The Astrophysical Journal, 858:114 (17pp), 2018 May 10 https://doi.org/10.3847/1538-4357/aabfed

Deep Learning Identifies High-z Galaxies in a Central Blue Nugget Phase in a Characteristic Mass Range

M. Huertas-Company[1,2,3], J. R. Primack[4], A. Dekel[4,5], D. C. Koo[6], S. Lapiner[5], D. Ceverino[7], R. C. Simons[8], G. F. Snyder[9], M. Bernardi[10], Z. Chen[11], H. Domínguez-Sánchez[10], C. T. Lee[1,4], B. Margalef-Bentabol[1], and D. Tuccillo[1,12]
[1] Sorbonne Université, Observatoire de Paris, Université PSL, CNRS, LERMA, F-75014, Paris, France; marc.huertas@obspm.fr
[2] Sorbonne Paris Cité, Université Paris Diderot, F-75013, France
[3] Institut Universitaire de France, France
[4] Physics Department, University of California, Santa Cruz, CA 95064, USA
[5] Racah Institute of Physics, The Hebrew University, Jerusalem 91904, Israel
[6] Department of Astronomy and Astrophysics, University of California, Santa Cruz, CA 95064, USA
[7] Institut fur Theoretische Astrophysik, Zentrum fur Astronomie, Universitiat Heidelberg, Albert-Ueberle-Str. 2 D-69120 Heidelberg, Germany
[8] Johns Hopkins University, 3400 N. Charles Street, Baltimore, MD 21218, USA
[9] Space Telescope Science Institute, 3700 San Martin Drive, Baltimore, MD 21218, USA
[10] Department of Physics and Astronomy, University of Pennsylvania, Philadelphia, PA 19104, USA
[11] Shanghai Key Lab for Astrophysics, Shanghai Normal University, 100 Guilin Road, 200234, Shanghai, People's Republic of China
[12] MINES Paristech, PSL Research University Centre for Mathematical Morphology, Fontainebleau, France
Received 2018 March 21; revised 2018 April 17; accepted 2018 April 17; published 2018 May 15

Abstract

We use machine learning to identify in color images of high-redshift galaxies an astrophysical phenomenon predicted by cosmological simulations. This phenomenon, called the blue nugget (BN) phase, is the compact star-forming phase in the central regions of many growing galaxies that follows an earlier phase of gas compaction and is followed by a central quenching phase. We train a convolutional neural network (CNN) with mock "observed" images of simulated galaxies at three phases of evolution— pre-BN, BN, and post-BN—and demonstrate that the CNN successfully retrieves the three phases in other simulated galaxies. We show that BNs are identified by the CNN within a time window of ∼0.15 Hubble times. When the trained CNN is applied to observed galaxies from the CANDELS survey at $z = 1$–3, it successfully identifies galaxies at the three phases. We find that the observed BNs are preferentially found in galaxies at a characteristic stellar mass range, $10^{9.2-10.3}\ M_{\odot}$ at all redshifts. This is consistent with the characteristic galaxy mass for BNs as detected in the simulations and is meaningful because it is revealed in the observations when the direct information concerning the total galaxy luminosity has been eliminated from the training set. This technique can be applied to the classification of other astrophysical phenomena for improved comparison of theory and observations in the era of large imaging surveys and cosmological simulations.

Key words: galaxies: bulges – galaxies: fundamental parameters – galaxies: high-redshift

1. Introduction

Over the past years, we have acquired a detailed view of the statistical properties of galaxies at different cosmic epochs, thanks in particular to large-scale imaging and spectroscopic surveys (e.g., SDSS; York et al. 2000; CANDELS; Koekemoer et al. 2011). However, establishing causal connections between galaxy populations remains an important challenge (e.g., Lilly & Carollo 2016). This is obviously because of the timescales involved, which do not allow observations to follow the evolution of individual galaxies, and because of the degenerate link between commonly used observables and astrophysical processes.

This is particularly true for the processes leading to morphological transformations of galaxies, which remain largely unconstrained despite the large quantities of available data. A fundamental question, how bulges form and grow in galaxies at different cosmic times, is still largely debated. One of the reasons is that morphological observables extracted from images are rather simplistic and have essentially remained unchanged for many years. The characterization of galaxies is essentially limited to the prominence of the bulge and disk components based on the measurement of the central density (e.g., Barro et al. 2017), a parametric decomposition (e.g., Sérsic 1968; Peng et al. 2002), or a ratio between enclosed light at different radii (e.g., Abraham et al. 1996). The interpretation of these observables to constrain an assembly history is a very degenerate problem; i.e., there are many different processes that can lead to the same observables.

Our community is about to generate unprecedentedly large imaging data sets (e.g., *Euclid*, LSST, *WFIRST*). Hydro-cosmological numerical simulations are also growing rapidly. It is thus timely to investigate alternative ways to extract a maximum amount of information from polychromatic images that might help break degeneracies with physics and improve the comparison between observations and simulated data sets. This is precisely the goal of this work. Ideally, one would like to have morphological measurements that directly correlate with some astrophysical process as predicted by theory and detected in simulations. That way, it would be possible to isolate objects from large surveys with a high probability of experiencing a physical process and enable a more complete follow-up. This is easily understandable for galaxy–galaxy mergers, since it is a relatively well-defined process associated with expected morphological features, at least at a first approximation. As a result, many efforts have been made to characterize merging galaxies from images (e.g., Conselice et al. 2000; Lotz et al. 2008) and calibrate their observability timescale to constrain the merger history (Lotz et al. 2008;

Snyder et al. 2017). In that respect, it is important to calibrate with simulations that closely match the properties of the observed samples. For example, as shown in Cibinel et al. (2015), the morphological signatures of mergers at $z > 1$ differ from those of mergers at $z \sim 0$, and parametric classifications that robustly identify low-z mergers fail at $z > 1$.

Generalizing to other processes is less obvious, since one needs to find the appropriate tracers from the multiwavelength pixel distribution. In recent years, there has been significant progress in the image-processing community with the emergence of the so-called unsupervised feature-learning techniques, or deep learning (DL). These algorithms allow the user to automatically extract observables (or features) from the pixel space without any a priori dimension reduction. As in many other disciplines, DL is rapidly being adopted in astronomy as well. It has been successfully used for several standard classification (e.g., Dieleman et al. 2015; Huertas-Company et al. 2015; Domínguez-Sánchez et al. 2018) and regression (e.g., Tuccillo et al. 2018) problems. We aim at investigating here an alternative way of using these advanced machine-learning techniques to extract more physically relevant features from images and help establish a more solid link between theory and observations.

In this exploratory proof-of-concept work, we explore whether DL can be used to detect a phenomenon dubbed blue nugget (BN), recently found in numerical simulations of high-redshift galaxies. Indeed, these cosmological simulations (Zolotov et al. 2015; Tacchella et al. 2016a, 2016b) reveal that a large fraction of the simulated galaxies undergo events of gaseous compaction, triggered, e.g., by mergers or counter-rotating inflowing streams, which leads to a central BN at a characteristic stellar mass of $10^{9.2–10.3}\ M_{\odot}$. The BN phase, in turn, triggers central gas depletion and central quenching of star formation, sometimes surrounded by an extended, freshly formed, gaseous, star-forming ring/disk. Most of the structural, kinematic, and compositional galaxy properties undergo significant transitions as the galaxy evolves through the BN phase (Ceverino et al. 2015; A. Dekel et al. 2018, in preparation). One way to investigate whether these gaseous compactions are frequent in the observed galaxies would be to directly detect features in the data (stellar distribution, in our case) unequivocally associated with the BN phase. This is what we attempt in this paper. One main advantage of high-resolution numerical simulations over, for example, semi-analytical models or low-resolution large-volume simulations is that we can use them to generate realistic observed simulated images for which the evolution history is known by construction (e.g., Snyder et al. 2015). One can therefore isolate a sample of simulated galaxies in the BN phase, as well as in the pre-BN or post-BN phases. In this work, we use state-of-the-art zoom-in cosmological simulations with high spatial resolution (Ceverino et al. 2014) to generate mock images as observed by *HST* of galaxies in a BN phase. We then use this data set to train deep neural nets and explore whether the network is able to automatically find morphological proxies associated with the different phases in the observed mock data. We then apply the trained network to observed CANDELS data.

The paper proceeds as follows. Sections 2 and 3 describe the simulations and data used in this work. The main methodology is discussed in Section 4. We show the main results of simulations and observations in Sections 5 and 6, respectively.

2. Simulations

2.1. Main Properties of the Simulations

We use a set of zoom-in hydro-cosmological simulations of 35 intermediate-mass galaxies, of which 31 are used in this work. The typical stellar mass of the simulated galaxies at $z \sim 2$ is $10^{10}\ M_{\odot}$, as shown in Table 2. This is part of the VELA simulation suite, which has been described and analyzed in several previous works (Ceverino et al. 2014, 2015; Tacchella et al. 2015; Zolotov et al. 2015; Tomassetti et al. 2016; Tacchella et al. 2016b). We refer the reader to the aforementioned works for a detailed description of the simulations. We summarize here only the most relevant properties. The initial conditions for the simulations are based on dark matter halos that were drawn from dissipationless *N*-body simulations. The simulations were run with the AdaptiveRefinement Tree (ART) code (Kravtsov et al. 1997; Kravtsov 2003; Ceverino & Klypin 2009), and the maximum resolution is 17–35 pc at all times, which is achieved at densities of $\sim 10^4$–$10^3\ \mathrm{cm}^{-3}$. The code includes several physical processes relevant for galaxy formation: gas cooling by atomic hydrogen and helium, metal and molecular hydrogen cooling, photoionization heating by the UV background with partial self-shielding, star formation, stellar mass loss, metal enrichment of the interstellar medium (ISM), and stellar feedback. In particular, the high spatial resolution allows tracing the cosmological streams that feed galaxies at high redshift, including mergers and smooth flows, and they resolve the violent disk instabilities (VDIs) that govern high-z disk evolution and bulge formation (Dekel et al. 2009). This is important for this work focused on the growth of bulges and the reason why this small set of simulations is preferred to larger but lower-resolution runs like Illustris. We recall that the gravitational softening for baryons in the Illustris series is of the order of $\sim$1 kpc, which means that any physical process that acts in smaller scales is unresolved. This is the case of the BN phase explored in this work.

However, as with all state-of-the-art numerical simulations, the VELA simulations suffer from several limitations specially related to subgrid physics. Like other simulations, the treatment of star formation and feedback processes still depends on uncertain recipes. The code assumes an star formation rate (SFR) efficiency per freefall time without following in detail the formation of molecules and the effect of metallicity on the SFR (Krumholz & Dekel 2012). Additionally, no active galactic nuclei (AGN) feedback is yet included in the run used in this work. As a result, the full quenching observed in the data is not reached in many galaxies by the end of the simulations at $z \sim 1$. Since we are more interested here in the BN phase that occurs when the galaxy is still star-forming, we do not expect that AGNs will have a big impact on our results. However a color mismatch between simulated and observed galaxies might be expected. Besides that, as shown in Ceverino et al. (2014) and Tacchella et al. (2016b), the SFRs, gas fractions, and stellar-to-halo mass ratios are all close to the constraints imposed by observations, providing a better match to observations than earlier simulations. The uncertainties and any possible remaining mismatches by a factor of order 2 are comparable to the observational uncertainties.

We stress that we are fully aware that the simulations present many limitations and that they are still very far from capturing all the complex physics of galaxy formation. This is mainly

Table 1
Explanation of the 19 Camera Orientations Used to Generate Mock 2D Images from the Simulations

Camera Number	Orientation
cam00/02	Angular momentum face-on (opposite directions)
cam01/03	Angular momentum edge-on (opposite directions)
cam04	Angular momentum 45°
cam05/06/07	Fixed to *x*-, *y*-, and *z*-axis simulation box
cam08-11	Random (same simulation coord. for all snapshots)
cam12-18	Fully random

why the present work needs to be considered a proof-of-concept work in that respect. However, we are at a stage at which we can produce fairly realistic galaxies that capture some of the physical processes governing the assembly history, and we have good reason to think that this will be improved in the future. This enables a comparison with observations in a more general way that we explore in this work.

2.2. Mock CANDELized Images

The full output of the simulation is saved at many time steps and analyzed at steps of $\Delta a = 0.01$ of the expansion factor, which roughly correspond to ∼100 Myr at $z \sim 2$. For every snapshot between $z \sim 4$ and $z \sim 1$, we generate mock 2D images as they would be observed by the *HST*. They are generated using the radiative transfer code SUNRISE[13] (Jonsson 2006; Jonsson & Primack 2010; Jonsson et al. 2010) by propagating the light of stars through the dust. We refer to Snyder et al. (2015) for details on the procedure followed.

