final results. The results of every test are shown in Appendix A. We discuss selected results in the sections that follow.

3.2.1 How important is colour to network performance?

Previous tests already give us reason to believe that colour is not an *essential* ingredient of satisfactory network success. For example,

Fig. 9 showed that even the SMFR networks (which do not use radiative transfer) achieve a reasonable MOP of 79.6 per cent on PHFR test images using only single-channel input. Similarly, the upper right panel of Fig. 6 showed that the colourblind, idealized SM network achieved 90.2 per cent MOP when tested on idealized PH test images – only a 5.8 per cent drop with respect to the results of the full-colour PH network.

Fig. 10 shows the confusion matrices for the idealized r -band (upper panel) and i -band (lower panel) PH networks that were tested on the respective r -band and i -band PH test images. For the idealized networks/data, the single-channel PH networks still achieve exceptional overall performances. The change in MOP, ΔMOP , for each of the single-channel networks is minor with respect to the 96.0 per cent performance of the three-channel, full-colour PH networks from Fig. 4: $\Delta\text{MOP}(r, i) = (-0.4, -0.5)$ per cent.

However, there is a potential problem with using the idealized images as a ‘representative’ scenario for examining the importance of colour: the possibility of interplay between realism and colour. Between the three realism levels (idealized, SEMIREAL, FULLREAL), networks trained using idealized images are least likely to exploit colour information because the low-surface brightness morphological features are most easily exploitable. In contrast, low-surface brightness morphological features are often hidden in the sky noise

or blurred by the PSF in the more realistic SEMIREAL and FULLREAL images. Consequently, a colour-sensitive network that is trained using these more realistic images is more likely to use colour to classify galaxies if the correlation between colour and merger stage is strong *and* morphological information is limited.

To evaluate the importance of colour in the more realistic image data, we compare the single-channel PHFR test results (see Appendix A) with the three-channel, colour-sensitive PHFR network results (the right-hand panel of Fig. 7, MOP= 87.1 per cent). The PHFR networks trained using individual bands have mild losses in MOP with respect to the three-channel networks: $\Delta\text{MOP}(r, i) = (-1.1, -2.3)$ per cent. While these are greater losses than in the idealized case, these results still demonstrate that colour is not an essential ingredient to network success and that the network is primarily targeting morphological features. However, without a training set that includes both red and blue discs, it is uncertain whether it is not simply this information that the network is exploiting to achieve the mildly higher performance in the full-colour data.

Lastly, we compare the timing histograms of both correctly and incorrectly classified galaxies for the single-channel networks (Fig. 10) and three-channel networks (Fig. 4). The results are qualitatively similar in every case – including the timing preference for isolated galaxies that are misclassified as pairs. The fact that the timing preference persists in the single-channel networks disqualifies colour as being a driver of preferential misclassification of these isolated galaxies as pairs – in particular, those corresponding to early, high SFR, and morphologically unstable snapshots from the isolated simulation runs. A network that does not have access to colour cannot exploit its relationship with star formation. Combined with our visual analysis of correctly and incorrectly classified isolated galaxies in Section 3.1.1 and Fig. 5, this result demonstrates that the networks are focusing on morphological features. However, it should still be noted that certain morphological properties can be significantly enhanced by star formation (for example, compactness). In our discussion of the limitations of our suite, we argue that this problem can be solved using cosmological training sets that include greater variety of isolated galaxies and merger properties (gas fractions, initial morphologies, etc.).

3.2.2 Does the bandpass make a difference?

Table 3 shows that the typical sky surface brightness uncertainty, $\langle\sigma_{\text{sky, Field}}\rangle$, in the SDSS r band is 1.5 mag arcsec⁻² fainter than in the i band. By comparing the performances in each of these bands individually, we determine the sensitivity of network performance to a modest change in imaging depth and the differences in the intrinsic brightnesses of targets in each band. Fig. 10 shows that the difference in MOP in the i band with respect to the r band is only -0.1 per cent in the idealized PH images, as expected. The idealized images contain no noise – so the only change between the r band and the i band is the brightnesses of stellar populations in each bandpass. The difference in performance broadens for networks trained on the more realistic single-channel PHFR images. The single-channel PHFR networks achieve MOPs of 86.0 per cent (r band) and 84.8 per cent (i band) with a minor difference in MOP of -1.2 per cent in the i band with respect to the r band. While this is a small change in performance, one must recall that source surface brightness in a bandpass diminishes rapidly with source redshift – by a factor of $(1+z)^{-5}$. Consequently, a difference of 1.5 mag arcsec⁻¹ may be a much greater hindrance to network

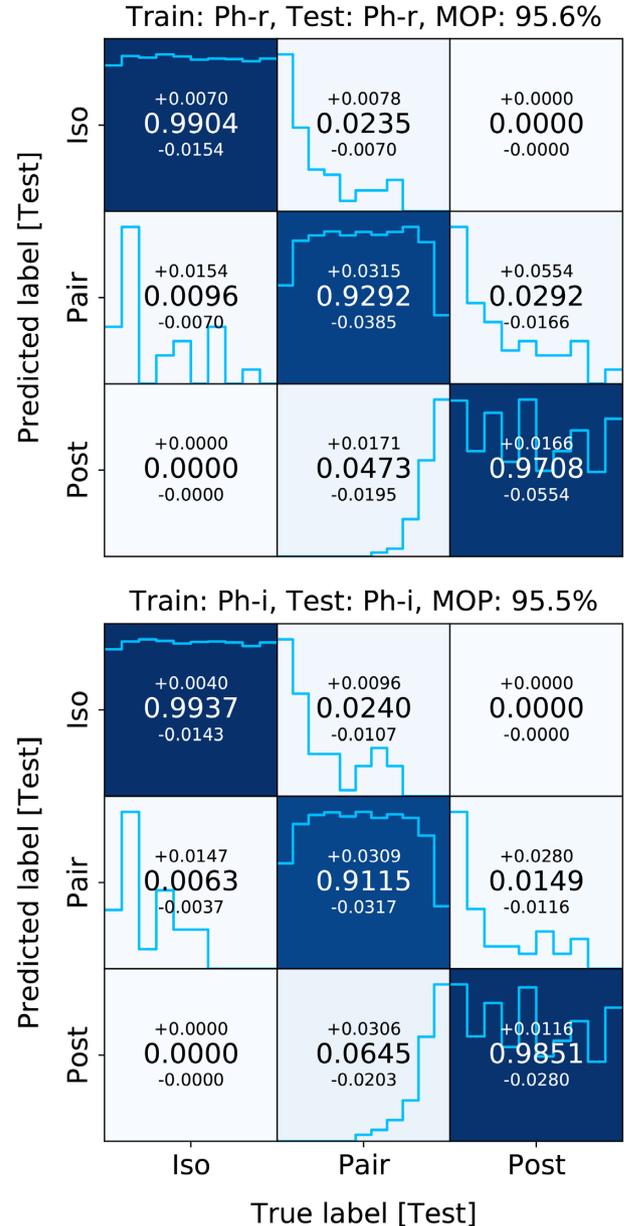


Figure 10. The importance of colour to network performance with idealized images. Network results shown in the upper and lower panels were trained and tested on single-channel r - and i -band idealized PH images, respectively. The networks achieve MOPs of 95.6 per cent (r band) and 95.5 per cent (i band) compared to 96.0 per cent when training and testing using idealized images from all three gri bands. The losses in performance when reducing to single-band photometry are minor.

performance for images of galaxies at the higher- z end of a realistic SDSS redshift distribution.

