



CEERS Key Paper. IX. Identifying Galaxy Mergers in CEERS NIRCam Images Using Random Forests and Convolutional Neural Networks

Caitlin Rose¹, Jeyhan S. Kartaltepe¹, Gregory F. Snyder², Marc Huertas-Company^{3,4,5}, L. Y. Aaron Yung², Pablo Arrabal Haro⁶, Micaela B. Bagley⁷, Laura Bisigello^{8,9}, Antonello Calabrò¹⁰, Nikko J. Cleri^{11,12}, Mark Dickinson⁶, Henry C. Ferguson², Steven L. Finkelstein⁷, Adriano Fontana¹⁰, Andrea Grazian¹³, Norman A. Grogin², Benne W. Holwerda¹⁴, Kartheik G. Iyer¹⁵, Lisa J. Kewley¹⁶, Allison Kirkpatrick¹⁷, Dale D. Kocevski¹⁸, Anton M. Koekemoer², Jennifer M. Lotz¹⁹, Ray A. Lucas², Lorenzo Napolitano^{10,20}, Casey Papovich^{11,12}, Laura Pentericci¹⁰, Pablo G. Pérez-González²¹, Nor Pirzkal²², Swara Ravindranath^{23,24}, Rachel S. Somerville²⁵, Amber N. Straughn²³, Jonathan R. Trump²⁶, Stephen M. Wilkins^{27,28}, and Guang Yang²⁹

- ¹ Laboratory for Multiwavelength Astrophysics, School of Physics and Astronomy, Rochester Institute of Technology, 84 Lomb Memorial Drive, Rochester, NY 14623, USA
- ² Space Telescope Science Institute, 3700 San Martin Drive, Baltimore, MD 21218, USA
- ³ Instituto de Astrofísica de Canarias, La Laguna, Tenerife, Spain
- ⁴ Universidad de la Laguna, La Laguna, Tenerife, Spain
- ⁵ Université Paris-Cité, LERMA—Observatoire de Paris, PSL, Paris, France
- ⁶ NSF's National Optical-Infrared Astronomy Research Laboratory, 950 North Cherry Avenue, Tucson, AZ 85719, USA
- ⁷ Department of Astronomy, The University of Texas at Austin, Austin, TX, USA
- ⁸ INAF, Istituto di Radioastronomia, Via Piero Gobetti 101, 40129 Bologna, Italy
- ⁹ Dipartimento di Fisica e Astronomia “G. Galilei,” Università di Padova, Via Marzolo 8, I-35131 Padova, Italy
- ¹⁰ INAF—Osservatorio Astronomico di Roma, via di Frascati 33, 00078 Monte Porzio Catone, Italy
- ¹¹ Department of Physics and Astronomy, Texas A&M University, College Station, TX 77843-4242, USA
- ¹² George P. and Cynthia Woods Mitchell Institute for Fundamental Physics and Astronomy, Texas A&M University, College Station, TX 77843-4242, USA
- ¹³ INAF—Osservatorio Astronomico di Padova, Vicolo dell'Osservatorio 5, I-35122, Padova, Italy
- ¹⁴ Physics & Astronomy Department, University of Louisville, Louisville, KY 40292, USA
- ¹⁵ Dunlap Institute for Astronomy & Astrophysics, University of Toronto, Toronto, ON M5S 3H4, Canada
- ¹⁶ Center for Astrophysics—Harvard & Smithsonian, 60 Garden Street, Cambridge, MA 02138, USA
- ¹⁷ Department of Physics and Astronomy, University of Kansas, Lawrence, KS 66045, USA
- ¹⁸ Department of Physics and Astronomy, Colby College, Waterville, ME 04901, USA
- ¹⁹ Gemini Observatory/NSF's National Optical-Infrared Astronomy Research Laboratory, 950 North Cherry Avenue, Tucson, AZ 85719, USA
- ²⁰ Dipartimento di Fisica, Università di Roma Sapienza, Città Universitaria di Roma—Sapienza, Piazzale Aldo Moro, 2, 00185, Roma, Italy
- ²¹ Centro de Astrobiología (CAB), CSIC-INTA, Ctra. de Ajalvir km 4, Torrejón de Ardoz, E-28850, Madrid, Spain
- ²² ESA/AURA Space Telescope Science Institute, USA
- ²³ Astrophysics Science Division, NASA Goddard Space Flight Center, 8800 Greenbelt Road, Greenbelt, MD 20771, USA
- ²⁴ Center for Research and Exploration in Space Science and Technology II, Department of Physics, Catholic University of America, 620 Michigan Avenue NE, Washington, DC 20064, USA
- ²⁵ Center for Computational Astrophysics, Flatiron Institute, 162 5th Avenue, New York, NY 10010, USA
- ²⁶ Department of Physics, 196 Auditorium Road, Unit 3046, University of Connecticut, Storrs, CT 06269, USA
- ²⁷ Astronomy Centre, University of Sussex, Falmer, Brighton BN1 9QH, UK
- ²⁸ Institute of Space Sciences and Astronomy, University of Malta, Msida MSD 2080, Malta
- ²⁹ Nanjing Institute of Astronomical Optics and Technology, Nanjing 210042, People's Republic of China

Received 2024 July 30; revised 2024 October 7; accepted 2024 October 22; published 2024 November 14

Abstract

A crucial yet challenging task in galaxy evolution studies is the identification of distant merging galaxies, a task that suffers from a variety of issues ranging from telescope sensitivities and limitations to the inherently chaotic morphologies of young galaxies. In this paper, we use random forests and convolutional neural networks to identify high-redshift JWST Cosmic Evolution Early Release Science Survey (CEERS) galaxy mergers. We train these algorithms on simulated $3 < z < 5$ CEERS galaxies created from the IllustrisTNG subhalo morphologies and the Santa Cruz SAM light cone. We apply our models to observed CEERS galaxies at $3 < z < 5$. We find that our models correctly classify $\sim 60\%$ – 70% of simulated merging and nonmerging galaxies; better performance on the merger class comes at the expense of misclassifying more nonmergers. We could achieve more accurate classifications, as well as test for a dependency on physical parameters such as gas fraction, mass ratio, and relative orbits, by curating larger training sets. When applied to real CEERS galaxies using visual classifications as ground truth, the random forests correctly classified 40%–60% of mergers and nonmergers at $3 < z < 4$ but tended to classify most objects as nonmergers at $4 < z < 5$ (misclassifying $\sim 70\%$ of visually classified mergers). On the other hand, the CNNs tended to classify most objects as mergers across all redshifts (misclassifying 80%–90% of visually classified nonmergers). We investigate what features the models find most useful, as well as the characteristics of false positives and false negatives, and also calculate merger rates derived from the identifications made by the models.

Unified Astronomy Thesaurus concepts: James Webb Space Telescope (2291); Galaxy mergers (608); Astronomical simulations (1857); Random Forests (1935); Convolutional neural networks (1938)

1. Introduction

Galaxy merger studies are important for understanding the evolution of galaxies over cosmic time, as mergers can affect multiple physical phenomena and processes such as star formation and active galactic nucleus activity (e.g., S. L. Ellison et al. 2008; D. R. Patton et al. 2011; K. L. Larson et al. 2016), the morphological evolution of spiral galaxies to elliptical galaxies (e.g., A. Toomre 1977; T. J. Cox et al. 2006; J. Kormendy et al. 2009; V. Rodriguez-Gomez et al. 2017), and the buildup of massive structures (e.g., S. Tacchella et al. 2015; A. Zolotov et al. 2015; L. Costantin et al. 2021, 2022). To better understand galaxy evolution, it is necessary to be able to accurately identify merging galaxies over a wide range of mass ratios (J. M. Lotz et al. 2010a, 2010b) and redshifts.

However, at high redshift, merger identification can be difficult. The limited sensitivity of modern telescopes can result in samples biased toward the most massive galaxies and major mergers (mass ratio $< 4:1$; e.g., S. L. Ellison et al. 2013; K. B. Mantha et al. 2018). Cosmological surface brightness dimming and poor angular resolution can result in the loss of faint, detailed structures. As rest-frame optical light is shifted toward the infrared, infrared telescopes like JWST are needed to observe the stellar population. Finally, high-redshift galaxies can inherently have irregular morphologies that are difficult to discern from merger signatures (e.g., A. Dekel et al. 2009; J. S. Kartaltepe et al. 2012, 2015).

Classic methods of identifying mergers—visual classifications (e.g., C. J. Lintott et al. 2008; J. S. Kartaltepe et al. 2015), the close pair method (e.g., L. Lin et al. 2004; J. S. Kartaltepe et al. 2007; E. Ventou et al. 2017; K. Duncan et al. 2019; E. A. Shah et al. 2020; Q. Duan et al. 2024), and quantitative morphology parameters (e.g., C. J. Conselice 2003; J. M. Lotz et al. 2004; P. E. Freeman et al. 2013)—all have limitations. The close pair method requires that both companions have accurate redshifts in order to identify pairs that are physically close in space and is most sensitive to galaxies that have not yet merged. Visual classifications are qualitative and subjective, as well as time-consuming. Finally, individual morphology parameters are each sensitive to different merger phases as well as noise (e.g., J. M. Lotz et al. 2010a, 2010b). Recently, studies have begun investigating machine learning methods for merger identification tasks, as they show promise for exploiting complex data (whether from multiband images or from precomputed morphology parameters) to better detect merger signatures in high-redshift images than classic methods (e.g., G. F. Snyder et al. 2019; L. Ferreira et al. 2022). Furthermore, the use of deep, high-resolution near-infrared JWST imaging will also help in the detection of merger features at high redshifts. One such JWST survey is the Cosmic Evolution Early Release Science Survey (CEERS; S. L. Finkelstein et al. 2022), which surveyed galaxies at $0.5 < z < 13$ over 100 arcmin^2 in the Extended Groth Strip (EGS) Hubble Space Telescope (HST) legacy field (M. Davis et al. 2007).

In C. Rose et al. (2023), we trained random forests to identify merging galaxies at $0.5 < z < 4$ in simulated CEERS images from IllustrisTNG using a suite of morphology parameters measured using Galaxy Analysis over Large Areas:

Parameter Assessment by Galfit-fitting Objects from SE (Galapagos-2; S. P. Bamford et al. 2011; B. Häußler et al. 2013; M. Vika et al. 2013) and statmorph (V. Rodriguez-Gomez et al. 2019), as well as merger history information from IllustrisTNG. We found that the forests correctly classified $\sim 60\%$ of mock CEERS mergers and nonmergers across all redshift bins. In this follow-up study, we investigate using convolutional neural networks trained directly on the mock images for the task of merger identification at high redshift, as well as test our trained models on real galaxies from CEERS (described in Section 2.1).

In recent years, many studies have trained neural networks to classify galaxies by morphology in general (e.g., S. Dieleman et al. 2015; M. Huertas-Company et al. 2018, 2019, 2024; P. H. Barchi et al. 2020; T.-Y. Cheng et al. 2020; M. Walmsley et al. 2022; M. Huertas-Company & F. Lanusse 2023) or to identify mergers specifically (e.g., S. Ackermann et al. 2018; C. Bottrell et al. 2019; W. J. Pearson et al. 2019a, 2019b; A. Čiprijanović et al. 2020, 2021; L. Ferreira et al. 2020, 2022; R. W. Bickley et al. 2021). However, most studies regarding merger identification train and test on galaxy samples at low redshift. At high redshift, W. J. Pearson et al. (2019a), A. Čiprijanović et al. (2020, 2021), and L. Ferreira et al. (2020, 2022) each demonstrate that convolutional neural networks show promise for accurately identifying high-redshift mergers. In particular, A. Čiprijanović et al. (2020) apply their convolutional neural network (“DeepMerge”) to $z = 2$ simulated Illustris-1 HST F814W and F160W images created in G. F. Snyder et al. (2019) and achieve an accuracy of 76% on the merger class for noisy data.

In this paper, we train DeepMerge³⁰ to identify galaxy mergers from simulated CEERS images from IllustrisTNG. Then we explore using our trained random forest from C. Rose et al. (2023) and DeepMerge algorithms to identify merging galaxies in real CEERS imaging. In Section 2, we describe both the simulated JWST CEERS images and observed images and how our data sets were curated. In Section 3, we describe training DeepMerge, as well as measuring quantitative parameters for observed galaxies. In Sections 4 and 5, we present our DeepMerge results on simulated data, then show our attempts to identify observed mergers using both DeepMerge and random forests. We summarize and conclude in Section 6.

2. Data

2.1. CEERS Observations and Data Reduction

CEERS (S. L. Finkelstein et al. 2022) is a JWST Early Release Science program (Proposal ID #1345) that observed the EGS deep field in 10 NIRCам pointings. The first four NIRCам pointings—hereafter CEERS1, CEERS2, CEERS3, and CEERS6—were taken on 2022 June 21 in seven filters (F115W, F150W, F200W, F277W, F356W, F410M, and F444W) with a typical exposure time of $2835 \text{ s exposure}^{-1}$. We use these four pointings in this work. A full description of CEERS and the data reduction steps can be found in S. L. Finkelstein et al. (2022) and M. B. Bagley et al. (2023),

³⁰ <https://github.com/deepskies/deepmerge-public>

respectively. To summarize, the CEERS NIRCам images were reduced with version 1.7.2 of the JWST Calibration Pipeline³¹ (H. Bushouse et al. 2022) with custom modifications and current NIRCам reference files.³² The images were processed through stages 1 and 2 of the pipeline, where reduction steps included detector-level corrections; removal of $1/f$ noise, wisps, and snowballs; and astrometric calibration. The images were distortion-corrected and combined into final mosaics in stage 3 of the JWST pipeline using the drizzle algorithm with inverse variance map weighting (S. Casertano et al. 2000; A. S. Fruchter & R. N. Hook 2002). The final mosaics for each pointing in each filter are background-subtracted and have pixel scales of $0''.03 \text{ pixel}^{-1}$.³³

In this work, we use the sample of 850 galaxies at $z > 3$ from the four CEERS pointings that were selected by J. S. Kartaltepe et al. (2023) as well as their visual classifications (described in Section 3.1).

