Draft version September 14, 2025
Typeset using LaTeX **twocolumn** style in AASTeX63

Beyond the Brightest: A Deep Learning Approach to Identifying Major and Minor Galaxy Mergers in CANDELS at $z \sim 1$

Aimee L. Schechter,[1] Aleksandra Ćiprijanović,[2,3,4] Xuejian Shen,[5,6] Rebecca Nevin,[2] Julia M. Comerford,[1] Aaron Stemo,[7] and Laura Blecha[8]

[1] *Department of Astrophysical and Planetary Sciences, University of Colorado, Boulder, CO 80309, USA*
[2] *Fermi National Accelerator Laboratory, Batavia, IL 60510, USA*
[3] *Department of Astronomy and Astrophysics, University of Chicago, Chicago, IL 60637, USA*
[4] *NSF-Simons AI Institute for the Sky (SkAI), Chicago, IL 60611, USA*
[5] *TAPIR, California Institute of Technology, Pasadena, CA 91106, USA*
[6] *Kavli Institute for Astrophysics and Space Research, Massachusetts Institute of Technology, Cambridge, MA 02139, USA*
[7] *Vanderbilt University, Department of Physics & Astronomy, 6301 Stevenson Center, Nashville, TN 37235, USA*
[8] *Department of Physics, University of Florida, Gainesville, FL 32611, USA*

ABSTRACT

Galaxy mergers play an important role in many aspects of galaxy evolution, therefore, more accurate merger identifications are paramount for achieving a complete understanding of galaxy evolution. As we enter the era of very large imaging surveys, we are able to observe mergers extending to even lower masses and higher redshifts. Despite low-mass galaxies being more common, many previous merger identification methods were mostly calibrated for high-mass, local galaxies, which are easier to identify. To prepare for upcoming surveys, we train a convolutional neural network (CNN) using mock *HST* CANDELS images at $z \sim 1$ created from the IllustrisTNG50 cosmological simulation. We successfully identify galaxy mergers between a wide range of galaxies ($10^8 M_\odot < M_\star < 10^{12.5} M_\odot$, and $\mu > 1:10$), achieving overall accuracy, purity, and completeness of $\sim 73\%$. We show, for the first time, that a CNN trained on this diverse set of galaxies is capable of identifying both major and minor mergers, early and late stage mergers, as well as nonmerging galaxies, similar to that of networks trained at lower redshifts and/or higher masses (with accuracies ranging between $66 - 80\%$). We discuss the inherent limits of galaxy merger identification due to orientation angle and explore the confounding variables, such as star formation, to consider when applying to real data. This network enables the exploration of the impact of previously overlooked mergers of high mass ratio and low stellar masses on galaxy evolution in CANDELS, and can be expanded to surveys from *JWST*, *Rubin*, *Roman*, and *Euclid*.

Keywords: galaxy evolution — galaxy mergers

1. INTRODUCTION

Galaxy mergers are one of the main avenues for galaxies to evolve from clumpy, high-redshift galaxies into the organized structures we see in the local universe, both with large morphological changes due to major mergers and building stellar mass through minor mergers (e.g. Toomre & Toomre 1972; Mihos & Hernquist 1996; Buitrago et al. 2013; Martin et al. 2018). Identifying their role in galaxy evolution is a key task for observational astronomy, especially in the era of high redshift science with *JWST* (Gardner et al. 2006), and large astronomical surveys and telescopes such as the Vera C. Rubin Observatory's (*Rubin*) Legacy Survey of Space and Time (LSST; Ivezić et al. 2019), Nancy Grace Roman Space Telescope (*Roman*; Akeson et al. 2019), and *Euclid* (Scaramella et al. 2022).

Though mergers are an influential process in galaxy structures, they can prove tricky to identify. A galaxy merger can take hundreds of Myr to a few Gyr, depending on factors such as mass ratio and orbital parameters (Lotz et al. 2008). Merger stage can also complicate merger identification, as the galaxies do not experience each stage for an equal amount of time, and thus some stages are more common than others. There

Corresponding author: Aimee Schechter
aimee.schechter@colorado.edu

is a range of morphologies associated with each merger stage, and thus the method used influences which mergers are found and which are missed, as certain methods are more sensitive to certain morphologies (Lotz et al. 2008).

Mergers experiencing their first pass (early stage) are easier to identify than mergers experiencing the coalescence of the nuclei (late stage). The human eye is historically trustworthy at identifying low-redshift, early-stage mergers, especially when multiple people visually inspect each image (e.g., Darg et al. 2010). Visual (i.e., "by-eye", performed by scientists) classification efforts have enabled huge amounts of science on the role of mergers, but they have some drawbacks. First, it requires many human hours. Some of this has been outsourced to citizen scientists with projects like Galaxy Zoo (Lintott et al. 2008). A method that does not require large groups of people spending hours of their time to classify galaxies is more desirable. Second, the human eye may struggle to capture a range of galaxy mergers. Visual classification studies are biased to early-stage, major (merger mass ratio $\mu \geq 1:4$) mergers, due to these systems having obvious morphological disturbances and/or signatures of multiple galaxies, such as two stellar bulges. Late-stage, minor mergers (mass ratio $\mu < 1:4$) may be missed since they can be mistaken for isolated galaxies, not highly disturbed, or even just visually overlapping galaxies.

These biases in visual classifications lead to the use of quantitative imaging predictors such as the concentration (C), asymmetry (A), and smoothness (clumpiness) (S) i.e., CAS parameters (Conselice 2003) and Gini-M_{20} (Lotz et al. 2004). These non-parametric morphology measurements are all based on various ways of measuring distributions of light in images, often separating quiescent galaxies from star-forming galaxies, and more defined morphologies from disturbed systems. Separating those types of morphologies is extremely useful in identifying mergers. The merger stage and image quality affect which method correctly identifies more mergers, as morphological asymmetry is more accurate for mergers in early stages, and Gini-M_{20} is more successful with mergers near the end of the process (Nevin et al. 2019; Wilkinson et al. 2024). However, all of these methods are calibrated on lower-z galaxies and may need to be adjusted for high redshift galaxies.

Galaxies and galaxy mergers at higher redshifts, z, look different than those in our local universe (Conselice 2014 and references therein). For example, galaxies at $z > 1$ are often clumpier and contain more gas and dust compared to low-redshift galaxies. The physical morphology of high-redshift galaxies could make it harder to identify mergers visually, since the clumpy, uneven structure could make these galaxies appear to be merging even when they are actually isolated. These galaxies are also often smaller than their present-day counterparts. Therefore, how we identify low-redshift mergers may need to be different from how we identify high-redshift mergers. There have been by-eye classification efforts out to $z \sim 7$(Kartaltepe et al. 2015; Simmons et al. 2017; Willett et al. 2017; Smethurst et al. 2025), though not all by-eye classifications can identify mergers.

Convolutional Neural Networks (CNNs) are a type of neural network for working with imaging data. Their ability to extract useful features from images (even those that might not be obvious to the human eye) makes them a promising tool with which to identify galaxy mergers. Many studies have successfully shown that CNNs outperform both by-eye identification and Gini-M_{20}/CAS at $z < 1$ (Bickley et al. 2021; Ackermann et al. 2018). By using machine learning rather than visual by-eye identification, we avoid biases humans may have of identifying only more obvious mergers, such as a merger between two large spiral galaxies. Additionally, CNNs are much faster than human identification, which is an increasingly important consideration as we prepare for large observational survey telescopes.

When considering CNNs as a merger identification tool, it is important to take into account that CNNs, as any other machine learning model, can only learn from the information contained in the training set. Therefore, which mergers astronomers decide to include in their training set dictates which masses, mass ratios, and merger stages the CNN will be most likely to find. CNNs have been commonly trained on samples of galaxies that are identified by eye, both by citizen scientists and by experts. Many papers have used the Galaxy Zoo data to train or test the performance of their network to identify galaxy morphologies (e.g., Dieleman et al. 2015; Domínguez Sánchez et al. 2018; Cavanagh et al. 2021). However, Bickley et al. (2021) showed that compared to CNNs, visual inspection can often lead to many different classifications of any given galaxy depending on the training of the individual performing the classification. Therefore, it is becoming more common to train networks on mock images of simulated galaxies because this ensures that the mergers/nonmergers used in the training set are known a-priori.

Mock images from the IllustrisTNG cosmological simulation suite (Marinacci et al. 2018; Naiman et al. 2018; Nelson et al. 2018; Pillepich et al. 2018; Springel et al. 2018) are a common choice for a training set due to TNG's wide range of galaxy morphologies and match

of observational galaxy properties at $z = 0$. Many papers choose to make mock images from TNG100-1 or TNG300-1, with the $\sim$ 100Mpc per side and $\sim$ 300Mpc per side boxes respectively, and use neural networks to classify mergers (Bickley et al. 2021; Bottrell et al. 2022; Ferreira et al. 2022; Avirett-Mackenzie et al. 2024; Margalef-Bentabol et al. 2024; Ferreira et al. 2024b). The highest spatial resolution run of TNG, TNG50 ($\sim$ 50Mpc per side box; Nelson et al. 2019; Pillepich et al. 2019), was used as the training set in Omori et al. (2023). The TNG100 and TNG300 boxes contain significantly more volume than the TNG50 box, and thus enable much larger training sets of mergers and non-mergers. These works show that machine learning combined with TNG can be successful at identifying mergers at both early and late stages at lower redshifts $z < 1$ and higher masses $M_\star \gtrsim 10^9 M_\odot$ for a variety of mocked imaging surveys. Pearson et al. (2019) also showed successful ML merger identification with with the EAGLE simulation (Crain et al. 2015; Schaye et al. 2015), and Domínguez Sánchez et al. (2023) with the Horizon-AGN simulation (Kaviraj et al. 2017).

Along with the use of neural networks, recent years have witnessed an explosion in the field of eXplainable aritificial intelligence (XAI), which seeks to promote interpretability in machine learning tools (see Arrieta et al. 2019 for a review). Pixel attribution methods such as saliency maps (Simonyan et al. 2013) and Gradient-Weighted Class Activation Mapping (Selvaraju et al. 2020) highlight specific pixels and regions of an input image that the CNN relied on for its final classification. Additionally, inspecting feature maps from hidden layers can also provide physical intuition into why the network makes the decisions it does. This makes CNNs less of a black box and more useful for astronomical purposes. Ćiprijanović et al. (2020) identified merging galaxies at $z = 2$ in the original Illustris (Vogelsberger et al. 2014) simulation with a CNN for the first time. They used XAI to show that when identifying mergers, their CNN focused on larger areas of the image that contained faint substructures of a galaxy, but when identifying non-mergers, the influential pixels were in a more compact area. This provided insight into what physical processes the galaxy images contained that the network observed to make a decision between merger and nonmerger.

Using a galaxy merger sample from IllustrisTNG50, we create mock galaxy images from the *HST* CANDELS survey (Koekemoer et al. 2011; Grogin et al. 2011). This survey has some overlapping wavelength coverage with *Roman*, but since the data are already public, it has the benefit of creating mock images with real backgrounds instead of simulated background noise. We want to study mergers around cosmic noon ($1< z <3$), the peak of morphological evolution, for which CANDELS provides a large sample of galaxies. CANDELS is $> 90\%$ complete at $z <3$ (Guo et al. 2013; Mantha et al. 2018).

In this paper, we use extremely realistic, fully radiatively transferred mock *HST* images from IllustrisTNG50 to classify mergers in optical and infrared filters at $1 \leq z \leq 1.5$. We include galaxies down to $M_\star = 10^8 M_\odot$, while still including mass ratios down to 1:10 and multiple merger stages. We aim to create a trustworthy classifier to find mergers at high-z and lower stellar masses with *JWST*, *Rubin*, *Roman*, and *Euclid* with the help of a high resolution simulation and XAI interpretation.

Our merger selection from TNG and mock image process is discussed in Section 2, the CNN architecture, performance evaluation metrics and model interpretability are discussed in Section 3, the performance of our CNN is discussed in Section 4, and a discussion of the limitations of our sample and model is in Section 5. We use the Planck cosmological parameters (Planck Collaboration et al. 2016), the same ones used by IllustrisTNG: a matter density $\Omega_m = \Omega_{dm} + \Omega_b = 0.3089$, baryonic density $\Omega_b = 0.0486$, cosmological constant $\Omega_\Lambda = 0.6911$, Hubble constant $H_0 = 100\,h$ km s^{-1} Mpc^{-1} with $h = 0.6774$, normalization $\sigma_8 = 0.8159$ and spectral index $n_s = 0.9667$.