Very briefly, a spectral energy distribution (SED) is assigned to every star particle in the simulation based on its mass, age, and metallicity. The dust density is assumed to be directly proportional to the metal density predicted by the simulations. We set a dust-to-metals mass ratio of 0.4 (e.g., Dwek 1998; James et al. 2002) and the dust grain size distribution from that updated by Draine & Li (2007). SUNRISE then performs dust radiative transfer using a Monte Carlo ray-tracing technique. As each multiwavelength ray emitted by every star particle and H II region (according to its SED) propagates through the ISM and encounters dust mass, its energy is probabilistically absorbed or scattered until it exits the grid or enters one of the viewing apertures (cameras). The output of this process is then the SED at each pixel in all cameras. For this run, we set 19 cameras, of which five are fixed with respect to the angular momentum vector of each galaxy, seven are fixed in the simulation coordinates, and the remaining seven are fully random at each time step. The camera numbers are summarized in Table 1.

Finally, from these data cubes, we create raw mock images by integrating the SED in each pixel over the spectral response functions of the CANDELS WFC3 filters (F160W, F125W, and F105W) in the observer frame. Images are then convolved with the corresponding *HST* point spread function (PSF) at a given wavelength. We finally add a random real-noise stamp taken from the CANDELS data. This ensures that the galaxies are simulated at the same depth as the real CANDELS data and the correlated noise from the *HST* pipeline is well reproduced. We call this process CANDELization.

[13] SUNRISE is freely available at https://bitbucket.org/lutorm/sunrise.

For each 3D snapshot ($\Delta a = 0.01$), we therefore generate 19 different 2D images corresponding to the 19 different camera orientations. The resulting data set corresponds to approximately ∼10,000 images in each of three filters. Even if the CANDELS filters probed the optical rest frame up to $z \sim 3$, we included galaxies up to $z \sim 4$, since the most intense compaction events tend to happen at higher redshift in the VELA simulations. Given that the gas fractions (stellar-to-halo mass relations) are slightly underestimated (overestimated) in the simulations, as previously stated, including a higher redshift is justified and increases the size of our training set. We have checked, however, that the main results of the paper remain unaltered if only galaxies up to $z \sim 3$ are used. We emphasize that the same procedure has been used to generate mock *JWST* galaxies in the different filters; therefore, a similar analysis to the one presented in this work can be applied to this data set in order to prepare *JWST* observations.

3. Data

We also use *HST* observational data to test our model in Section 6. We use CANDELS images in the three infrared filters (F105W, F125W, and F160W) from the two GOODS fields (North and South; Grogin et al. 2011; Koekemoer et al. 2011). The selection is based on the morphological catalog presented in Huertas-Company et al. (2015), which is essentially a selection of the brightest galaxies (F150W < 24.5) from the official CANDELS catalogs (Guo et al. 2013; Barro et al. 2017). This is required to have enough signal-to-noise ratio (S/N) to measure morphological information from images. For this work, we select only galaxies in the redshift range 1–3 to match the simulated redshift range. As shown in Huertas-Company et al. (2016), the sample is mass complete down to 10^9 $M_\odot$ at $z \sim 1$ and 10^{10} at $z \sim 3$. We restrict our analysis to galaxies more massive than 10^9 $M_\odot$ to have enough statistics and match the typical stellar masses from the simulations. The sample might therefore suffer from incompleteness at high redshift. This is not critical for the illustrative purpose of this work.

In addition to the reduced images, we also use official CANDELS redshifts (Dahlen et al. 2013), which are a combination of photometric redshifts computed with several codes and spectroscopic when available. Stellar masses and star formation rates from SED fitting are also used. Stellar masses are computed through SED fitting using the best redshift adopting a Chabrier (2003) initial mass function (IMF). The SFRs are computed by combining IR and UV rest-frame luminosities (Kennicutt 1998; Bell et al. 2005) with a Chabrier (2003) IMF (see Barro et al. 2011 for more details). The following relation was used: $\mathrm{SFR_{UV+IR}} = 1.09 \times 10^{-10}(L_{\mathrm{IR}} + 3.3L_{2800})$. Total IR luminosities are obtained using Chary & Elbaz (2001) templates fitting MIPS 24 μm fluxes and applying a *Herschel*-based recalibration. For galaxies undetected in 24 μm, SFRs are estimated using rest-frame UV luminosities (Wuyts et al. 2011). We also compute, for the selected data set, the central mass density (Σ_1) following the methodology of Barro et al. (2017).

4. Methods: Training the Network

4.1. Training Set: Using the Simulation Metadata to Label Images

The final goal is to train a deep neural network to identify, from the mock images, the BN phase (and, consequently, the pre- and post-BN phases as well). As put forward by a previous

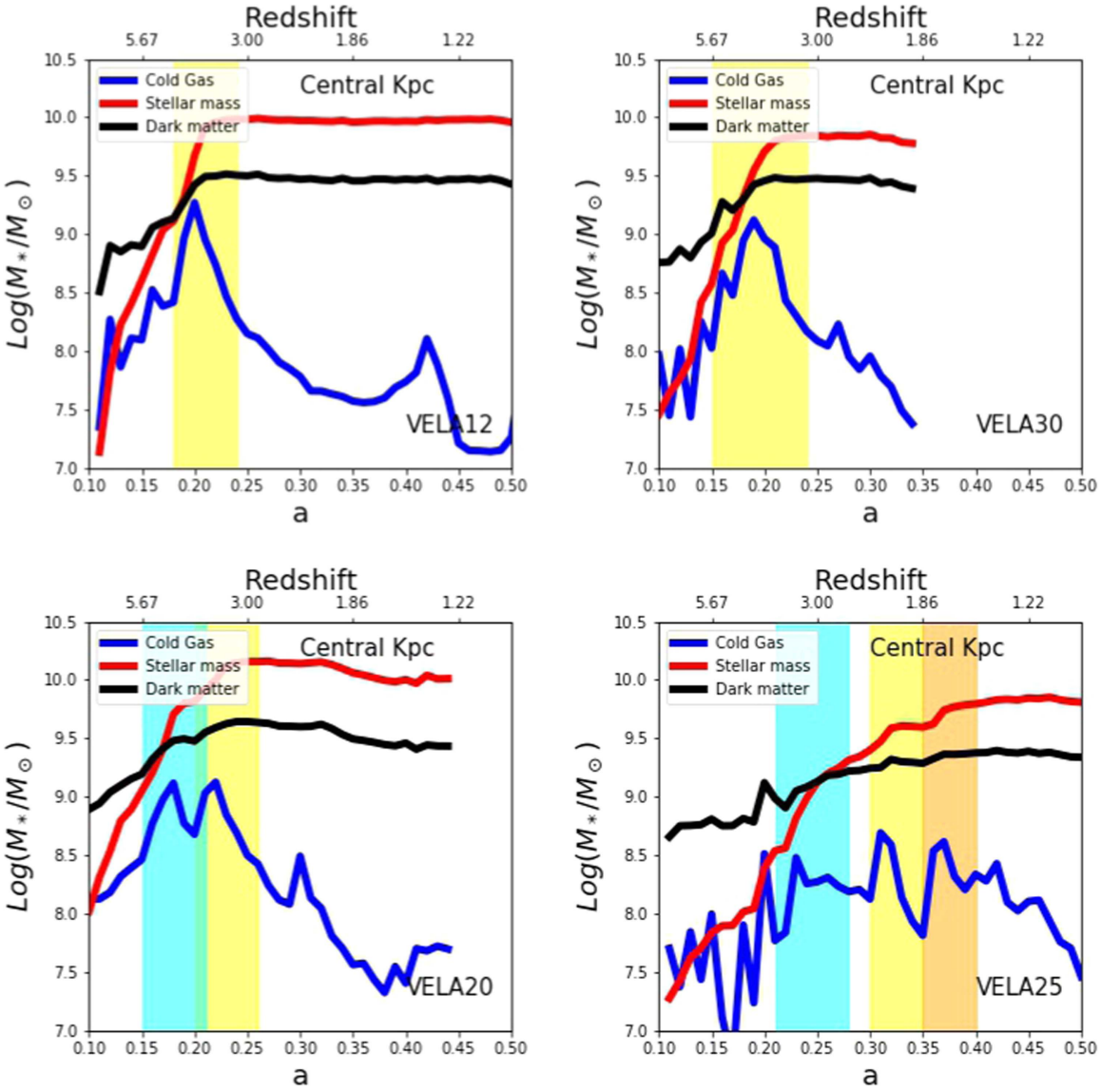

Figure 1. Definition of the different phases. Both the cold gas and the stellar mass in the central kpc are used to define the BN phase. The blue and red lines show the evolution of the cold gas and stellar mass in the central kpc as a function of time. The black line is the dark matter mass (adapted from Zolotov et al. 2015). The yellow shaded region shows the main BN event as defined in this work (see text for details). The second and third events are shown in cyan and orange, respectively. The ranking refers to the intensity of the event and not to the time of occurrence (see text for details). Each panel shows a different galaxy. The top panels show clear examples of massive galaxies with one unique BN phase. The bottom panels show more complex cases with more than one BN event.

analysis of the same simulated data set (Zolotov et al. 2015), almost all the simulated galaxies seem to evolve in three characteristic phases. They go from diffuse to compact star-forming objects through wet gas compaction and then quench in the central regions and build a central bulge that will, in most cases, rebuild a surrounding stellar disk. We notice that the intensity of the compaction depends on stellar mass, and while most of the simulations go through a BN phase, only the most massive become compact star-forming galaxies.

As part of the training set, we first define these three phases for all the galaxies in the simulation. The identification of the three phases is performed on an individual basis for each galaxy using the gas density evolution in the central galactic regions, as explained in Zolotov et al. (2015) and A. Dekel et al. (2018, in preparation). Basically, we identify the peak of the BN phase as the time at which the gas density in the central kpc is maximum. We define the end of the BN phase when the central stellar density stops increasing, which is a signature that star formation has been quenched in the center of the galaxy. The onset of the BN phase is considered to be when the central gas density starts to increase toward the BN peak. Naturally, this is more complicated than selecting the peak. In our current approach, the selection is done by eye using the 2D projection of the gas density to confirm the choice. Figure 1 shows the cold gas and stellar mass evolution in the central kpc for some galaxies for illustrative purposes. We also show the dark matter content in the central kpc. The key takeaway from these plots is that compaction is not always well defined and that it comes in many different flavors. There are, for instance, some clean cases, such as VELA12, in which there is a single peak of the gas mass. However, there are other cases, such as VELA25, for which the peak is not so pronounced and identifying the boundaries of the BN phase is not obvious and somewhat arbitrary. Notice also that many galaxies experience several BN

Table 2
Summary of the BN Phases for All Simulated Galaxies Used in This Work

Simulation	$z^1_{\rm onset}$	$z^1_{\rm post}$	$z^2_{\rm onset}$	$z^2_{\rm post}$	$z^3_{\rm onset}$	$z^3_{\rm post}$	$\mathrm{Log}M_*/M_\odot$ $z = z_{\rm comp}$	$\mathrm{Log}M_*/M_\odot$ $z = 2$
VELA01	1.86	1.38	⋯	⋯	⋯	⋯	10.05	9.39
VELA02	1.70	1.00	⋯	⋯	⋯	⋯	9.72	9.32
VELA03	3.00	1.94	1.27	0.96	⋯	⋯	9.47	9.70
VELA04	2.23	1.63	1.50	1.17	⋯	⋯	9.18	9.07
VELA05	1.38	1.08	⋯	⋯	⋯	⋯	9.47	9.09
VELA06	5.25	3.17	2.57	1.86	⋯	⋯	9.60	10.42
VELA07	3.55	2.57	4.88	3.35	⋯	⋯	10.39	10.83
VELA08	2.23	1.50	0.96	0.69	⋯	⋯	9.79	9.79
VELA09	4.00	3.00	1.63	1.33	⋯	⋯	9.73	10.09
VELA10	3.17	2.13	1.44	1.13	⋯	⋯	9.59	9.83
VELA11	4.00	2.85	2.12	1.70	⋯	⋯	9.67	10.05
VELA12	4.56	3.17	⋯	⋯	⋯	⋯	9.98	10.33
VELA13	2.85	2.03	⋯	⋯	⋯	⋯	9.76	10.06
VELA14	2.33	1.56	⋯	⋯	⋯	⋯	10.26	10.19
VELA15	2.70	2.13	1.70	1.38	⋯	⋯	9.70	9.77
VELA17	7.33	3.55	3.76	2.57	⋯	⋯	9.63	⋯
VELA19	9.00	4.56	2.70	2.13	⋯	⋯	9.75	⋯
VELA20	4.00	2.85	5.67	3.76	⋯	⋯	10.33	10.62
VELA21	3.55	2.57	4.88	3.35	7.33	4.56	10.51	10.65
VELA22	4.88	3.55	⋯	⋯	⋯	⋯	10.02	10.67
VELA25	2.33	1.86	3.76	2.57	1.86	1.50	9.89	9.91
VELA26	3.17	2.13	5.25	3.55	⋯	⋯	9.82	10.25
VELA27	2.23	1.70	3.35	2.57	⋯	⋯	9.90	10.01
VELA28	1.63	1.22	⋯	⋯	⋯	⋯	9.71	9.51
VELA30	5.67	3.17	⋯	⋯	⋯	⋯	9.87	10.25
VELA32	7.33	4.00	⋯	⋯	⋯	⋯	9.71	10.52
VELA33	4.88	3.00	3.35	2.45	2.33	1.78	9.61	10.73
VELA34	3.00	1.78	4.26	2.70	⋯	⋯	10.06	10.32

Note. For each galaxy, we show the redshift(s) at which the BN phase(s) were identified to start ($z_{\rm onset}$) and end ($z_{\rm post}$). We also indicate the stellar mass of the galaxy when the main BN phase occurs, as well as the stellar mass at $z = 2$. A dash (−) means that the simulation did not run until $z \sim 2$.

phases, as, for example, discussed in Tacchella et al. (2016a). In this work, we define a maximum of three BN phases for each galaxy, as shown in Table 2. Table 2 summarizes the redshifts of the BN phase of all galaxies analyzed. This is to say that the network that will be trained needs to somehow capture this heterogeneity in the process. It is important to keep this in mind when analyzing the results.