2.2. Simulated CEERS Images

The simulated mock CEERS images used in this work were created by C. Rose et al. (2023). The images were constructed from IllustrisTNG100-1 (F. Marinacci et al. 2018; J. P. Naiman et al. 2018; D. Nelson et al. 2018; A. Pillepich et al. 2018; V. Springel et al. 2018) and the Santa Cruz SAM light cone (R. S. Somerville et al. 2021; L. Y. A. Yung et al. 2022). We refer the reader to C. Rose et al. (2023) for further details on the creation of the pristine, noiseless images. The images are available in six filters (F115W, F150W, F200W, F277W, F356W, and F444W) and cover $\sim 100 \text{ arcmin}^2$ (with pixel scales of $0''.03 \text{ pixel}^{-1}$) and contain over 100,000 galaxies to mimic the size of the complete CEERS mosaic of the EGS field. The images were convolved with model point-spread functions (PSFs), then noise was added on top to mimic CEERS noise (estimated using the JWST Exposure Time Calculator; K. M. Pontoppidan et al. 2016).³⁴

These images are accompanied by catalogs with simulation information such as redshift, star formation rate, and stellar mass. Additionally, the merger history catalogs for IllustrisTNG galaxies (V. Rodriguez-Gomez et al. 2015; D. Nelson et al. 2019) were also available. These catalogs give the IllustrisTNG snapshot numbers for each galaxy's most recent past merger and next future merger (both major, with mass ratio $\leq 4:1$, and minor, with mass ratio between $4:1$ and $10:1$). This merger history information makes IllustrisTNG an ideal data set for training and testing machine learning algorithms to identify galaxy mergers at different stages. From the full simulated mosaic, we obtained morphology measurements for 5438 galaxies with $z > 3$. Figure 1 shows the redshift and mass distribution for these objects in gray. This sample also peaks at $z \sim 3$ and extends out to $z \sim 9$. This sample also contains more low-mass galaxies than in the observed CEERS sample. While A. Pillepich et al. (2018) note that the IllustrisTNG galaxy stellar mass function (GSMF) is in better agreement with observations than the Illustris GSMF, especially at $z \leq 1$, IllustrisTNG may still overproduce low-mass galaxies at higher redshifts. Stellar masses for the observed CEERS galaxies were

computed by J. S. Kartaltepe et al. (2023) using the EGS photometry (M. Stefanon et al. 2017) from CANDELS (N. A. Grogin et al. 2011; A. M. Koekemoer et al. 2011) and estimated redshifts (D. Kodra et al. 2023).

It is important to note the downside of training machine learning models on simulated data: that simulated data may not perfectly mimic reality. As A. Ćiprijanović et al. (2020) note, there are several ways in which simulated data can have persistent differences from real data. These include approximations in physical modeling due to our lack of understanding, approximations or shortcuts used to reduce the computing power and time needed to produce the simulations, and imperfect modeling of noise from the telescope, its instruments, and the sky.

3. Methods

3.1. Visual Classifications of Observed Galaxies

Visual classifications were reported by J. S. Kartaltepe et al. (2023) and were carried out by members of the CEERS team. Images of the galaxies in our sample were hosted on the Zooniverse project builder.³⁵ Volunteer classifiers viewed stamps of the galaxies in seven JWST NIRCам filters and three HST ACS/WFC3 filters, as well as red, green, blue (RGB) stamps and JWST segmentation map stamps (see Figure 14 in the Appendix of J. S. Kartaltepe et al. 2023) from SExtractor (SE v2.23.2; E. Bertin & S. Arnouts 1996). Each object was classified by three volunteer classifiers who followed a morphology classification scheme based on that used in J. S. Kartaltepe et al. (2015). The five tasks in this scheme ask volunteer classifiers to select main morphology classes, interaction classes, and structural and quality flags and, if the classifier wishes, to leave comments about the object. The main morphology class options were disk, spheroid, irregular/peculiar, point source/unresolved, and unclassifiable/junk. There were four interaction class options.

(1) *Merger*. Galaxies (that are single objects) that appear to have experienced a merger due to the presence of structures like tidal tails, double nuclei, or other asymmetries. All mergers should have irregular/peculiar selected as their main classification, but not all galaxies classified as irregulars are mergers.

(2) *Interaction within SE segmentation map*. Two or more galaxies that appear to be interacting within the same segmentation map. There must be clear evidence of an interaction (tidal arms, bridges, etc.).

(3) *Interaction beyond SE segmentation map*. Two or more galaxies that appear to be interacting that have their own distinct segmentation maps.

(4) *Noninteracting companion*. The main galaxy has a nearby companion yet has no evidence of tidal interaction.

The structural and quality flags question included flags to mark galaxies with merger features such as tidal tails and double nuclei. Figure 2 shows examples of galaxies from the different interaction classes.

3.2. Morphology Parameters for Observed Galaxies

SE (E. Bertin & S. Arnouts 1996) was used to detect $z > 3$ galaxies in each of the four pointings. Weighted F150W+F200W+F277W+F356W images were used as the detection images. We ran SE in both “cold” and “hot” mode, in order to

³¹ jwst-pipeline.readthedocs.io

³² jwst-crds.stsci.edu, [jwst 0989.pmap](https://jwst-0989.pmap), [jwst nircam 0232.imap](https://jwst-nircam-0232.imap).

³³ The publicly released images are available at <https://ceers.github.io/releases.html> and on MAST via doi: [10.17909/z7p0-8481](https://doi.org/10.17909/z7p0-8481).

³⁴ The mock CEERS images are publicly available at https://ceers.github.io/ancillary_data.html.

³⁵ <https://www.zooniverse.org/lab>

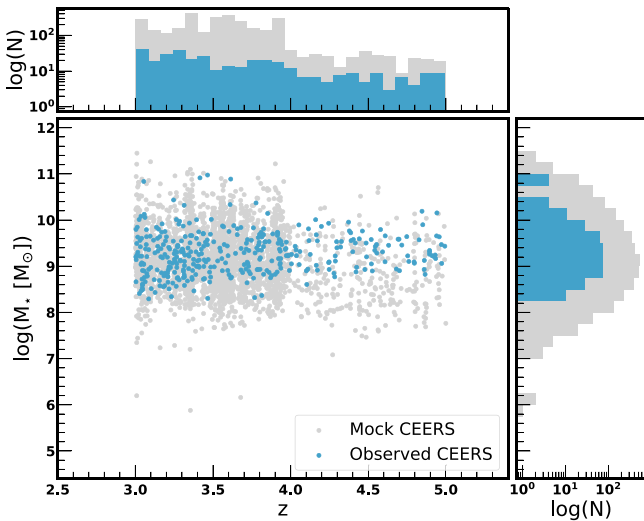


Figure 1. Stellar mass vs. redshift for objects from the mock CEERS sample at $3 < z < 5$ (gray) and from the observed CEERS sample with $3 < z < 5$ (blue). Above and to the right are the distributions of redshift and stellar mass, respectively. These samples do include the S/N and flag cuts discussed in Sections 3.3 and 4.1.

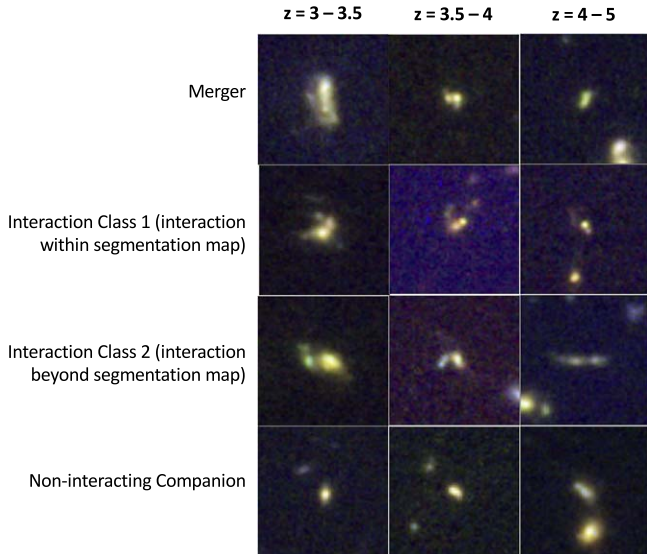


Figure 2. Examples of CEERS galaxies (F150W+F200W+F356W stamps) for each visually classified interaction class at $3.5 < z < 4$, $3.5 < z < 4$, and $4 < z < 5$.

deblend overlapping galaxies and detect faint galaxies, respectively. The “cold” and “hot” detections were then combined following the logic of the Galapagos-2 program (S. P. Bamford et al. 2011; M. Barden et al. 2012; B. Häußler et al. 2013; M. Vika et al. 2013).

Galapagos-2 is an IDL program that runs the multi-wavelength Sérsic fitting program *GalfitM* (C. Y. Peng et al. 2002, 2010; C. Y. Peng 2012) on large survey images. Information from the output SE catalog was used as initial guesses for the Sérsic fits from Galapagos-2. Empirical PSFs made from stacking stars (S. L. Finkelstein et al. 2022) were used for these fits. We ran Galapagos-2 on all seven filters simultaneously to obtain Sérsic indices and related size and magnitude information. In addition to the output Sérsic model catalog, Galapagos-2 also returns input image,

model, and Sérsic residual stamps for each galaxy in the catalog.

We also ran the nonparametric morphology program *statmorph* (V. Rodriguez-Gomez et al. 2019) separately on each filter for each of the four pointings for the $z > 3$ objects in the Galapagos-2 catalog in order to obtain morphology parameters such as concentration (C), asymmetry (A), and clumpiness (S) (M. A. Bershadsky et al. 2000; C. J. Conselice et al. 2000; C. J. Conselice 2003), and the Gini coefficient (G) and the moment of light (M_{20}) (R. G. Abraham et al. 2003; J. M. Lotz et al. 2004). The output SE segmentation map was used as input to *statmorph* to aid in calculating these parameters, as well as the PSFs from S. L. Finkelstein et al. (2022).

Finally, it has been shown that asymmetric or unusual features indicative of mergers can be more prominent in Sérsic residuals (e.g., K. B. Mantha et al. 2019). Therefore, we also run *statmorph* on the Sérsic residuals output by Galapagos-2. To do so, we add an offset of $+1 \text{ counts s}^{-1}$ to every pixel of all residual images, since *statmorph* requires that input images have positive flux, which is not necessarily true for residual images. This offset does affect *statmorph*’s measurements, but the effect will be consistent across all residual images.

We begin with the catalog of visual classifications for 850 objects at $z > 3$ from J. S. Kartaltepe et al. (2023). After combining all of the above catalogs (matching by CANDELS ID number or by R.A. and decl.), we create a sample of 798 observed $z > 3$ CEERS galaxies that have successful morphology measurements from both the *statmorph* and Galapagos catalogs.

3.3. Training Random Forests on Simulated Galaxies

In C. Rose et al. (2023), we trained and tested the random forest algorithm (T. K. Ho 1995; L. Breiman 2001) on the simulated images described in Section 2.2. To do so, we ran Galapagos-2 and *statmorph* on the mock CEERS data in a similar fashion as described in Section 3.2. These morphology parameters were used as the input “features” for the forests, which were the single Sérsic index n , the two-component Sérsic indices n_{bulge} and n_{disk} , and various *statmorph* parameters (outlined in C. Rose et al. 2023) in all six filters for both science and residual images. From the merger history catalogs, we convert the IllustrisTNG snapshot numbers to ages of the Universe in Gyr, then calculate the difference in time between the galaxy of interest and the last and next merger events. The resulting time since the last merger (both major and minor) and the time until the next (both major and minor) are used as the “labels” for the random forests. “Mergers” are objects that had or will have a merging event within $\pm 250 \text{ Myr}$, in accordance with G. F. Snyder et al. (2019), as merger signatures from outside that time frame will have likely vanished. “Nonmergers” are objects that either have a merging event outside that time frame or have not had merging events at all. In C. Rose et al. (2023), we restrict the mock CEERS data set to have *statmorph* signal-to-noise (S/N) per pixel ($S/N_{\text{F150W}} > 3$ and $\text{Flag}_{\text{Galfit}} = 2.0$ (which indicates successful *Galfit* measurements)). Then, we split the final sample of 40,391 mock galaxies into seven redshift bins from $0.5 < z < 4$ and trained a separate forest for each bin, using `BalancedRandomForestClassifier()` from Python’s *imblearn* package as the random forest algorithm

Table 1Number of Simulated CEERS Mergers in Each Redshift Bin after Imposing $S/N > 3$ and $\text{Flag}_{\text{GalFit}} = 2.0$

Redshift Bin	3–3.5	3.5–4	4–5
Total number of galaxies	2039	2367	3105
Number of mergers	668	839	1476

and `keras-tuner` (T. O’Malley et al. 2019) for hyperparameter optimization.

Following the same procedure in C. Rose et al. (2023), we additionally train a random forest for $z = 4\text{--}5$. In Section 4.3, we describe the outcome of training and testing random forests for merger identification on the mock CEERS data set. In Section 4.5, we apply the $3 < z < 3.5$, $3.5 < z < 4$, and $4 < z < 5$ random forests to observed CEERS galaxies.

Additionally, we compute the Spearman correlations between the morphology parameters and the merger timescales using `spearmanr` from the `SciPy` package in order to understand which features might be most important for the random forests. The Spearman correlation describes the relationship between variables, where $+1$ or -1 indicates perfect positive or negative correlations, respectively, and 0 indicates no correlation. For each of the redshift bins studied in this work ($3 < z < 3.5$, $3.5 < z < 4$, and $4 < z < 5$), the strongest correlations ranged from $\sim|0.1|$ to $|0.2|$. For example, the most correlated features for the $4 < z < 5$ bin were F356W concentration with a correlation coefficient of -0.21 , F200W residual flux (within an elliptical aperture) with a correlation coefficient of 0.11 , F444W asymmetry with a correlation coefficient of 0.27 , and F356W residual flux (within an elliptical aperture) with a correlation coefficient of 0.07 .

3.4. Training the DeepMerge Algorithm on Simulated Galaxies

An alternative to using segmentation maps and morphology parameters is to train a machine learning algorithm directly on the image data. Here, we explore using the DeepMerge³⁶ convolutional neural network (A. Čiprijanović et al. 2020) to identify mock CEERS mergers. Convolutional neural networks have special “convolutional” layers that convolve the input image with a kernel in order to extract features from the input image while preserving spatial information from the image. DeepMerge is a relatively simple network containing only three convolutional layers.

As with the random forest analysis, we use the same mock CEERS sample (with $S/N_{\text{F115W}} > 3$ and $\text{Flag}_{\text{GalFit}} = 2.0$). This time, we restrict our mock CEERS sample to three redshift bins at $3 < z < 3.5$, $3.5 < z < 4$, and $4 < z < 5$ (see Table 1). We limit ourselves to these three redshift bins, since we will apply our networks to our observed CEERS galaxies for which visual classifications are only available for $z > 3$. We cut out stamps of size 75 by 75 pixels of each galaxy. The time since last merger and time until next merger were again used to determine the “labels” for DeepMerge.