4 DISCUSSION

4.1 The importance of realism

In Sections 3.1.3 and 3.1.4 and Figs 7 and 8, we showed that adding full realism to synthetic training images (including realistic skies, resolution, and crowding by nearby sources) is a necessary condition of strong network performance in identifying and characterizing

galaxy interactions in realistic images. Indeed, we have shown that every ingredient of full observational realism is essential and that omitting any ingredient leads to systematic misclassifications in testing. Without exposure to contamination by nearby sources in the training images, networks systematically preferred to assign images to the pair class. In particular, the systematic misclassification of fully realistic images as pairs persists even when the training images have both realistic skies and resolution but lack crowding effects.

Similarly, the tests in which we train on idealized PH or SM images and test on PHSR or SMSR images (orange and blue bars in the two middle panels of Fig. 9) reveal that realistic skies and resolution are also vital (also see Fig. C1 of Appendix C). The sky and resolution effects bury and wash away the low-surface brightness morphological features that made idealized PH or SM networks successful on images of their own types (e.g. tidal tails, bridges, and shells). Consequently, networks that are trained using idealized images perform very poorly on test images that contain realistic skies and resolution – even without crowding effects.¹⁰ Combined with our results showing the independent biases that arise from excluding crowding effects in training images, the importance of each component of a FULLREAL image is clear.

Indeed, in Section 3.1.5, we demonstrated that the level of realism is even more important than whether the synthetic images originate from radiative transfer or from STELLARMAP images. The bottom right panel of Fig. 9 shows that the only networks with performances that approach those of the PHFR networks on the PHFR test images are the SMFR networks – which exploit neither colour nor any higher order features available from radiative transfer. Ultimately, we want to be able to construct training images from the current state-of-the-art cosmological hydrodynamical simulations, which will be used to train networks that can then identify and characterize real galaxy interactions by merger stage. Only these simulations can provide the necessary scope, diversity, and accuracy in galaxy properties (e.g. morphologies, masses, merger characteristics, orbital properties, gas fractions, etc.) to form training sets that are sufficiently generalizable to real galaxies and interactions. Consequently, training sets generated from these simulations will necessarily comprise very large numbers of synthetic images. However, in Section 3.1.2, we noted that constructing training sets on these scales using radiative transfer corresponds to potentially enormous computational and data management expenses. Therefore, the results that (1) radiative transfer is secondary to realism and (2) one can avoid radiative transfer using SMFR images for a modest compromise in performance are of great importance to the primary science application of the methods we present.

Lastly, with more diverse data (for example, data that are not so tightly temporally correlated) realism may become an even more important factor in generating large samples of images and training networks that effectively handle crowding effects. For the goals of this paper, it was sufficient to insert each projection into a single FOV and increase the size of our data sets by augmentation. However, one could produce a much larger sample of images and train a network that may be even less sensitive to crowding effects by inserting *each* projection of a synthetic image into, for example, $N = 10$ unique fields. This would expose the network to a greater

diversity of crowding effects for each target. As long as the training set is sufficiently large and each target gets an equal amount of unique insertions, the network will not overfit to a particular target. This would simultaneously ensure that no overfitting can ever occur due to particular configurations of targets and projected objects in the image FOV.

4.2 Limitations of the suite

As we have previously stated, the *specific* networks we trained in this study would have limited application to real data. This is primarily because our simulations are not cosmological. Despite the fact that the FIRE merger suite covers a broad range of mass ratios and orbital properties, the range is small in comparison to the parameter spaces of galaxy and merger properties encompassed by the observable Universe or by the current state-of-the-art cosmological hydrodynamical simulations (Patton et al. in preparation; Blumenthal et al., in preparation). Consequently, we know that our training data are not generalizable to real data. Only cosmological simulations can provide the necessary scope to construct training data that are *both* similar and diverse enough to train networks that can be applied to real data.

However, generating networks that can be applied to real data is not a goal of this work. The goals are (1) to provide the methodology with which CNNs, trained and calibrated using hydrodynamical simulations, can be used to identify mergers and predict merger stage in *realistic* images and (2) to assess the importance of realism in the synthetic training images. The experiments we used to accomplish these goals did not require cosmological simulations. Indeed, our experiments correspond to scenarios in which we know that the training data are fully generalizable to the test data, as desired, because both training and test data are drawn from the same merger suite – just different parts of it. Ultimately, whether we use training data from a cosmological simulation or from our suite would make no qualitative difference to our results regarding the importance of realism. However, it will be important to assess whether the reduction in intrinsic simulation resolution in the cosmological simulations has an additional effect on performance. We will address this question in a follow-up study.

Another limitation of the suite used in this work is that all of the galaxies used therein are relatively gas-rich discs, and hence not representative of the full morphological complexity seen in the real Universe. This morphology bias would undoubtedly be an issue if we were to apply our networks to real data, but is not a limitation for the goals of this work. However, observations and numerical simulations alike show that mergers between gas-rich discs induce central star formation in galaxies during the pair phase and in the merger remnant. Therefore, the relationship between merger stage and star formation may be exploited by networks that are sensitive to properties related to central star formation. While we have demonstrated that eliminating colour sensitivity makes an insignificant difference to network performance in Section 3.2.1, we do not rule out a morphological connection with high central SFRs – such as with CAS Concentration index (Bershady, Jangren & Conselice 2000; Conselice 2003) or Gini coefficient (Abraham, van den Bergh & Nair 2003; Lotz et al. 2004). However, the increases in central surface brightnesses from recent star formation are associated with the low M/L of young stellar populations formed in the bursts. So, while the PHOTOMETRY-based images and networks are more liable to exploit such connections between recent central star formation and morphology, the STELLARMAP-based images will be largely insensitive to the morphological effects of recent

¹⁰Note that this simultaneously demonstrates that the realism *should* be survey specific. However, the results of Domínguez Sánchez et al. (2019) for morphological classifications demonstrate that it may be possible to use a *transfer learning* approach – in which CNNs optimized for one survey can be adapted to another using a small sample of images from the target survey.

star formation because they are *completely* insensitive to M/L ratio. Fig. 9 showed that the SMFR network performs nearly as well as the PHFR network. Indeed, an MOP of 95.8 per cent is achieved with networks trained and tested on the idealized SM images compared to the 96.0 per cent achieved by the networks trained and tested on idealized PH images. Therefore, while a connection between central morphology (as induced by central star formation) and merger stage may exist in the PHOTOMETRY-based images, it is not essential to network performance. Additionally, the capacity to exploit such a connection would be expected to be further suppressed in a more homogeneous galaxy sample (such as from a cosmological simulation) with mergers between galaxies that are red, blue, gas-rich, gas-poor, and everything between.