As can be seen in Table 2, in the simulations, the BN phase tends to happen at a characteristic galaxy stellar mass $\sim 10^{9.2-10.3}\ M_\odot$ (e.g., Zolotov et al. 2015). Given the existing correlation between mass and luminosity, this implies that there is a brightness gradient between pre-BN, BN, and post-BN, with pre-BN galaxies being generally fainter than post-BN galaxies. The difference in luminosity also implies a difference in S/N when the *HST* noise is added. A DL approach, as the one used in this work, has the unique power to automatically extract the optimal tracers from the data to minimize the classification error. It also implies a risk, since the network can potentially use any available information. In our case, given the properties of the training set, there is a potential risk that the network uses the S/N difference existing between the different phases to classify. Since we do not want the network to learn based on S/N but rather learn the characteristic features of the BN phase, we artificially shuffle the magnitudes of the galaxies given to the network. To do so, before adding the noise (see Section 2.2), we associate a random magnitude to all snapshots in the F160W filter (19–25, in order to match the CANDELS distribution). This way, galaxies in the different phases have similar luminosities and S/N distributions. By doing so, the characteristic mass information is also washed out, preventing the network from using that information to learn. We will discuss the effect of this choice in Section 6. We remark that all other properties are kept unchanged. It includes, obviously, the spatial distribution of pixels that measure the degree of compactness and the relative luminosities in each filter that are correlated with the SFR.

We thus use this three-class classification (pre-BN, BN, and post-BN) to associate a unique label to every simulated image. Pre-BN includes all galaxies before experiencing any compaction event, i.e., with a redshift larger than the maximum of ($z^1_{\rm onset}$, $z^3_{\rm onset}$, $z^3_{\rm onset}$). Galaxies in the BN phase are the ones with redshifts between $z^y_{\rm onset}$ and $z^y_{\rm post}$, with $y = 1, 2, 3$. Finally, all remaining images are labeled as post-BN. So, galaxies with several compaction events are classified as post-BN between two events. As a result of this labeling process, every mock image has an associated label corresponding to its evolutionary phase. The final data set consists, therefore, of ∼10,000 labeled images with the simulation assembly history that will be used to train and test a convolutional neural network (CNN) model.

Figure 2 shows some random example stamps of galaxies in the three phases in the *HST*/WFC3 F160W filter. Pre-BN galaxies generally look smaller, and post-BN galaxies tend to have a diffuse disk structure. However, no obvious visual difference is apparent. This underlines the challenge of this

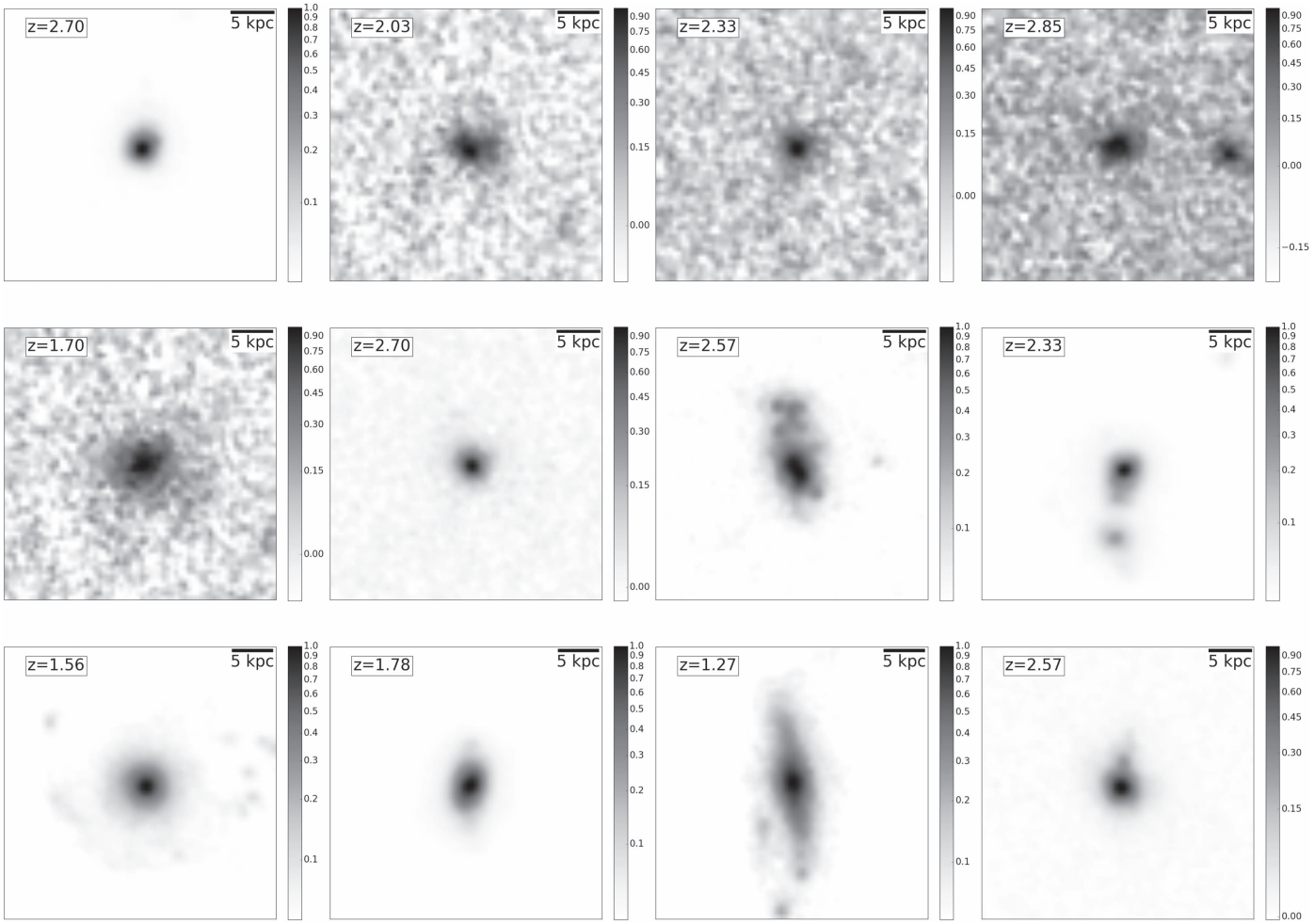

Figure 2. Random examples of simulated F160W CANDELized images in the three phases discussed in this work. The image size is $3\farcs8 \times 3\farcs8$. The top row shows pre-BN galaxies, the middle row shows galaxies in the BN phase, and the bottom row shows post-BN objects. The images have been rescaled so that they span the same range of luminosities in the three phases.

work, which is to train a CNN capable of distinguishing between the different phases.

4.2. Architecture

We use a very simple sequential CNN architecture with only three convolutional layers followed by two fully connected layers implemented in Keras[14] with a Theano back end (Figure 3). The main reason to use a relatively shallow network is the limited size of the training set. The architecture is inspired by previous configurations that were successful in detecting strong lenses in space-based images (Metcalf et al. 2018) and also for classical morphological classification (Domínguez-Sánchez et al. 2018). We then add two fully connected layers to perform the classification. The last layer has a *softmax* activation function to ensure that the three outputs behave like probabilities and add to one. We use a *categorical crossentropy* as loss function, and the model is optimized with the *adam* optimizer.

The network is fed with images (fits format) of fixed size (64×64 pixels), with three channels corresponding to the three main NIR CANDELS filters (F160W, F125W, and F105W). We also tried to include bluer filters (F850LP), but the results did not change significantly. For simplicity in this illustrative work, we used the three redder filters, since the pixel scale is the same and hence no interpolation is required. In principle, the number of filters could be increased without any significant modification of the methodology. The input size is a trade-off between properly probing the galaxy outskirts ($\sim$30 kpc in the redshift range 1–3) and having a small enough number of input parameters to prevent overfitting.

In addition to this, we also use standard techniques to avoid overfitting at first order. First, after each convolutional layer, we apply a 50% dropout. Second, we include a Gaussian noise layer at the entrance of the network to avoid the model learning from the noise pattern, given that our training set is small. Finally, we use real-time data augmentation. Images are randomly rotated (within 45°), flipped, and slightly off-centered by 5 pixels maximum at every iteration so that the network never sees exactly the same image.

4.3. Training and Validation Strategy

One obvious limitation we face in this work is that our training data set is made up of only $\sim$28 galaxies. Even though we increase the number of available images by using different camera orientations, as well as data augmentation, there is a potential risk that the network learns how to identify the

[14] https://keras.io/

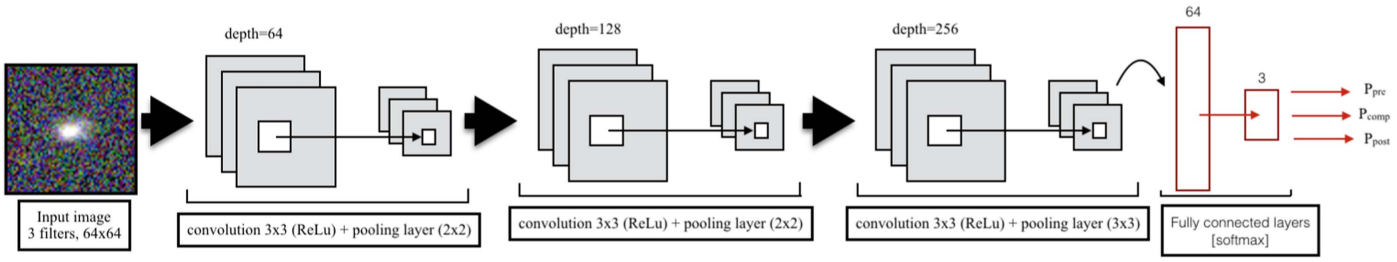

Figure 3. Architecture of the deep network used for classification in this work. The network is a standard and simple CNN configuration made of three convolutional layers followed by pooling and dropout.

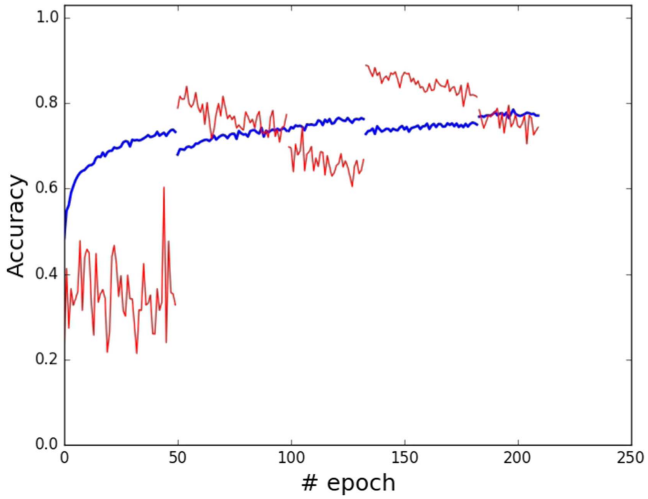

Figure 4. Learning history resulting from the strategy described in the text. The blue solid lines show the accuracy of the training set, and the red solid lines show the accuracy of the validation set. Every 50 epochs, the validation and training data sets are modified, which explains the discontinuities. The accuracy of the validation is generally unstable because it is only made of two galaxies. See text for details.

different phases for this particular set of galaxies without generalizing. To avoid this situation, we have designed a custom training strategy that slightly differs from the classical approaches typically used in machine learning.