We use stratified sampling to split our data into training, validation, and test sets. The ratio of the train/validation and test split was 0.18 , which was smaller than the split used for the random forests in C. Rose et al. (2023), in order to divert more examples to the training set. The training and validation split was 0.3 . Due to the low numbers of objects in each redshift bin,

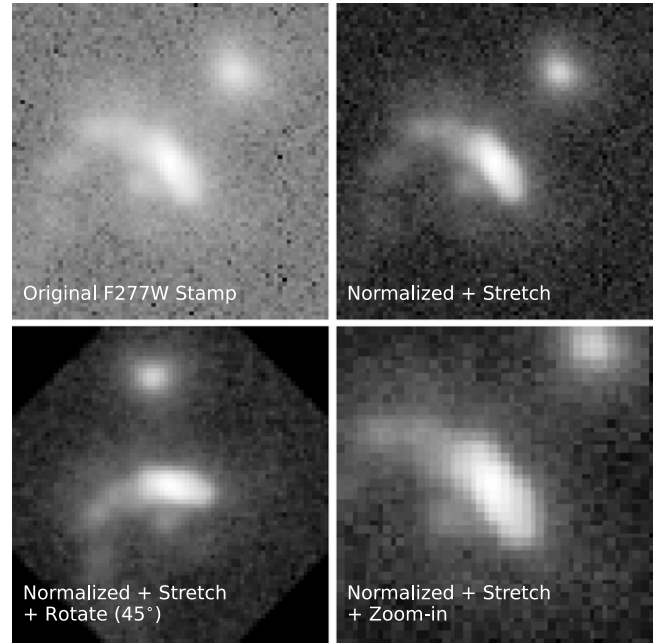


Figure 3. Example of a normalized, stretched, and augmented $3 < z < 3.5$ galaxy merger. For this redshift bin, normalization was based on the F115W filter. After normalizing and stretching, the extended structures of this merger are more distinct from the background.

we employ augmentation to create a larger training set. This included flipping and rotating the images, as well as zooming in and zooming out while maintaining a consistent stamp size of 75 by 75 pixels. For each object, we provide DeepMerge with all six filters (F115W, F150W, F200W, F277W, F356W, and F444W). We also normalize all images based on one of the six filters and then stretch the images using an `arcsinh` stretch to emphasize low surface brightness features. Figure 3 shows an example of normalizing, stretching, and augmenting a $3 < z < 3.5$ galaxy merger. Our data sets are imbalanced since there are more nonmergers than mergers. In order to help prevent our classifier from becoming biased toward the nonmerger class, we maintain balanced classes during training by randomly undersampling the nonmerger class, which decreases the size of the nonmerger class. The validation and test sets are left unbalanced to better represent proportions of real-world data.

We start with the architecture of DeepMerge and train using our data. We employ multiple strategies in order to properly train the DeepMerge networks. In addition to normalizing, augmenting, and balancing the training set, we also test the following.

1. Using different ratios for the split between the training, validation, and test sets.
2. Using different subsets of augmentations (e.g., only rotations).
3. Using only three filters, or only one rest-frame filter, rather than all six available filters.
4. Reducing the number of layers within the network, in order to reduce the complexity of the network.
5. Shifting the location of the dropout layers within the network. Dropout layers mask contributions from some neurons to the next layer in an effort to prevent overfitting.

³⁶ <https://github.com/deepskies/deepmerge-public>

6. Using rectified linear unit activation rather than sigmoid activation in the final layer. These activation functions determine how the result of the last layer will be mapped to a class probability.
7. Using different patience values (ranging from 5 to 100) for early stopping,³⁷ which is a method that stops training once a given metric (e.g., validation loss) has stopped improving in order to combat overfitting (e.g., L. Prechelt 2012).
8. Tracking the difference between training and validation accuracy, rather than validation loss, for early stopping.

Most of the above strategies resulted in networks that performed very well in one class but very poorly in the other. We also attempted to narrow our merger definition to include only major mergers, but the small number of labeled mergers in the final data sets (with the $3 < z < 3.5$ bin having only 376 mergers) resulted in networks that classified all objects as nonmergers. We also attempted to include objects with any $\text{Flag}_{\text{Galfit}}$ (about ~ 10 objects in each redshift bin), as this flag only pertains to the quality of the Galapagos/Galfit measurements that are not relevant here, but the inclusion of these objects also did not improve the performance of the networks. The reason for this may be that these objects were typically found in very crowded regions, which may have caused DeepMerge to focus on the surrounding galaxies.

For all redshift bins, normalizing and arcsinh stretching the images resulted in improvements. We use all available filters. For the $3.5 < z < 4$ and $4 < z < 5$ redshift bins, we reduce the number of layers in DeepMerge to two sets of Convolution2D, BatchNormalization, MaxPooling2D, and Dropout layers rather than three sets. For the $3.0 < z < 3.5$ redshift bin, we also reduce the number of layers to two sets of Convolution2D, BatchNormalization, and MaxPooling2D; additionally, we use only one Dropout layer, which we shifted to just before DeepMerge’s Dense layers. We train for 500 epochs with a batch size of 32. For the $3 < z < 3.5$ redshift bin, we normalize with respect to the F115W filter; for the $3.5 < z < 4$ redshift bin, we normalize with respect to the F150W filter; and for the $4 < z < 5$ redshift bin, we again normalize with respect to the F150W filter. For each redshift bin, we perform the following augmentations on the training set: flip up–down, flip left–right, flip up–down and left–right, rotate at angles from 45° to 270° , zoom in by a factor of 1.5, and zoom out by a factor of 0.75. For the $3 < z < 3.5$ bin, for early stopping, we set patience to 50 and monitor the validation loss; for the $3.5 < z < 4$ redshift bin, we set patience to 5 and monitor the ratio between validation and training accuracy so that training stops once this ratio is less than 0.8; for the $4 < z < 5$ redshift bin, we set patience to 50 and monitor the validation loss. It does not appear that training was more difficult for the $3.5 < z < 4$ bin, since some experiments tracking validation loss and using larger patience values yielded similar results. It seems more likely that some difference in the initial random state of the network caused it to slightly favor the strategy tracking validation and training accuracy.

In Section 4.4, we describe the best outcome of training and testing the DeepMerge network for merger identification on the mock CEERS data set at $3 < z < 5$. In Section 4.6, we apply DeepMerge to observed CEERS galaxies.

Table 2

Number of Observed CEERS Mergers in Each Redshift Bin after Imposing $S/N > 3$ and $\text{Flag}_{\text{Galfit}} = 2.0$

Redshift Bin	3–3.5	3.5–4	4–5	>5
Total number of galaxies	183	99	87	43
Group 1 mergers	38	23	15	6
Group 2 mergers	71	43	43	17
Group 3 mergers	28	20	18	11
Group 4 mergers	92	46	55	20

4. Analysis

4.1. Visual Classification Results for Observed CEERS Galaxies

We construct four possible “merger” definitions based on the visual classifications. As described in Section 3.1, classifiers can select a main morphology class, an interaction class, and structure and quality flags. Our different merger definitions are motivated by these different categories.

Group 1 (G1). At least two out of the three classifiers assigned the merger class, interaction within segmentation map class, or interaction beyond segmentation map class.

Group 2 (G2). At least two out of the three classifiers assigned the irregular main morphology class, since many merging galaxies appear irregular, although not all irregular galaxies are mergers.

Group 3 (G3). At least two out of the three classifiers assigned the tidal tails or double nuclei structure flags, which are merger signatures.

Group 4 (G4). One out of the three classifiers assigned the merger class, interaction within segmentation map class, or interaction beyond segmentation map class.

For groups 1, 2, and 3, we examine objects for which at least two out of three classifiers indicated that the object was potentially a merger, which means classifiers were overall more confident that these objects actually are mergers. For group 4, we examine objects where only one classifier indicated that a given object had merger signatures. All objects from group 1 will also be in this group, as well as additional objects for which classifiers were less confident about their merger status.

We start with the 798 objects that result from combining our morphology catalogs in Section 3.2. Since we restricted our mock CEERS data set to have $S/N_{\text{F115W}} > 3$ and $\text{Flag}_{\text{Galfit}} = 2.0$, we do the same to our real CEERS data set. Doing so results in a final sample of 369 galaxies at $3 < z < 5$. Table 2 shows the total number of objects from this restricted real CEERS data set, as well as the number of mergers from each group, in our three redshift bins. In this paper, we focus on merger definitions using groups 1, 2, and 4, since we found that most group 3 objects are also in the other groups. Only two objects were assigned the tidal tails structure flag without also being assigned an interaction class by at least one person.

4.2. Classical Techniques: Performance on Observed CEERS Galaxies

Past studies have often used classical techniques such as G versus M_{20} (J. M. Lotz et al. 2004, 2008) to identify mergers. In such plots, the merger discriminating line is

$$G > -0.14M_{20} + 0.33, \quad (1)$$

³⁷ https://keras.io/api/callbacks/early_stopping/

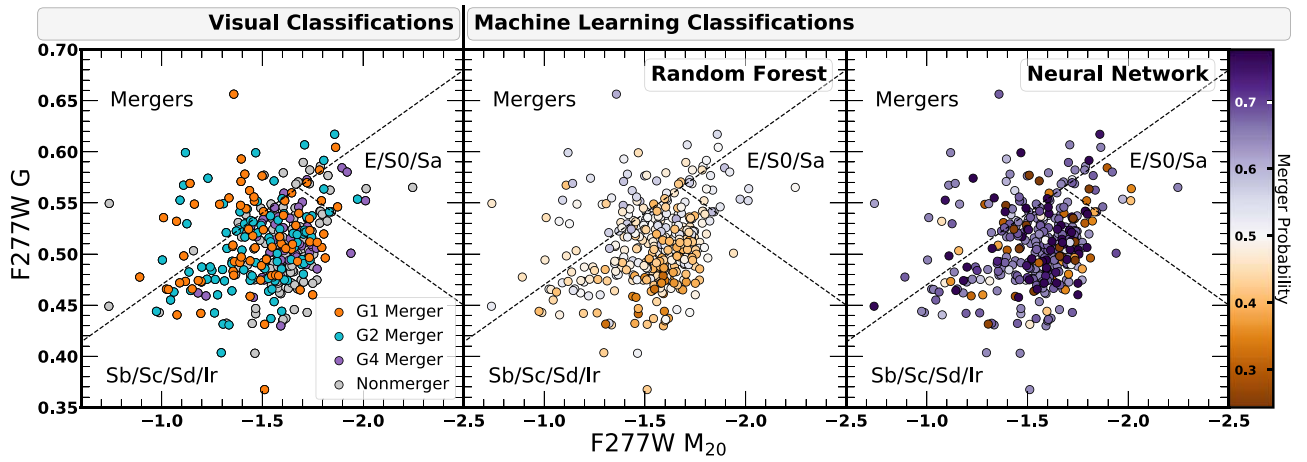


Figure 4. $F277W\ G-M_{20}$ space for observed CEERS galaxies from $3 < z < 5$. In the left panel, observed galaxies are color-coded by visual classifications as defined in Section 4.1 (*group 1/G1*: at least two out of three volunteer classifiers assigned an interaction class; *group 2/G2*: at least two out of three classifiers assigned the Irregular morphology class; *group 4/G4*: one of out three classifiers assigned an interaction class). In the middle and right panels, galaxies are color-coded by the merger probability output by the random forest or neural network, respectively.

defined by J. M. Lotz et al. (2008). The first panel of Figure 4 shows G versus M_{20} for the observed CEERS data set, color-coded by visual classification (groups 1, 2, and 4 have been abbreviated as G1, G2, and G4 in the legend). G versus M_{20} does not appear to effectively separate visually classified mergers and nonmergers, which is to be expected since $G-M_{20}$ is not sensitive to all stages of a merger. The fraction of correctly classified mergers in each group, based on the merger discriminating line, is only 27.6% for group 1, 21.7% for group 2, and 19.2% for group 4. Of the objects in the merger region above the merger discriminating line, only 35.6% are actually group 1 mergers, 57.6% are actually group 2 mergers, and 62.7% are actually group 4 mergers. Most objects tend to fall in the late-type region of the plot. This performance of the $G-M_{20}$ method is similar to that in Figure 6 of C. Rose et al. (2023) for the simulated CEERS data, where only 19%–23% of the simulated mergers were correctly classified, and only 4%–50% of predicted mergers were actually mergers.

$G-M_{20}$ has been shown to be unable to select a pure sample of mergers in other studies. For example, C. J. Conselice et al. (2008) measured G and M_{20} for visually classified HST Ultra Deep Field galaxies in several redshift bins from $0.4 < z < 3$ and find that, besides peculiars/mergers, most other galaxy types were actually also found above the merger discriminating line. In their last redshift bin of $2.5 < z < 3$, they find that $\sim 93\%$ of peculiars are found in the merger region, and of all objects in the merger region, $\sim 67\%$ are peculiars. W. J. Pearson et al. (2019a) plot $G-M_{20}$ for $0.002 < z < 0.15$ KiDS-GAMA galaxies with visual classifications from Galaxy Zoo. They report no specific numbers for the fraction of mergers selected by $G-M_{20}$ but find that most mergers and many nonmergers appear in the merger region. However, they then also include visual classifications from D. W. Darg et al. (2010a, 2010b) and find that most D. W. Darg et al. (2010a, 2010b) mergers lie in the nonmerger region. J. S. Kartaltepe et al. (2010) plot $G-M_{20}$ for $0.01 < z < 3.5$ visually classified COSMOS galaxies and also do not report specific numbers but find that most (major and minor) mergers appear in the nonmerger region. They explored several automated methods besides $G-M_{20}$ and find that no technique is sensitive

to all merger stages. G. F. Snyder et al. (2019) determined $G-M_{20}$ for simulated Illustris-1 galaxies, with mergers defined by a timescale of ± 250 Myr, and also did not report specific numbers but found that diagnostics including $G-M_{20}$ select a highly incomplete sample of mergers, as most mergers appeared in the nonmerger region.

The second and third panels show the merger classifications predicted by the random forest (Section 4.5) and the DeepMerge network (Section 4.6), respectively. These panels are both color-coded by the color bar at the far right, which represents the merger probability output by either the random forest or neural network. Probabilities higher than 0.5 mean the object was classified as a merger (and colored purple or light purple). Probabilities lower than 0.5 mean the object was classified as a nonmerger (and colored orange or yellow). See Section 5.1 for further discussion of these two panels.