4.3 Overfitting

As explained in the previous section, we know that our networks are limited to the set of merger scenarios encompassed by our suite with respect to galaxy and merger properties that would be present in a representative volume of the Universe. However, for evaluating the importance of realism, this limitation is immaterial because all we needed was a reasonably sized training set that includes typical merger features and test sets to which the training data are known to be generalizable. In contrast, a bias that would not be desirable is one that might arise from the construction of our images – such as camera angles or orientation. For example, in Fig. 5, the correctly classified isolated galaxy image in the 3rd row, 2nd column of the left-hand panel is the same galaxy and snapshot as the one that is incorrectly classified as a pair in the 3rd row, 1st column of the right-hand panel. The only difference between these images is a slight change in zoom and rotation. A high sensitivity of predicted class to orientation is a common characteristic of overfitting – where a network learns to exploit properties of the training data that are not generalizable to test data.

While CNNs with max-pooling layers are architecturally invariant to translation, they are not rotationally invariant by default and require large and diversified training data to achieve *learned* rotational invariance (see chapter 9, fig. 9.9 of Goodfellow, Bengio & Courville 2016 for an intuitive example). We apply rotational, translational, and zoom augmentation to all of our data sets in an effort to (1) increase our data size and (2) achieve rotationally invariant networks. Given that every image in the augmented training data (including all possible orientations) contributes equally to network optimization, we find it unlikely that our networks are classifying based on orientation. However, another example from Fig. 5 is the correctly classified inclined disc in the 3rd row, 4th column of the left-hand panel and its incorrectly classified counterpart in the 1st row, 2nd column of the right-hand panel. Both images correspond to the same galaxy and inclination – only the incorrectly classified one is from a much later snapshot and is rotated. Despite the visual similarity between these targets, the network confidently classifies these images as isolated and pair, respectively. Although Fig. 4 shows that such misclassifications are rare, this high sensitivity between isolated and pair classifications, without obvious visible justification, may arise from our class definitions.

4.4 Class definitions

By using hydrodynamical simulations to train networks, we attempt to eliminate as much subjectivity as possible for merger-stage classifications. The advantage of this strategy is that, based on a set of simple quantitative definitions for each class, one is always

optimizing network performance on the absolute truth. However, the definitions themselves are one remaining source of subjectivity that cannot be avoided in supervised learning. The beginning of the post-merger class requires a definition of coalescence that also defines the end of the pair class. The beginning of the pair phase also requires a definition. We defined the pair phase as beginning 100 Myr before first pericentric passage. Was our choice to use this temporal criterion appropriate? What were the consequences?

Fig. 4 and the right-hand panel of Fig. 7 show that pairs that are misclassified as isolated are preferentially *early* pairs. The clear consequence of our definition is that galaxies in the early pair phase are indistinguishable from isolated galaxies because no galaxies in these early pairs have experienced visible disturbances resulting from gravitational interaction with their companions. Subsequently, this definition is also a likely culprit for the seemingly spurious misclassifications of a few isolated galaxies shown in Fig. 5 that were discussed in the last section. However, for the purposes of this paper, our definition happened to be beneficial (see Section 3.1.1). The fact that the networks had difficulty distinguishing early pairs (by our definition of the pair phase) from isolated galaxies was evidence that the networks were behaving intuitively. Meanwhile, since the majority of images from the pair class do not resemble isolated galaxies (all those except for the early pairs), the networks still accurately classified most pairs in the test images.

Ultimately, we propose that reduced continuity between the isolated and pair images through an alternative definition of the pair phase would lead to better network performance and fewer misclassifications in these classes (for example, starting the pair phase at first pericentre). However, testing the sensitivity of performance to alternative definitions for each class is beyond the scope of this work. There is a large parameter space to be explored. The time or spatial separation at which a galaxy’s properties start to be affected by an interaction and persist after coalescence is sensitive to the masses, morphologies, mass ratio, and orbital properties at hand (e.g. Lotz et al. 2008, 2010a,b; Ji et al. 2014; Nevin et al. 2019). None the less, we highlight that a few key advantages of calibrating networks using simulations are that, for a given set of class definitions, one can (1) train networks that make completely reproducible predictions and (2) evaluate the biases associated with these definitions. So, while *our* class definitions resulted in some confusion between early pairs and the isolated class, these definitions can be easily changed and optimized to improve performance.

5 SUMMARY

CNNs are becoming a popular tool for identifying galaxy mergers in large surveys. In this paper, we use galaxy merger simulations to train CNNs that identify mergers and predict merger stage. We assess the importance of producing realistic images from simulations to the performance of CNNs, in order to guide future applications of this method.

We train and calibrate a set of CNNs using synthetic images generated from a suite of hydrodynamical binary merger simulations (Moreno et al. 2019) run with the FIRE-2 physical model (Hopkins et al. 2018). Training networks on simulations offers the significant benefit of foreknowledge of interaction stage and, therefore, optimization targets that are not biased by factors such as image quality or personal subjectivity. We examine the importance of adding realistic ingredients to the synthetic images. To do so, networks are trained using two types of galaxy images, stellar maps, and dust-inclusive radiatively transferred images, each with three levels of observational realism: (1) no observational effects

(idealized images), (2) realistic sky and PSF (semirealistic images), and (3) insertion into a real sky image (fully realistic images) (see Section 2.2 and summary in Table 2). Each image data set covers the same set projections and simulation snapshots and is divided into isolated, pair, and post-merger classes. In our main handshake experiment (see Section 3.1 and Fig. 3), we test each network on data of every other type. Each network is also tested on data of the same type upon which it was trained but that the network never sees during training. The PHFR data – in which the synthetic images are injected into real survey fields – are the most realistic representation of real observations. Therefore, the PHFR test data are used to evaluate how well networks trained on images of a particular type would handle real data (see Section 4.2 for an important discussion on the limitations of *this suite* for applications to real data). The results of our main handshake experiment are:

(i) [Section 3.1.1] **Networks trained on idealized images (SM and PH) classify images of the same type with 96.0 per cent accuracy** (Fig. 4 and the upper left panel of Fig. 6). Misclassifications behave predictably. Early pairs are difficult to distinguish from isolated galaxies. Recent post-mergers are difficult to distinguish from pairs nearing coalescence. Isolated galaxies and post-mergers are *never* confused for one another.