2. DATA

To quantify the role of mergers in processes at cosmic noon ($1 < z < 3$) such as galaxy morphological evolution, cosmic star formation, and black hole activity, a reliable classifier that can separate mergers from non-mergers at low masses, including among peculiar (not spiral or elliptical) galaxies is necessary. Therefore, we aim to build a CNN that is able to identify both major and minor mergers, mergers at early and late stages, and mergers between galaxies with $M_\star > 10^8 M_\odot$ at $z \sim 1$.

We create a dataset of mock *HST* imaging with galaxies in the $1 \leq z \leq 1.5$ redshift range. This redshift range is the highest redshift where spiral and elliptical galaxies are each about as common as peculiar galaxies (which would include mergers; Buitrago et al. 2013). Above this redshift, the universe has a different makeup than it does locally, with peculiar galaxies being much more prevalent, especially among low mass galaxies (e.g., Ferreira et al. 2023). Additionally, we know minor mergers are an important avenue for galaxies to build up stellar mass over cosmic time, and that low mass galaxies are more common at higher redshifts (Martin et al. 2018). However, peculiar galaxies dominate at $M_\star < 10^{9.5} M_\odot$ (Conselice et al. 2008).

The *HST* CANDELS survey is the basis for our mock images due to its high spatial resolution, coverage around the peak of cosmic star formation, and, importantly, existing catalogs with which to compare our results and build additional trust in our model before applying to newer surveys. In order to train a successful network we need a highly realistic training set that includes galaxy mergers with these mass ratios, merger stages, and stellar masses. We use galaxies from the TNG50 cosmological simulation as our training set. We want to use high-resolution simulations and images like those from TNG50 (with a median spatial resolution of ∼0.1kpc, which is comparable to or better than the resolution of CANDELS images at $z > 0.2$) for the merger identification process to be sure the CNN can separate small, clumpy, isolated galaxies from merging galaxies. Using mock images from TNG50 will also provide a strong training set since we will know the true classification of each galaxy, allowing our tool to be better trained at identifying mergers close to cosmic noon. We create mock images in three *HST* filters (F814W, F160W, and F606W) for a three channel CNN. This section describes the merger selection process from the simulation, and how we go from the simulated galaxies to fully realistic mock images.

2.1. *Cosmological Simulation and Galaxy Selection*

IllustrisTNG is a suite of cosmological magnetohydrodynamical simulations spanning $0 < z < 127$ with improved physics from the original Illustris simulation (Vogelsberger et al. 2014). It comes in three box sizes, the largest being TNG300 at roughly 300Mpc/side and the smallest, highest resolution run, TNG50, at roughly 50Mpc/side (Nelson et al. 2019; Pillepich et al. 2019). All volumes have both "dark matter only" and "baryonic physics" runs. The IllustrisTNG model includes physical processes such as gas cooling (primordial and metalline), star formation, supermassive black hole formation and mergers, chemical enrichment from supernovae, stellar and black hole feedback, and cosmic magnetic field evolution, all of which impact galaxy morphologies and evolution. This self-consistent simulation enables us to investigate how morphologies change over time and how this may affect merger identification techniques at different redshifts.

IllustrisTNG has "full" and "mini" snapshots. Both snapshots trace all of the same subhalos through cosmic time and include the full 50 co-moving Mpc/side box. A "full" snapshot contains all of the physics that TNG calculates, while the mini snapshots do not calculate physics for all particle fields. Since we aim to run radiative transfer, we only can only create images from snapshots that are "full", which are necessary to make the mock images described in Section 2.2. The redshift bin centers of $z = 1$ and $z = 1.5$ are both "full" snapshots. Notable examples of outputs that are only in the "full" snapshots are: magnetic fields, neutral hydrogen density, dark matter density, and stellar metallicity.

The SubLink assembly history traces a subhalo back in time and connects it to subhalos in previous snapshots (higher redshifts), that it evolved from. For a given snapshots, we identify all subhalos with a stellar mass greater than 1000 times the baryonic mass resolution of TNG50, which is $8.5 \times 10^4 M_\odot$. We trace the subhalos across cosmic time using the merger trees and merger definitions from Rodriguez-Gomez et al. (2015), which link the identities of a given subhalo to its progenitor and descendant subhalos at the previous and following snapshot, respectively. We define a merger as a subhalo that has both a first and a next progenitor at the previous snapshot in time, where both progenitors share a descendant history with the given subhalo and the first progenitor is the subhalo with the more massive assembly history. We additionally require that the stellar mass ratio of the first and next progenitor be greater than 0.1 to include minor and major mergers.

We select mergers from TNG50 in two redshift bins centered at $z = 1$ and $z = 1.5$. To identify mergers (and nonmergers) for each redshift bin, we begin by identifying all snapshots that are within 250 Myr of the bin center. At these redshifts, this is one mini snapshot on each side of the central full snapshot. There are a few hundred independent mergers in each redshift bin. We then find mass-matched nonmerging galaxies by adapting the matching scheme from Bickley et al. (2021), as described in Schechter et al. (2025). Each merging galaxy is matched with a nonmerging galaxy in the same snapshot by searching for a galaxy within a factor of $e^{0.1}$ and expanding the threshold by an exponential factor of 1.5 if a match is not found. No nonmergers are used twice as matched galaxies. A nonmerger must have not have merged within the past 2Gyr to be an eligible match.

After following this procedure, relative to the central full snapshot, we have galaxies that merged at the previous snapshot back in time (a mini snapshot), galaxies that merge at the center full snapshot, and we have galaxies that will merge at the next snapshot forward in time (a mini snapshot). Though we identify some merging one snapshot earlier or later, we trace them and take their image from the "full" snapshot bin center instead of the mini snapshot in which they truly merge. That gives us different merger stages, as we are imaging some galaxies pre-merging and post-merging. They are

the true mergers one snapshot forward and backward in time respectively, but are imaged at these pre or post phases in full central snapshot. For our analysis, we consider the pre-merging galaxies our "pre-coalescence" or "early-stage" galaxy merger sample since the merger trees still identify two subhalos, and both the true merger galaxies in the central snapshot and the post-merging galaxies as "post-coalescence" or "late-stage" mergers as the merger trees only identify one subhalo in both of these stages. In our case, because we are using theoretical definitions from the TNG merger trees, these may not match exactly to observational merger stages such as "first passage" or "nearing coalescence".

The resultant sample contains 260 mergers and 260 mass-matched nonmergers at $z = 1$ and 446 mergers and 446 mass-matched nonmergers at $z = 1.5$. We combine the two redshift bins into one dataset, and the distributions of stellar mass (left figure), merger stage, and mass ratio (middle figure) are seen in Figure 1. The stellar mass ranges from $10^{7.9} - 10^{12.5} M_{\odot}$. The vast majority of our galaxies have lower masses $M_{\star} < 10^{9.5} M_{\odot}$, which could make the classification task more challenging for the CNN. However, we want to utilize the lower stellar mass galaxies as well to get a full picture of the role mergers overall play in galaxy evolution and cosmic star formation, which the high resolution of TNG50 enables us to do. All merger classes (pre-coalescence, merger, and post-coalescence) are combined into one class of "merger" for the CNN. The middle plot on Figure 1 is a stacked histogram, so the overall distribution is for the entire merger class. The shading displays how the stages are broken down relative to definitions in Section 2.1. The specific star formation rates (sSFRs) are shown in Figure 1 (right plot), showing that the merging distriubtion peaks at a slightly higher sSFR than the non-merging distribution. We show sSFR instead of SFR to remove the mass dependence of the star formation main sequence. While our sample is mass matched, it is not matched in SFR, due to the small box size. Matching in both SFR and mass required expanding the mass match threshold many times, so we elect to use a close mass match and no SFR match.

2.2. *Realistic Mock Images*

We recognize that higher spatial resolution and deeper imaging at $z \sim 1$ exists with *JWST*. We choose to use CANDELS in this work in order to compare our upcoming merger catalog (which will be the focus of our future work) to existing merger catalogs created by non-ML methods to build trust in our CNN. It also gives us a baseline at optical wavelengths to compare to when applying to *Rubin* and *Roman*. Lastly, in order to not just identify mergers but draw statistical conclusions about their role in galaxy evolution, we require a large sample of observed galaxies, ideally with known stellar masses and star formation rates. CANDELS provides a large field, which *JWST* does not, with existing value-added catalogs.

To train the CNN on IllustrisTNG50 images for this goal, we must first make them as similar to real *HST* images as possible. Bottrell et al. (2019) investigated how important realistic training images are to the final classifications of a CNN while using one to not only identify mergers but also predict the merger stage. They found that as long as a network is exposed to background noise, other sources, and the spatial resolution of the telescope in training, it is able to predict mergers accurately when given real data. Notably, this paper discovered that realistic environments are more important to accurate predictions than radiative transfer in making mock images to train a CNN. Nevin et al. (2019) showed detailed steps to create SDSS mock images to use for machine learning and quantitative galaxy identification methods. To create mock *HST* images, we will be adapting the method used by Nevin et al. (2019).

2.2.1. *Radiative Transfer with* SKIRT

The first step to create realistic mocks is to post-process TNG50 galaxies through full Monte Carlo continuum dust radiative transfer calculations, using the publicly available code SKIRT (version 9; Baes et al. 2011; Baes & Camps 2015; Camps & Baes 2015, 2020). We use the prescription described in Vogelsberger et al. (2020) and Shen et al. (2020, 2022); Shen et al. (2024) , as done in Nevin et al. (2019).

In this prescription, stellar particles in the simulations are assigned intrinsic emission using the stellar population synthesis method. Specifically, we adopt the Flexible Stellar Population Synthesis (FSPS) code (Conroy et al. 2009; Conroy & Gunn 2010) to model the intrinsic spectral energy distributions (SEDs) of old stellar particles with $t_{\rm age} > 10\,{\rm Myr}$ (using the MILES spectral library and MIST isochrone library, this choice defines solar metallicity, $Z_{\odot}$) and the MAPPINGS-III SED library (Groves et al. 2008) to model those of young stellar particles with $t_{\rm age} < 10\,{\rm Myr}$. The MAPPINGS-III SED library self-consistently considers the dust attenuation in the birth clouds of young stars, which cannot be properly resolved in the simulations. We employ a K-D tree algorithm to calculate the smoothing length enclos-

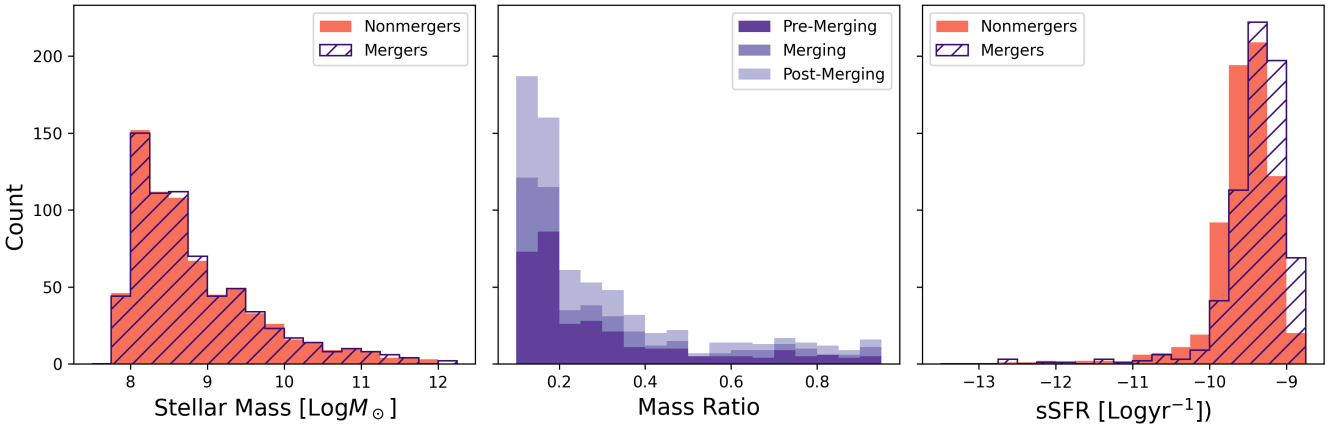

Figure 1. *Left:* Distributions of merging and nonmerging galaxies' stellar masses. *Center:* Stacked histogram of merger mass ratios in the merging sample color coded by merger stage as defined in Section 2.1. *Right:* Distributions of sSFR for mergers and nonmergers.

ing 64 [1] nearest stellar particles for all stellar particles within the galaxy. Given the spatial location and the smoothing length values of stellar particles, SKIRT then creates a photon source distribution and emissivity profile through the entire space by interpolating over these kernels. At the beginning of radiative transfer calculations, photon packages are randomly released based on the source distribution characterized in this way. Wavelengths from the Lyman edge out to optical and IR wavelengths included in CANDELS filters are covered. The total number of released photon packages is set to be $N_p = 10^{10}$. We also use an active galactic nuclei (AGN) template[2] to assign emission to the black hole particles. The AGN emission is a broken power law (Schartmann et al. 2005) characterized by the mass accretion rate and radiative efficiency of supermassive black holes (SMBHs) in the simulation.