Among the 28 galaxies, we use 24 galaxies for training (i.e., ∼9000 images), two for real-time validation during the training, and two additional completely independent galaxies for testing at the end of the training process. It is important to keep in mind that, when we say two galaxies, it does not mean two images. Each galaxy corresponds to the full assembly history of the galaxy between $z = 4$ and 1, with 19 images at each time step. Therefore, the test and validation sets contain ∼1000 galaxies each.

We then train for a maximum of 250 epochs. The novelty is that every 50 epochs, we move two galaxies from the training set to the validation sample and add the validation galaxies to the training. This helps the network not to overfit on the first sample of 24 galaxies while training for enough epochs to enable convergence. The two test galaxies are obviously never used in the process. Finally, in order to have more than two galaxies to test the classification accuracy, we repeat the training procedure just described five times, using two different galaxies for the test sample at every run. The final test data set thus contains 10 galaxies, classified with five slightly different models. Figure 4 illustrates the learning history parameterized by the evolution of the accuracy as a function of the number of epochs of one of the five runs for illustration purposes. We plot the accuracy computed on the training and validation data sets. As expected, the training curve monolithically increases and reaches roughly an accuracy of 80%. Notice, however, some small discontinuities every 50 epochs, corresponding with the modification of the training set. The fact that the discontinuity is small suggests that small modifications of the training sample do not significantly alter the network performance. In other words, there is no overfitting. The validation curve shows a particular behavior. This, again, is a consequence of the adopted training strategy. Every 50 epochs, there is a clearly noticeable jump. The break is larger than that for the training because the validation is only made of two galaxies and the sample is fully changed every 50 epochs. So the break somehow reflects the accuracy variation between galaxies, which can go from 100% for some galaxies to ∼60%. As previously stated, compaction is not a very well-defined process, and some galaxies have complex assembly histories with multiple BN phases. The red curve in Figure 4 also presents large jumps between epochs. This is also most probably a consequence of the size and redundancy of the sample. Given that there are 19 images per snapshot, a change in the classification of a few snapshots implies big changes in the accuracy value.

5. Results

In this section, we analyze the classification accuracy. For that purpose, we use the test data set (10 galaxies) that was not used in the training process (see Section 4) throughout the section.

5.1. Detection of BNs

We first analyze the average accuracy of the trained model to detect pre-BNs, post-BNs, and BNs. The global accuracy, defined as the fraction of images correctly classified, computed on the test data set is ∼70%, which means that 30% of the objects are misclassified. This is certainly not very high. Recall, however, that there is a significant amount of redundancy in the test set. It is helpful to look into more detail to better understand what is going on before drawing conclusions. We first compute a standard confusion matrix showing the relation between input and output classes (Figure 5) for different probability thresholds. At the lower probability threshold (0.5), most of the confusion comes from true pre-BNs (or post-BNs) that are predicted as BNs. This is probably because, as previously stated, the compaction event is not very well defined. The duration and intensity strongly depend on the galaxy. As expected, the degree of contamination decreases when a higher probability threshold is used to select galaxies.

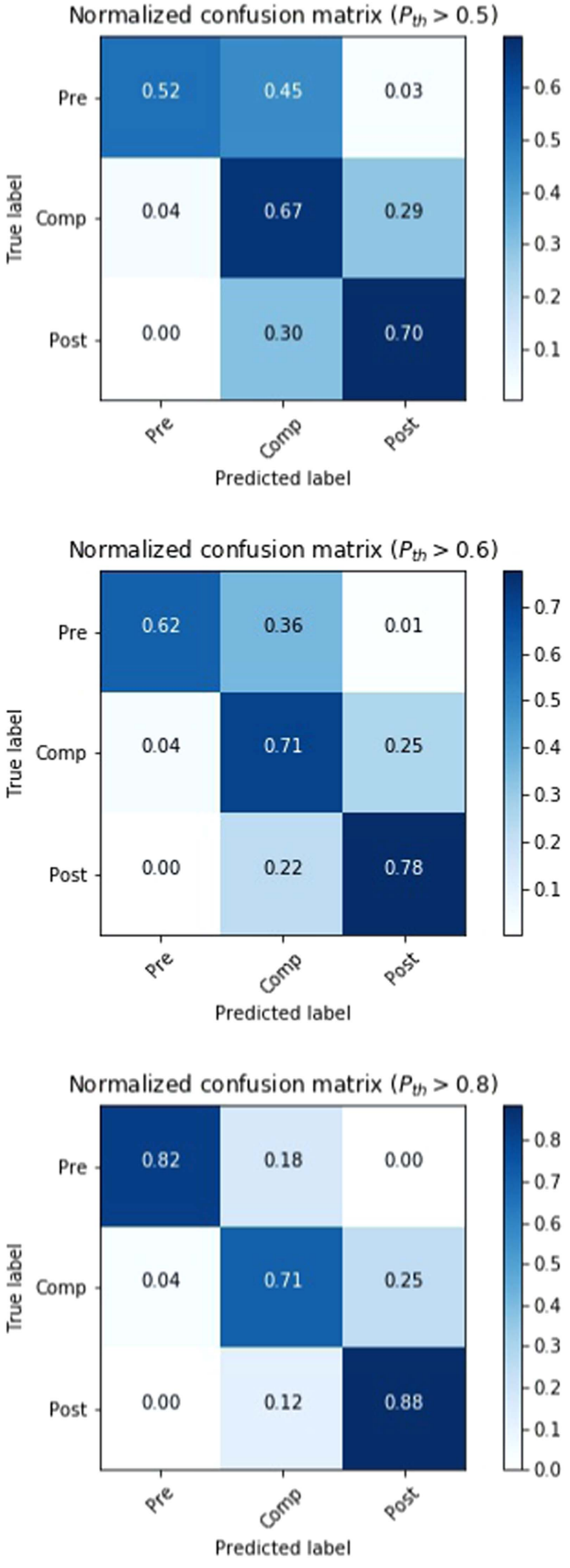

Figure 5. Normalized confusion matrix of the three-label classification on a test data set not used for training or validation. The *y*-axis shows the true class from the simulation metadata, and the *x*-axis is the predicted class. From top to bottom, we show the effect of increasing the probability threshold to select the galaxies belonging to a given class.

At the highest threshold (0.8), 25% of true BNs are predicted to be post-BNs. In fact, one should keep in mind that the test set contains snapshots in steps of $\Delta_a = 0.01$. A galaxy might be misclassified as post-BN just before the compaction event ends, for example, or where there are multiple compactions closely followed in time, reducing the accuracy of the classification. However, the classification might still be meaningful.

To investigate this further, in Figure 6, we plot the output probabilities for a subset of individual galaxies from the test sample as a function of time. In this figure, the lines show the average probability over all camera orientations at a given snapshot. The shaded regions show the scatter due to different camera orientations. For illustration purposes, we show three cases with increasing complexity. The first example (VELA30) has a single intense BN phase. VELA11 is less massive and has two events of smaller intensity. Finally, VELA08 is a low-mass galaxy with a very weak compaction. These three examples bracket the diversity of assembly histories the network needs to capture. As can be seen, there is a good correlation between the evolution of the probability values and the evolutionary phase. We observe that, typically, the probability of pre-BNs tends to decrease before the compaction event, while the compaction probability increases. Toward the end of the BN phase, the probability of post-BNs increases. This is true even for galaxies with complex assembly histories. This result indicates two main things. First, it shows that the machine has learned somehow that there is a sequential order between the three phases. This is not obvious, since all images were randomly included in the training process with random luminosities and, as seen in Table 2, the BN phase can happen at very different redshifts and have very different durations. Second, it shows that despite the relatively low global accuracy, the confusion seems to essentially come from the snapshots taken at the transition phases. This is important because it means that when the machine misclassifies, it is not fully random. The misclassification, therefore, is a reflection of the difficulty in defining the different phases. It is also worth noticing that the scatter due to different camera orientations is generally not large (∼0.1–0.2 in terms of probability). It suggests a mild impact of the projection in the classification accuracy.

5.2. Impact of Camera Orientation

We investigate this further in Figure 7, which shows the confusion matrix divided by camera orientation. Despite some statistical fluctuations, no significant differences are appreciated, as already suggested by the results shown in Figure 6. This is also quantified in Figure 8, which shows the global accuracy as a function of the camera number (see Table 1 for an explanation of the different numbers). The figure confirms that there is no systematic trend with the orientation. The global accuracy increases equally for all cameras when the probability threshold is increased.

5.3. Calibration of Observability Timescales

In fact, in view of applying the model to real data, probably the most interesting property to investigate is whether we can calibrate the observability timescales of the features learned by the classifier. In other words, what is the typical time window in which the network detects BNs? This is important because it allows us to better interpret the classification in terms of an evolutionary sequence and to compute a BN rate from the

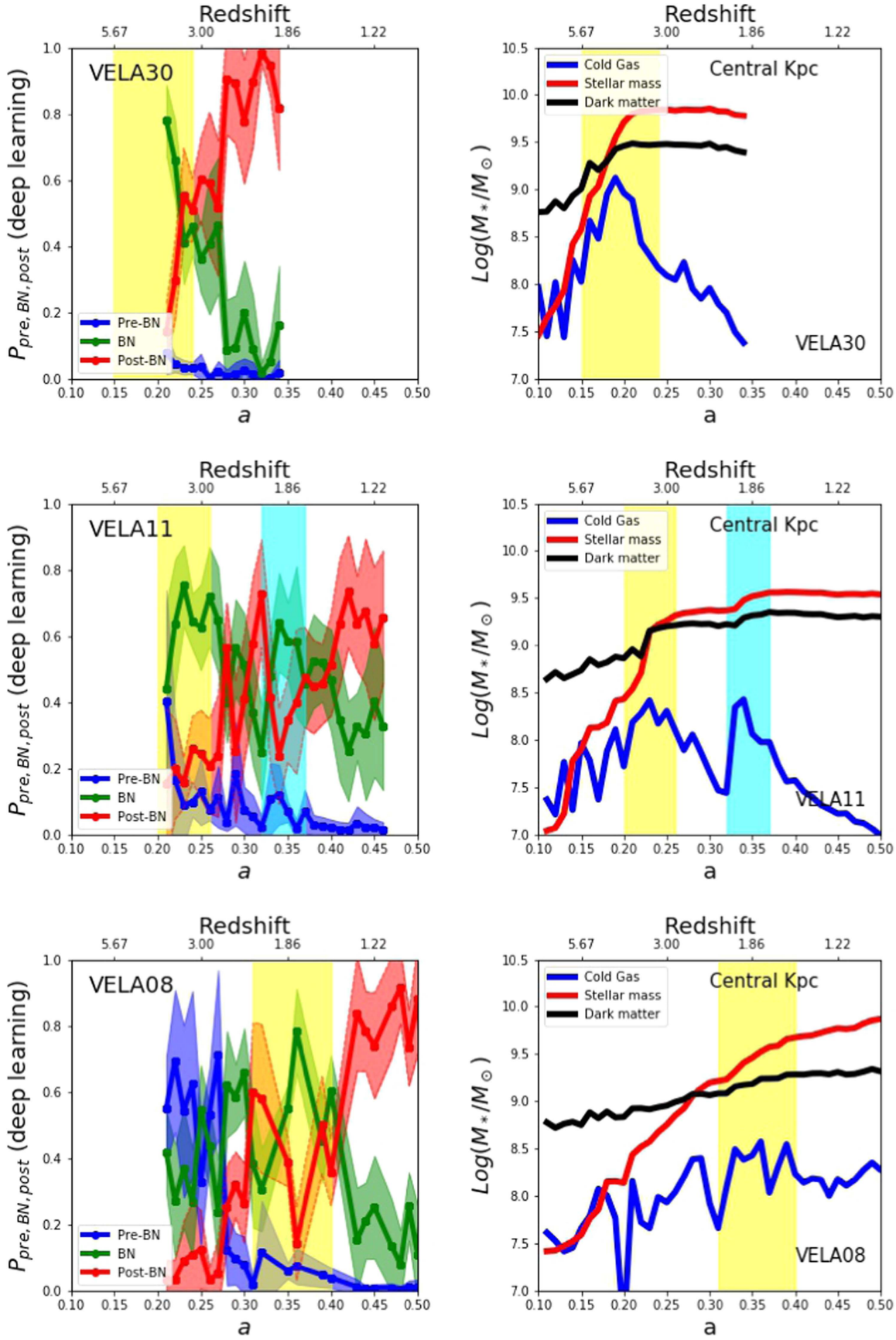

Figure 6. Examples of predictions on a test sample of increasing complexity. The left column shows the mean probability of being pre-BN (blue line), BN (green line), and post-BN (red line) predicted by the CNN. The shaded regions around the lines indicate the scatter due to different camera orientations. The right column shows the input simulation metadata used to define the phases, as in Figure 1. The yellow and cyan shaded regions show the primary and secondary BN phases.