4.3. Random Forests: Performance on Simulated CEERS Galaxies

Here we summarize the results from training random forests on mock CEERS galaxies from IllustrisTNG in C. Rose et al. (2023). Across all redshift bins, the forests accurately classified about $\sim 60\%$ of merging and nonmerging galaxies (according to our merger definition of ± 250 Myr). The forests performed better than both a random classifier and classical techniques. For lower redshift bins, rest-frame asymmetry features were more important, and for higher redshift bins, rest-frame bulge and clump features were more important to the forests for merger identification. The feature importances were calculated by the random forest and accessed via the `feature_importances_` attribute. The forests misclassified nonmergers that had segmentation maps contaminated with emission from nearby or background galaxies or had merger signatures persisting from outside the chosen merger time frame. The forests also misclassified mergers that appeared visually undisturbed. Finally, the merger fraction and merger rate calculated using random-forest-selected mergers were underestimated compared to previous works that used Illustris-1.

In this work, we additionally trained a random forest for $z = 4-5$. This forest correctly classified 59% of nonmerging galaxies and 67% of merging galaxies. The confusion matrix

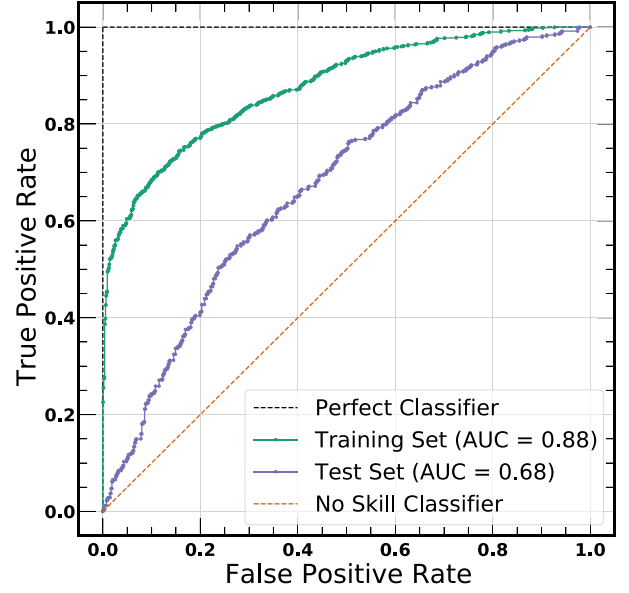
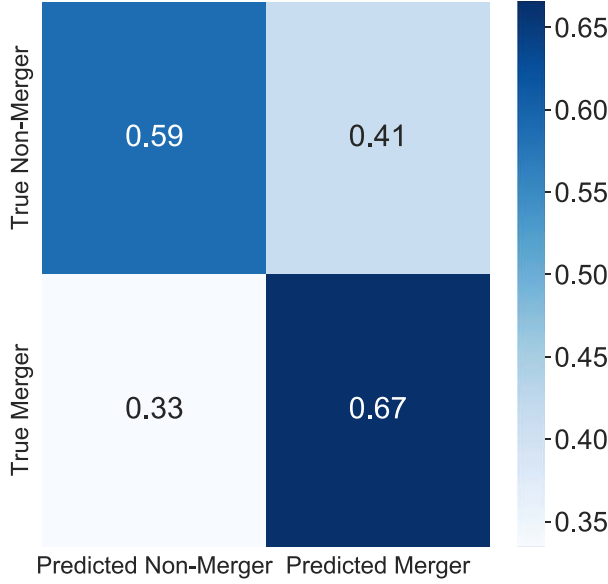


Figure 5. Left: random forest confusion matrix for $4 < z < 5$ objects from the test set. The diagonal shows the fraction of objects correctly classified for each class. Right: ROC curve for the $4 < z < 5$ random forest. The training set (green) and test set (purple) curves lie in the region of “good” classifiers between the perfect (black) and no skill (red) classifiers.

and receiver operating characteristic (ROC) curve are shown in Figure 5, which indicate similar performance to the forests in C. Rose et al. (2023).

4.4. DeepMerge: Performance on Simulated CEERS Galaxies

Figure 6 shows the confusion matrices and ROC curves resulting from applying our trained DeepMerge models to the simulated CEERS test set. Our DeepMerge networks correctly classify $\sim 60\%$ – 70% of merging and nonmerging galaxies. The ROC curves show that the networks perform better than a random classifier. As in C. Rose et al. (2023), we categorize the test set results into four classes.

1. True positives (TPs): the number of true mergers correctly classified by the random forest.
2. False positives (FPs): the number of nonmergers incorrectly classified as mergers.
3. True negatives (TNs): the number of correctly classified nonmergers.
4. False negatives (FNs): the number of true mergers incorrectly classified as nonmergers.

There are several metrics used to quantify the performance of the neural networks.

1. TP rate (TPR), also known as *recall* or *completeness*:

$$\text{TPR} = \text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \quad (2)$$

2. FP rate (FPR), also known as *fall out*:

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}. \quad (3)$$

3. Positive predictive value (PPV), also known as *precision*:

$$\text{PPV} = \text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \quad (4)$$

4. F1 score, the harmonic mean of precision (P) and recall (R):

$$F_1 = 2 \frac{P \times R}{P + R} = \frac{\text{TP}}{\text{TP} + \frac{1}{2}(\text{FP} + \text{FN})}. \quad (5)$$

Since our simulated CEERS data sets are imbalanced, the F1 score is a more useful metric than accuracy for comparing the performance of our models. Table 3 compares the F1 scores of the random forests and DeepMerge networks for the simulated data. Based on the F1 scores, the DeepMerge networks technically performed better than the random forests in each redshift bin, particularly the $3.5 < z < 4$ bin.

As in A. Ćiprijanović et al. (2020), we investigate the use of Gradient-weighted Class Activation Mapping (Grad-CAM) maps (R. R. Selvaraju et al. 2020). These mappings, calculated from the gradient of the predicted class of the input image with respect to the activations in the last convolutional layer, allow us to validate that the network is “looking” at the correct regions in the input image. However, for our complex data with six filters, the Grad-CAMs were not directly interpretable and showed a six-by-six pattern of activations all around the input image. Therefore, we investigate Grad-CAMs from one of the previous tests described in Section 3.4, where we use only one rest-frame filter rather than all six. Figure 7 shows some examples of Grad-CAMs from the $z = 4\text{--}5$ bin trained using only the F356W filter. For this test, the network correctly classified 62% of simulated CEERS nonmergers and 53% of simulated CEERS mergers. As shown in Figure 7, some Grad-CAMs from this F356W filter test show that the network does in fact activate around the edges of the central galaxies and around the edges of neighbors, as we would expect, but other Grad-CAMs only show activations in a few random pixels. The distributions of the F356W magnitude and S/N per pixel indicate that objects with these “bad” Grad-CAMs are not necessarily more likely to have a fainter magnitude and lower S/N. The peak magnitude for objects with “bad” Grad-CAMs

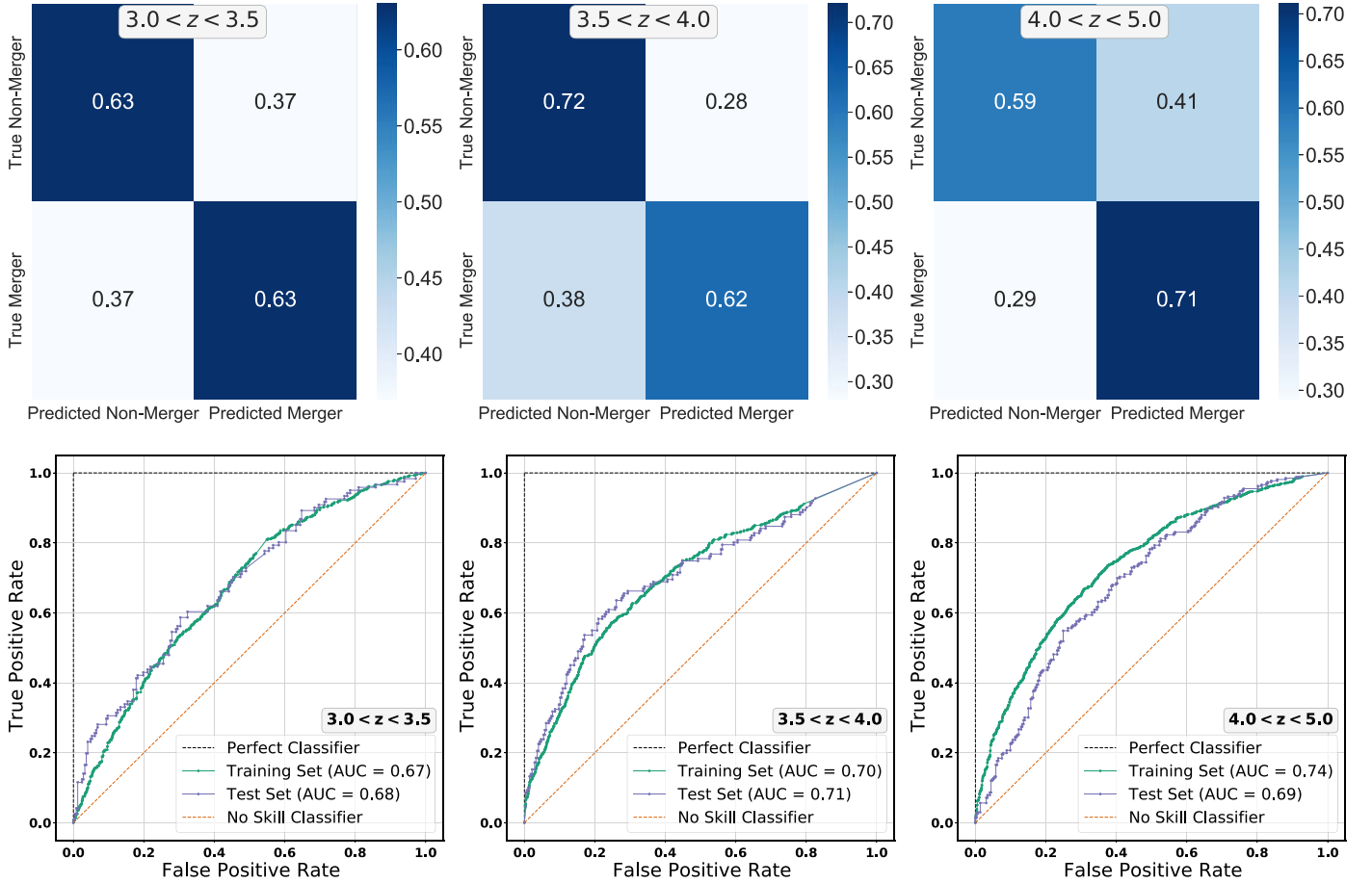


Figure 6. Top: DeepMerge confusion matrices for the simulated CEERS test set for each redshift bin. Bottom: corresponding ROC curves for the simulated CEERS test set for each redshift bin. The training set (green) and test set (purple) curves lie in the region of “good” classifiers between the perfect (black) and no skill (red) classifiers.

Table 3

Comparison of F1 Scores for the Random Forests (RF) and DeepMerge Neural Networks (NN) on the Simulated CEERS Test Sets

Redshift Bin	3–3.5	3.5–4	4–5
RF nonmergers	0.67	0.65	0.62
NN nonmergers	0.70	0.75	0.64
RF mergers	0.52	0.54	0.63
NN mergers	0.53	0.59	0.66

is 28.8 ± 1.2 , while the peak magnitude for objects with “good” Grad-CAMs is 28.1 ± 1.5 . Of objects with “bad” Grad-CAMs, 55.7% were correctly classified (either TPs or TNs), and 44.4% were incorrectly classified (either FPs or FNs). Of the objects with “good” Grad-CAMs, 57.6% were correctly classified and 42.4% were incorrectly classified.

4.5. Random Forests: Performance on Observed CEERS Galaxies

After training on the simulated images, we apply the random forest algorithm to observed CEERS galaxies, selected in Section 4.1, at the following redshift bins: $3 < z < 3.5$, $3.5 < z < 4$, and $4 < z < 5$. Figure 8 shows the confusion matrix for the $3 < z < 3.5$ redshift bin for group 1 (Figure 13 in the Appendix shows all the group 1, 2, and 4 confusion matrices for each redshift bin). Here, we use the visual

classifications as the true merger labels for the comparison. Since the visual classifications are subjective and volunteer classifiers may misclassify objects or disagree on classifications, the true merger labels may not be correct. In summary, the forests generally did not perform as well on the observed galaxies as they did on the simulated galaxies in the test sets. The only data set where the random forest performed better than on the test set was group 1 at $3 < z < 3.5$, where the forest correctly classified 63% of nonmergers and 66% of mergers. The forests performed the worst in the highest redshift bin, where only 25%–30% of mergers were correctly classified. Table 4 shows how the precision, recall, and F1 score change as a function of redshift for these groups.

4.6. DeepMerge: Performance on Observed CEERS Galaxies

After training on the simulated images, we apply our DeepMerge networks to observed CEERS galaxies at the same redshift bins of $3 < z < 3.5$, $3.5 < z < 4$, and $4 < z < 5$. Figure 9 shows the confusion matrix for the $3 < z < 3.5$ redshift bin for group 1 (Figure 14 in the Appendix shows all the group 1, 2, and 4 confusion matrices for each redshift bin). In summary, the networks did not perform as well on the observed galaxies as they did on the simulated galaxies in the test sets. At $3 < z < 3.5$ and $4 < z < 5$, the networks classified ~15%–20% of nonmergers correctly and ~70%–80% of mergers correctly across all visual classification groups. The networks performed the worst at $3.5 < z < 4$, with only 6%–9%

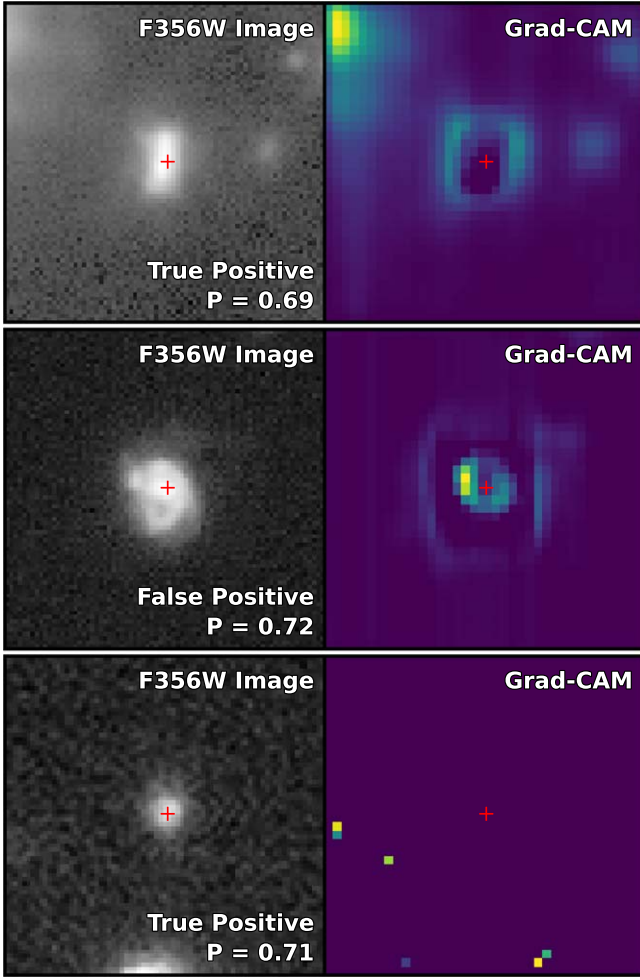


Figure 7. Examples of simulated CEERS galaxies and their corresponding Grad-CAMs from the single-filter DeepMerge test using F356W images. The bottom left corner of the F356W image stamps shows the categorization (e.g., TP, FP, etc.) of this object and the DeepMerge merger probability. “True Positive” indicates that the galaxy was correctly classified as a merger by DeepMerge. “False Positive” indicates that the galaxy was incorrectly classified as a merger by DeepMerge and is actually a nonmerger according to our merger definition from Section 3.3. The red plus signs show the location of the center of each image.

of nonmergers correctly classified and $\sim 85\%$ of mergers correctly classified for all visual classification groups. Table 4 shows how the precision, recall, and F1 score change as a function of redshift for these groups.