The emitted photon packages will further interact with the dust in the interstellar medium (ISM). To determine the distribution of dust in the ISM, we consider cold star-forming gas cells (star-forming or with temperature < 8000 K) from the simulations and calculate the metal mass distribution based on their metallicities. We assume that dust is traced by metals in the ISM and adopt a constant dust-to-metal ratio among all galaxies at a fixed redshift. In Vogelsberger et al. (2020), the dust-to-metal ratios at different redshifts has been calibrated based on the galaxy rest-frame UV luminosity functions at $z = 2 - 10$ (e.g., Ouchi et al. 2009; McLure et al. 2009; J. Bouwens et al. 2016; Finkelstein 2016; Oesch et al. 2018). For galaxies below this redshift range, we use a Milky Way dust-to-metal value of 0.4 (Dwek 1998). We then turn the metal mass distribution into a dust mass distribution with the dust-to-metal ratio and map the dust distribution onto an adaptively-refined grid. The grid is refined with an Octree algorithm to maintain the fraction of dust mass within each grid cell to be smaller than 2×10^{-6} (Saftly et al. 2014). The maximum refinement level is also adjusted to match the numerical resolution of the simulations. Besides, we assume a Draine et al. (2007) dust mixture of amorphous silicate and graphitic grains, including varying amounts of Polycyclic Aromatic Hydrocarbons (PAHs) particles, which can reproduce the averaged extinction properties of the Milky Way.

Ultimately, after photons fully interacted with dust in the galaxy and escaped, they are collected by six simulated detectors 10 Mpc away from the simulated galaxy along the positive (or negative) x,y,z-directions of the simulation coordinates. These six viewpoints will enlarge our training, validation, and test set sizes. We use a detector size that scales with redshift; the field of view is $100(1+z)^{-1}$ kpc and 512 pixels per side, which is a physical size of 30 kpc at $z = 1$ and 24 kpc at $z = 1.5$. We have selected the box size to scale with galaxy size, which also scales as $(1+z)^{-1}$ (Bouwens et al. 2004). The flux in each pixel, as well as the integrated SED of the galaxy, are then recorded. Any galaxies that encountered errors or memory issues during the radiative transfer processes were discarded.

2.2.2. *Filter, Rebin, and PSF*

From the radiative transferred images, we produce three-color (F814W, F160W, and F606W), filtered im-

[1] It is an empirical choice here for adaptive softening and morphological characterizations are not sensitive to the number of neighbors used (e.g. Torrey et al. 2015; Rodriguez-Gomez et al. 2019).

[2] https://skirt.ugent.be/skirt8/class_quasar_s_e_d.html

ages using the Python code `SEDPY` (Johnson 2021). All five CANDELS fields have coverage in these bands. Additionally, we avoid using the bluest bands available to encourage the CNN to learn overall morphology rather than focus on star forming clumps. Following filtering, the images are convolved with the point spread function (PSF) of each given *HST* band, which we simulate using `TinyTim` (Krist et al. 2011). We implement cosmological surface brightness dimming by a factor of $(1+z)^{-3}$ (in frequency space). We repeat the process for each CANDELS filter, as each filter has different standard flux values and background noise (Koekemoer et al. 2011).

2.2.3. *Realistic Environments*

A key step to making the images fully realistic is ensuring the galaxies are placed in the simulated image as they would be in *HST* observations, including in realistic field environments (Bottrell et al. 2019). We incorporate background sources and noise from the CANDELS mosaics by creating cutouts from the CANDELS mosaics that have no sources at the center. We then overlay our mock images on top of that section of sky. Therefore, the noise is truly that of CANDELS imaging, and there are real background CANDELS galaxies and foreground stars in our images, an essential piece for our network to be able to distinguish from the central mergers. All of the main steps to make an example mock image in the F814W filter can be seen in Figure 2.

2.3. *Final Dataset Creation*

We split up our dataset into training (70%), validation (15%), and test sets (15%). Each class is split by the ratios for the training/validation/test set above, then combined into the final training/validation/test set following the split. This ensures each set is split equally between mergers and nonmergers, so that neither outcome dominates the network's final decisions. In reality, there are far more nonmergers than mergers in the universe. However, in order for the CNN to truly learn the difference between mergers and nonmergers, we do not want to bias its decision-making by providing an imbalanced dataset. Recall from Section 2.2.1 that each galaxy is viewed from six viewpoints (as if from each face of a cube). When splitting the data, we make sure that all viewpoints of a given galaxy are always included in the same set, so that none of the viewpoints of galaxies included in the training set can appear when validating and testing the network. Each image is log normalized to have values between 0 and 1, typical for a CNN input. We employ the log stretch which helps faint features, such as tidal tails, to become more visible.

After the mock image process, our images are size 202×202 pixels at $z = 1$, and 154×154 pixels at $z = 1.5$. ResNet18 (see Section 3.1 for information on the CNN model) takes images of a size 224×224 pixels, so we use the resize operation from SKIMAGE to reach this size, as done in Bickley et al. (2024). This algorithm is stretching the images with interpolation, so we maintain the spatial resolution information of the pixels in the original mock image. To additionally enhance our training set size, we use data augmentation on both the mergers and nonmergers. Each galaxy image can be randomly rotated up to 30°, flipped horizontally or flipped vertically. The edges of the rotated images are filled with zeros ensuring that all images have the same shape. We include both the original image and the augmented image in training. Following the mock image and data augmentation processes, the final dataset consists of: training set with 5,940 mergers and 5,916 nonmergers, a validation dataset with 630 mergers and 624 nonmergers, and a test set of 630 mergers and 630 nonmergers.

3. METHODS

3.1. *Convolutional Neural Network*

A CNN is a type of neural network specifically designed to work with images. By convolving different filters with an input image, it is able to extract features such as edges and shapes from the input image. While even simpler CNN architectures can achieve very high accuracies (above 99%) when trained to classify everyday objects and animals (e.g., cars, chairs, horses, dogs etc.) using very large benchmark datasets such as MNIST (Deng 2012), classification tasks in astronomy can be much harder. Training data sets in astronomy are often much smaller (for example due to the lack of many high-fidelity labels) and include classes that look somewhat similar (e.g., mergers and nonmergers both look like galaxies), which makes training CNNs much harder.

Even with data augmentation, our dataset is on the smaller side, therefore, we use transfer learning to obtain smoother loss and accuracy curves. We use the ResNet18 model (He et al. 2015) with pre-trained weights from Zoobot2.0.2 `zoobot-encoder-resnet18` (Walmsley et al. 2023). Zoobot models are trained on millions of galaxy images from Galaxy Zoo using labels provided by citizen scientists. Since Zoobot was trained on galaxies as opposed to images of everyday objects, its weights provide a better starting point than models trained on everyday objects for our goal of classifying mergers. Using Zoobot's `FinetuneableZoobotClassifier` class we set `num_classes = 2` to force a binary classification: galaxy merger or nonmerger. We train our CNN with the Adam optimizer (Kingma & Ba 2017), initial learning rate of

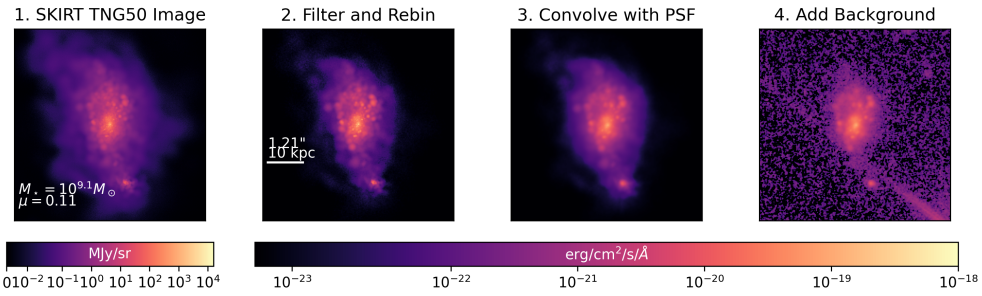

Figure 2. Main steps to create a mock F814W CANDELS image from the radiative transferred TNG50 data: 1) The left-most panel shows the image just after it has been processed by SKIRT; 2) Next, we apply an *HST* F814W filter to the image so we are no longer seeing all wavelengths of light; we also rebin the image to the same pixel scale as the CANDELS mosaics; 3) Next we convolve with the PSF of the telescope to replicate what this galaxy would look like if observed by *HST*. 4) Lastly we add real CANDELS backgrounds to include real CANDELS noise and background sources. This process will supply a more realistic training set, and is a crucial piece to building a robust network that can classifiy CANDELS galaxies.

10^{-5}, exponential learning rate decay of 0.5 and cross-entropy loss. Only the dense layer weights are updated during training. We initiate early stopping during training if the validation set loss does not decrease by at least 0.0005 after 5 epochs. The weights of the model from the epoch with the lowest validation loss are saved and used as the best model to obtain predictions on the test data. Training takes up to two hours and fifty four minutes depending on the random seed with one Nvidia_a100 GPU. We train with three different random seeds to ensure model stability.

3.2. *Model Performance and Calibration*

To assess the performance of the CNN, we start with standard metrics, such as confusion matrices and Receiver Operating Characteristic (ROC) curves. When discussing these metrics in the context of this study, we consider merger the positive class and nonmerger the negative class. Therefore a true positive (TP) is a merger correctly identified as a merger, and a true negative (TN) is a nonmerger correctly identified as a nonmerger. A false positive (FP) is a nonmerger misidentified as a merger, and a false negative (FN) is a merger misidentified as a nonmerger.

A confusion matrix tracks how many false positives and false negatives the network found, along with the accurate predictions. In our case, a false positive would be an isolated galaxy identified as a merger, and a false negative would be a merger identified as a nonmerger. If the network performed perfectly, the confusion matrix would have values only on the diagonal, with zeros everywhere else, indicating that there were no false positives or false negatives. The rows of the confusion matrix correspond to the actual class of the galaxy (true merger or nonmerger) and each column corresponds to the network's predicted class.

Some metrics to analyze how well the network is performing are accuracy, purity, and completeness.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$\text{Purity} = \frac{TP}{TP + FP} \tag{2}$$

$$\text{Completeness} = \frac{TP}{TP + FN} \tag{3}$$

Accuracy is a measurement of the fraction of the time the network is making the correct prediction overall, whereas purity and completeness are specific to the positive (merger) class. Purity and completeness are also sometimes known as precision and recall. Purity can be thought of as the percentage of predicted mergers that are correctly classified (e.g., how often a predicted merger is actually a merger). Completeness is the fraction of mergers that are correctly retrieved (e.g., out of the true mergers how many are correctly classified). Here, a low completeness would mean we are missing many mergers by classifying them as nonmergers. Often an increase in purity, can lead to a decrease in completeness, and vice versa. In the case of our CNN, having high purity and completeness would mean low contamination in our merger sample, since we would be classifying most galaxies correctly (with the purity of our merger sample being particularly important).

ROC curves are another tool to analyze the success of a CNN. This type of curve shows the true positive rate (completeness) against the false positive rate. The goal is to increase the area under the curve, meaning that there are more true positives and fewer false positives. A model in which the network is guessing randomly each time will have an ROC curve with a diagonal line with a slope of 1. A perfect network has area under the curve

of 1, and a completely random one has area under the curve of 0.5.

A trustworthy neural network must be both accurate and appropriately confident in its predictions. To examine the confidence of a neural network, we use two calibration tools. First, we use the Brier Score (Brier 1950). The Brier score is a measurement of "how correct" a prediction is. For binary classification, it is defined as:

$$\text{Brier Score} = \frac{1}{N}\sum_{t=1}^{N}\left(p_t - o_t\right)^2, \tag{4}$$

where N is the number of galaxies in the test dataset, p_t is the probability assigned to each galaxy by the neural network (the value following the softmax activation function, which always lies between 0 and 1), and o_t is the final class (1 or 0 corresponding to merger or nonmerger with a cutoff score of 0.5) assigned to each galaxy. A Brier score is always between 0 and 1 because it is a mean of squared differences between output probabilities and predicted classes. A lower Brier score indicates a network that is making accurate and well-calibrated predictions, as there is less of a difference between the probability and the final classification.