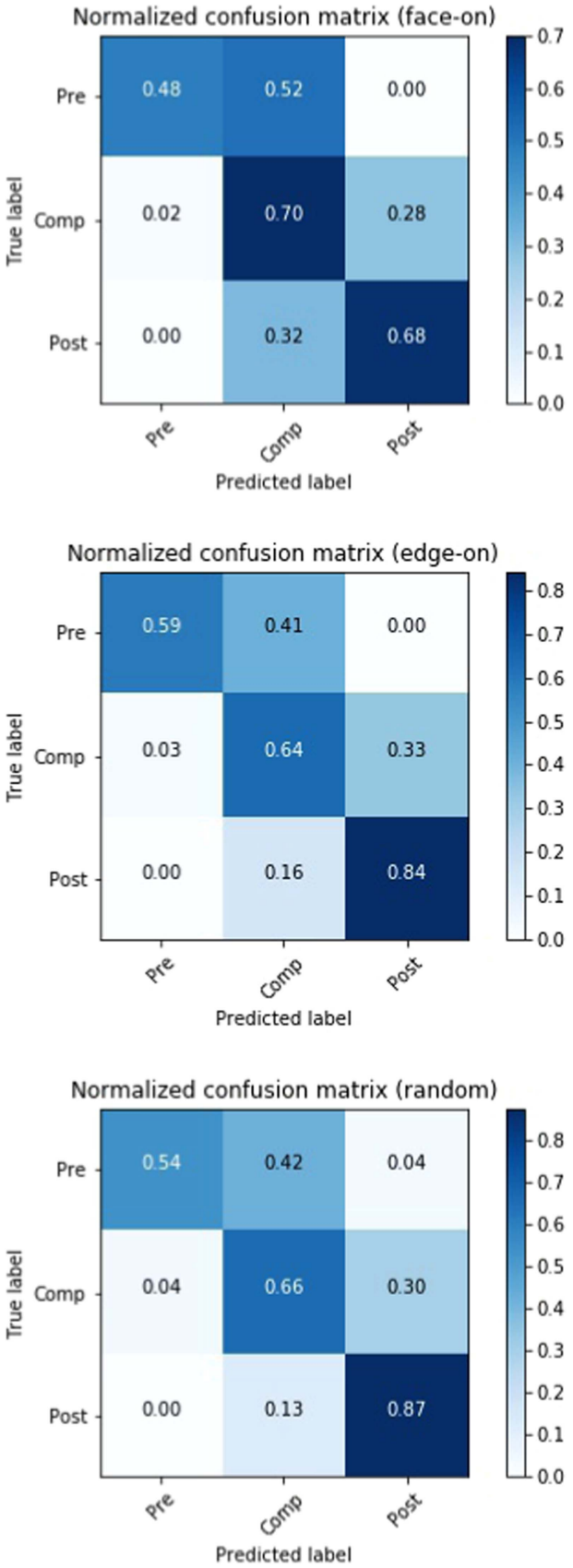

Figure 7. Same as Figure 5, but the confusion matrix is shown for different camera orientations. Top: face-on (cam00/02); middle: edge-on (cam01/03); bottom: random (cam13+).

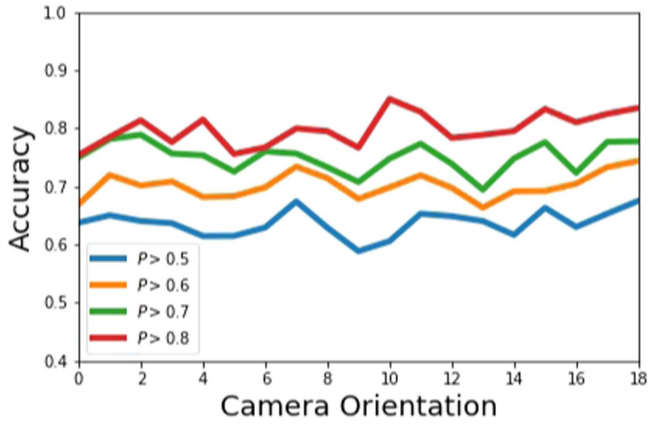

Figure 8. Measured accuracy on the test data set as a function of the camera orientation. The numbers indicate the orientation (see Table 1). The different colors indicate different probability thresholds, as labeled. The accuracy does not depend on the camera orientation.

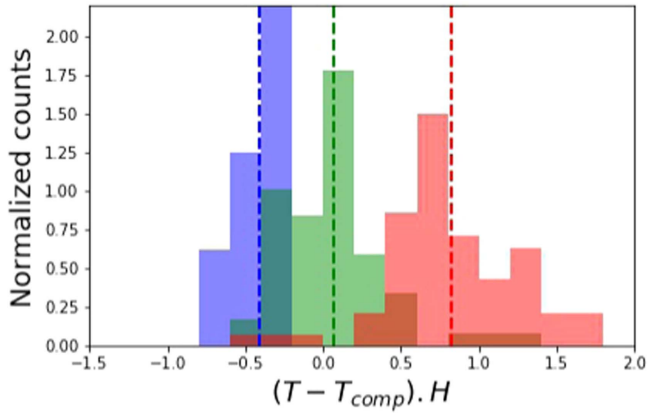

Figure 9. Observability of the BN phase with the calibrated classifier. The histograms show the distributions of time (relative to the Hubble time at the time of the peak of the BN phase). The blue, green, and red histograms show the pre-BN, BN, and post-BN phases. The dashed vertical lines show the average values for each class with the same color code. Despite some overlap, the classifier is able to establish temporal constraints on the different phases. The darker regions indicate overlapping histograms.

observations, as usually done for mergers. To do so, we take the test sample and classify all galaxies in the three classes according to the output probabilities. We simply add each image to the class of maximum probability and require that the probability value is larger than 0.5. We then compute, for each galaxy, the time difference with the main BN phase (computed as a fraction of the Hubble time at the BN peak, i.e., $1/H(t)$, $H(t)$ being the Hubble constant). Figure 9 shows the histograms for the three classes. We confirm that the three classes tend to probe a different regime, although with some overlap, as expected from the results of the previous sections. Pre-BN galaxies are, on average, selected $\sim 0.40/H(t)$ before the event, and post-BN galaxies are typically observed $\sim 0.80/H(t)$ after the compaction. The galaxies classified are centered on the BN phase (0.05 ± 0.3 Hubble times).

Although there is some overlap between the different histograms, it is worth noticing that all galaxies that passed the BN phase by more than half a Hubble time are classified as post-BN galaxies. Also, there are no galaxies classified as BN or pre-BN objects for which the event is more than ~ 0.5

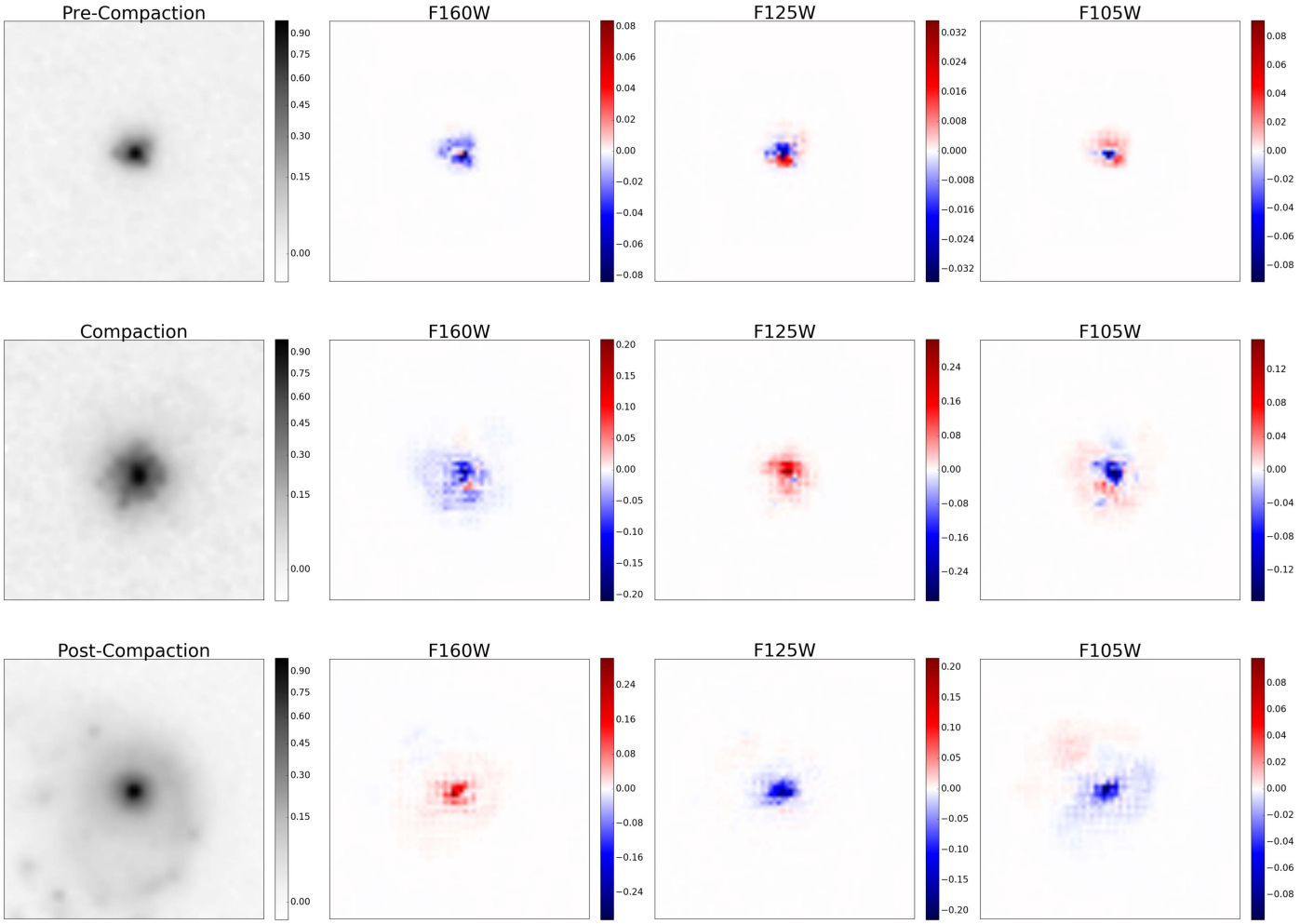

Figure 10. Integrated gradient output of the model. Each row shows a galaxy in a different stage (pre-BN, BN, post-BN). The left column is the original image, and the second, third, and fourth columns show the integrated gradients for the different filters. The network automatically detects the pixels belonging to the galaxy and uses all of them to make the decisions.

Hubble times away. This means that our classifier can indeed establish some temporal constraints regarding the BN phase based only on the stellar distributions. This is extremely important because it shows that there is a temporal sequence implied in the classification. So when applied to real data, one can more easily interpret the results in terms of evolution, as will be discussed in Section 6.

5.4. Inside the Network

An important caveat of the machine-learning approach presented above is that it somehow behaves as a black box. It is thus difficult to precisely determine what features the machine is using to decide the output classification. This is a general problem for all DL applications. However, there exist more and more network interrogation techniques that identify the pixels in the input image that most contributed to the final classification. One recent method is called integrated gradients (Sundararajan et al. 2017). It is based on the measurement of the differences between gradients computed by the network in an input image as compared to a test image (usually a blank image with only zeros). We tested this method in our model and computed the integrated gradients for some of the galaxies. Figure 10 shows one example for each class. The interpretation is not straightforward. However, some useful information can be extracted from this exercise. It is interesting to see that the model automatically segments all the pixels belonging to the galaxy and takes the decision based on all the galaxy pixels. It also means that it understood that there is no information in the noise and confirms that the model is not overfitting on the noise pattern. Also, as pointed out in previous works, after the BN phase, a gaseous disk is often built in the simulations (Zolotov et al. 2015; Tacchella et al. 2016b). The bottom panels of the figure clearly show that the machine detects the diffuse disk component even if faint and probably uses this information to make the decision concerning the post-BN and sometimes the BN phase. For galaxies in the BN phase, the relevant pixels are more concentrated in the center, since the galaxies are generally more compact as the obvious signature of this phase. It is also worth noticing that the gradient tends to have values of different sign in the center and outskirts, as if the machine was using a difference in flux between the center and the outskirts to classify. This is expected, since the compaction event is, by definition, accompanied by a burst of central star formation, and the sSFR profiles evolve from decreasing to rising, indicating quenching outside-in in the pre-BN phase and inside-out in the post-BN phase (Tacchella et al. 2016b). The model is capturing