Based on the confusion matrices, the DeepMerge networks tend to incorrectly classify more objects as mergers than the random forests. In addition to the strategies employed in Section 3.4, we additionally tried the following strategies during training to prevent underfitting and improve performance on the observed CEERS data set.

1. Giving DeepMerge only one rest-frame filter to train on or summing the six filters together into one image in order to reduce the complexity of the problem.
2. Reducing the number of layers in the network in order to reduce the complexity of the network.
3. Adjusting the probability threshold for merger classification (default 0.5), such that only objects with merger probabilities higher than 0.55, 0.6, or 0.65 were identified as mergers.

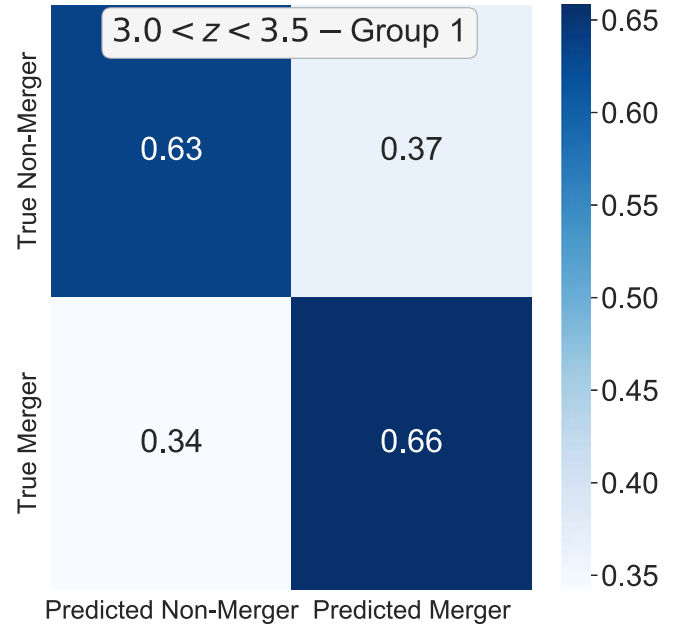


Figure 8. Random forest confusion matrix for observed CEERS galaxies at $3 < z < 3.5$.

The results of these tests did not significantly improve performance on the observed CEERS data set. While some tests did result in confusion matrices that were slightly more balanced, the networks still either classified almost all objects as mergers or almost all objects as nonmergers.

5. Discussion

5.1. Classical Techniques: Observed CEERS Galaxies with Machine Learning Classifications

The second and third panels of Figure 4 show G versus M_{20} for the observed CEERS data set, color-coded by the merger probabilities predicted by the random forests and the DeepMerge network, respectively. Objects were classified as mergers if the merger probability was greater than 0.5 and classified as nonmergers if the merger probability was less than 0.5. As we saw in Section 4.2, the fraction of correctly classified mergers in each group, based on the merger discriminating line, ranged from only 19.2% to 27.6%. Thus, the machine learning algorithms were able to correctly classify more mergers than the $G-M_{20}$ method, and Figure 4 shows that many of these mergers are below the $G-M_{20}$ merger discriminating line, as expected given that $G-M_{20}$ only selects mergers at a narrow portion of the merger phase. Therefore, the algorithms are able to identify mergers that the $G-M_{20}$ method misses due to $G-M_{20}$'s limited sensitivity. A visual inspection of the second panel of Figure 4 reveals a potential trend where objects with very low probabilities from the random forest are far below the merger discriminating line, while objects with higher probabilities are closer to or above the line. A visual inspection of the third panel of Figure 4 shows that most galaxies were classified as mergers with relatively high probabilities, and there appears to be no clear trend in relation to the merger discriminating line.

5.2. Feature Importance

In C. Rose et al. (2023), we calculated the relative importance of the features (morphology parameters) given to

Table 4
Precision, Recall, and F1 Score for Merger Identification of Observed CEERS Galaxies for Both Random Forests and DeepMerge

			$3.0 < z < 3.5$		$3.5 < z < 4$		$4 < z < 5$	
			Nonmerger	Merger	Nonmerger	Merger	Nonmerger	Merger
Random forest	Group 1	Precision	0.88	0.32	0.71	0.19	0.84	0.24
		Recall	0.63	0.66	0.38	0.48	0.82	0.27
		F1 score	0.74	0.43	0.50	0.27	0.83	0.25
	Group 2	Precision	0.64	0.42	0.51	0.40	0.57	0.76
		Recall	0.60	0.46	0.38	0.53	0.91	0.30
		F1 score	0.62	0.44	0.43	0.46	0.70	0.43
	Group 4	Precision	0.59	0.63	0.56	0.48	0.41	0.82
		Recall	0.68	0.53	0.43	0.61	0.91	0.25
		F1 score	0.63	0.58	0.49	0.54	0.57	0.39
DeepMerge network	Group 1	Precision	0.72	0.19	0.64	0.22	0.73	0.14
		Recall	0.16	0.76	0.09	0.83	0.22	0.60
		F1 score	0.26	0.31	0.16	0.34	0.34	0.22
	Group 2	Precision	0.53	0.37	0.45	0.42	0.41	0.46
		Recall	0.15	0.79	0.09	0.86	0.20	0.70
		F1 score	0.24	0.50	0.15	0.56	0.27	0.56
	Group 4	Precision	0.44	0.49	0.27	0.43	0.23	0.58
		Recall	0.15	0.80	0.06	0.83	0.16	0.69
		F1 score	0.23	0.61	0.09	0.57	0.19	0.63

Note. *Group 1 (G1) mergers* are observed galaxies for which at least two out of three volunteer classifiers assigned an interaction class. *Group 2 (G2) mergers* are observed galaxies for which at least two out of three volunteer classifiers assigned the Irregular morphology class. *Group 4 (G4) mergers* are observed galaxies for which at least one out of three volunteer classifiers assigned an interaction class.

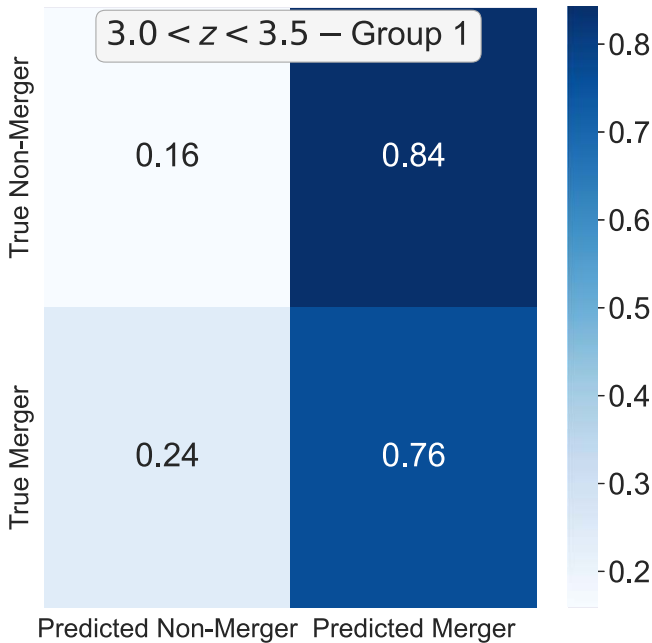


Figure 9. DeepMerge neural network confusion matrix for observed CEERS galaxies at $3 < z < 3.5$.

the random forests, which shows how useful a given feature is for identifying simulated mergers. We found that at lower redshift, bluer filters and asymmetry features were more useful, while at higher redshift, redder filters and bulge and clump features were more useful. Here we calculate feature importance for the $4 < z < 5$ random forest for simulated galaxies, which are (in order of importance)

1. F277W multimode (M)—0.030,
2. F444W moment of light (M_{20})—0.029,
3. F356W Gini (G)—0.029,
4. F444W concentration (C)—0.028, and
5. F356W concentration (C)—0.024.

The numerical importance values are also listed. Note that since the importances are normalized, the numerical values are small due to the many features included. This continues the trend seen in Figure 11 in C. Rose et al. (2023), where redder filter and bulge and clump features are more important at higher redshift.

Of the most correlated features identified by the Spearman correlation in Section 3.3, we see that only one of those features (F356W concentration) was present in the top five most important features calculated by the random forest, indicating that the Spearman correlation is not necessarily a good predictor of the random forest results.

It is important to note that morphology parameters are known to be sensitive to different merger stages. For example, G. F. Snyder et al. (2015) find that, when simulating a galaxy as observed with HST from $z \sim 3$ to $z \sim 1$ (with a merger at $z = 1.6$), the $G-M_{20}$ merger statistic is sensitive to the early phases of the merger, while the MID statistics (multimode, intensity, and deviation; P. E. Freeman et al. 2013) are more sensitive after coalescence. Future studies with larger data sets could benefit from investigating random forest feature importance as a function of merger stage, in addition to redshift.

5.3. Galaxy Categorization Results

Figure 10 shows the categorization of the mock CEERS test set at $3.5 < z < 4$ into TPs, FPs, TNs, and FNs as output by

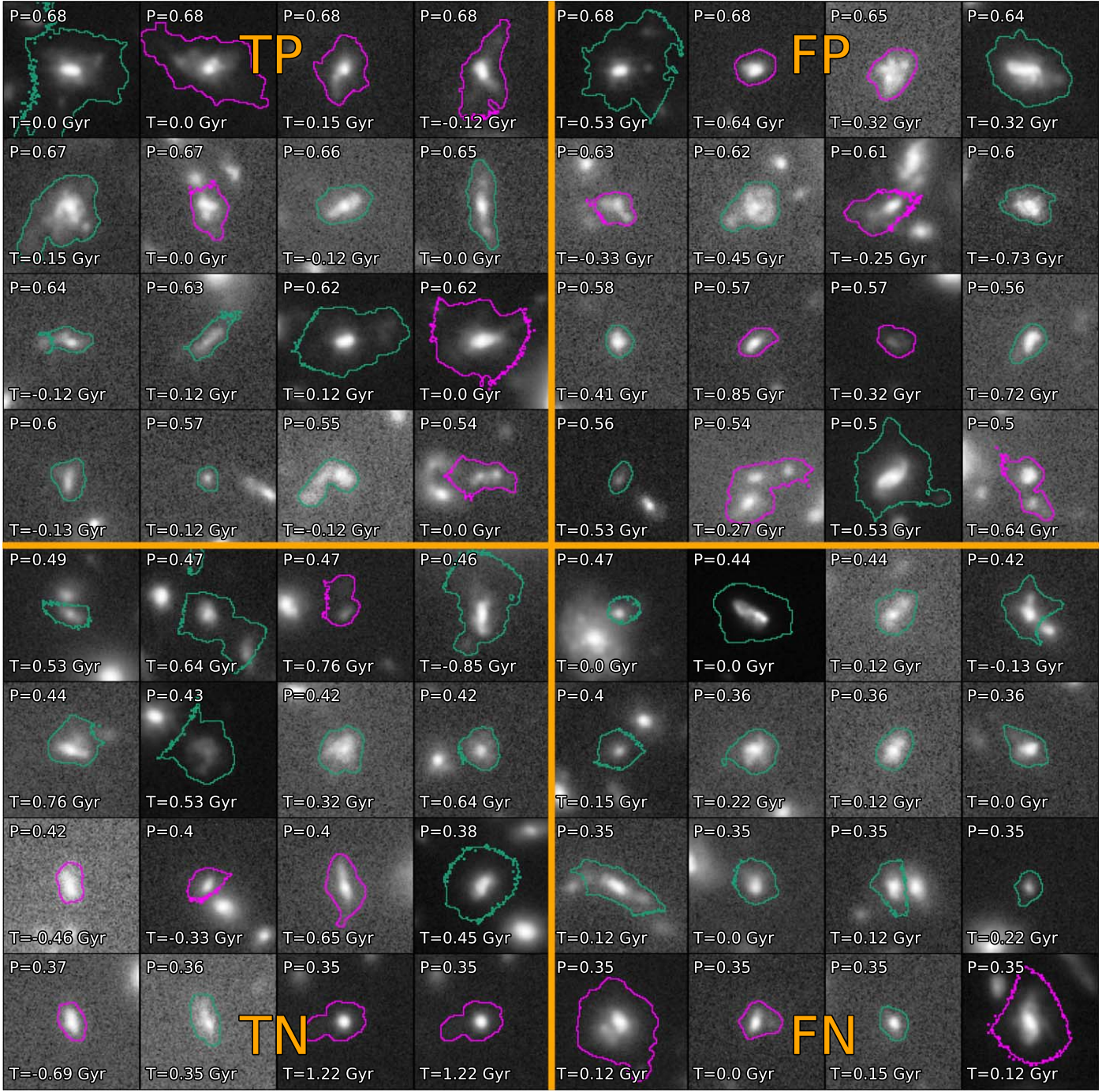


Figure 10. Examples of TP, FP, TN, and FN (as defined in Section 4.4) simulated CEERS galaxies in the F356W filter from the $3.5 < z < 4$ redshift bin. Each stamp is $3'' \times 3''$. Within each stamp, the merger probability output by the DeepMerge network is in the upper left, and the timescale since or until the most recent merger (major or minor) is in the bottom left. The outlines show the segmentation map, color-coded by major (magenta) and minor (green) mergers.

DeepMerge. The top left-hand corner of each stamp gives the probability of the object being a merger as determined by the DeepMerge network. The stamps are arranged in order of decreasing probability, so the horizontal orange line effectively represents the 0.5 probability threshold between merger and nonmerger classifications. In each redshift bin, the probabilities range from ~ 0.3 to 0.7 , which is similar to the range of probabilities output by the random forest.