The second calibration metric we use is the Expected Calibration Error (ECE; Naeini et al. 2015). ECE measures a weighted average of the difference between the accuracy and predicted probabilities. ECE is defined by splitting the dataset into M equally spaced bins and calculating a weighted average of the difference between the accuracy and output probabilities in each bin.

$$\text{ECE} = \sum_{m=1}^{M}\frac{|B_m|}{N}\left|\text{acc}\left(B_m\right) - p\left(B_m\right)\right|, \tag{5}$$

where N is the total number of galaxies in the test set, B_m is the number of galaxies in the m^{th} bin, acc is the accuracy, or percentage of correctly classified galaxies, in each bin, and p is the predicted probability (output from softmax) of the CNN in each bin. A lower ECE indicates a smaller difference between the accuracy and probability in each bin. This is the desired outcome for a well-calibrated network. We can visualize this difference through a reliability diagram (DeGroot & Fienberg 1983; Niculescu-Mizil & Caruana 2005). These diagrams show the accuracy in each bin as a function of the probability in each bin. A perfectly calibrated model would have no difference, or "gap", between the accuracy and probability, and thus would plot a 1:1 line. The difference from a 1:1 line shows the miscalibration of the network.

3.3. *CNN Interpretability*

Going beyond the more standard metrics, we use Gradient-weighted Class Activation Mapping (Grad-CAM; Selvaraju et al. 2020), an XAI technique that highlights sections of an input image that are important to the final prediction. Grad-CAM uses gradients from the final convolutional layers to highlight these influential regions so that they contain spatial information before it is lost in the fully connected layers. Understanding where in the image our CNN is basing its decisions can help us build trust in the model, both showing when it perceives an image "correctly", and to see why it makes an incorrect decision when it does. Often, the decisions make sense when we can see what the CNN focused on. Grad-CAM can be activated to either the merger or nonmerger class for any input image. Activating a specific class shows which pixels are influential to that class, and not to other classes.

Finally we apply Uniform Manifold Approximation and Projection (UMAP; McInnes et al. 2020), a dimensionality reduction method. We use it to visualize the high dimensional latent space of a fully connected layer in the CNN in only two dimensions. Galaxies that the network deems similar will lie close to each other in UMAP space, and galaxies that the network deems different from each other will lie far apart. We can use the UMAP distribution along with the physical quantities associated with each galaxy (e.g., stellar mass, merger stage, SFR) to investigate how the network related different galaxies.

4. RESULTS

We present our CNN training metrics and results when the model is applied to our test set. We train our model with 3 different random seed initializations and present the mean and standard deviations of different performance metrics for our test set in Table 1. Accuracy, purity, and completeness are all $\sim 73\%$, with the Brier score and ECE being low and indicating good calibration overall of our models. When presenting performance of an individual model (in the text and figures in the following sections), we use Seed 626, since that seed had the highest overall accuracy.

4.1. *Model Performance and Calibration*

The loss and accuracy curves for one of our models (Seed 626) can be seen in Figure 3. We use the weights from epoch 43, the epoch with the lowest validation accuracy before overfitting occurred, as out final model. The confusion matrix for this model our test set is shown in Figure 4. The overall accuracy of this model is 73%, with a completeness also of 73%. The majority of our

Accuracy	Purity	Completeness	Brier Score	ECE	AUC
$73.02 \pm 0.41\%$	$74.0 \pm 0.01\%$	$72.0 \pm 0.01\%$	0.19 ± 0.01	0.08 ± 0.03	0.8 ± 0.01

Table 1. Accuracy, Purity, Completeness, Brier Score, ECE, and AUC for our model. The values shown are the mean and standard deviation of the three random seeds.

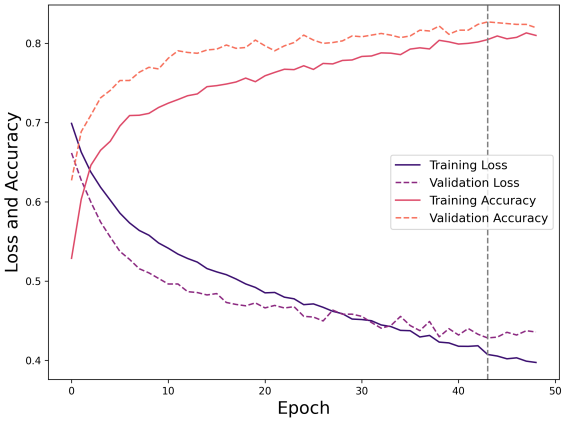

Figure 3. The loss and accuracy curves for our network. The small data set size leads to the bumpier curves, especially in the validation set shown in the orange dashed line. Though the training set curves in solid purple continued to improve, the validation set curves plateaued, so we implemented early stopping to avoid overfitting. We use the weights from epoch 43 as our best model, noted by the grey dashed line.

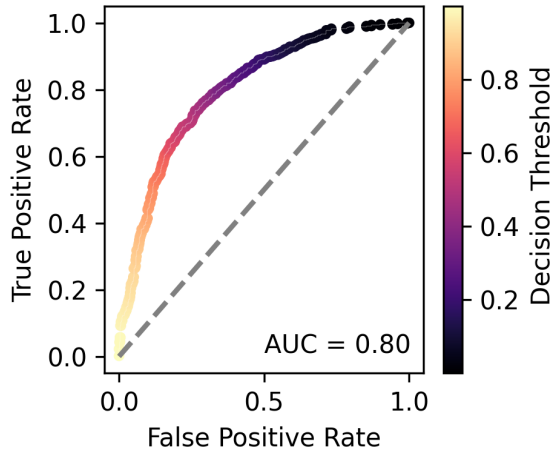

Figure 5. The receiver operating characteristic curve for our test set in Seed 626. The better the network performs, the closer the AUC is to 1. The grey dashed line indicates a network that randomly guesses each time, with an accuracy of 0.5.

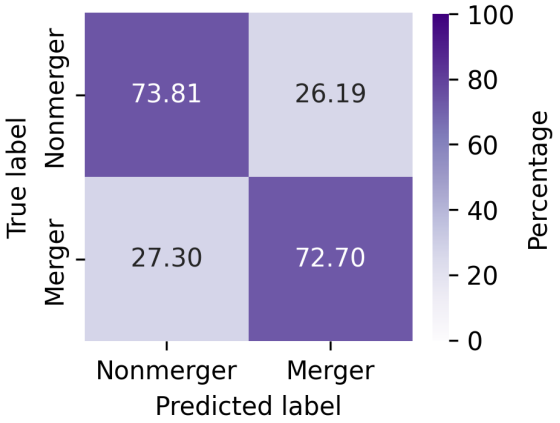

Figure 4. The confusion matrix of our Seed 626 network showing that we do classify the majority of galaxies correctly. The darker squares along the diagonal show the galaxies classified correctly.

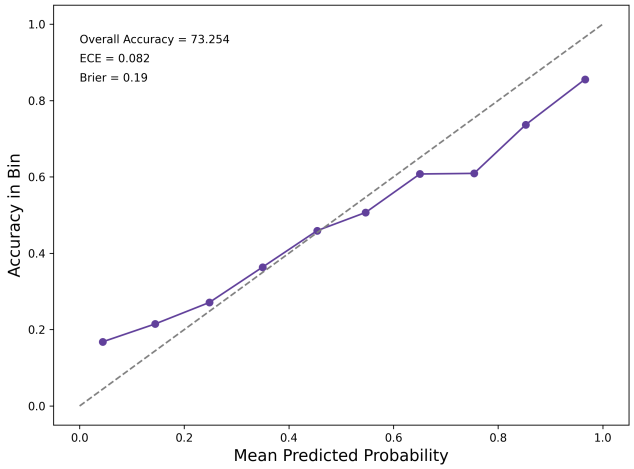

Figure 6. Calibration curve of our Seed 626 network with the test set data. The Brier score and ECE, in addition to the overall accuracy, are denoted in the top center. A perfectly calibrated network would have all bins lying along the 1:1 dashed line.

galaxies are classified correctly, but about a quarter of each class are not. We discuss possible sources of misclassifications in Section 4.3.

Our network's ROC curve for Seed 626 is shown in Figure 5. It performs better than a random classifier, seen by our purple line above the grey dashed (random) line. The area under the curve AUC = 0.80. We use the default decision threshold of 0.5 (i.e. a predicted probability of more than 0.5 is a merger, and equal to or less than 0.5 is a nonmerger) throughout our analysis. The colorbar in Figure 5 shows what changing that decision threshold does to the true and false positive rates.

In addition to simple metrics like accuracy, we also implement two calibration metrics and plot a calibration curve for our CNN in Figure 6. The Brier score of our network is 0.19, and the ECE is 0.08. Overall, the network is not severely miscalibrated.

4.2. *Effects of Orientation Angle, Merger Mass Ratio, and Stellar Mass*

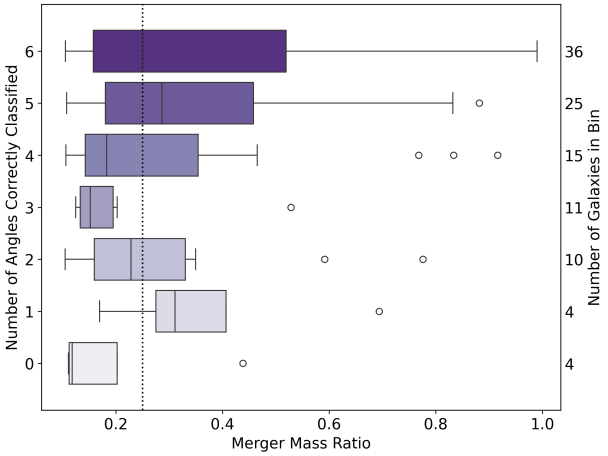

Figure 7. Box plot showing the number of angles (out of 6 possible angles) for which each merger was correctly classified as a function of the merger mass ratio in Seed 626. A box plot extends from the first quartile to the third quartile of the data, with the solid line marking the median. The whiskers extend to 1.5 times the range of the first to third quartile. The empty circles are any points not included in the range of the whiskers. The black dotted line marks $\mu = 0.25$, the divider between a minor and major merger. The bottom, lightest box shows that very few galaxies are misidentified from every angle.

To understand where the network performs well and where it does not, we seek to understand the effects of orientation angle, specifically in tandem with the merger mass ratio and stellar mass. Even a merger that is clear from one angle, perhaps a face-on disk with clear tidal tails, may look like a nonmerger from another. Each galaxy is viewed from 6 angles, so we now look at how many angles each galaxy in the test set was correctly classified from. For the entire sample as a whole, the mean and standard deviation of the number of angles correct per galaxy for our highest performing random seed is 4.4 ± 1.8, with a median of 5. However, considering how varied the galaxies in our sample can be, we break this down into more detail for the best model (Seed 626).

We start by examining only the mergers. We show in Figure 7 the number of angles for which a galaxy was correctly identified relative to its merger mass ratio. The distributions in these bins (the number of correctly classified angles) are shown with box plots, in order to demonstrate the spread. Box plots show the middle 50% of data inside the box, with the whiskers extending to the farthest datapoint within 1.5 times that range. The outliers on these plots are galaxies with a mass ratio much higher than the bulk of the galaxies in that bin, simply indicating that mass ratio is more rare.

Encouragingly, only four merging galaxies were classified incorrectly every time (the lowest bin on Figure 7). Almost every merger that was classified as a nonmerger from all six viewpoints was a minor merger. Despite that, we also see that plenty of minor mergers are identified often, if not from every angle. We conclude that observation angle does affect our ability to identify mergers, especially minor mergers. If observation angle was not a factor, we would only see galaxies being correctly identified from no angles or every angle. Instead, we see many galaxies that are correctly classified at most angles, but not all. This gives us a better understanding of how many mergers we may be missing when these networks are applied to real data, assuming we miss a similar percentage due to observation angle.

We next examine the relationship between the stellar mass of a galaxy and how many angles the network could correctly classify it from. In Figure 8, we look at both mergers (top figure) and nonmergers (bottom figure). It may be initially surprising that many high mass galaxies are often misclassified, since a larger mass could make a merger easier to see, but a large, spheroidal galaxy undergoing a merger, especially a minor one, could easily completely obscure its companion from some angles. Additionally, we note that there are far more low-mass galaxies than high-mass galaxies in our sample. Nonmergers with $M_\star > 10^{10} M_\odot$ tend to be classified correctly more often than the high mass mergers, as seen by the longer whiskers on the top box on the nonmergers' plot. Nonmergers are classified correctly across all masses, but in contradiction to the mergers, if they are misclassified it tends to be at low stellar masses. Many merger studies are only able to consider high mass galaxies, so it is encouraging that with enough spatial resolution, we do have tools to identify at least a fraction of lower mass mergers, even at $z \sim 1$.