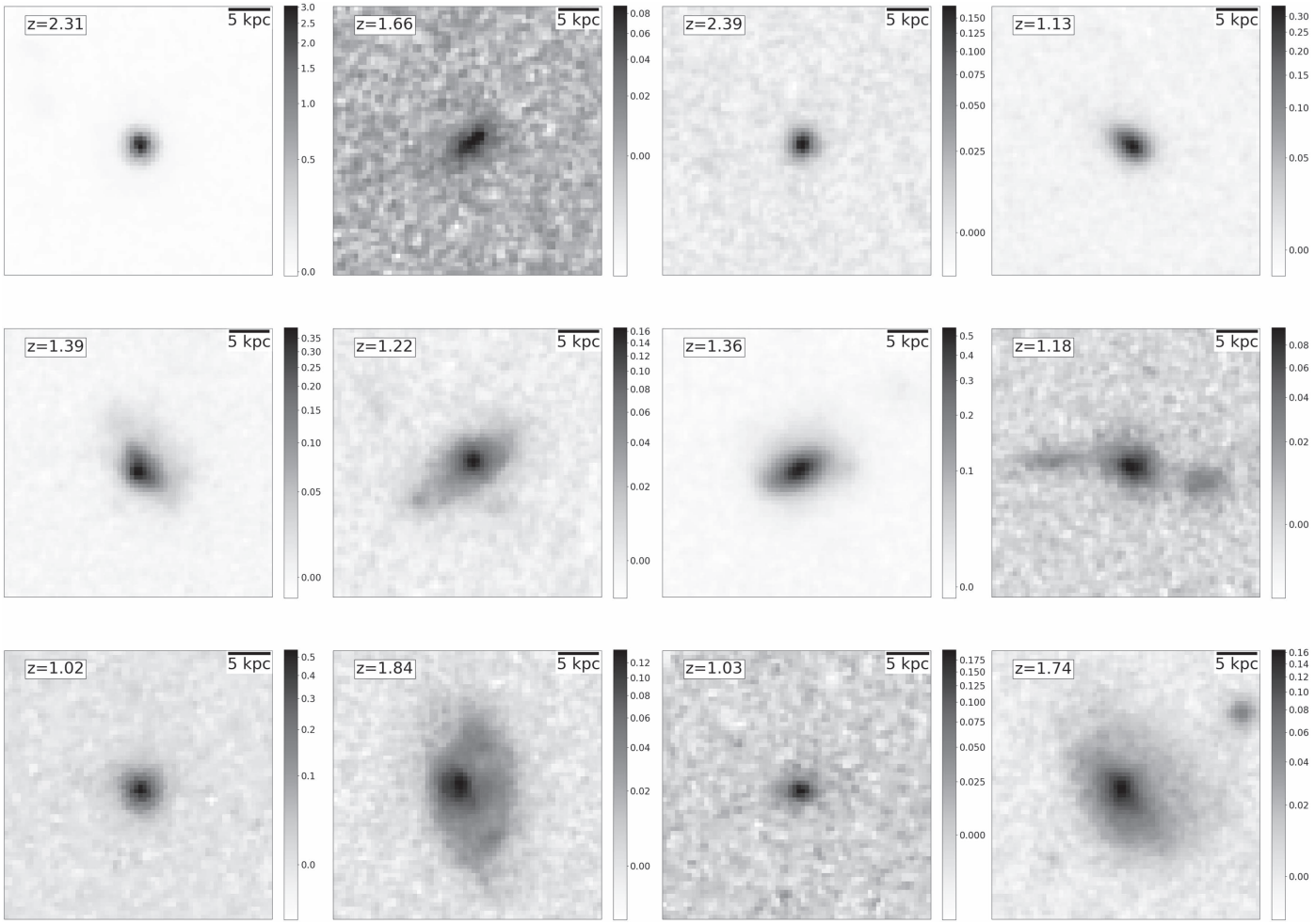

Figure 11. Random examples of F160W CANDELS images in the three phases discussed in this work. The image size is $3\farcs8 \times 3\farcs8$. The top row shows pre-BN galaxies, the middle row shows galaxies in the BN phase, and the bottom row shows post-BN objects.

all these correlations automatically. This is the strength of the presented methodology. Although the information that can be extracted from integrated gradients is quite limited at this stage, it is reasonable to think that interrogation techniques will become more advanced, and therefore there is potentially information that can be learned from postprocessing of the model outputs in the future.

6. Identifying BNs in the Observations

We now apply the model to the *HST*/CANDELS sample presented in Section 3. We simply cut stamps around the selected galaxies in the three infrared filters (F160W, F125W, and F105W) and classify them into three classes using the trained models. Since 10 models were produced (see Section 4), we use each of them to classify all galaxies. Each real galaxy has, therefore, 10 different classifications using slightly different models. We then compute the average probability to increase the robustness of the classification. We stress that there is a general good agreement between the different models that confirms that the classification does not strongly depend on the specific subset of simulated galaxies used for training. The typical scatter in the probability values is of the order of $\sim$0.1.

The first thing to notice is that the classification applied to real data returns objects with high probability values in the three classes. The fraction of galaxies with all probabilities lower than 0.5 is only 2% of the total sample. It means that the model found galaxies that resemble the galaxies in the simulation with high confidence. This reflects that the simulated galaxies are fairly similar to the observed ones and that the network found characteristic features learned in the simulations in the CANDELS observations. Figure 11 shows some example stamps of observed galaxies in the three phases. It is not obvious to establish what would happen if galaxies from the training were very different from real data sets. This will be explored in future work. In order to have a first idea of how the network would behave when confronted with very different objects, we perform a simple exercise. We take the real observed galaxies from CANDELS and first randomly shuffle the central pixels of the galaxies, then shuffle all the pixels in the galaxies (inner+outskirts). This creates two fake data sets with different degrees of perturbation, which are given to the network. Figure 12 shows the probability distributions for the three classes when these fake data sets are given. The figure shows that the first effect of shuffling the center is that the number of galaxies with a compaction probability larger than 0.5 almost drops to zero. This is somehow expected, as most of the compaction features are naturally seen in the central parts. It confirms that the network is significantly using this information to classify. Since the probabilities need to add up

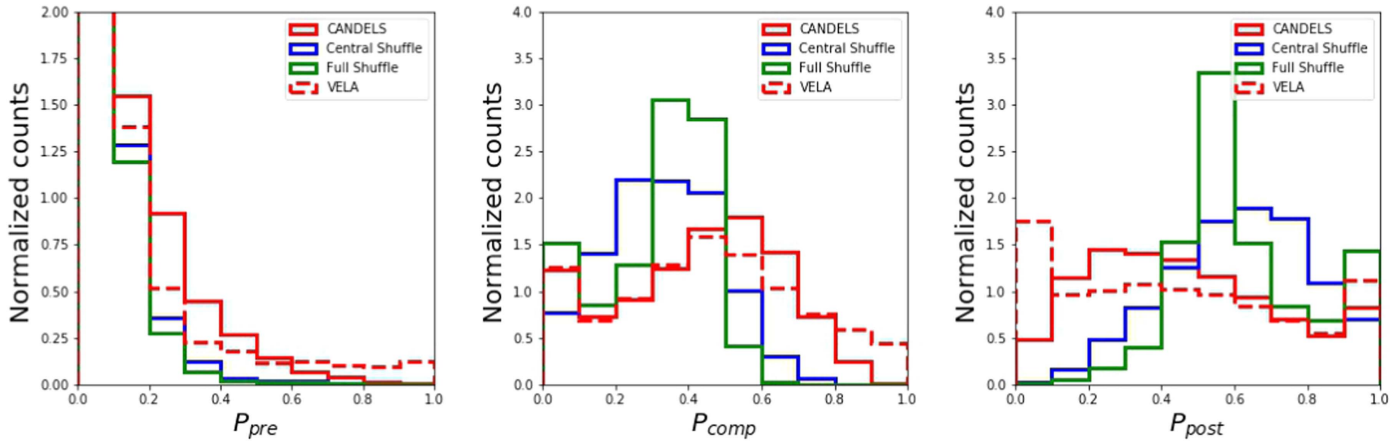

Figure 12. Impact of shuffling the pixels on the output probability distributions. From left to right, we show the pre-BN, BN, and post-BN probability distributions. The red solid lines show the distribution for the original CANDELS images. The blue (green) lines show the same distributions when the central (outskirts+central) pixels are shuffled. For reference, we also show the distribution for the simulated galaxies in the test data set with a red dashed line. Shuffling the pixels tends to narrow the distributions around a probability value of ∼0.5.

to 1, central shuffling also provokes an increase in the number of galaxies with a large probability of post-BNs. Given that post-BNs tend to be more extended, the fact of shuffling the central pixels pushes the network to boost the post-BN probability, since it focuses on the outer pixels. However, the values remain low (∼0.6), indicating a moderate level of confidence. When both outskirts and inner pixels are shuffled, both probability distributions, BN and post-BN, significantly narrow and peak at ∼0.4–0.5, meaning that the network is not able to clearly assign galaxies to classes. This exercise shows that the probability distributions somehow reflect the similarity between the simulations and the observations. We notice, however, that even in the shuffled images, there is a fraction of galaxies with high post-BN probabilities. A visual inspection shows that these are bright galaxies for which the shuffling has pushed bright pixels toward very large distances. The network most likely interprets this as a very extended disk.

The fact that the distributions on CANDELS galaxies resemble the ones obtained on the test simulated sample (red solid/dashed lines in Figure 12) suggests, therefore, that simulated and observed galaxies look similar to the network. This allows us to push the analysis a bit further by exploring the properties of galaxies in the three phases (BN, post-BN, and pre-BN) in the observations.

6.1. A Characteristic Mass Range for the BN Phase

In Figure 13, we first look at the stellar mass distributions of CANDELS galaxies in the three different phases. Recall that the simulations used for training stop at $z \sim 1$, so we only show galaxies above this redshift in the observations. The abundance of galaxies in different phases strongly depends on stellar mass. Pre-BN galaxies tend to increase at low stellar masses ($M_*/M_\odot < 10^{9-9.5}$), and post-BN galaxies dominate at large stellar masses ($M_*/M_\odot > 10^{10.5}$). The BNs are most frequent at intermediate masses and peak at $\sim 10^{9.2-10.3}$. Interestingly, the position of the peak seems to be relatively independent of redshift, with a small tendency to move toward lower masses at lower redshifts. We notice that at this characteristic stellar mass, the CANDELS data set is affected by incompleteness, as indicated by the vertical line in the plots. This should not affect the result in the sense that there are no reasons to think that post-BN galaxies are more difficult to detect.

This characteristic mass for compaction is a prediction from the VELA simulations, as first reported in Zolotov et al. (2015) and Tomassetti et al. (2016) and reflected in Table 2 (see also Tacchella et al. 2016a and A. Dekel et al. 2018, in preparation). The fact that it appears clearly in the observations confirms that the network is automatically extracting the correlations existing in the simulations. It is worth recalling that the luminosity has been removed from the training set, which ensures that the network is not classifying based on luminosity that is directly correlated with the stellar mass. The network is necessarily using other information, such as spatial distribution, shape, or color, to identify the different phases. The characteristic mass naturally emerges in the observations. The network successfully identifies a population that resembles simulated galaxies experiencing compaction in the feature space learned, and these galaxies tend to be near a characteristic stellar mass similar to the characteristic mass seen in the simulations.

For comparison purposes, we also show in the Appendix the resultant mass distributions in the observations when the luminosity is left in the training set. The results are similar, confirming that luminosity is not the main parameter used by the network. There is a tendency to find more pre-BN galaxies, however. We speculate that this is because the algorithm uses some S/N-related information if available. Since pre-BNs are generally fainter, they also have lower S/Ns in the observed mock images, so the network will tend to classify fainter objects as pre-BN. It highlights both the strengths and risks of the DL approach, in the sense that all information is taken into account in our unsupervised learning.