In Figure 10, the bottom left-hand corner of each stamp shows the merger timescale for each galaxy, which includes both major and minor mergers. A positive timescale indicates

the time since a past merger, and a negative timescale indicates the time until a future merger. Galaxies can have both past and future mergers, so whichever timescale is smallest is shown here. TPs and FNs will have merger timescales of < 0.25 Gyr, while FPs and TNs will have merger timescales of > 0.25 Gyr. Finally, the segmentation map outlines are color-coded by whether the given timescale corresponds to a major or minor merger.

Like the FPs in C. Rose et al. (2023), the FPs in Figure 10 tend to have irregular segmentation maps due to neighboring or background galaxies or asymmetric features that are possibly

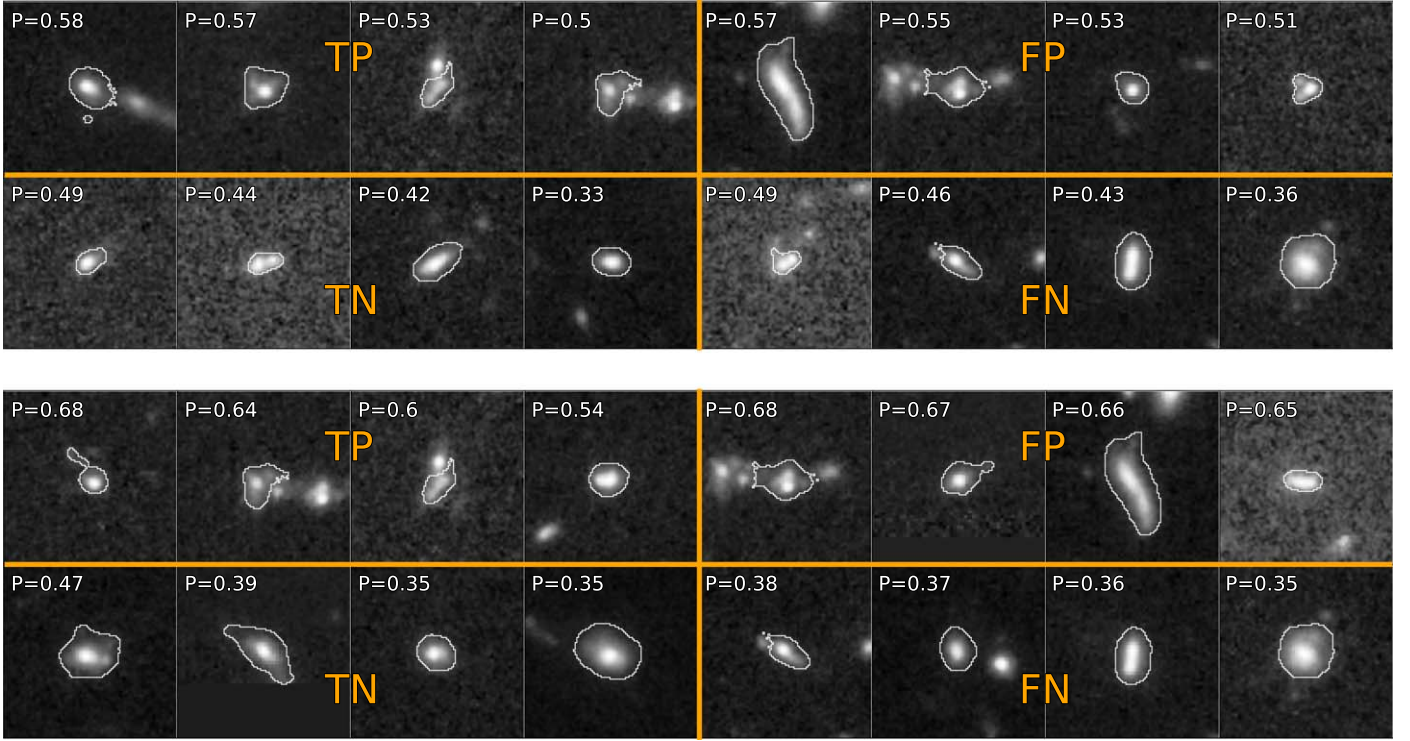


Figure 11. Examples of TP, FP, TN, and FN observed CEERS galaxies (from group 1) in the F356W filter from the $3.5 < z < 4$ redshift bin categorized by the random forest (top) or the DeepMerge network (bottom). Each stamp is $3'' \times 3''$. Within each stamp, the merger probability output by the respective algorithm is in the upper left. The white outlines are segmentation map outlines.

due to a merging event outside of our merger timescale of ± 250 Myr. The average past and future timescales of TNs are 0.72 ± 0.30 Gyr and 0.62 ± 0.24 Gyr, respectively. The average past and future timescales of FPs are 0.49 ± 0.17 Gyr and 0.41 ± 0.19 Gyr, respectively. This indicates that FPs are slightly more likely to be closer in time to a merging event than other nonmergers. On the other hand, the FNs appear less disturbed, possibly due to minor mergers that would have a minor effect on morphology. The fraction of minor mergers among TP and FN in the simulated CEERS test set is 42.9% and 52.4%, respectively, which indicates that FNs are slightly more likely to be minor mergers. This contrasts with the random forest results in C. Rose et al. (2023), where FPs were not more likely to be closer in time to a merging event and FNs were not more likely to be minor mergers.

As with the simulated CEERS images, we also categorize the observed CEERS images into TPs, FPs, TNs, and FNs. Figure 11 shows this categorization resulting from the random forest and the DeepMerge network. In each case, the “true” label is based on group 1 visual classification (whether or not two or more people said the object was a merger or interaction). As with Figure 10, the probability of the object being a merger as determined by each algorithm is given in the upper left-hand corner of each stamp, and the stamps are arranged in decreasing probability.

One reason for the poorer performance of these algorithms on the observed data is that the random forest and the DeepMerge network were both trained on data where the ground-truth labels were determined by the Illustris past or future merger timescales within ± 250 Myr, while the ground-truth labels for the observed CEERS galaxies were based on visual classifications. It is possible that if the algorithms could

have been trained on simulated data with labels based on visual classifications or applied to observed data where the labels were somehow determined by merger timescales, then the algorithms might perform better. Another reason is that the observed galaxies’ visual classifications are subjective and were determined by only three classifiers. In some cases, it appears that a different visual classification could be appropriate. For example, the random forest’s second FP in Figure 11 (or DeepMerge’s first FP in Figure 11) could in fact be a merger. Likewise, DeepMerge’s second FN in Figure 11 could be a nonmerger. If these objects had been assigned the opposite visual classification, then the algorithms would have correctly classified them.

5.4. Merger Rates

Figure 12 shows the simulated CEERS and observed CEERS merger rates. In each panel, the black line is the theoretical Illustris merger rate derived by V. Rodriguez-Gomez et al. (2015) assuming a merger timescale of 0.5 Gyr. The gray stars represent the major merger rate from Q. Duan et al. (2024) using the close pair method on several JWST surveys, including CEERS. The gray line is their best-fit model to the JWST data. The second and third panels show this work’s simulated CEERS merger rate as determined by the random forests (Section 4.3) and the DeepMerge network (Section 4.4), respectively, and are in good agreement with each other.

For the DeepMerge-selected simulated CEERS mergers, we calculate the merger rate in each of the three redshift bins. The merger rate was calculated following the same procedure as in C. Rose et al. (2023), first using the following equation to

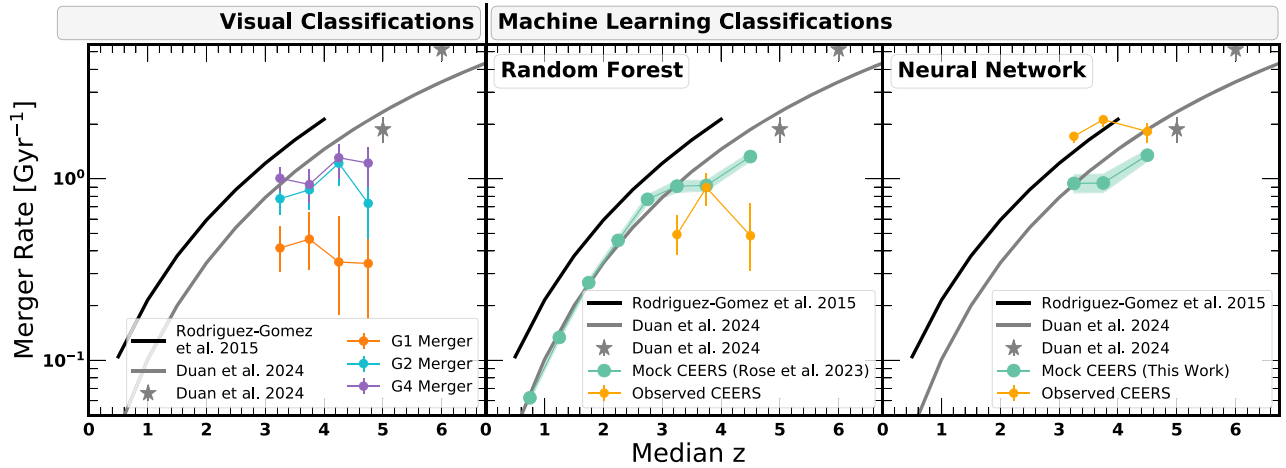


Figure 12. Merger rate based on visual classifications (left panel), random forest classifications (middle panel), and neural network classifications (right panel) as compared to the theoretical Illustris merger rate (black line) derived from V. Rodríguez-Gomez et al. (2015) and the major merger rate (gray line: power-law + exponential model; gray stars: binned JWST data) of Q. Duan et al. (2024) derived from the close pair method. The green shaded regions and the observed galaxy error bars indicate the binomial 95% confidence interval.

calculate the merger fraction:

$$f_{\text{merger}}(\text{NN}) = \frac{N_{\text{NN}}}{N} \frac{\text{PPV}}{\text{TPR}} \left\langle \frac{M}{N} \right\rangle. \quad (6)$$

Then we divide the merger fraction by our merger timescale of 0.5 Gyr to obtain the merger rate. Here, N_{NN} is the number of mergers determined by the DeepMerge network and N is the total number of objects in our sample in each redshift bin. PPV is the positive predictive value, or precision, of the classifier. TPR is the TP rate, or recall, of the classifier. PPV/TPR is meant to correct for the classifier’s known incompleteness and purity based on the training set. $\langle M/N \rangle$ is the average number of merging events per true merger and accounts for the possibility that mergers can experience more than one merging event within the ± 250 Myr timescale used to determine the labels for the neural network. The DeepMerge-simulated CEERS merger rate is shown as the green line in the right panel of Figure 12 and is in good agreement with the random-forest-simulated CEERS merger rate.

The first panel of Figure 12 shows the observed CEERS merger rate for each visual classification group. This was calculated by first determining the merger fraction using simply

$$f_{\text{merger}}(\text{obs}) = \frac{N_{\text{obs}}}{N}. \quad (7)$$

We then divide the merger fraction by our timescale of 0.5 Gyr used for the simulated data. Here, N_{obs} is the number of mergers based on our visual classification groups, and N is the total number of objects in our observed CEERS sample in each redshift bin. The second and third panels also give the observed CEERS merger rate based on the machine learning outputs. These merger rates were calculated as in Equation (6), where PPV/TPR is from the respective machine learning algorithm and $\langle M/N \rangle$ is calculated from the simulated data. Like the simulated merger rates, this work’s observed merger rates in the first and second panels are underestimated compared to the theoretical Illustris merger rate. However, the DeepMerge observed merger rate is higher than the simulated CEERS merger rates, the theoretical Illustris merger rate, and the

observed merger rates in the first panel due to the fact that DeepMerge overestimated the number of mergers. The random forest observed merger rate is in better agreement with the observed merger rates in the first panel. In addition, the simulated merger rates continue to increase at $z > 4$ (in agreement with Q. Duan et al. 2024), while this work’s observed merger rates appear relatively flat. This would contradict the observed merger rate of Q. Duan et al. (2024), which increased out to $z \sim 6$.

5.5. Comparison with Previous Works

As discussed in Section 1, few studies apply convolutional neural networks to the task of merger identification at high redshift. We compare our work to W. J. Pearson et al. (2019a), A. Čiprijanović et al. (2020), and L. Ferreira et al. (2020). The DeepMerge network is introduced in A. Čiprijanović et al. (2020) to identify $z = 2$ simulated Illustris-1 HST mergers. They have access to four different “camera” perspectives, which they use to create a sample of 8930 unique galaxies. They then augment the merger class to create a more balanced sample of 8120 mergers and 7306 nonmergers before they split their data into train and test sets. Their merger class consists of mergers with a stellar mass ratio of at least 0.1 and within a time frame of ± 250 Myr. They find that DeepMerge achieves an accuracy of 76% when training and testing on noisy data. They report correctly classifying 78% of mergers and 73% of nonmergers in their noisy test set. In A. Čiprijanović et al. (2021), they use domain adaptation techniques to apply their trained network to real data; however, this is done at $0.005 < z < 0.1$ using data from the Galaxy Zoo project (C. J. Lintott et al. 2008, 2011).

L. Ferreira et al. (2020) use simulated IllustrisTNG300-1 CANDELS major mergers (up to $z \sim 3$) within ± 0.3 Gyr and nonmergers. Additionally, they impose a limit on the separation between each pair of galaxies. With these criteria, they curate a sample of $\sim 30,000$ distinct major merger candidates. For the nonmergers, they impose a timescale limit of ± 0.5 Gyr (i.e., the galaxy must have a merger timescale *outside* this time frame). Using these criteria, they build up a random sample of nonmergers that equals the number of mergers in each redshift

bin. They train two classifiers—one to identify mergers versus nonmergers and a second to further subdivide mergers into pre- and postmergers—and are able to correctly classify 87% and 94% of pre- and postmergers, respectively. They apply their methodology to a sample of 3759 real CANDELS galaxies with visual classifications given by J. S. Kartaltepe et al. (2015) but do not report an accuracy for this test. They do note that since their validation sample is balanced while the CANDELS sample is not, the performance of the model on simulated data will not translate to the same performance of the model on real data.

Finally, W. J. Pearson et al. (2019a) both train and test their convolutional neural network on observed CANDELS galaxies up to $z < 4$ in the F814W, F125W, and F160W bands using visual classifications from J. S. Kartaltepe et al. (2015). They construct training samples of 694 objects in each class in each of their five redshift bins from $0 < z < 4$. They also augment their images to create a larger training set. They report a final accuracy of 81.8% after testing on the full redshift range of $0 < z < 4$.