We next ask the question, does our network perform equally well on each type of merger in our merging sample? We include multiple merger stages in our sample. The earliest mergers are still recognized as two separate subhalos by the TNG merger trees, and the latest stage mergers are after coalescence into one subhalo. We also include both major and minor mergers. Our breakdown of merger type and accuracy can be seen in Table 2.

Accuracy					
All Mergers	Major	Minor	Early Stage	Late Stage	Nonmergers
$71.85 \pm 0.97\%$	$75.77 \pm 0.82\%$	$68.3 \pm 0.66\%$	$79.63 \pm 0.74\%$	$66.0 \pm 0.01\%$	$74.0 \pm 0.01\%$

Table 2. The accuracy for our model, broken down by different types of galaxies. The values shown are the mean and standard deviation from our three random seeds.

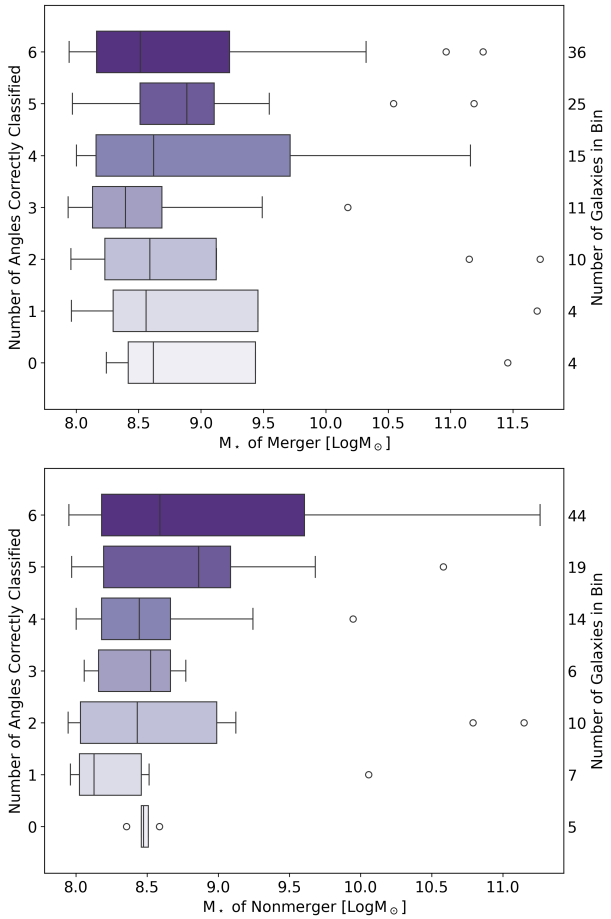

Figure 8. Box plot similar to Figure 7 showing the number of angles for which a galaxy was classified correctly for a range of stellar masses. The distributions for the mergers are shown on top, and the nonmergers on the bottom. A higher stellar mass did not automatically mean the galaxy was easier to classify, as seen by the outlier circles. The high mass nonmergers were easier to classify than the high mass mergers, seen by fewer cirlces on the bottom right of the bottom plot than the top.

One would expect the major mergers to be easier for the network to identify than the minor mergers, as major mergers tend to lead to large-scale disruptions of morphology. Minor mergers, on the other hand, may not have much of an effect on the morphology of the larger galaxy. However, identifying minor mergers is a crucial step in understanding how galaxies evolve and grow (e.g., Newman et al. 2012; Kaviraj 2014; Martin et al. 2018. We find that our network performs similarly on major mergers and minor mergers. This is optimistic for our chances at identifying minor mergers going forward, even at $z > 1$. CNNs prove to be a useful tool in identifying minor mergers that the human eye and nonparametric methods may miss.

The network identified early stage mergers more accurately than late stage mergers. We expect early stage mergers that still have two obvious bulges and may maintain some organized structure before they truly encounter the other galaxy in the merger to be easier to classify. However, multiple papers have found that CNNs can be successful at finding post-merger galaxies at $z < 1$ (e.g., Bickley et al. 2021; Ferreira et al. 2024b; Bickley et al. 2024). This points towards a combination of machine learning and more traditional methods like spectroscopic close pairs (e.g., Duncan et al. 2019) to be a promising way to identify all mergers, even at high redshift. The benefit to CNNs is that photometry is faster and cheaper than spectra. Additionally, CNNs will be key in analyzing the volume of data coming from large imaging survey telescopes.

4.3. *Understanding Misclassifications*

In Figure 9 we show example galaxies and their Grad-CAMs set up in the same layout as a confusion matrix. For each example, the input galaxy image is on the left, the Grad-CAM with class "merger" activated is in the center, and the Grad-CAM with class "nonmerger" activated is on the right. The true positives are on the diagonal, with the misclassifications on the off-diagonals. When discussing Figure 9, activating a class refers to asking the network to highlight which pixels are important for that class. We can see that in most cases when activating the Grad-CAM for the unpredicted class, the edges tend to be highlighted, and when we activate the Grad-CAM for the predicted class, it tends to highlight the galaxy at the center. This is promising, showing that even when the network makes an incorrect classification, it is often still making its decision on the physical features of the galaxy. The Grad-CAMs are not different enough between classes to draw conclusions about specific image features, but do offer confidence that the network has learned not to focus on background noise or sources, as seen in the bottom figure of the top left true

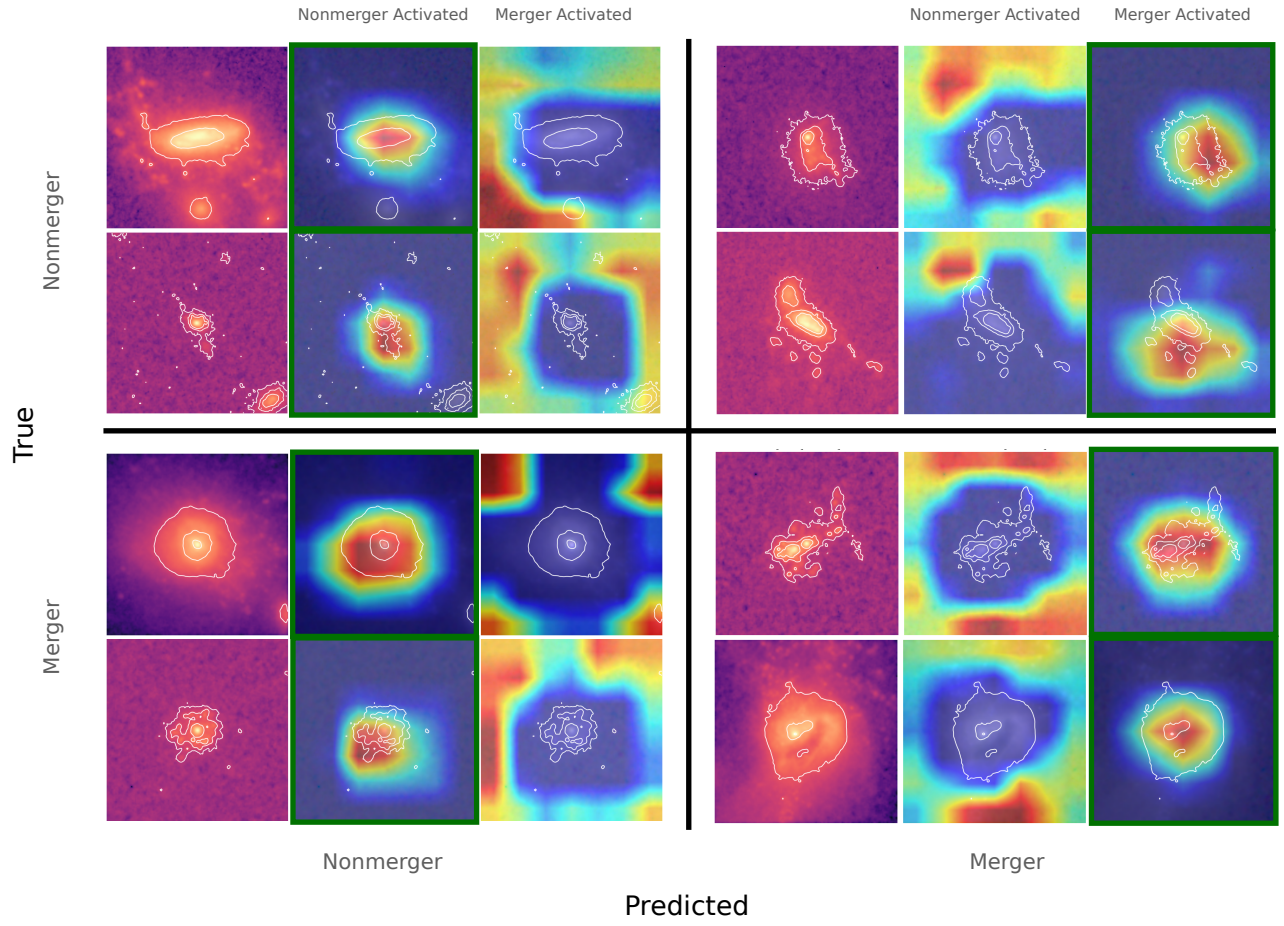

Figure 9. Visual confusion matrix of galaxies in the test set. True negatives are in the upper left, and true postives are in the lower right. The contours are 3σ and 5σ overlayed simply to guide the eye to where the structure is in the Grad-CAM images. The left image is the input image, the middle is the Grad-CAM with class "merger" activated, and the right is the Grad-CAM with class "nonmerger" activated. The class that the network predicted is boxed in green. The Grad-CAMs show that the network focuses on the galaxy when making its prediction, but does not highlight specific morphological features.

negative quadrant, an important quality of successful merger identification using CNNs (Bottrell et al. 2019).

We can also see in this image that the misclassifications often make sense to the human eye. The false negatives are seen in the lower left. The top image has a central structure but is overall quite smooth, with no obvious secondary bulge. The false negative quadrant shows a lack of extended features, with only one bright nucleus. Though the galaxy appears to have some structure from the contours the single bright nucleus and clean background make sense for why this merger was misclassified as a nonmerger. The top right corner represents false positives. Both of the galaxies shown here exhibit extended features and clumpy structures. Inspecting by eye, it is easy to see how the network thought both of these galaxies could be mergers with the multiple bright patches and clumpy structures.

UMAP (McInnes et al. 2020) is a non-linear, dimensionality reduction technique that we use to visualize the latent space of a CNN. On Figure 10, we show UMAPs of the test set images with colors representing values of different physical quantities of the galaxies. The mergers (circles) primarily reside on the left of the UMAP, and nonmergers (triangles) on the right. However, there is a lot of overlap in the middle. There is a clump of points sitting to the bottom of the main distribution, in which both mergers and nonmergers reside. When inspecting these galaxies by eye, we find they happen to be images with no background sources in the CANDELS background cutout and appear very smooth. It makes sense that these few galaxies are dissimilar to the rest

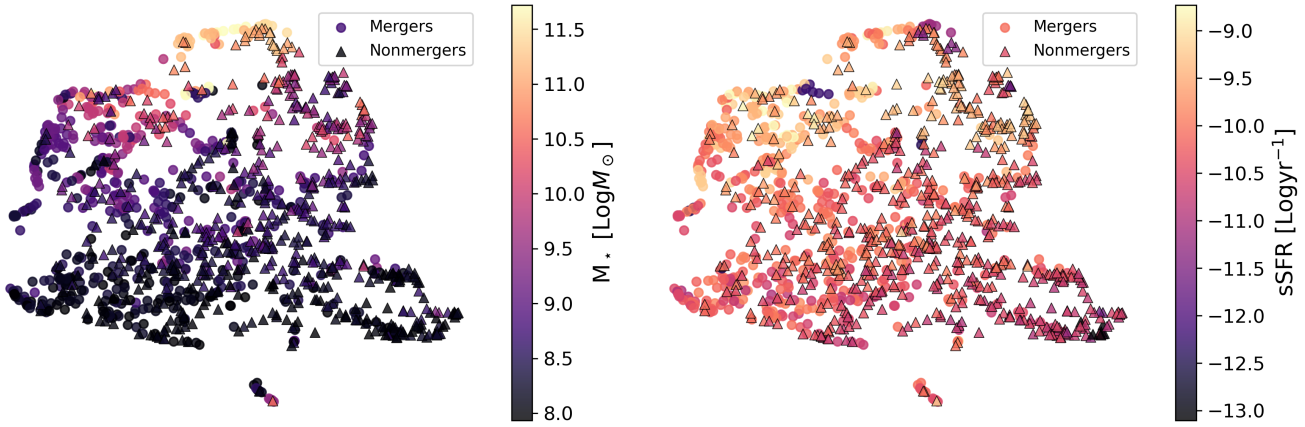

Figure 10. UMAPs of the test set color-coded by stellar mass (*left*), and specific star formation rate *right*. The true nonmergers are triangles and the true mergers are circles. We exclude axes because the important information in a UMAP is in the relative distance between points and clustering patterns, not in absolute distances. There are clear trends with both stellar mass and star formation rate, showing that the network picks up on these quantities even with no input information about them.

of the dataset which includes background galaxies and noisy cutouts.