(ii) [Section 3.1.2] **SM images can be used in place of more computationally expensive (but more realistic) images produced with radiative transfer at a modest cost in performance.** Networks trained on idealized SM images classify idealized PH images with 90.2 per cent accuracy (upper right panel of Fig. 6).

(iii) [Section 3.1.3] **PH and PHSR networks – of which neither are exposed to training images that include contamination by nearby sources – systematically classify PHFR images as belonging to the pair class.** Networks trained on idealized PH or semirealistic PHSR images (realistic skies and resolution) both perform very poorly (48.9 and 56.2 per cent accuracies, respectively) on the PHFR images (left-hand and centre panels of Fig. 7).

(iv) [Section 3.1.4] **As long as networks are exposed to all ingredients of realism in training (skies, resolution, and crowding) they can learn to efficiently handle these effects in test images.** While PH and PHSR networks fail to handle realistic images, networks trained on PHFR images classify PHFR test images with 87.1 per cent accuracy (right-hand panel of Fig. 7). Additionally, (a) there is no clear systematic preference towards classifying images as pairs and (b) PHFR networks are even more accurate on PHSR test images (88.4 per cent) than PHFR test images (see Fig. 8).

(v) [Section 3.1.5] **Realism is more important than whether the images originate from radiative transfer or from maps of stellar mass.** Networks trained on SMFR images classify PHFR test images with 79.6 per cent accuracy (lower right panel Fig. 9). Indeed, these are the only networks other than the PHFR networks that achieve reasonable performance on the PHFR test images.

We perform a secondary handshake experiment aimed at characterizing the roles of colour and depth to network performance (see Section 3.2). Single-channel networks are trained on the *r*- and *i*-band images, individually, taken from the PH, PHSR, and PHFR data sets. These tests eliminate the possibility for networks to exploit colour information and allow us to compare results for networks trained on images in bands of varying photometric depths. The main results of these tests are:

(i) [Section 3.2.1] **Networks trained without colour incur very mild penalties to performance with respect to colour-sensitive networks.** The performances of the single-

channel *r*- and *i*-band PH (PHFR) networks are 95.6 per cent (86.0 per cent) and 95.5 per cent (84.8 per cent), respectively, compared to 96.0 per cent (87.1 per cent) with the full-colour networks (see Fig. 10). These results demonstrate that, while the colour-sensitive networks *can* exploit colour information, colour is not a *necessary* ingredient for high network performance.

(ii) [Section 3.2.2] **The difference in average photometric depth between the *r*- and *i* bands (~ 1.5 mag arcsec⁻²) yields a small difference in the performances of networks trained on each band individually** (see Fig. 10). However, in this study, we do not match the redshift distribution of SDSS galaxies and instead insert galaxies at the median redshift of galaxies in the DR14 MaNGA galaxy sample. Therefore, these differences might be expected to be larger for training and test data that include galaxies that are more distant or have lower intrinsic brightnesses.

The pertinent applications of this work are: (1) to train networks using realistic synthetic images from cosmological simulations and (2) to use a model trained on cosmological simulations to identify and characterize interactions in the real Universe. The most important feature of cosmological simulations in this respect is that mergers *and* isolated galaxies that are selected from a statistically representative simulation will cover a larger range of morphologies, masses, gas fractions, etc. This diversity will be a necessary component of a training set that can be expected to perform well on real test data.

ACKNOWLEDGEMENTS

CB acknowledges the support of a National Sciences and Engineering Research Council of Canada (NSERC) Graduate Scholarship. MHH and HT contributed equally to this research. MHH acknowledges the receipt of a Vanier Canada Graduate Scholarship. SLE and LS gratefully acknowledge support under the Canadian Discover Grants Programme. The data used in this paper were, in part, generated and hosted using facilities supported by the Scientific Computing Core at the Centre for Computational Astrophysics, a division of the Simons Foundation. The computations in this research were enabled in part by support provided by Compute Canada (www.computecanada.ca). The numerical simulations in this paper were run on the Odyssey cluster supported by the FAS Division of Science, Research Computing Group at Harvard University. Support for JM is provided by the NSF (AST Award Number 1516374), and by the Harvard Institute for Theory and Computation, through their Visiting Scholars Program.

Funding for the SDSS IV has been provided by the Alfred P. Sloan Foundation, the U.S. Department of Energy Office of Science, and the Participating Institutions. SDSS-IV acknowledges support and resources from the Center for High-Performance Computing at the University of Utah. The SDSS web site is www.sdss.org.

SDSS-IV is managed by the Astrophysical Research Consortium for the Participating Institutions of the SDSS Collaboration including the Brazilian Participation Group, the Carnegie Institution for Science, Carnegie Mellon University, the Chilean Participation Group, the French Participation Group, Harvard-Smithsonian Center for Astrophysics, Instituto de Astrofísica de Canarias, The Johns Hopkins University, Kavli Institute for the Physics and Mathematics of the Universe (IPMU)/University of Tokyo, the Korean Participation Group, Lawrence Berkeley National Laboratory, Leibniz Institut für Astrophysik Potsdam (AIP), Max-Planck-Institut für Astronomie (MPIA Heidelberg), Max-Planck-Institut für Astrophysik (MPA Garching), Max-Planck-Institut für

Extraterrestrische Physik (MPE), National Astronomical Observatories of China, New Mexico State University, New York University, University of Notre Dame, Observatório Nacional/MCTI, The Ohio State University, Pennsylvania State University, Shanghai Astronomical Observatory, United Kingdom Participation Group, Universidad Nacional Autónoma de México, University of Arizona, University of Colorado Boulder, University of Oxford, University of Portsmouth, University of Utah, University of Virginia, University of Washington, University of Wisconsin, Vanderbilt University, and Yale University.