Each of these studies reports better performance than we report in this work, which could be due to a number of differences. First, A. Ćiprijanović et al. (2020) and L. Ferreira et al. (2020) use Illustris-1 and IllustrisTNG300-1, respectively, while we use IllustrisTNG100-1. L. Ferreira et al. (2020) impose additional criteria to define their merger and nonmerger samples. Perhaps most importantly, both A. Ćiprijanović et al. (2020) and L. Ferreira et al. (2020) are able to curate a large training set, even before augmentation. All three studies also use balanced training, validation, and test sets, whereas our validation and test sets are left unbalanced, since real data will be unbalanced. Additionally, A. Ćiprijanović et al. (2020) augment their data prior to splitting their data into training, validation, and test sets, and it is unclear if they ensure that all augmentations of a given galaxy end up in the same subset. Otherwise, the training, validation, and test sets would be correlated, which could inflate the accuracy of the network. L. Ferreira et al. (2020) do not specifically comment on the performance of their first classifier, which separates mergers from nonmergers. W. J. Pearson et al. (2019a) do not specifically report performance on their nonmerger class. Thus, these papers may have focused on optimizing performance to correctly identify most mergers, rather than focusing on the purity of the merger and nonmerger classifications, similarly to G. F. Snyder et al. (2019). Finally, A. Ćiprijanović et al. (2020) and L. Ferreira et al. (2020) study galaxies at $z < 3$ with HST images, while we push to $3 < z < 5$ with JWST. W. J. Pearson et al. (2019a) do extend out to $z = 4$, but their training set consists of data from the full range of $0 < z < 4$.

6. Conclusions

In this work, we investigate using the DeepMerge convolutional neural network algorithm to identify merging galaxies at $3 < z < 5$ in simulated CEERS images constructed from IllustrisTNG and the Santa Cruz SAM. Additionally, we calculate morphology parameters for observed CEERS galaxies using the morphology programs *Galapagos-2* and *statmorph*. We apply the trained random forests from C. Rose et al. (2023), a $4 < z < 5$ random forest trained in this work, and the DeepMerge neural network trained in this work to observed CEERS galaxies and verify using visual

classifications from J. S. Kartaltepe et al. (2023). We find the following results.

1. The $4 < z < 5$ random forest performed similarly to those in C. Rose et al. (2023), correctly classifying 59% of nonmerging galaxies and 67% of merging galaxies.
2. The DeepMerge network correctly classified $\sim 60\%$ – 70% of simulated merging and nonmerging galaxies. Although Grad-CAMs from our tests using six filters were not clearly interpretable, Grad-CAMs from our tests using one filter showed that the networks usually activated around the edges of galaxies, as expected.
3. The random forests and neural networks did not perform as well when applied to real CEERS galaxies. In particular, the neural network tended to classify most objects (60%–80%) as mergers. This could be due to a number of reasons, such as the different methods of defining mergers for the simulated data and the JWST CEERS observations.
4. As in C. Rose et al. (2023), DeepMerge FPs tend to appear to have close neighbors or asymmetric features, and FNs appear less disturbed. The FPs appear to be more likely to be closer in time to merging events than the TNs, and FNs appear more likely to be minor mergers than TPs. The categorization of observed CEERS galaxies shows that positive classifications do look more irregular and asymmetric and negative classifications do look more undisturbed, but also that visual classifications are subjective.
5. Merger rates from (a) observed visual classifications only, (b) random-forest-selected observed merger classifications, and (c) DeepMerge-selected simulated merger classifications generally agree with each other and with the random-forest-selected simulated merger rate from C. Rose et al. (2023). The merger rate calculated using DeepMerge-selected observed merger classifications is comparatively overestimated due to the fact that DeepMerge misclassified almost all observed nonmergers.

6.1. Future Work

One clear area of improvement is curating a larger sample of simulated mergers in order to increase the number of objects in the training set, which could help the network better learn merger signatures. This could be done by remaking the pristine simulated images using the larger IllustrisTNG300-1 simulation (as in L. Ferreira et al. 2020) or using another large-volume, high-resolution simulation. A larger sample would also enable the performance as a function of other physical parameters that can impact the observed merger signatures, such as the gas fraction of the galaxies, their merger ratio, and their relative orbits. For Illustris300-1, the trade-off is that Illustris300-1 has a lower resolution than Illustris100-1, so additional study would be needed to understand the effect of the lower resolution on the visibility of merger features. Another potential route to creating larger training sets is to use deep learning models such as generative adversarial networks (I. Goodfellow et al. 2014) to generate more simulated galaxies, as in M. J. Smith & J. E. Geach (2019).

Another area of improvement would be investigating the use of transfer learning techniques to improve performance on the observed simulated CEERS data. Domain transfer could help mitigate the differences between simulated data and real

observations, such as the issues mentioned in Section 2.2. A. Ćiprijanović et al. (2021) show that using transfer learning techniques improves the performance of their simulation-trained networks on observed data by up to $\sim 20\%$. Finally, it would be interesting to train and test machine learning techniques directly on observed CEERS galaxies (similar to W. J. Pearson et al. 2019a) once more visual classifications become available. It would also be interesting to compare the Grad-CAM maps to the regions of the images for which the morphology parameters are sensitive in order to better understand the features that the neural network has learned.

Acknowledgments

Support for this work was provided by NASA through grants JWST-ERS-01345.015-A and HST-AR-15802.001-A awarded by the Space Telescope Science Institute, which is operated by the Association of Universities for Research in Astronomy, Inc., under NASA contract NAS 5-26555. This research is based in part on observations made with the NASA/ESA Hubble Space Telescope obtained from the Space Telescope Science Institute, which is operated by the Association of Universities for Research in Astronomy, Inc., under NASA contract NAS 5-26555. This work is based in part on observations made with the NASA/ESA/CSA James Webb Space Telescope. The data were obtained from the Mikulski Archive for Space Telescopes at the Space Telescope Science

Institute, which is operated by the Association of Universities for Research in Astronomy, Inc., under NASA contract NAS 5-03127 for JWST. These observations are associated with program #1345.

The authors acknowledge Research Computing at the Rochester Institute of Technology for providing computational resources and support that have contributed to the research results reported in this publication (Rochester Institute of Technology 2023).

The authors acknowledge the Texas Advanced Computing Center (TACC) at the University of Texas at Austin for providing HPC resources that have contributed to the research results reported within this paper (<http://www.tacc.utexas.edu>).

Facility: JWST (NIRCam).

Software: SExtractor (E. Bertin & S. Arnouts 1996), Galapagos-2 (B. Häußler et al. 2013), statmorph (V. Rodriguez-Gomez et al. 2019), DeepMerge (A. Ćiprijanović et al. 2020).

Appendix

Confusion Matrices for Observed CEERS Galaxies

Figure 13 shows the results of applying the trained random forests to observed CEERS galaxies in Section 4.5. Figure 14 shows the results of applying the trained DeepMerge neural networks to observed CEERS galaxies in Section 4.6.

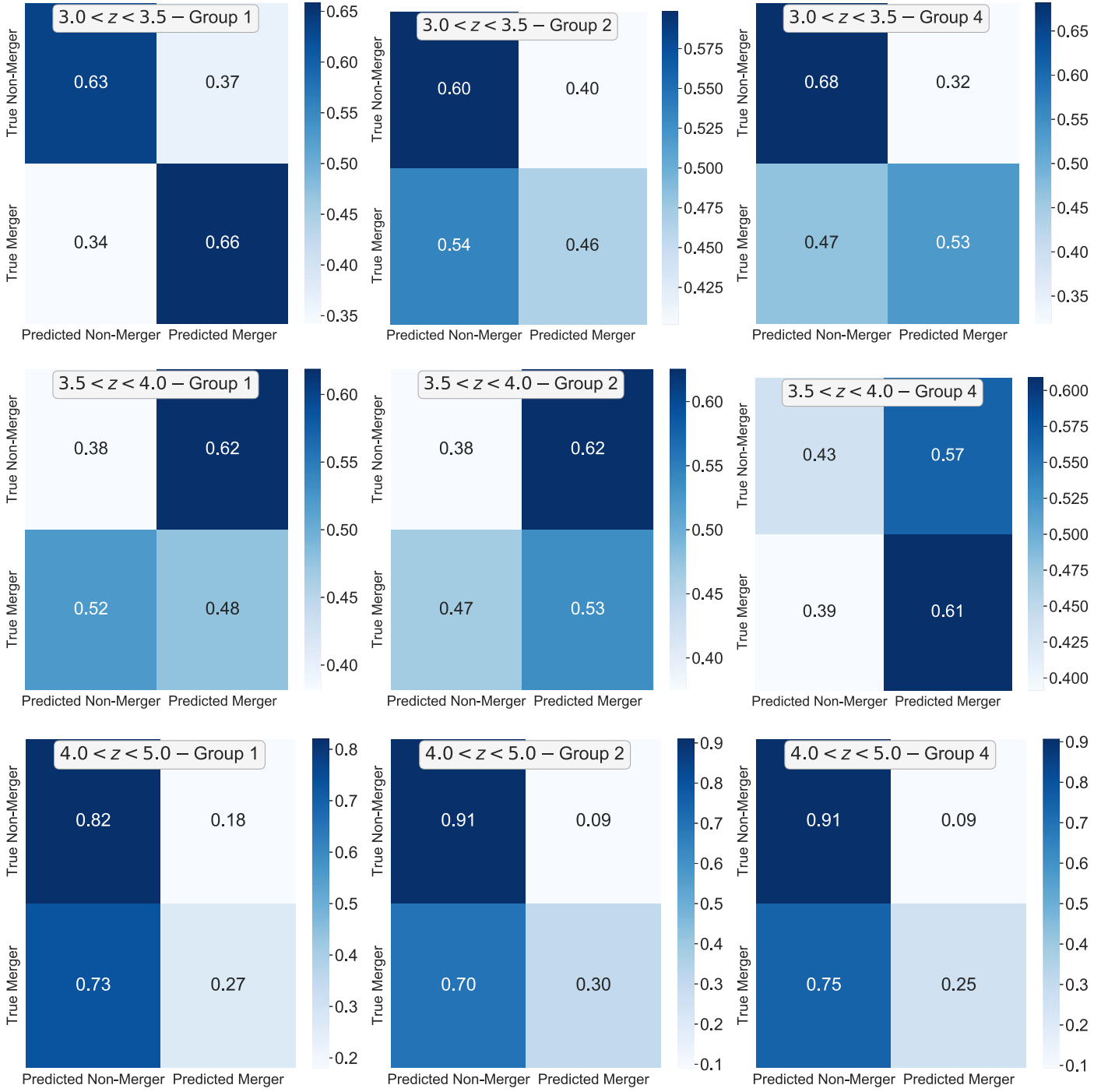


Figure 13. Random forest confusion matrices for observed CEERS galaxies at $3 < z < 3.5$ (top row), $3.5 < z < 4$ (middle row), and $4 < z < 5$ (bottom row).

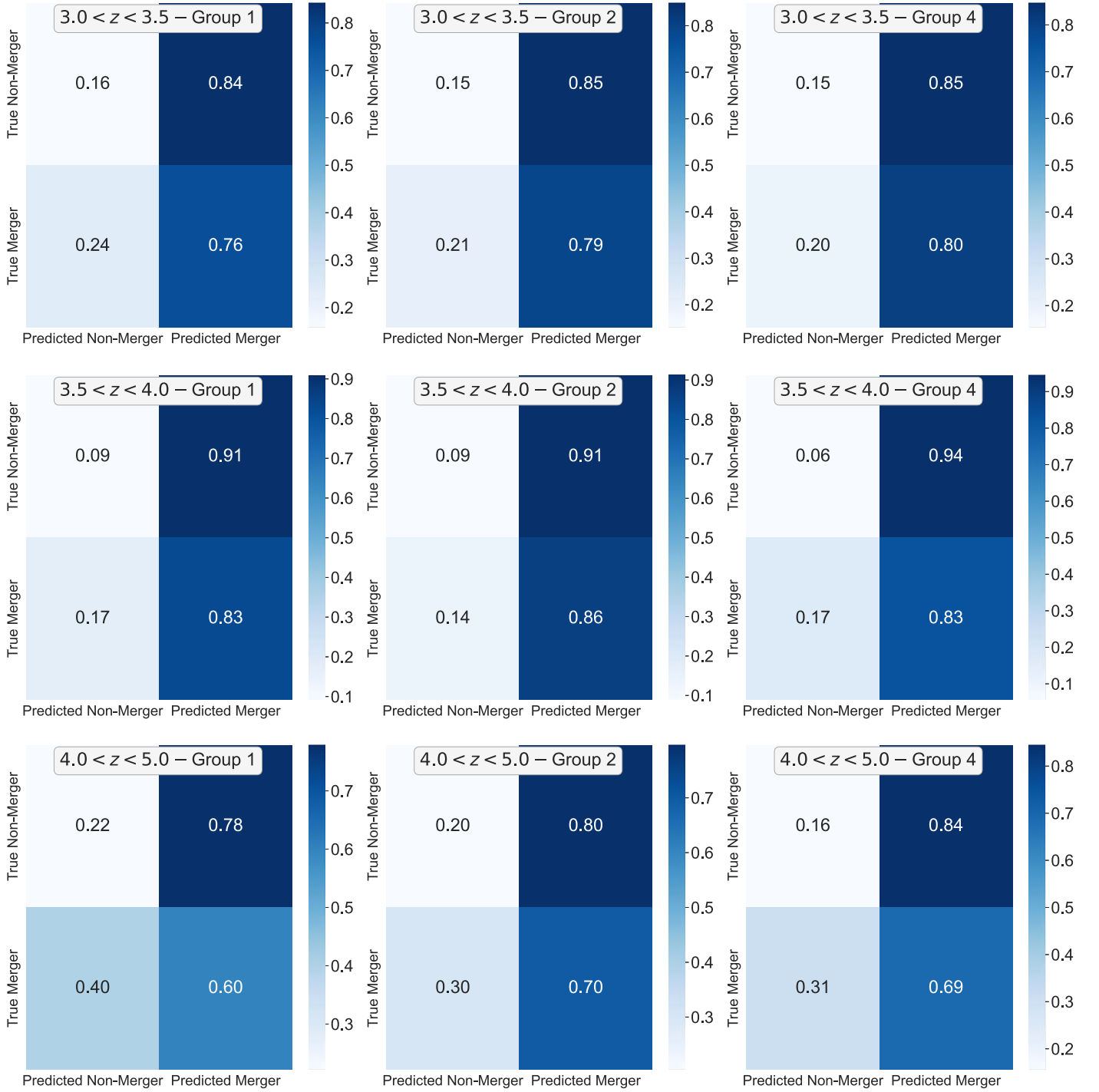


Figure 14. DeepMerge neural network confusion matrices for observed CEERS galaxies at $3 < z < 3.5$ (top row), $3.5 < z < 4$ (middle row), and $4 < z < 5$ (bottom row).