We see a clear gradient in the UMAP when colored by the stellar mass of each galaxy (left plot on Figure 10). The low-mass galaxies are on the bottom of the UMAP, with stellar mass increasing towards the top. No stellar mass information was provided during training, but the network was able to recognize this physically meaningful quantity. There is again a clear gradient in the UMAP when colored by specific star formation rate (right plot of Figure 10). Similarly to stellar mass, the lower sSFR galaxies are on the bottom, and sSFR increases towards the top, even though no sSFR information was input to the model. The exception is extremely low sSFR, which is more scattered throughout. Because of this gradient, we speculate some of the nonmergers misclassified as mergers may be due to high, clumpy star formation, likely in the nonmergers seen in the top left. There were no obvious trends for UMAPs relative to the merger stage or merger mass ratio.

To further investigate our speculation from UMAP, we plot the specific star formation rates in Figure 11. The nonmergers incorrectly classified as mergers had a higher mean sSFR than the correctly classified nonmergers (dotted and dashed lines on the left plot, respectively). The mergers incorrectly classified as nonmergers had a lower mean sSFR than the correctly classified mergers (dotted and dashed lines on the right plot, respectively). This agrees with what the UMAPs showed: the trend with sSFR could account for many of the misclassifications. The network may have learned that mergers tend to have higher sSFRs than nonmergers, or picked up on a feature in the image correlated to sSFR, leading to this result.

5. DISCUSSION

5.1. *Comparison to Other Studies*

Galaxy merger identification is an inherently difficult task. Many methods have been developed for this task, from visual identification to non-parametric methods such as CAS and Gini-M_{20}. These non-parametric methods are designed to quantify the distribution of light in an image, for example, where the light is concentrated and the range of galaxy brightness throughout the image, into a single value. The threshold value separates mergers and nonmergers. These statistics alone can only capture $\sim 50\%$ of mergers, but when combined can be incredibly powerful (Nevin et al. 2019; Snyder et al. 2019; Wilkinson et al. 2024).

For example, Nevin et al. (2019) successfully classify mock SDSS mergers and nonmergers with linear discriminant analysis (LDA). They achieve accuracies of 85% for major mergers and 81% for minor mergers. Wilkinson et al. (2024) builds on this result, creating mock galaxy images at $z < 0.2$ with spatial resolutions varying from lower than SDSS to higher than *Rubin* ten year co-adds, applying non-parametric methods, and additionally combining them with LDA and random forest methods. Even for their pristine imaging (before downgrading resolution with atmospheric blurring and sky noise), single non-parametric statistics provide maximum completeness (number of true mergers identified correctly/number of total mergers in the sample) of only $\sim$55%. This increases to 73% with an LDA, and 86% with a random forest. Random forests have also been used for higher-redshift samples. At a redshift of $z = 4$, Snyder et al. (2019) achieves a $\sim$70% completeness for mock *HST* and *JWST* images, roughly

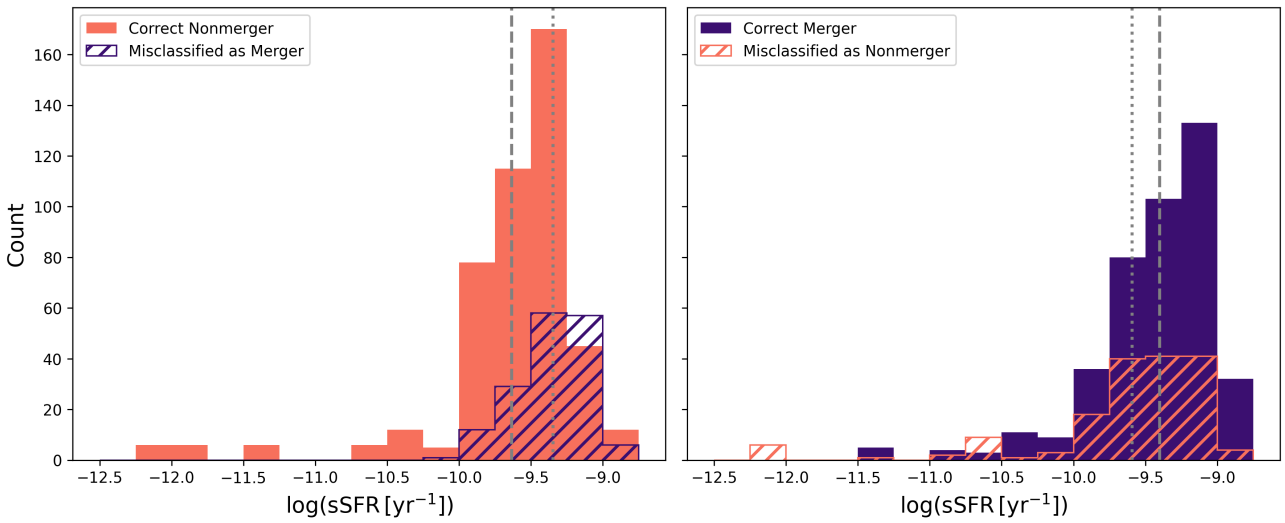

Figure 11. *Left:* sSFRs of nonmergers that were classified correctly as nonmergers in orange and incorrectly classified as mergers in purple. *Right:* sSFRs of mergers that were correctly classified as mergers in purple and incorrectly as nonmergers in orange. The dashed and dotted lines represent the mean of the correct and missclassified distributions, respectively. The incorrectly classified nonmergers have higher sSFRs than the correctly classified images and vice versa for the incorrectly classified mergers.

twice that when only using two non-parametric statistics (Gini - M_{20} or C-A). All three of these works use galaxies with $M_\star \sim 10^{10} M_\odot$. From $0.5 < z < 4$, the random forest in Rose et al. (2023) applied to galaxies with $10^5 M_\odot < M_\star < 10^{12} M_\odot$ attains an accuracy of $\sim$60% on mock *JWST* CEERS imaging. They extend this result to $4 < z < 5$ in Rose et al. (2024) and correctly classify 59% of nonmergers and 67% of mergers.

Many works have shown that CNNs are successful at classifying galaxy mergers and often outperform non-parametric methods and LDA and random forest at higher redshifts. With $M_\star > 10^{10} M_\odot$ at $z \sim 0.01$, Pearson et al. (2019) classifies mergers and nonmergers in simulated SDSS images from the EAGLE simulation, achieving only 65.2% accuracy. They discuss that the simulation includes a more complete sample of mergers, not just those easy to identify by eye. Therefore, it is by default a harder task for the CNN than identifying mergers in SDSS observations that were also visually identified, where they achieved 95.1% accuracy.

Mergers and nonmergers around cosmic noon were first classified with a CNN in Ćiprijanović et al. (2020), who used mock images at $z = 2$ from the original Illustris simulation. They showed that adding noise to images does decrease accuracy (79% accuracy on pristine images and 76% on noisy images), but noise is a key aspect for realistic mock images. Their Grad-CAMs highlighted larger areas of the image for the merger class and focused on more compact regions for nonmergers. Relevant for this work, Ćiprijanović et al. (2020) uses the same 500Myr merging window as our work here. However, their stellar mass lower limit is $M_\star = 10^{9.5} M_\odot$. They note that most of their misclassifications are coming from the low mass end of their sample. Our sample stretches to even lower stellar masses, so stellar mass may be responsible for some of our misclassifications (see Section 4.3).

CANDELS galaxies with $M_\star > 10^{10} M_\odot$ have been classified out to $z = 3$ (Ferreira et al. 2020). That mass limit was specifically chosen to use with the largest box size of TNG, TNG300, in order to maximize training set size. They achieve 87% accuracy on pre-mergers, 78% on post-mergers, 94% on nonmergers. It is mentioned that the spatial resolution in CANDELS is higher than that of some of their mock images, due to the simulation's spatial resolution. Our work uses the smaller box, but higher resolution run of TNG50, and thus we stretch to galaxies with stellar masses down to 100 times smaller. That wider mass range comes with a drop in overall accuracy.

Pushing past cosmic noon, Rose et al. (2024) classifies mergers and nonmergers in CEERS with mock images from TNG100. Different from other works, their test set of mock images is unbalanced, with far more nonmergers than mergers, to better represent the real universe. They split their data into three redshift bins between $3 < z < 5$ and achieve $\sim 60 - 70\%$ accuracies in all of them. Their mass range stretches to even lower masses than this work, with galaxies included at $M_\star > 10^7 M_\odot$. The Grad-CAMs of these mock *JWST* galaxies do not show clear patterns when all six filters are included. They note they see evidence of the network focusing on the galaxy in some single filter Grad-CAM

images, but the activation in other images are seemingly random.

5.2. *Why Do CNNs Struggle with Merger Identification?*

We showed that CNNs can find non local ($z \sim 1$) mergers at low mass ratios ($\mu > 1:10$) and low stellar masses ($M_\star > 10^8 M_\odot$) when enough training examples are provided. TNG50 provides a high resolution training set enabling this classification. While the network in this paper outperforms any previous CNN at this low mass range, it still misclassifies some galaxies. We now explore what makes merger identification an inherently difficult task, especially at $z > 1$. First we discuss the morphology of galaxies in combination with image resolution and depth. We then focus on the effect of viewing orientation and the inherent ceiling on merger identification.

5.2.1. *Impact of Morphology and Image Quality*

While a task like identifying two stellar bulges at low redshift and high masses may be trivial for a CNN, distinguishing between clumpy star-forming regions and two stellar bulges at higher redshifts and low masses is not so trivial. Additionally, tidal features can be key to distinguishing a merger from a clumpy, isolated galaxy, but these features can be dim, especially compared to the brightnesses of star-forming regions and stellar bulges. Normalization of the images with a log stretch does help this issue, ensuring that all pixel values are between 0 and 1 and the values are well dispersed within that range. With high spatial resolution data from *Roman*, *Rubin*, *JWST*, and *Euclid*, merger identification at high-z should become easier, provided we create accurate training sets, and trust our ML algorithms. Ensuring that the network can understand the difference in these brightness scales and the physical features that they tie to requires many training images. In our case, TNG50 did not include enough mergers at $M_\star > 10^{10} M_\odot$ for the network to understand how to classify high-mass mergers. Conversely, Margalef-Bentabol et al. (2024) did not include as many $M_\star \sim 10^9 M_\odot$ galaxies as $M_\star > 10^{10} M_\odot$ galaxies, and their completeness scores were highly mass dependent (see their Figure 8). Training a network with realistic mock images improves performance on real data (Bottrell et al. 2019). In addition to supplying a deep network with high-quality mock images, it is also necessary to have an adequate number of training examples spanning different stellar masses, merger stages, and mass ratios.

5.2.2. *Impact of Viewing Angle on Accuracy*

Orientation angle has a known effect on classifying the morphology of galaxies. Extending this to mergers, the orientation angle can make it easier or more difficult to visually identify a companion galaxy or merger. Bickley et al. (2024) suggests that there may be an upper limit on the accuracy of a CNN for merger identification around 90 percent, partially due to orientation. They note that at some level, there is no way to distinguish a clumpy nonmerger from a merger, or to view every galaxy from a favorable angle. Wilkinson et al. (2024) conduct a thorough investigation of the limitations of their merger identification scheme by viewing angle. They classify a $3 \times 10^{10} M_\odot$ major merger ($\mu = 0.26$) viewed at 648 angles. They find for this easy-to-identify major merger at $z = 0.05$ that $\sim$8% of the viewing angles are unable to identify the merger, even with their well trained LDA and random forest algorithms.

This might be an inherent limit, and indeed our accuracy does not exceed $\sim$90%. When we account for the fact that our sample is composed of lower mass (by up to 100 times) and higher redshift (which has an effect on image resolution) galaxies and mergers with larger mass ratios (i.e. minor mergers), we should expect an even lower accuracy upper limit. If this upper limit applies to $z < 1$ galaxies, then the upper limit will undoubtedly be even lower than 90% at $z = 1 - 1.5$.

5.2.3. *Impact of Star Formation*

As seen in Figure 10, the CNN recognizes star formation to be important to its final classifications. With the multiband images that the CNN is fed, it can learn that a bluer galaxy, and thus a galaxy undergoing recent star formation, is more likely to be a merger than a nonmerger. This could be seen as a drawback, and Bottrell et al. (2019) use one band to specifically avoid the CNN classifying bluer galaxies as mergers and instead force it to classify based on morphology. However, we argue this trend with star formation rate shows that CNN is learning something physically meaningful: a bluer galaxy is more likely to be a merger than a nonmerger. Indeed, the mass-matched mergers in our TNG50 dataset, on average, have higher star formation rates than the nonmergers (Schechter et al. 2025). Therefore, it makes sense that the CNN would identify a bluer galaxy as a merger more often.