An analogous behavior is also seen in Figure 14, where the redshift evolution of the fractions of galaxies in the three phases at fixed stellar mass is shown. Both plots are complementary. As expected, the redshift evolution strongly depends on stellar mass. The galaxies that are more frequently potentially in the BN phase in the CANDELS redshift range are in the stellar mass range of $10^{9.2} < M_*/M_\odot < 10^{10.3}$. The massive compact star-forming galaxies identified in previous works might be the high-mass tail of the BN population. More massive galaxies indeed tend to be in the post-compaction phase at all redshifts. This means that if one wants to observe

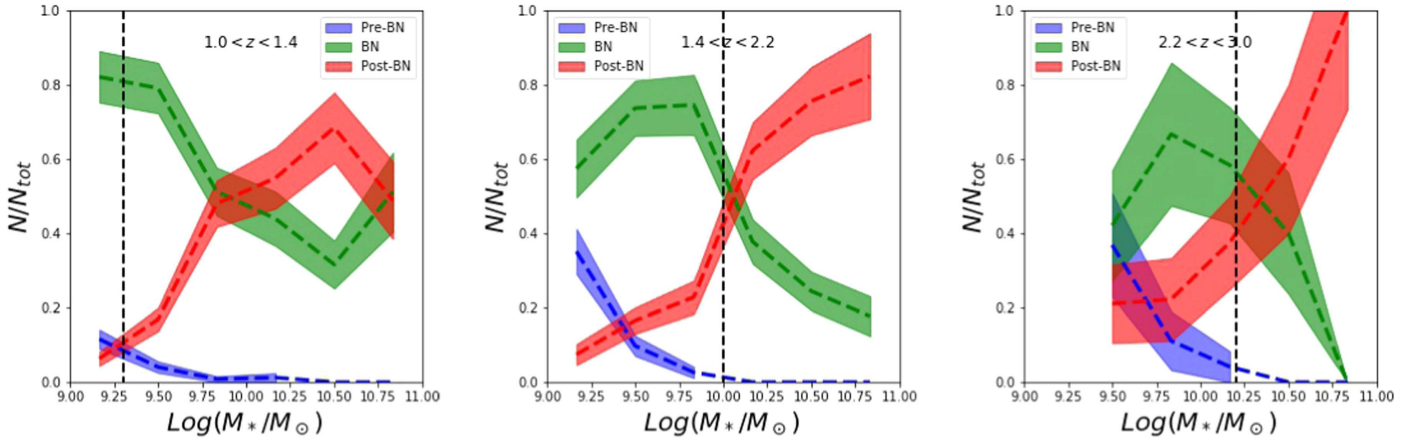

Figure 13. Stellar mass distributions of CANDELS galaxies in pre-BNs (blue lines), BNs (green lines), and post-BNs (red lines) for different redshift bins, as labeled. Galaxies in the BN phase typically peak at stellar masses of $10^{9.2-10.3}$, as predicted by the simulations. In more detail, the BN range is 9.5–10.3 in the high-z bin, 9.25–10.0 in the middle-z bin, and a smaller mass in the low-z bin. This possible redshift dependence may or may not be significant. The vertical dashed lines show the mass completeness limits from Huertas-Company et al. (2016). The peak is generally below the completeness limit. This should not significantly impact the presence of the peak unless post-BN galaxies are more difficult to detect at these masses, which is unlikely.

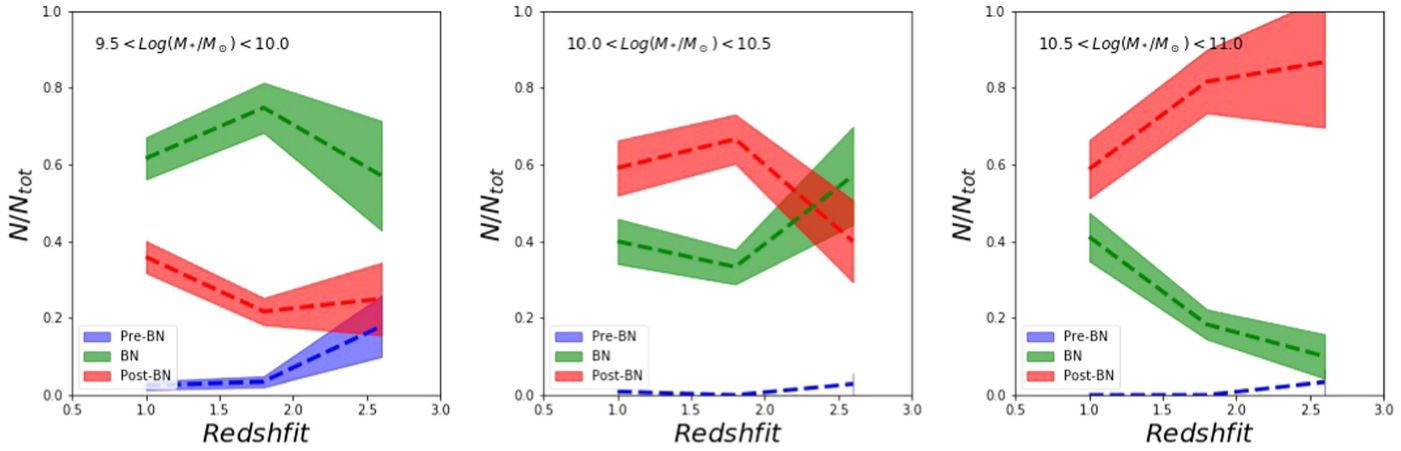

Figure 14. Redshift evolution of the fractions of CANDELS galaxies in pre-BNs (blue lines), BNs (green lines), and post-BNs (red lines) for different stellar mass bins, as labeled. In the redshift range of CANDELS ($1 < z < 3$), BNs dominate at a characteristic stellar mass of $\sim 10^{9.2-10.4}\ M_{\odot}$, as predicted by the simulations.

the progenitors of these most massive galaxies in the process of compaction, it is required to probe $\sim 10^{9.5}\ M_{\odot}$ galaxies at $z > 3$. That will be straightforward with *JWST*.

6.2. The L Shape in sSFR versus M_*

The previous section has shown that the network successfully identifies a characteristic galaxy stellar mass range for the BN phase in the CANDELS data. This is remarkable given the known limitations of the simulations (see Section 2) and suggests that there are important similarities between simulated and observed galaxies.

In future work, we plan to analyze in more detail how the different classes relate to other physical properties. As a preliminary step, we attempt a first look at the sSFRs and central mass densities (Σ_1; Barro et al. 2017) of galaxies in the pre-BN, BN, and post-BN phases. This is motivated because in the simulations, the compaction, BN, and quenching sequence put the galaxy into a characteristic L-shaped track in sSFR–Σ_1 with the BN phase at the turning point (e.g., Zolotov et al. 2015). This L shape is similar to the observed distribution (Barro et al. 2013, 2017).

In Figure 15, we show the sSFR–Σ_1 plane for pre-BN, BN, and post-BN galaxies in CANDELS as defined by the CNN trained on the simulations. As previously reported, galaxies form a characteristic L-shaped distribution in the plane.

At first approximation, the median position (large dots in the figure) of the pre-BN, BN, and post-BN galaxies is different and crudely follows the expected evolutionary sequence from the simulations. Pre-BN galaxies tend to be in the main sequence and have low central density values, while post-BN galaxies have lower specific star formation rates and larger central densities. The BN galaxies lie in between. Given the observability timescales calibrated in Section 5.3, this suggests that there is an evolutionary sequence in the plane and that galaxies tend to move from left to right. We observe, however, that there is also significant overlap between the different phases in the three quadrants of the sSFR–Σ_1 diagram. For example, several galaxies are classified as post-BN while they have low Σ_1 values. Also, there is mixing of low- and high-sSFR compact galaxies that is not fully consistent with the distinction between the BN and post-BN phases in the simulations. For comparison, we show the same plot for the VELA simulations, which shows a clearer separation, namely a stronger correlation between the three phases as defined based

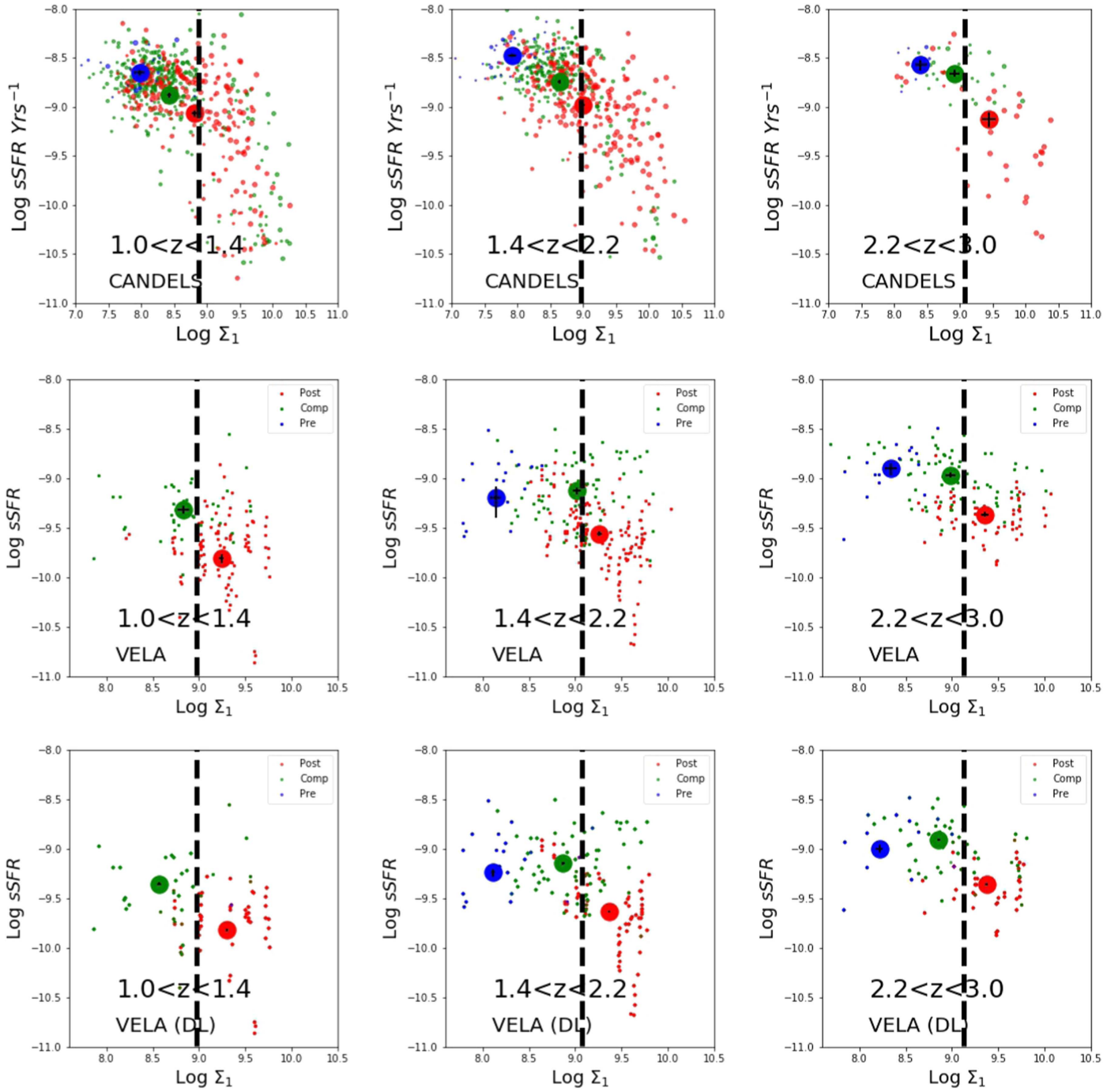

Figure 15. Distribution of pre-BN (blue dots), BN (green dots), and post-BN (red dots) galaxies with $M_*/M_\odot > 10^9$ in the sSFR–Σ_1 plane. The large dots show the average positions, and the black error bars are the 68% confidence intervals obtained through bootstrapping. The top row shows the distribution of CANDELS galaxies. The middle row shows the simulated galaxies with the phase defined from the assembly history. The bottom row shows the same simulated galaxies when the phase is determined through DL. The vertical black dashed lines in the top row show the location of the quiescent ridgeline at a stellar mass of $10^{10}\,M_\odot$ computed by Barro et al. (2017).

on the gas/SFR distribution and the distribution to three quadrants in the sSFR–Σ_1 diagram as derived from the stellar distribution.

We emphasize that the main purpose of this work is to illustrate the methodology. We thus keep for future work a detailed investigation of the reasons for this increased confusion in CANDELS. One possible explanation resides in the definition of the BN phase used for training. We recall that several galaxies in the simulation present complex assembly histories, with many wet compaction events of different intensities (see Figure 1). A similar behavior is also reported in Tacchella et al. (2016a); i.e., compaction and quenching events confine the galaxy to the main sequence until a major BN event that is followed by long-term quenching as a result of a hot massive halo. Therefore, according to our labeling of the training set explained in Section 4.1, galaxies can still be considered post-BN (see, for example, VELA11 in Figure 6) between several events that

could also contribute to explaining the overlap we see in CANDELS. A way to explore the effects of minor compaction events would be to train a network with only major compactions and see how the classification changes. To do that, a larger and more diverse training set is needed and also at higher redshift, in the *JWST* range, where major events tend to happen in the simulations. We keep this for future work.

7. Summary and Conclusions

We have explored a new approach to classify galaxy images using DL calibrated on numerical simulations. The general methodology first consists of generating mock images of galaxies, reproducing the observing conditions from hydro-cosmological simulations, which are then labeled based on the known evolution of gas, SFR, and stars. The images are then fed to an unsupervised feature-learning machine that automatically learns the features to detect a given evolution pattern. We have applied the method for detecting the characteristic BN phase as seen in cosmological simulations, near a critical mass and preferentially at high redshifts, following a wet compaction process and followed by central quenching. We have used for that purpose a suite of high-resolution zoom-in hydro-numerical simulations of intermediate-mass galaxies in the redshift range $1 < z < 3$. We have shown that a simple CNN is able to detect galaxies in the BN phase with $\sim$80% accuracy within a time window of ± 0.2 Hubble times and hence establish temporal constraints in the data. The described methodology presents several key advantages over more traditional approaches. First of all, it does not require any image preprocessing. Only the pixel distributions are fed into the network, which automatically extracts the relevant information. This does not, however, prevent combining the automatically extracted features with other standard measurements, such as colors or sizes. Moreover, there is no need of an a priori assumption of the optimal observables for a given physical process. The procedure will automatically extract the best tracers if present in the data.