REFERENCES

- Abazajian K. N. et al., 2009, *ApJS*, 182, 543
- Abraham R. G., van den Bergh S., Nair P., 2003, *ApJ*, 588, 218
- Ackermann S., Schawinski K., Zhang C., Weigel A. K., Turp M. D., 2018, *MNRAS*, 479, 415
- Baes M., Verstappen J., De Looze I., Fritz J., Saftly W., Vidal Pérez E., Stalevski M., Valcke S., 2011, *ApJS*, 196, 22
- Barnes J. E., Hernquist L., 1992, *ARA&A*, 30, 705
- Barton E. J., Geller M. J., Kenyon S. J., 2000, *ApJ*, 530, 660
- Berg T. A. M., Simard L., Mendel Trevor J., Ellison S. L., 2014, *MNRAS*, 440, L66
- Bershady M. A., Jangren A., Conselice C. J., 2000, *AJ*, 119, 2645
- Bertin E., Arnouts S., 1996, *A&AS*, 117, 393
- Bluck A. F. L., Conselice C. J., Buitrago F. o., Grützbauch R., Hoyos C., Mortlock A., Bauer A. E., 2012, *ApJ*, 747, 34
- Bluck A. F. L. et al., 2019, *MNRAS*, 485, 666
- Blumenthal K. A., Barnes J. E., 2018, *MNRAS*, 479, 3952
- Blumenthal G. R., Faber S. M., Primack J. R., Rees M. J., 1984, *Nature*, 311, 517
- Bottrell C., Torrey P., Simard L., Ellison S. L., 2017a, *MNRAS*, 467, 1033
- Bottrell C., Torrey P., Simard L., Ellison S. L., 2017b, *MNRAS*, 467, 2879
- Bottrell C., Simard L., Mendel J. T., Ellison S. L., 2019, *MNRAS*, 486, 390
- Buda M., Maki A., Mazurowski M. A., 2017, preprint ([arXiv:1710.05381](https://arxiv.org/abs/1710.05381))
- Bundy K. et al., 2015, *ApJ*, 798, 7
- Bushouse H. A., 1987, *ApJ*, 320, 49
- Camps P., Baes M., 2015, *Astron. Comput.*, 9, 20
- Cardie C., Howe N., 1997, in ICML. Morgan Kaufmann Publishers Inc., San Francisco CA, p. 57
- Carlberg R. G., Pritchett C. J., Infante L., 1994, *ApJ*, 435, 540
- Casteels K. R. V. et al., 2013, *MNRAS*, 429, 1051
- Casteels K. R. V. et al., 2014, *MNRAS*, 445, 1157
- Chan P. K., Stolfo S. J., 1998, in KDD. AAAI Press, New York, NY, p. 164
- Chollet F. et al., 2015, *Keras: The Python Deep Learning Library*, <https://keras.io>
- Conselice C. J., 2003, *ApJS*, 147, 1
- Cox T. J., Jonsson P., Somerville R. S., Primack J. R., Dekel A., 2008, *MNRAS*, 384, 386
- Darg D. W. et al., 2010, *MNRAS*, 401, 1043
- Dieleman S., Willett K. W., Dambre J., 2015, *MNRAS*, 450, 1441
- Di Matteo T., Springel V., Hernquist L., 2005, *Nature*, 433, 604
- Doi M. et al., 2010, *AJ*, 139, 1628
- Domínguez Sánchez H., Huertas-Company M., Bernardi M., Tuccillo D., Fischer J. L., 2018, *MNRAS*, 476, 3661
- Domínguez Sánchez H. et al., 2019, *MNRAS*, 484, 93
- 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., Patton D. R., Hickox R. C., 2015, *MNRAS*, 451, L35
- Ellison S. L., Viswanathan A., Patton D. R., Bottrell C., McConnachie A. W., Gwyn S., Cuillandre J.-C., 2019, *MNRAS*, 487, 2491
- Faucher-Giguère C.-A., Lidz A., Zaldarriaga M., Hernquist L., 2009, *ApJ*, 703, 1416
- Fukushima K., 1980, *Biol. Cybern.*, 36, 193
- Goodfellow I., Bengio Y., Courville A., 2016, *Deep Learning*. MIT Press, Cambridge, MA
- Goulding A. D. et al., 2018, *PASJ*, 70, S37
- Groves B., Dopita M. A., Sutherland R. S., Kewley L. J., Fischera J., Leitherer C., Brandl B., van Breugel W., 2008, *ApJS*, 176, 438
- Grzymala-Busse J. W., Goodwin L. K., Grzymala-Busse W. J., Zheng X., 2004, in Pal S. K., Polkowski L., Skowron A., eds, *Rough-neural Computing*. Springer, Berlin, Heidelberg, p. 543
- Haixiang G., Yijing L., Shang J., Mingyun G., Yuanyue H., Bing G., 2017, *Expert Syst. Appl.*, 73, 220
- Hani M. H., Sparre M., Ellison S. L., Torrey P., Vogelsberger M., 2018, *MNRAS*, 475, 1160
- Hausen R., Robertson B., 2019, preprint ([arXiv:1906.11248](https://arxiv.org/abs/1906.11248))
- Hayward C. C., Hopkins P. F., 2017, *MNRAS*, 465, 1682
- He K., Zhang X., Ren S., Sun J., 2015, preprint ([arXiv:1512.03385](https://arxiv.org/abs/1512.03385))
- Hernquist L., 1989, *Nature*, 340, 687
- Hernquist L., 1990, *ApJ*, 356, 359
- Hernquist L., 1992, *ApJ*, 400, 460
- Hezaveh Y. D., Perreault Levasseur L., Marshall P. J., 2017, *Nature*, 548, 555
- Hopkins P. F., 2015, *MNRAS*, 450, 53
- Hopkins P. F., 2017, preprint ([arXiv:1712.01294](https://arxiv.org/abs/1712.01294))
- Hopkins P. F., Quataert E., 2010, *MNRAS*, 407, 1529
- Hopkins P. F., Hernquist L., Cox T. J., Kereš D., 2008a, *ApJS*, 175, 356
- Hopkins P. F., Cox T. J., Kereš D., Hernquist L., 2008b, *ApJS*, 175, 390
- Hopkins P. F., Hernquist L., Cox T. J., Dutta S. N., Rothberg B., 2008c, *ApJ*, 679, 156
- Hopkins P. F., Cox T. J., Hernquist L., Narayanan D., Hayward C. C., Murray N., 2013, *MNRAS*, 430, 1901
- Hopkins P. F. et al., 2018, *MNRAS*, 480, 800
- Huertas-Company M. et al., 2015, *ApJS*, 221, 8
- Huertas-Company M. et al., 2019, *MNRAS*, 489, 1859
- Jacobs C. et al., 2019, *MNRAS*, 484, 5330
- Ji I., Peirani S., Yi S. K., 2014, *A&A*, 566, A97
- Johnson S. D., Chen H.-W., Mulchaey J. S., 2015, *MNRAS*, 449, 3263
- Jonsson P., 2006, *MNRAS*, 372, 2
- Jonsson P., Groves B. A., Cox T. J., 2010, *MNRAS*, 403, 17
- Kartaltepe J. S. et al., 2015, *ApJS*, 221, 11
- Keel W. C., Kennicutt R. C. J., Hummel E., van der Hulst J. M., 1985, *AJ*, 90, 708
- Kewley L. J., Geller M. J., Barton E. J., 2006, *AJ*, 131, 2004
- Koss M., Mushotzky R., Veilleux S., Winter L., 2010, *ApJ*, 716, L125
- Krizhevsky A., Sutskever I., Hinton G. E., 2012, in Pereira F., Burges C. J. C., Bottou L., Weinberger K. Q., eds, *Advances in Neural Information Processing Systems 25*. Curran Associates Inc., USA, p. 1097
- Kroupa P., 2001, *MNRAS*, 322, 231
- Lacey C., Cole S., 1993, *MNRAS*, 262, 627
- LeCun Y., Boser B., Denker J. S., Henderson D., Howard R. E., Hubbard W., Jackel L. D., 1989, *Neural Comput.*, 1, 541
- LeCun Y., Bengio Y. et al., 1995, *The Handbook of Brain Theory and Neural Networks*, Vol. 3361. MIT Press, Cambridge, MA
- LeCun Y., Bottou L., Bengio Y., Haffner P., 1998, *Proc. IEEE*, 86, 2278
- Lecun Y., Bengio Y., Hinton G., 2015, *Nature*, 521, 436
- Leitherer C. et al., 1999, *ApJS*, 123, 3
- Lin L. et al., 2004, *ApJ*, 617, L9
- Lin L. et al., 2007, *ApJ*, 660, L51
- López-Sanjuan C. et al., 2011, *A&A*, 530, A20
- López-Sanjuan C. et al., 2013, *A&A*, 558, A135
- Lotz J. M., Primack J., Madau P., 2004, *AJ*, 128, 163
- Lotz J. M., Jonsson P., Cox T. J., Primack J. R., 2008, *MNRAS*, 391, 1137
- Lotz J. M., Jonsson P., Cox T. J., Primack J. R., 2010a, *MNRAS*, 404, 575
- Lotz J. M., Jonsson P., Cox T. J., Primack J. R., 2010b, *MNRAS*, 404, 590
- Lotz J. M., Jonsson P., Cox T. J., Croton D., Primack J. R., Somerville R. S., Stewart K., 2011, *ApJ*, 742, 103
- Lupton R., Gunn J. E., Ivezić Z., Knapp G. R., Kent S., 2001, in Harnden F. R., Jr, Primini F. A., Payne H. E., eds, *ASP Conf. Ser. Vol. 238, Astronomical Data Analysis Software and Systems X*. Astron. Soc. Pac., San Francisco, p. 269