We confirm in Figure 11 (left plot) that the nonmergers that are incorrectly classified have a higher mean sSFR than the nonmergers that are correctly classified. The right plot on Figure 11 shows that mergers misclassified as nonmergers have a lower mean sSFR than the mergers correctly classified, especially with a few low

sSFR galaxies causing a longer tail of the missclassified distribution.

In order to confirm that color information is not hindering our model, we additionally ran two versions of a greyscale CNN. One model was fed a single image that summed the flux from all three filters in preprocessing, and thus did not have the breakdown of astronomical filters than the three channel CNN did. The other was given a single filter F814W image. Both greyscale CNNs had $\sim 10\%$ lower overall accuracies than our main three channel model. More nonmergers were misclassified as mergers in the greyscale CNNs, leading to the decrease in accuracy. Additionally, both greyscale models still saw a gradient with sSFR in the latent space, implying that the CNN was able to pick up on sSFR even without astronomical filter information, purely from morphology. Taking filters away did not decrease a reliance on star formation, and appears to have removed some key features the model used when separating mergers from nonmergers.

5.3. *Looking Ahead With Applications to Real Data*

In this work we used fully radiative transferred images, as it is important to train on mock images as similar to real observations as possible (Bottrell et al. 2019). However, no matter how careful the mock image procedure is, there will always be differences between mock images and real data such as noise, image artefacts, and morphology of galaxies depending on the simulation used for training. Ćiprijanović et al. (2021) applied domain adaptation techniques, which enable the network to find and utilize similar features from the training domain (mock images of simulated galaxies) and the target domain (real observations) i.e., domain-invariant features. These techniques help overcome the limitation any differences in the mock images and real observations. Since radiative transfer is expensive, some papers opt to not use radiative transfer with only a small reduction in accuracy (e.g., Bottrell et al. 2019). At $z > 1$ we argue it is important to use radiative transfer as dust is impacting our observations and AGN, which can be very dusty, are more common. In an upcoming paper, we will explore using domain adaptation in place of full radiative transfer for merger classification with CNNs in order to save computational time.

As we broach the high-z merger universe, we want to build trust in our use of AI and understand where AI cannot provide all of the answers. Using XAI techniques is crucial as we are looking at imaging where there is no obvious correct answer. Additionally, understanding which galaxies cause miscalibration can provide insight into which tasks ML is well-suited for and for which tasks more classical methods may still be preferred. For example, combining spectroscopic pairs for early-stage mergers with CNNs for late-stage mergers can give a more complete catalog of all mergers in a sample (Ferreira et al. 2024a). This also makes the CNN's task easier, because it only has to learn what a late-stage merger looks like, instead of generalizing to any stage of merger. Knowing which tools to apply to which problems is crucial as we enter an era of big data in astronomy.

UMAPs show us that the network is sensitive to the stellar mass and star formation rates of the galaxies. In the future, we could potentially improve performance by combining galaxies with $M_\star < 10^{9.5} M_\odot$ from TNG50 with a sample of galaxies with $M_\star > 10^{9.5} M_\odot$ from the larger simulation box size, TNG100, to create a more mass-balanced and larger training set. This would be similar to the treatment in Margalef-Bentabol et al. (2024), where authors combine TNG100 and TGN300 galaxies for their training set. Our nonmerger sample is currently only mass-matched due to the small box size of TNG50. To match in any other quantity required widening the mass match threshold too much, due to the small number of galaxies with $M_\star > 10^{10} M_\odot$. However, if we draw high-mass galaxies from TNG100, we could find SFR and mass matched nonmergers to try to reduce misclassifications. This would also provide a larger training set overall, which would be helpful since our current training set is small by ML standards. With a training set of similar numbers of major and high-mass mergers as minor and low-mass mergers, we could potentially improve the distinction between mergers and nonmergers for all subcategories of galaxies.

6. CONCLUSION

We trained a CNN on mock *HST* CANDELS images to identify galaxy mergers at $1 \leq z \leq 1.5$. We used the highest spatial resolution simulation in the TNG suite, TNG50, enabling more detailed structures of galaxies and galaxies of lower masses to be included in our training set. With mergers in stages from pre- to post-coalescence, $M_\star > 10^8 M_\odot$, and $\mu > 1:10$, we successfully classified less obvious galaxy mergers close to cosmic noon. Our main results are as follows:

1. Our network was 73% accurate, and was able to identify major mergers about 76% and minor mergers about 68% of the time.

2. Early-stage mergers (two clear galaxies) were identified about 10% more often than late-stage mergers (around coalescence).

3. Orientation angle matters when searching for mergers, as there are some angles where the merger is unlikely to be correctly identified.

4. In order for a single model to accurately classify galaxies in a wide range of stellar masses (e.g., $\sim 10^8 M_\odot - 10^{12} M_\odot$), the training set must include sufficient examples from the entire mass range (even if that is not representative of the observed galaxy mass distribution).

5. CNNs can learn that higher star formation is likely to be in a merger, even with greyscale images. SFR- and mass-matched training sets are needed to confidently classify high star-forming, nonmerging galaxies or mergers between low star-forming galaxies.

Accurately identifying galaxy mergers is a key step in understanding how galaxies build stellar mass and evolve in morphologies over time. To do this with large imaging surveys we need trustworthy ML algorithms. Understanding how to build better training sets and where high-z merger identification is failing is crucial as we step into the era of *JWST*, *Roman*, and *Rubin*.

ACKNOWLEDGMENTS

A.L.S. would like to thank Michelle Ntampaka for helpful discussions in the early stages of this work.

A.L.S. and J.M.C. acknowledge support from NASA's Astrophysics Data Analysis program, grant number 80NSSC21K0646, and NSF AST-1847938.

XS acknowledges the support from the NASA theory grant JWST-AR-04814.

The work of A.S. was supported by the National Science Foundation MPS-Ascend Postdoctoral Research Fellowship under grant No. 2213288.

L.B. acknowledges support from NASA Astrophysics Theory program, grant 80NSSC22K0808, and NSF AAG 2307171.

A.Ć: This work was produced by Fermi Forward Discovery Group, LLC under Contract No. 89243024CSC000002 with the U.S. Department of Energy, Office of Science, Office of High Energy Physics. The United States Government retains and the publisher, by accepting the work for publication, acknowledges that the United States Government retains a nonexclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this work, or allow others to do so, for United States Government purposes. The Department of Energy will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan).

We acknowledge the use of LLMs in troubleshooting code and occasional rewording of parts of a sentence. The main code and complete sentences are original work.

We acknowledge the Deep Skies Lab as a community of multi-domain experts and collaborators who've facilitated an environment of open discussion, idea generation, and collaboration. This community was important for the development of this project.

Author Contributions: The following authors contributed in different ways to the manuscript.
Schechter: Writing manuscript, mock image creation post radiative transfer, all ML code and analysis
Ćiprijanović: editing manuscript, mentoring and overseeing ML code and analysis
Shen: Radiative transfer, writing section 2.2.1
Nevin: Editing manuscript, creating TNG merger catalog, mentoring and overseeing ML code and analysis
Comerford: Editing manuscript, mentoring and overall direction of paper
Stemo: Assistance with mock image creation
Blecha: Initial ideas and direction of paper

REFERENCES

Ackermann, S., Schawinski, K., Zhang, C., Weigel, A. K., & Turp, M. D. 2018, Monthly Notices of the Royal Astronomical Society, 479, 415, doi: 10.1093/mnras/sty1398

Akeson, R., Armus, L., Bachelet, E., et al. 2019, The Wide Field Infrared Survey Telescope: 100 Hubbles for the 2020s. https://arxiv.org/abs/1902.05569

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., et al. 2019, Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI, arXiv, doi: 10.48550/arXiv.1910.10045

Avirett-Mackenzie, M. S., Villforth, C., Huertas-Company, M., et al. 2024, Monthly Notices of the Royal Astronomical Society, 528, 6915, doi: 10.1093/mnras/stae183

Baes, M., & Camps, P. 2015, Astronomy and Computing, 12, 33, doi: 10.1016/j.ascom.2015.05.006
Baes, M., Verstappen, J., De Looze, I., et al. 2011, The Astrophysical Journal Supplement Series, 196, 22, doi: 10.1088/0067-0049/196/2/22
Bickley, R. W., Wilkinson, S., Ferreira, L., et al. 2024, Monthly Notices of the Royal Astronomical Society, 534, 2533, doi: 10.1093/mnras/stae2246
Bickley, R. W., Bottrell, C., Hani, M. H., et al. 2021, Monthly Notices of the Royal Astronomical Society, 504, 372, doi: 10.1093/mnras/stab806
Bottrell, C., Hani, M. H., Teimoorinia, H., Patton, D. R., & Ellison, S. L. 2022, Monthly Notices of the Royal Astronomical Society, 511, 100, doi: 10.1093/mnras/stab3717
Bottrell, C., Hani, M. H., Teimoorinia, H., et al. 2019, Monthly Notices of the Royal Astronomical Society, 490, 5390, doi: 10.1093/mnras/stz2934
Bouwens, R. J., Illingworth, G. D., Blakeslee, J. P., Broadhurst, T. J., & Franx, M. 2004, The Astrophysical Journal, 611, L1, doi: 10.1086/423786
Brier, G. W. 1950, Monthly Weather Review, 78, 1, doi: 10.1175/1520-0493(1950)078⟨0001:VOFEIT⟩2.0.CO;2
Buitrago, F., Trujillo, I., Conselice, C. J., & Häußler, B. 2013, Monthly Notices of the Royal Astronomical Society, 428, 1460, doi: 10.1093/mnras/sts124
Camps, P., & Baes, M. 2015, Astronomy and Computing, 9, 20, doi: 10.1016/j.ascom.2014.10.004
—. 2020, SKIRT 9: redesigning an advanced dust radiative transfer code to allow kinematics, line transfer and polarization by aligned dust grains, arXiv, doi: 10.48550/arXiv.2003.00721
Cavanagh, M. K., Bekki, K., & Groves, B. A. 2021, Monthly Notices of the Royal Astronomical Society, 506, 659, doi: 10.1093/mnras/stab1552
Conroy, C., & Gunn, J. E. 2010, The Astrophysical Journal, 712, 833, doi: 10.1088/0004-637X/712/2/833
Conroy, C., Gunn, J. E., & White, M. 2009, The Astrophysical Journal, 699, 486, doi: 10.1088/0004-637X/699/1/486
Conselice, C. J. 2003, The Astrophysical Journal Supplement Series, 147, 1, doi: 10.1086/375001
—. 2014, Annual Review of Astronomy and Astrophysics, 52, 291, doi: 10.1146/annurev-astro-081913-040037
Conselice, C. J., Rajgor, S., & Myers, R. 2008, Monthly Notices of the Royal Astronomical Society, 386, 909, doi: 10.1111/j.1365-2966.2008.13069.x
Crain, R. A., Schaye, J., Bower, R. G., et al. 2015, Monthly Notices of the Royal Astronomical Society, 450, 1937, doi: 10.1093/mnras/stv725
Darg, D. W., Kaviraj, S., Lintott, C. J., et al. 2010, Monthly Notices of the Royal Astronomical Society, 401, 1043, doi: 10.1111/j.1365-2966.2009.15686.x
DeGroot, M. H., & Fienberg, S. E. 1983, Journal of the Royal Statistical Society. Series D (The Statistician), 32, 12, doi: 10.2307/2987588
Deng, L. 2012, IEEE Signal Processing Magazine, 29, 141, doi: 10.1109/MSP.2012.2211477
Dieleman, S., Willett, K. W., & Dambre, J. 2015, Monthly Notices of the Royal Astronomical Society, 450, 1441, doi: 10.1093/mnras/stv632
Domínguez Sánchez, H., Huertas-Company, M., Bernardi, M., Tuccillo, D., & Fischer, J. L. 2018, Monthly Notices of the Royal Astronomical Society, 476, 3661, doi: 10.1093/mnras/sty338
Domínguez Sánchez, H., Martin, G., Damjanov, I., et al. 2023, Monthly Notices of the Royal Astronomical Society, 521, 3861, doi: 10.1093/mnras/stad750
Draine, B. T., Dale, D. A., Bendo, G., et al. 2007, The Astrophysical Journal, 663, 866, doi: 10.1086/518306
Duncan, K., Conselice, C. J., Mundy, C., et al. 2019, The Astrophysical Journal, 876, 110, doi: 10.3847/1538-4357/ab148a
Dwek, E. 1998, ApJ, 501, 643, doi: 10.1086/305829
Ferreira, L., Conselice, C. J., Duncan, K., et al. 2020, The Astrophysical Journal, 895, 115, doi: 10.3847/1538-4357/ab8f9b
Ferreira, L., Conselice, C. J., Kuchner, U., & Tohill, C.-B. 2022, The Astrophysical Journal, 931, 34, doi: 10.3847/1538-4357/ac66ea
Ferreira, L., Ellison, S. L., Patton, D. R., et al. 2024a, Galaxy evolution in the post-merger regime I – Most merger-induced in-situ stellar mass growth happens post-coalescence, arXiv. http://arxiv.org/abs/2410.06356
Ferreira, L., Conselice, C. J., Sazonova, E., et al. 2023, The Astrophysical Journal, 955, 94, doi: 10.3847/1538-4357/acec76
Ferreira, L., Bickley, R. W., Ellison, S. L., et al. 2024b, Monthly Notices of the Royal Astronomical Society, 533, 2547, doi: 10.1093/mnras/stae1885
Finkelstein, S. L. 2016, Publications of the Astronomical Society of Australia, 33, e037, doi: 10.1017/pasa.2016.26
Gardner, J. P., Mather, J. C., Clampin, M., et al. 2006, SSRv, 123, 485, doi: 10.1007/s11214-006-8315-7
Grogin, N. A., Kocevski, D. D., Faber, S. M., et al. 2011, The Astrophysical Journal Supplement Series, 197, 35, doi: 10.1088/0067-0049/197/2/35