ORCID iDs

Caitlin Rose <https://orcid.org/0000-0002-8018-3219>
 Jeyhan S. Kartaltepe <https://orcid.org/0000-0001-9187-3605>
 Gregory F. Snyder <https://orcid.org/0000-0002-4226-304X>
 Marc Huertas-Company <https://orcid.org/0000-0002-1416-8483>
 L. Y. Aaron Yung <https://orcid.org/0000-0003-3466-035X>
 Pablo Arrabal Haro <https://orcid.org/0000-0002-7959-8783>
 Micaela B. Bagley <https://orcid.org/0000-0002-9921-9218>
 Laura Bisigello <https://orcid.org/0000-0003-0492-4924>

Antonello Calabrò <https://orcid.org/0000-0003-2536-1614>
 Nikko J. Cleri <https://orcid.org/0000-0001-7151-009X>
 Mark Dickinson <https://orcid.org/0000-0001-5414-5131>
 Henry C. Ferguson <https://orcid.org/0000-0001-7113-2738>
 Steven L. Finkelstein <https://orcid.org/0000-0001-8519-1130>
 Adriano Fontana <https://orcid.org/0000-0003-3820-2823>
 Andrea Grazian <https://orcid.org/0000-0002-5688-0663>
 Norman A. Grogin <https://orcid.org/0000-0001-9440-8872>
 Benne W. Holwerda <https://orcid.org/0000-0002-4884-6756>
 Kartheik G. Iyer <https://orcid.org/0000-0001-9298-3523>
 Lisa J. Kewley <https://orcid.org/0000-0001-8152-3943>

Allison Kirkpatrick  <https://orcid.org/0000-0002-5537-8110>
 Dale D. Kocevski  <https://orcid.org/0000-0002-8360-3880>
 Anton M. Koekemoer  <https://orcid.org/0000-0002-6610-2048>
 Jennifer M. Lotz  <https://orcid.org/0000-0003-3130-5643>
 Ray A. Lucas  <https://orcid.org/0000-0003-1581-7825>
 Lorenzo Napolitano  <https://orcid.org/0000-0002-8951-4408>
 Casey Papovich  <https://orcid.org/0000-0001-7503-8482>
 Laura Pentericci  <https://orcid.org/0000-0001-8940-6768>
 Pablo G. Pérez-González  <https://orcid.org/0000-0003-4528-5639>
 Nor Pirzkal  <https://orcid.org/0000-0003-3382-5941>
 Swara Ravindranath  <https://orcid.org/0000-0002-5269-6527>
 Rachel S. Somerville  <https://orcid.org/0000-0002-6748-6821>
 Amber N. Straughn  <https://orcid.org/0000-0002-4772-7878>
 Jonathan R. Trump  <https://orcid.org/0000-0002-1410-0470>
 Stephen M. Wilkins  <https://orcid.org/0000-0003-3903-6935>
 Guang Yang  <https://orcid.org/0000-0001-8835-7722>

References

- 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
- Bagley, M. B., Finkelstein, S. L., Koekemoer, A. M., et al. 2023, *ApJL*, **946**, L12
- Bamford, S. P., Häußler, B., Bamford, A., et al. 2011, in ASP Conf. Ser. 442, Astronomical Data Analysis Software and Systems XX, ed. I. N. Evans et al. (San Francisco, CA: ASP), 479
- Barchi, P. H., de Carvalho, R. R., Rosa, R. R., et al. 2020, *A&C*, **30**, 100334
- Barden, M., Häußler, B., Peng, C. Y., McIntosh, D. H., & Guo, Y. 2012, *MNRAS*, **422**, 449
- Bershady, M. A., Jangren, A., & Conselice, C. J. 2000, *AJ*, **119**, 2645
- Bertin, E., & Arnouts, S. 1996, *A&AS*, **117**, 393
- Bickley, R. W., Bottrell, C., Hani, M. H., et al. 2021, *MNRAS*, **504**, 372
- Bottrell, C., Hani, M. H., Teimoorinia, H., et al. 2019, *MNRAS*, **490**, 5390
- Breiman, L. 2001, *Mach. Learn.*, **45**, 5
- Bushouse, H., Eisenhamer, J., Dencheva, N., et al. 2022, JWST Calibration Pipeline, v1.8.2, Zenodo, doi:10.5281/zenodo.7325378
- Casertano, S., de Mello, D., Dickinson, M., et al. 2000, *AJ*, **120**, 2747
- Cheng, T.-Y., Conselice, C. J., Aragón-Salamanca, A., et al. 2020, *MNRAS*, **493**, 4209
- Ćiprijanović, A., Kafkes, D., Downey, K., et al. 2021, *MNRAS*, **506**, 677
- Ćiprijanović, A., Snyder, G. F., Nord, B., & Peek, J. E. G. 2020, *A&C*, **32**, 100390
- Conselice, C. J. 2003, *ApJS*, **147**, 1
- Conselice, C. J., Bershady, M. A., & Jangren, A. 2000, *ApJ*, **529**, 886
- Conselice, C. J., Rajgor, S., & Myers, R. 2008, *MNRAS*, **386**, 909
- Costantin, L., Pérez-González, P. G., Méndez-Abreu, J., et al. 2021, *ApJ*, **913**, 125
- Costantin, L., Pérez-González, P. G., Méndez-Abreu, J., et al. 2022, *ApJ*, **929**, 121
- Cox, T. J., Dutta, S. N., Di Matteo, T., et al. 2006, *ApJ*, **650**, 791
- Darg, D. W., Kaviraj, S., Lintott, C. J., et al. 2010a, *MNRAS*, **401**, 1043
- Darg, D. W., Kaviraj, S., Lintott, C. J., et al. 2010b, *MNRAS*, **401**, 1552
- Davis, M., Guhathakurta, P., Konidaris, N. P., et al. 2007, *ApJL*, **660**, L1
- Dekel, A., Sari, R., & Ceverino, D. 2009, *ApJ*, **703**, 785
- Dieleman, S., Willett, K. W., & Dambre, J. 2015, *MNRAS*, **450**, 1441
- Duan, Q., Conselice, C. J., Li, Q., et al. 2024, arXiv:2407.09472
- Duncan, K., Conselice, C. J., Mundy, C., et al. 2019, *ApJ*, **876**, 110
- Ellison, S. L., Mendel, J. T., Patton, D. R., & Scudder, J. M. 2013, *MNRAS*, **435**, 3627
- Ellison, S. L., Patton, D. R., Simard, L., & McConnachie, A. W. 2008, *AJ*, **135**, 1877
- Ferreira, L., Conselice, C. J., Duncan, K., et al. 2020, *ApJ*, **895**, 115
- Ferreira, L., Conselice, C. J., Kuchner, U., & Tohill, C.-B. 2022, *ApJ*, **931**, 34
- Finkelstein, S. L., Bagley, M. B., Haro, P. A., et al. 2022, *ApJL*, **940**, L55
- Freeman, P. E., Izbicki, R., Lee, A. B., et al. 2013, *MNRAS*, **434**, 282
- Fruchter, A. S., & Hook, R. N. 2002, *PASP*, **114**, 144
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., et al. 2014, in Advances in Neural Information Processing Systems, Vol. 27, ed. Z. Ghahramani et al. (Red Hook, NY: Curran Associates, Inc.), https://proceedings.neurips.cc/paper_files/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf
- Grogin, N. A., Kocevski, D. D., Faber, S. M., et al. 2011, *ApJS*, **197**, 35
- Häußler, B., Bamford, S. P., Vika, M., et al. 2013, *MNRAS*, **430**, 330
- Ho, T. K. 1995, in Proc. 3rd Int. Conf. on Document Analysis and Recognition, Vol. 1 (Piscataway, NJ: IEEE), 278
- Huertas-Company, M., Iyer, K. G., Angeloudi, E., et al. 2024, *A&A*, **685**, A48
- Huertas-Company, M., & Lanusse, F. 2023, *PASA*, **40**, e001
- Huertas-Company, M., Primack, J. R., Dekel, A., et al. 2018, *ApJ*, **858**, 114
- Huertas-Company, M., Rodriguez-Gomez, V., Nelson, D., et al. 2019, *MNRAS*, **489**, 1859
- Kartalpe, J. S., Dickinson, M., Alexander, D. M., et al. 2012, *ApJ*, **757**, 23
- Kartalpe, J. S., Mozena, M., Kocevski, D., et al. 2015, *ApJS*, **221**, 11
- Kartalpe, J. S., Rose, C., Vanderhoof, B. N., et al. 2023, *ApJL*, **946**, L15
- Kartalpe, J. S., Sanders, D. B., Le Floch, E., et al. 2010, *ApJ*, **721**, 98
- Kartalpe, J. S., Sanders, D. B., Scoville, N. Z., et al. 2007, *ApJS*, **172**, 320
- Kodra, D., Andrews, B. H., Newman, J. A., et al. 2023, *ApJ*, **942**, 36
- Koekemoer, A. M., Faber, S. M., Ferguson, H. C., et al. 2011, *ApJS*, **197**, 36
- Kormendy, J., Fisher, D. B., Comell, M. E., & Bender, R. 2009, *ApJS*, **182**, 216
- Larson, K. L., Sanders, D. B., Barnes, J. E., et al. 2016, *ApJ*, **825**, 128
- Lin, L., Koo, D. C., Willmer, C. N. A., et al. 2004, *ApJL*, **617**, L9
- Lintott, C., Schawinski, K., Bamford, S., et al. 2011, *MNRAS*, **410**, 166
- Lintott, C. J., Schawinski, K., Slosar, A., et al. 2008, *MNRAS*, **389**, 1179
- Lotz, J. M., Davis, M., Faber, S. M., et al. 2008, *ApJ*, **672**, 177
- 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., Primack, J., & Madau, P. 2004, *AJ*, **128**, 163
- Mantha, K. B., McIntosh, D. H., Brennan, R., et al. 2018, *MNRAS*, **475**, 1549
- Mantha, K. B., McIntosh, D. H., Ciaschi, C. P., et al. 2019, *MNRAS*, **486**, 2643
- Marinacci, F., Vogelsberger, M., Pakmor, R., et al. 2018, *MNRAS*, **480**, 5113
- Naiman, J. P., Pillepich, A., Springel, V., et al. 2018, *MNRAS*, **477**, 1206
- Nelson, D., Pillepich, A., Springel, V., et al. 2018, *MNRAS*, **475**, 624
- Nelson, D., Pillepich, A., Springel, V., et al. 2019, *MNRAS*, **490**, 3234
- O'Malley, T., Bursztin, E., Long, J., et al. 2019, Keras Tuner, GitHub, <https://github.com/keras-team/keras-tuner>
- Patton, D. R., Ellison, S. L., Simard, L., McConnachie, A. W., & Mendel, J. T. 2011, *MNRAS*, **412**, 591
- Pearson, W. J., Wang, L., Alpaslan, M., et al. 2019a, *A&A*, **631**, A51
- Pearson, W. J., Wang, L., Trayford, J. W., Petrillo, C. E., & van der Tak, F. F. S. 2019b, *A&A*, **626**, A49
- Peng, C. Y. 2012, Galfit User's Manual, <https://users.obs.carnegiescience.edu/peng/work/galfit/README.pdf>
- Peng, C. Y., Ho, L. C., Impey, C. D., & Rix, H.-W. 2002, *AJ*, **124**, 266
- Peng, C. Y., Ho, L. C., Impey, C. D., & Rix, H.-W. 2010, *AJ*, **139**, 2097
- Pillepich, A., Nelson, D., Hernquist, L., et al. 2018, *MNRAS*, **475**, 648
- Pontoppidan, K. M., Pickering, T. E., Laidler, V. G., et al. 2016, *Proc. SPIE*, **9910**, 991016
- Prechelt, L. 2012, in Early Stopping—But When?, ed. G. Montavon, G. B. Orr, & K.-R. Müller (Berlin: Springer), 53
- Rochester Institute of Technology 2023, Research Computing Services (Rochester, NY: Rochester Institute of Technology)
- Rodriguez-Gomez, V., Genel, S., Vogelsberger, M., et al. 2015, *MNRAS*, **449**, 49
- Rodriguez-Gomez, V., Sales, L. V., Genel, S., et al. 2017, *MNRAS*, **467**, 3083
- Rodriguez-Gomez, V., Snyder, G. F., Lotz, J. M., et al. 2019, *MNRAS*, **483**, 4140
- Rose, C., Kartaltepe, J. S., Snyder, G. F., et al. 2023, *ApJ*, **942**, 54
- Selvaraju, R. R., Cogswell, M., Das, A., et al. 2020, *Int. J. Comput. Vis.*, **128**, 336
- Shah, E. A., Kartaltepe, J. S., Magagnoli, C. T., et al. 2020, *ApJ*, **904**, 107
- Smith, M. J., & Geach, J. E. 2019, *MNRAS*, **490**, 4985
- Snyder, G. F., Lotz, J., Moody, C., et al. 2015, *MNRAS*, **451**, 4290
- Snyder, G. F., Rodriguez-Gomez, V., Lotz, J. M., et al. 2019, *MNRAS*, **486**, 3702
- Somerville, R. S., Olsen, C., Yung, L. Y. A., et al. 2021, *MNRAS*, **502**, 4858
- Springel, V., Pakmor, R., Pillepich, A., et al. 2018, *MNRAS*, **475**, 676
- Stefanon, M., Yan, H., Mobasher, B., et al. 2017, *ApJS*, **229**, 32
- Tacchella, S., Carollo, C. M., Renzini, A., et al. 2015, *Sci*, **348**, 314
- Toomre, A. 1977, in Evolution of Galaxies and Stellar Populations, ed. B. M. Tinsley, D. C. Larson, & R. B. Gehret (New Haven, CT: Yale Univ. Observatory), 401
- Ventou, E., Contini, T., Bouché, N., et al. 2017, *A&A*, **608**, A9
- Vika, M., Bamford, S. P., Häußler, B., et al. 2013, *MNRAS*, **435**, 623
- Walmsley, M., Lintott, C., Geron, T., et al. 2022, *MNRAS*, **509**, 3966
- Yung, L. Y. A., Somerville, R. S., Ferguson, H. C., et al. 2022, *MNRAS*, **515**, 5416
- Zolotov, A., Dekel, A., Mandelker, N., et al. 2015, *MNRAS*, **450**, 2327