Lupton R. H., Ivezić Z., Gunn J. E., Knapp G., Strauss M. A., Yasuda N., 2002, in Tyson J. A., Wolff S., eds, Proc. SPIE Conf. Ser. Vol. 4836, Survey and Other Telescope Technologies and Discoveries. SPIE, Bellingham, p. 350

Lupton R. H., Ivezić Ž., Gunn J. E., Knapp G. R., Strauss M. A., 2012, The Photo-Lite Draft, Plus Other Notes at RHL's Web Site. Available at: <http://www.astro.princeton.edu/rhl/photo-lite.pdf>

Mac Namee B., Cunningham P., Byrne S., Corrigan O. I., 2002, *Artif. Intell. Med.*, 24, 51

Martin C. L., 2005, *ApJ*, 621, 227

Martin G., Kaviraj S., Devriendt J. E. G., Dubois Y., Pichon C., 2018, *MNRAS*, 480, 2266

Mendel J. T., Simard L., Palmer M., Ellison S. L., Patton D. R., 2014, *ApJS*, 210, 3

Mihos J. C., Hernquist L., 1994, *ApJ*, 431, L9

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

Mo H. J., Mao S., White S. D. M., 1998, *MNRAS*, 295, 319

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

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

Moster B. P., Naab T., White S. D. M., 2013, *MNRAS*, 428, 3121

Naab T., Burkert A., 2003, *ApJ*, 597, 893

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

Negroponte J., White S. D. M., 1983, *MNRAS*, 205, 1009

Nelson D. et al., 2018, *MNRAS*, 475, 624

Nevin R., Blecha L., Comerford J., Greene J., 2019, *ApJ*, 872, 76

Noguchi M., 1991, *MNRAS*, 251, 360

Ntampaka M. et al., 2019, *ApJ*, 876, 82

Oke J. B., Gunn J. E., 1983, *ApJ*, 266, 713

Patton D. R., Atfield J. E., 2008, *ApJ*, 685, 235

Patton D. R. et al., 2002, *ApJ*, 565, 208

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

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

Patton D. R., Qamar F. D., Ellison S. L., Bluck A. F. L., Simard L., Mendel J. T., Moreno J., Torrey P., 2016, *MNRAS*, 461, 2589

Pawlik M. M., Wild V., Walcher C. J., Johansson P. H., Villforth C., Rowlands K., Mendez-Abreu J., Hewlett T., 2016, *MNRAS*, 456, 3032

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

Perez J., Michel-Dansac L., Tissera P. B., 2011, *MNRAS*, 417, 580

Pillepich A. et al., 2018, *MNRAS*, 475, 648

Radivojac P., Chawla N. V., Dunker A. K., Obradovic Z., 2004, *J. Biomed. Inform.*, 37, 224

Ribli D., Pataki B. Á., Csabai I., 2019, *Nat. Astron.*, 3, 93

Robertson B., Bullock J. S., Cox T. J., Di Matteo T., Hernquist L., Springel V., Yoshida N., 2006, *ApJ*, 645, 986

Robotham A. S. G. et al., 2014, *MNRAS*, 444, 3986

Rodriguez-Gomez V. et al., 2015, *MNRAS*, 449, 49

Rodriguez-Gomez V. et al., 2019, *MNRAS*, 483, 4140

Rupke D. S., Veilleux S., Sanders D. B., 2005a, *ApJS*, 160, 115

Rupke D. S., Veilleux S., Sanders D. B., 2005b, *ApJ*, 632, 751

Rupke D. S. N., Kewley L. J., Barnes J. E., 2010a, *ApJ*, 710, L156

Rupke D. S. N., Kewley L. J., Chien L. H., 2010b, *ApJ*, 723, 1255

Saintonge A. et al., 2016, *MNRAS*, 462, 1749

Satyapal S., Ellison S. L., McAlpine W., Hickox R. C., Patton D. R., Mendel J. T., 2014, *MNRAS*, 441, 1297

Schaye J. et al., 2015, *MNRAS*, 446, 521

Simard L., Mendel J. T., Patton D. R., Ellison S. L., McConnachie A. W., 2011, *ApJS*, 196, 11

Simmons B. D. et al., 2017, *MNRAS*, 464, 4420

Smith B. J., Struck C., Hancock M., Appleton P. N., Charmandaris V., Reach W. T., 2007, *AJ*, 133, 791

Snyder G. F., Rodriguez-Gomez V., Lotz J. M., Torrey P., Quirk A. C. N., Hernquist L., Vogelsberger M., Freeman P. E., 2019, *MNRAS*, 486, 3702