We have then applied the trained model to observed galaxy multicolor images from the CANDELS survey observed with *HST* in the same redshift range and classified them into three main classes: pre-BN, BN, and post-BN.

The key results are as follows.

1. The network finds galaxies with a high probability of being in the three classes, indicating a similarity between simulated and observed galaxies.
2. The classification recovers a characteristic stellar mass for the BN phase of $\sim 10^{9.2-10.3}\ M_{\odot}$ mostly independent of redshift. More massive compact galaxies are found to be preferentially in the post-BN class, so they are compatible with having gone through the BN phase more than 0.5 Hubble times before the time of observation.
3. Pre-BN, BN, and post-BN galaxies occupy different regions in the sSFR–Σ_1 plane, suggesting an evolutionary sequence in the plane as predicted by the simulations. There is, however, some degree of confusion, i.e., post-BN galaxies with low central densities that will be investigated in future work.

In particular, one important point that will be addressed in forthcoming works is the impact of the specific set of simulations used for training. Despite the similarities between simulations and observations suggested in Section 6.1, the VELA simulations used in this work might still be too limited to adequately represent the entire CANDELS data set, not only because of the lack of AGNs but also because the sample is small and covers a limited mass range. Additionally, the assumptions regarding the subgrid astrophysics are not well constrained by theory or observations, as discussed in Section 2. To further investigate the impact of these limitations, we plan to enlarge our training sets by using new available simulated data sets with the same VELA initial conditions but different subgrid astrophysics, as well as other independent simulated data sets including AGNs.

The presented methodology could then be adapted to other robust physical processes captured in simulations and could constitute a useful tool to better compare future imaging surveys with forthcoming simulations.

The authors are grateful to Google for the unrestricted gift given to the University of California Santa Cruz to carry out the project "deep learning for Galaxies" that greatly contributed to making this work possible. We also appreciate helpful discussions with Sander Dielemann, Daniel Freedman, Eric Hayashi, and Jon Shlens at Google. We also thank Frédéric Bournaud for refereeing this work and providing interesting suggestions. This work was partly supported by the grants France-Israel PICS, US-Israel BSF 2014-273, and NSF AST-1405962. J.R.P. acknowledges support from *HST*-AR-14578.001-A. A.D. also acknowledges support from GIF I-1341-303.7/2016, DIP STE1869/2-1 GE625/17-1, and I-CORE PBC/ISF 1829/12. M.H.C. acknowledges support from the ANR ASTROBRAIN. D.C. has been funded by the ERC Advanced Grant, STARLIGHT: Formation of the First Stars (project number 339177). The VELA simulations were performed at the National Energy Research Scientific Computing Center (NERSC) at Lawrence Berkeley National Laboratory and at NASA Advanced Supercomputing (NAS) at the NASA Ames Research Center.

Appendix
The Effect of Luminosity

In the training set used in this work, the magnitudes of the galaxies in the different phases were randomly changed. This is to ensure that all galaxies have similar S/Ns and the network does not learn based on that. As a matter of fact, since the pre-BN galaxies in the simulations are found at higher redshift and have lower stellar masses than post-BN galaxies, they will be more noisy in the CANDELized images. The network might therefore use this information. To check the effect of this in the final classification, we show in Figure 16 the same stellar mass distributions of galaxies in the three different phases in CANDELS as in Figure 13 but obtained with a training set without randomizing the magnitudes. As can be seen, the distribution is similar, i.e., a BN peak at a characteristic stellar mass. However, the code tends to find more pre-BN galaxies at low mass. This is because it is learning some information from the S/N distribution. This exercise shows the strength of the DL approach, since it demonstrates that the network uses all available information. However, it highlights the risks too. One needs to control the information that should not be used by the net.

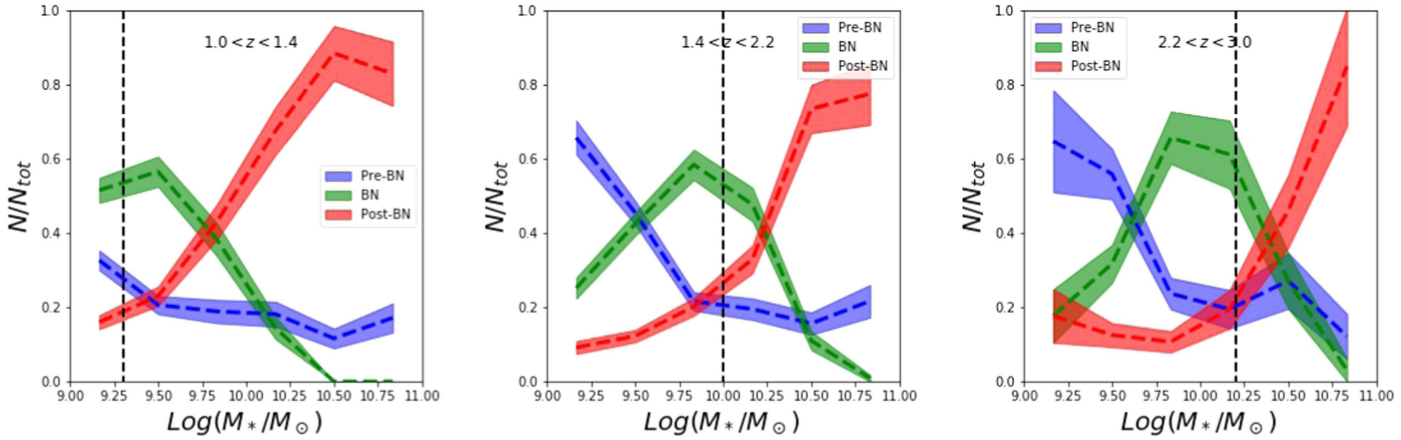

Figure 16. Stellar mass distributions of CANDELS galaxies in pre-BNs (blue lines), BNs (green lines), and post-BNs (red lines) for different redshift bins, as labeled. The classification is performed with a training set including the luminosity information (see text for details). Galaxies in the BN phase typically peak at stellar masses of $10^{9.2-10.3}$ at all redshifts. The vertical dashed lines show the completeness limits from Huertas-Company et al. (2016).

ORCID iDs

J. R. Primack https://orcid.org/0000-0001-5091-5098
D. C. Koo https://orcid.org/0000-0003-3385-6799
D. Ceverino https://orcid.org/0000-0002-8680-248X
R. C. Simons https://orcid.org/0000-0002-6386-7299

References

Abraham, R. G., Tanvir, N. R., Santiago, B. X., et al. 1996, MNRAS, 279, L47
Abramson, L. E., Gladders, M. D., Dressler, A., et al. 2016, ApJ, 832, 7
Barro, G., Faber, S. M., Koo, D. C., et al. 2017, ApJ, 840, 47
Barro, G., Faber, S. M., Pérez-González, P. G., et al. 2013, ApJ, 765, 104
Barro, G., Kriek, M., Pérez-González, P. G., et al. 2016, ApJL, 827, L32
Barro, G., Pérez-González, P. G., Gallego, J., et al. 2011, ApJS, 193, 13
Barro, G., Faber, S. M., Pérez-González, P. G., et al. 2014, ApJ, 791, 52B
Bell, E. F., Papovich, C., Wolf, C., et al. 2005, ApJ, 625, 23
Buitrago, F., Trujillo, I., Conselice, C. J., et al. 2008, ApJL, 687, L61
Carollo, C. M., Bschorr, T. J., Renzini, A., et al. 2013, ApJ, 773, 112
Ceverino, D., & Klypin, A. 2009, ApJ, 695, 292
Ceverino, D., Klypin, A., Klimek, E. S., et al. 2014, MNRAS, 442, 1545
Ceverino, D., Primack, J., & Dekel, A. 2015, MNRAS, 453, 408
Chabrier, G. 2003, PASP, 115, 763
Chary, R., & Elbaz, D. 2001, ApJ, 556, 562
Cibinel, A., Le Floc'h, E., Perret, V., et al. 2015, ApJ, 805, 181
Conselice, C. J., Bershady, M. A., & Jangren, A. 2000, ApJ, 529, 886
Dahlen, T., Mobasher, B., Faber, S. M., et al. 2013, ApJ, 775, 93
Dekel, A., Sari, R., & Ceverino, D. 2009, ApJ, 703, 785
Dieleman, S., Willett, K. W., & Dambre, J. 2015, MNRAS, 450, 1441
Domínguez Sánchez, H., Huertas-Company, M., Bernardi, M., Tuccillo, D., & Fisher, J. L. 2018, MNRAS, 476, 3661
Draine, B. T., & Li, A. 2007, ApJ, 657, 810
Dwek, E. 1998, ApJ, 501, 643
Elbaz, D., Daddi, E., Le Borgne, D., et al. 2007, A&A, 468, 33
Grogin, N. A., Kocevski, D. D., Faber, S. M., et al. 2011, ApJS, 197, 35
Guo, Y., Ferguson, H. C., Giavalisco, M., et al. 2013, ApJS, 207, 24
Huertas-Company, M., Bernardi, M., Pérez-González, P. G., et al. 2016, MNRAS, 462, 4495
Huertas-Company, M., Gravet, R., Cabrera-Vives, G., et al. 2015, ApJS, 221, 8
Huertas-Company, M., Mei, S., Shankar, F., et al. 2013, MNRAS, 428, 1715
James, A., Dunne, L., Eales, S., & Edmunds, M. G. 2002, MNRAS, 335, 753
Jonsson, P. 2006, MNRAS, 372, 2
Jonsson, P., Groves, B. A., & Cox, T. J. 2010, MNRAS, 403, 17
Jonsson, P., & Primack, J. R. 2010, NewA, 15, 509
Kennicutt, R. C., Jr. 1998, ApJ, 498, 541
Koekemoer, A. M., Faber, S. M., Ferguson, H. C., et al. 2011, ApJS, 197, 36
Kravtsov, A. V. 2003, ApJL, 590, L1
Kravtsov, A. V., Klypin, A. A., & Khokhlov, A. M. 1997, ApJS, 111, 73
Krumholz, M. R., & Dekel, A. 2012, ApJ, 753, 16
Lilly, S. J., & Carollo, C. M. 2016, ApJ, 833, 1
Lotz, J. M., Davis, M., Faber, S. M., et al. 2008, ApJ, 672, 177
Lotz, J. M., Jonsson, P., Cox, T. J., & Primack, J. R. 2008, MNRAS, 391, 1137
Metcalf, R. B., Meneghetti, M., Avestruz, C., et al. 2018, arXiv:1802.03609
Nelson, E. J., Tadaki, K.-i., Tacconi, L. J., et al. 2018, arXiv:1801.02647
Newman, A. B., Ellis, R. S., Bundy, K., & Treu, T. 2012, ApJ, 746, 162
Peng, C. Y., Ho, L. C., Impey, C. D., & Rix, H.-W. 2002, AJ, 124, 266
Sérsic, J. L. 1968, Atlas de Galaxias Australes (Cordoba, Argentina: ESO)
Snyder, G. F., Lotz, J., Moody, C., et al. 2015, MNRAS, 451, 4290
Snyder, G. F., Lotz, J. M., Rodriguez-Gomez, V., et al. 2017, MNRAS, 468, 207
Sundararajan, M., Taly, A., & Yan, Q. 2017, arXiv:1703.01365
Tacchella, S., Carollo, C. M., Renzini, A., et al. 2015, Sci, 348, 314
Tacchella, S., Dekel, A., Carollo, C. M., et al. 2016a, MNRAS, 457, 2790
Tacchella, S., Dekel, A., Carollo, C. M., et al. 2016b, MNRAS, 458, 242
Tomassetti, M., Dekel, A., Mandelker, N., et al. 2016, MNRAS, 458, 4477
Trujillo, I., Feulner, G., Goranova, Y., et al. 2006, MNRAS, 373, L36
Tuccillo, D., Huertas-Company, M., Decencière, E., et al. 2018, MNRAS, 475, 894
van Dokkum, P. G., Franx, M., Kriek, M., et al. 2008, ApJL, 677, L5
Wuyts, S., Förster Schreiber, N. M., van der Wel, A., et al. 2011, ApJ, 742, 96
York, D. G., Adelman, J., Anderson, J. E., Jr., et al. 2000, AJ, 120, 1579
Zolotov, A., Dekel, A., Mandelker, N., et al. 2015, MNRAS, 450, 2327