Groves, B., Dopita, M. A., Sutherland, R. S., et al. 2008, The Astrophysical Journal Supplement Series, 176, 438, doi: 10.1086/528711

Guo, Y., Ferguson, H. C., Giavalisco, M., et al. 2013, The Astrophysical Journal Supplement Series, 207, 24, doi: 10.1088/0067-0049/207/2/24

He, K., Zhang, X., Ren, S., & Sun, J. 2015, Deep Residual Learning for Image Recognition, arXiv, doi: 10.48550/arXiv.1512.03385

Ivezić, \., Kahn, S. M., Tyson, J. A., et al. 2019, The Astrophysical Journal, 873, 111, doi: 10.3847/1538-4357/ab042c

J. Bouwens, R., Aravena, M., Decarli, R., et al. 2016, The Astrophysical Journal, 833, 72, doi: 10.3847/1538-4357/833/1/72

Johnson, B. D. 2021, bd-j/sedpy: sedpy v0.2.0, Zenodo, doi: 10.5281/zenodo.4582723

Kartaltepe, J. S., Mozena, M., Kocevski, D., et al. 2015, The Astrophysical Journal Supplement Series, 221, 11, doi: 10.1088/0067-0049/221/1/11

Kaviraj, S. 2014, Monthly Notices of the Royal Astronomical Society, 440, 2944, doi: 10.1093/mnras/stu338

Kaviraj, S., Laigle, C., Kimm, T., et al. 2017, Monthly Notices of the Royal Astronomical Society, 467, 4739, doi: 10.1093/mnras/stx126

Kingma, D. P., & Ba, J. 2017, Adam: A Method for Stochastic Optimization, arXiv, doi: 10.48550/arXiv.1412.6980

Koekemoer, A. M., Faber, S. M., Ferguson, H. C., et al. 2011, The Astrophysical Journal Supplement Series, 197, 36, doi: 10.1088/0067-0049/197/2/36

Krist, J. E., Hook, R. N., & Stoehr, F. 2011, 8127, 81270J, doi: 10.1117/12.892762

Lintott, C. J., Schawinski, K., Slosar, A., et al. 2008, Monthly Notices of the Royal Astronomical Society, 389, 1179, doi: 10.1111/j.1365-2966.2008.13689.x

Lotz, J. M., Jonsson, P., Cox, T. J., & Primack, J. R. 2008, Monthly Notices of the Royal Astronomical Society, 391, 1137, doi: 10.1111/j.1365-2966.2008.14004.x

Lotz, J. M., Primack, J., & Madau, P. 2004, The Astronomical Journal, 128, 163, doi: 10.1086/421849

Mantha, K. B., McIntosh, D. H., Brennan, R., et al. 2018, Monthly Notices of the Royal Astronomical Society, 475, 1549, doi: 10.1093/mnras/stx3260

Margalef-Bentabol, B., Wang, L., Marca, A. L., et al. 2024, Astronomy & Astrophysics, 687, A24, doi: 10.1051/0004-6361/202348239

Marinacci, F., Vogelsberger, M., Pakmor, R., et al. 2018, Monthly Notices of the Royal Astronomical Society, 480, 5113, doi: 10.1093/mnras/sty2206

Martin, G., Kaviraj, S., Devriendt, J. E. G., Dubois, Y., & Pichon, C. 2018, Monthly Notices of the Royal Astronomical Society, 480, 2266, doi: 10.1093/mnras/sty1936

McInnes, L., Healy, J., & Melville, J. 2020, UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, arXiv, doi: 10.48550/arXiv.1802.03426

McLure, R. J., Cirasuolo, M., Dunlop, J. S., Foucaud, S., & Almaini, O. 2009, Monthly Notices of the Royal Astronomical Society, 395, 2196, doi: 10.1111/j.1365-2966.2009.14677.x

Mihos, J. C., & Hernquist, L. 1996, The Astrophysical Journal, 464, 641, doi: 10.1086/177353

Naeini, M. P., Cooper, G., & Hauskrecht, M. 2015, Proceedings of the AAAI Conference on Artificial Intelligence, 29, doi: 10.1609/aaai.v29i1.9602

Naiman, J. P., Pillepich, A., Springel, V., et al. 2018, Monthly Notices of the Royal Astronomical Society, 477, 1206, doi: 10.1093/mnras/sty618

Nelson, D., Pillepich, A., Springel, V., et al. 2018, Monthly Notices of the Royal Astronomical Society, 475, 624, doi: 10.1093/mnras/stx3040

—. 2019, Monthly Notices of the Royal Astronomical Society, 490, 3234, doi: 10.1093/mnras/stz2306

Nevin, R., Blecha, L., Comerford, J., & Greene, J. 2019, The Astrophysical Journal, 872, 76, doi: 10.3847/1538-4357/aafd34

Newman, A. B., Ellis, R. S., Bundy, K., & Treu, T. 2012, The Astrophysical Journal, 746, 162, doi: 10.1088/0004-637X/746/2/162

Niculescu-Mizil, A., & Caruana, R. 2005, in Proceedings of the 22nd international conference on Machine learning, ICML '05 (New York, NY, USA: Association for Computing Machinery), 625–632, doi: 10.1145/1102351.1102430

Oesch, P. A., Bouwens, R. J., Illingworth, G. D., Labbé, I., & Stefanon, M. 2018, The Astrophysical Journal, 855, 105, doi: 10.3847/1538-4357/aab03f

Omori, K. C., Bottrell, C., Walmsley, M., et al. 2023, Astronomy and Astrophysics, 679, A142, doi: 10.1051/0004-6361/202346743

Ouchi, M., Mobasher, B., Shimasaku, K., et al. 2009, The Astrophysical Journal, 706, 1136, doi: 10.1088/0004-637X/706/2/1136

Pearson, W. J., Wang, L., Trayford, J. W., Petrillo, C. E., & Tak, F. F. S. v. d. 2019, Astronomy & Astrophysics, 626, A49, doi: 10.1051/0004-6361/201935355

Pillepich, A., Nelson, D., Hernquist, L., et al. 2018, Monthly Notices of the Royal Astronomical Society, 475, 648, doi: 10.1093/mnras/stx3112

Pillepich, A., Nelson, D., Springel, V., et al. 2019, Monthly Notices of the Royal Astronomical Society, 490, 3196, doi: 10.1093/mnras/stz2338

Planck Collaboration, Ade, P. A. R., Aghanim, N., et al. 2016, Astronomy and Astrophysics, 594, A13, doi: 10.1051/0004-6361/201525830

Rodriguez-Gomez, V., Genel, S., Vogelsberger, M., et al. 2015, Monthly Notices of the Royal Astronomical Society, 449, 49, doi: 10.1093/mnras/stv264

Rodriguez-Gomez, V., Snyder, G. F., Lotz, J. M., et al. 2019, MNRAS, 483, 4140, doi: 10.1093/mnras/sty3345

Rose, C., Kartaltepe, J. S., Snyder, G. F., et al. 2023, The Astrophysical Journal, 942, 54, doi: 10.3847/1538-4357/ac9f10

—. 2024, The Astrophysical Journal, 976, L8, doi: 10.3847/2041-8213/ad8dd4

Saftly, W., Baes, M., & Camps, P. 2014, A&A, 561, A77, doi: 10.1051/0004-6361/201322593

Scaramella, R., Amiaux, J., Mellier, Y., et al. 2022, Astronomy &; Astrophysics, 662, A112, doi: 10.1051/0004-6361/202141938

Schartmann, M., Meisenheimer, K., Camenzind, M., Wolf, S., & Henning, T. 2005, Astronomy & Astrophysics, 437, 861, doi: 10.1051/0004-6361:20042363

Schaye, J., Crain, R. A., Bower, R. G., et al. 2015, Monthly Notices of the Royal Astronomical Society, 446, 521, doi: 10.1093/mnras/stu2058

Schechter, A. L., Genel, S., Terrazas, B., et al. 2025, The Astrophysical Journal, 989, 149, doi: 10.3847/1538-4357/ade791

Selvaraju, R. R., Cogswell, M., Das, A., et al. 2020, International Journal of Computer Vision, 128, 336, doi: 10.1007/s11263-019-01228-7

Shen, X., Vogelsberger, M., Nelson, D., et al. 2022, Monthly Notices of the Royal Astronomical Society, 510, 5560, doi: 10.1093/mnras/stab3794

—. 2020, Monthly Notices of the Royal Astronomical Society, 495, 4747, doi: 10.1093/mnras/staa1423

Shen, X., Vogelsberger, M., Borrow, J., et al. 2024, MNRAS, 534, 1433, doi: 10.1093/mnras/stae2156

Simmons, B. D., Lintott, C., Willett, K. W., et al. 2017, Monthly Notices of the Royal Astronomical Society, 464, 4420, doi: 10.1093/mnras/stw2587

Simonyan, K., Vedaldi, A., & Zisserman, A. 2013, Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps, arXiv, doi: 10.48550/arXiv.1312.6034

Smethurst, R. J., Simmons, B. D., Géron, T., et al. 2025, Monthly Notices of the Royal Astronomical Society, doi: 10.1093/mnras/staf506

Snyder, G. F., Rodriguez-Gomez, V., Lotz, J. M., et al. 2019, Monthly Notices of the Royal Astronomical Society, 486, 3702, doi: 10.1093/mnras/stz1059

Springel, V., Pakmor, R., Pillepich, A., et al. 2018, Monthly Notices of the Royal Astronomical Society, 475, 676, doi: 10.1093/mnras/stx3304

Toomre, A., & Toomre, J. 1972, The Astrophysical Journal, 178, 623, doi: 10.1086/151823

Torrey, P., Snyder, G. F., Vogelsberger, M., et al. 2015, MNRAS, 447, 2753, doi: 10.1093/mnras/stu2592

Vogelsberger, M., Genel, S., Springel, V., et al. 2014, Monthly Notices of the Royal Astronomical Society, 444, 1518, doi: 10.1093/mnras/stu1536

Vogelsberger, M., Nelson, D., Pillepich, A., et al. 2020, Monthly Notices of the Royal Astronomical Society, 492, 5167, doi: 10.1093/mnras/staa137

Walmsley, M., Allen, C., Aussel, B., et al. 2023, Journal of Open Source Software, 8, 5312, doi: 10.21105/joss.05312

Wilkinson, S., Ellison, S. L., Bottrell, C., et al. 2024, Monthly Notices of the Royal Astronomical Society, 528, 5558, doi: 10.1093/mnras/stae287

Willett, K. W., Galloway, M. A., Bamford, S. P., et al. 2017, Monthly Notices of the Royal Astronomical Society, 464, 4176, doi: 10.1093/mnras/stw2568

Ćiprijanović, A., Snyder, G. F., Nord, B., & Peek, J. E. G. 2020, Astronomy and Computing, 32, 100390, doi: 10.1016/j.ascom.2020.100390

Ćiprijanović, A., Kafkes, D., Downey, K., et al. 2021, Monthly Notices of the Royal Astronomical Society, 506, 677, doi: 10.1093/mnras/stab1677