Sol Alonso M., Michel-Dansac L., Lambas D. G., 2010, *A&A*, 514, A57

Sparre M., Springel V., 2016, *MNRAS*, 462, 2418

Springel V., 2000, *MNRAS*, 312, 859

Springel V., Di Matteo T., Hernquist L., 2005, *MNRAS*, 361, 776

Strickland D. K., Heckman T. M., 2009, *ApJ*, 697, 2030

Teimoorinia H., Ellison S. L., 2014, *MNRAS*, 439, 3526

Thor P. M. D., Ellison S. L., Simard L., Sánchez S. F., Antonio B., 2019, *MNRAS*, 482, L55

Toomre A., 1977, in Tinsley B. M., Larson D., Campbell R. B. G., eds, Proc. Conf. Yale Univ., Evolution of Galaxies and Stellar Populations. Yale Univ. Observatory, New Haven, p. 401

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

Torrey P., Cox T. J., Kewley L., Hernquist L., 2012, *ApJ*, 746, 108

Torrey P., Hopkins P. F., Faucher-Giguère C.-A., Vogelsberger M., Quataert E., Kereš D., Murray N., 2017, *MNRAS*, 467, 2301

Veilleux S. et al., 2013, *ApJ*, 776, 27

Walmsley M., Ferguson A. M. N., Mann R. G., Lintott C. J., 2019, *MNRAS*, 483, 2968

White S. D. M., Rees M. J., 1978, *MNRAS*, 183, 341

Wong K. C. et al., 2011, *ApJ*, 728, 119

Woo J.-H., Son D., Bae H.-J., 2017, *ApJ*, 839, 120

Zschaechner L. K. et al., 2016, *ApJ*, 832, 142

Zubko V., Dwek E., Arendt R. G., 2004, *ApJS*, 152, 211

APPENDIX A: SINGLE-BAND PHOTOMETRY RESULTS

Figs A1 and A2 show the results for the single-channel handshake experiment in which the *r*-band and *i*-band images from PH, PHSR, and PHFR data sets are used to train single-channel, colour-insensitive networks that are then applied to images of each type.



Figure A1. Confusion matrices for the single-channel handshake experiment in which the *r*-band images from PH, PHSR and PHFR data sets are used to train colour-insensitive networks that are then applied to images of each type. For reference, Fig. 4 describes the information displayed by each individual confusion matrix in detail.



Figure A2. Same as Fig. A2 but for networks trained and tested on i -band images from the PH, PHSR, and PHFR data sets.

APPENDIX B: CORRELATIONS BETWEEN GALAXY IMAGES

In this appendix, we investigate the possibility that our snapshot sampling cadence for each interaction (e.g. see Fig. 1) was too fine – resulting in galaxy images that could be strongly correlated. For our data sets, strong correlations between images in neighbouring snapshots could result in overfitting to the training data and, consequently, loss of generalizability to test data from other data sets (e.g. train: PH, test: PHFR). Meanwhile, strong correlations between

images in tests where a network is applied to test data from the same data set (e.g. train: PH, test: PH) may lead to correlated training/test data and result in erroneously high test accuracies. Since our main investigation is focused on the sensitivity of network performance to realism, it is crucial to characterize (and preferably rule out) the sensitivity of the network performance to our snapshot selection cadence.

Fig. B1 shows PH (upper panels) and corresponding PHFR images (lower panels) for five neighbouring snapshots from our sampling of the fiducial G2G3e orbit 1 merger at a fixed camera angle. The snapshots are selected from near first apocentre ($t \approx 0.75$ Gyr in Fig. 1) where the rate of morphological evolution would be expected to be at a minimum during the pair phase and, therefore, are most likely to be visually correlated. It should be noted that the smaller object in the images is not the companion (which is outside the FOV) but is a tidal dwarf that is produced in this particular interaction. The PH images in the upper panels of Fig. B1 show that there is still visual evolution in galaxy structure (both inner and outer) during this relatively calm part of the pair phase. The PHFR images in the lower panels are even less visibly correlated due to the varying sky levels, resolution, and contamination by additional sources. Correlations between the images should therefore stand to be most problematic in the idealized images that do not incorporate any additional realism. However, since these visual assessments are subjective, we also devised a test that quantitatively investigates the possibility of correlations between images from neighbouring snapshots.

If images from neighbouring snapshots are strongly correlated, then randomly sampling the test data from a data set (e.g. splitting SM into training/validation/test data) will result in training, validation, and test data that are correspondingly correlated. There are two experiments that could be performed to investigate whether such correlations may be influencing our results: (1) reduce the sampling cadence (e.g. discard every second snapshot from our initial selection) and (2) reserve all images from a full merger simulation for testing and train/validate on the remaining images.

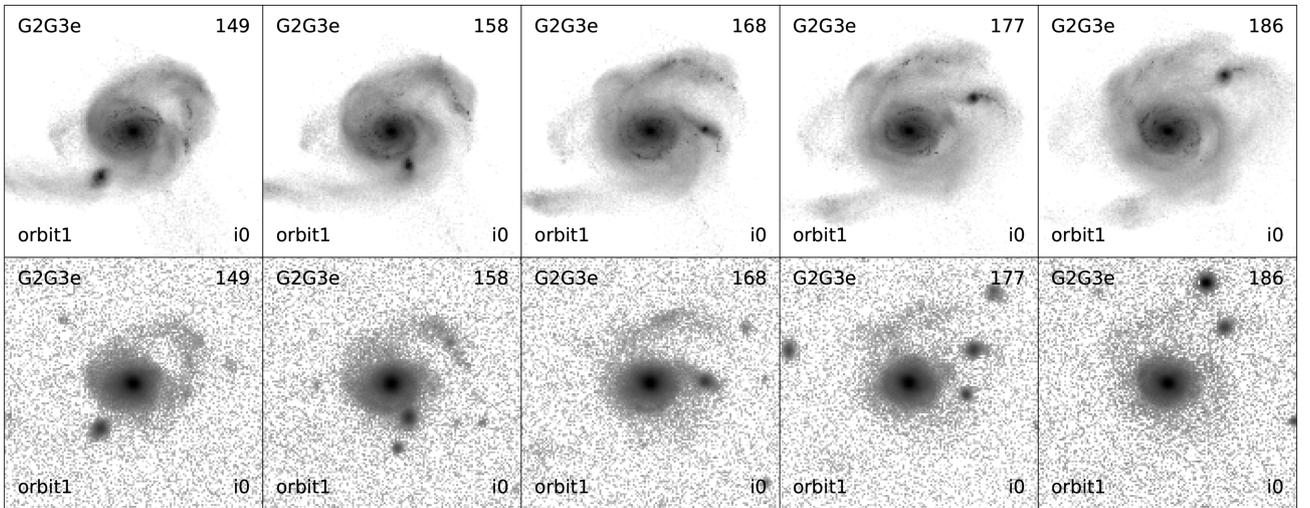


Figure B1. PH (top row) and corresponding PHFR (bottom row) images for five neighbouring snapshots at fixed camera angle from our sampling of the G2G3e orbit 1 merger near first apocentre ($t \approx 0.75$ Gyr in Figs 1 and B2). If there were strong correlations between images from neighbouring snapshots, they would be most likely to occur here, at first apocentre in the merger sequence – since the rate of morphological evolution should be lowest relative to the more rapid changes at first pericentre and beyond second pericentre. Visually, inner and outer structures of the galaxy both evolve appreciably in this sequence of images. These visual differences are apparent in both the PH and PHFR images. Fig. B2 shows the results of a more quantitative and robust experiment that demonstrates that images such as these are not so strongly correlated that they are influencing our main results.

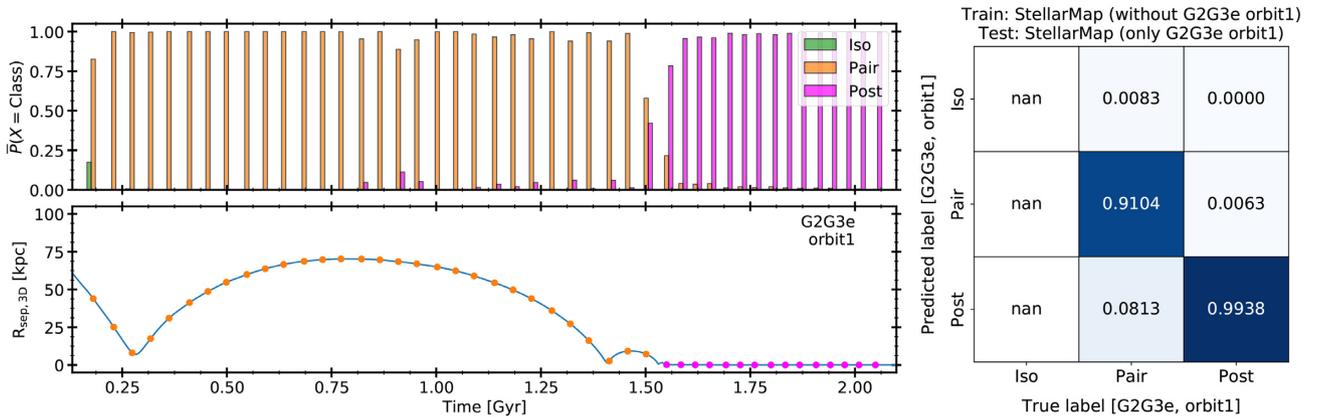


Figure B2. Classification scores along fiducial G2G3e orbit 1 interaction sequence (left-hand panels) and confusion matrix (right-hand panel) for the experiment described in Appendix B. The experiment is designed to show whether the sampling cadence for each merger yields images that are too strongly correlated – resulting in test images that are too similar to the training/validation images. Applying a network to images from an *entire merger* that the network never saw during training would reveal whether such correlations are affecting our results. The images from the fiducial G2G3e orbit 1 merger were removed from the SM data set and reserved for testing. An SM network was then trained using the remaining data SM and applied to the G2G3e orbit 1 test images. The right-hand panel shows the confusion matrix for this test. These results are consistent with the results shown in the upper left panel of Fig. 6 (train: SM, test: SM) – demonstrating that possible correlations between images from neighbouring snapshots are not affecting our results in the main experiments. The left-hand panels show the classification scores as coloured bars for each snapshot in the G2G3e orbit 1 test sequence, averaged over all camera angles at each snapshot, $\bar{P}(X = \text{Class})$. The results of this test show that: (1) the snapshot sampling cadence is not affecting our networks’ performances even in phases where the rate of morphological evolution is expected to be minimal (e.g. near apocentre) and (2) there is a continuous transition between class scores at the temporal class boundaries (e.g. between pair and post-merger at $t \approx 1.54$ Gyr).

The problem with (1) is that this approach would reduce our data volume by an integer factor – which could lead to overfitting from sparseness in the data set and make it impossible to assign any reduction in accuracy to either this sparseness or image correlations. Test (2) does not suffer from this problem because, for a given merger, it only reduces the data volume by relatively small factor of $1/23$. Poor classification performance on the images from the merger that were reserved for testing (particularly for phases where the rate of morphological evolution is expected to be small) would confirm that the snapshot selection is too fine and yields correlations between images that would be affecting our results. The second test would simultaneously show whether a network is generalizable to images from a merger that it has never seen before (albeit, within the limitations outlined in Section 4.2). For these reasons, we carry out the second test.

Fig. B2 shows the results of our investigation into possible correlations between images from neighbouring snapshots. The investigation was performed using the SM data set. We removed all images from the fiducial G2G3e orbit 1 merger (Fig. 1) from the SM data set and reserved them as a test set. We also removed images from the G2G3 merger in the mass ratio suite (which all use ‘e’ orbit 1 initial conditions). This merger is not identical to the fiducial simulation (due to the chaotic nature of galaxy mergers) but has the same initial conditions and is consequently removed and discarded to eliminate any possibility that the training/validation data include images that may be correlated with the test images from the fiducial G2G3e orbit 1 merger simulation. We trained a network on the remaining data with a (70, 30) per cent training/validation split. Only the images for the fiducial merger were used as test data. The right-hand panel of Fig. B2 shows the confusion matrix for this test. This test set contains no images with the Iso class because those images are drawn from separate isolated simulation runs as outlined in Section 2.1.3. Consequently, all elements in the first column are

NaNs. The performance on the G2G3e orbit 1 test images, after removing all images from the training/validation data that might be correlated to these test images, is consistent with the results shown in the upper left panel of Fig. 6. In other words, the network runs equally well on a merger for which it has seen zero images (this test), as mergers for which ~ 70 per cent of their images are included in the training data (upper left panel of Fig. 6). The results of this test demonstrate that possible correlations between images in neighbouring snapshots do not have a significant role in the successes of our networks.

The upper left panel of Fig. B2 shows the normalized classification scores as coloured bars for each snapshot in the G2G3e orbit 1 test images, averaged over all camera angles at each snapshot, $\bar{P}(X = \text{Class})$. The lower left panel shows the radial separation sequence for this merger with each selected snapshot and its true class indicated with coloured circles as in Fig. 1. This classification sequence plot demonstrates the two main results of this appendix: (1) correlations between images from neighbouring snapshots are not a significant contributor to the accuracies reported in our main analyses (even near first apocentre, where such correlations would be expected to be strongest) and (2) the changes to classification scores as one approaches a temporal class boundary are continuous and not choppy or sporadic (e.g. between the pair and post-merger phase at $t \approx 1.54$ Gyr).

APPENDIX C: MAIN HANDSHAKE RESULTS

Fig. C1 shows the combined confusion matrices for every test in the main handshake experiment. Each row corresponds to a type of training data. Each column corresponds to a type of test data. As in Fig. 4, each matrix shown combines 10 individual tests performed with different random allocations of training, validation, and test images.

