The Hubble Sequence at z ∼ 0 in the IllustrisTNG simulation with deep learning

Marc Huertas-Company

Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

MNRAS **489,** 1859–1879 (2019)
Advance Access publication 2019 August 30

doi:10.1093/mnras/stz2191

The Hubble Sequence at $z \sim 0$ in the IllustrisTNG simulation with deep learning

Marc Huertas-Company,[1,2]★ Vicente Rodriguez-Gomez,[3] Dylan Nelson,[4] Annalisa Pillepich,[5] Connor Bottrell,[6] Mariangela Bernardi,[7] Helena Domínguez-Sánchez,[7] Shy Genel,[8,9] Ruediger Pakmor,[10] Gregory F. Snyder[11] and Mark Vogelsberger[12]

[1]*Departamento de Astrofísica, Instituto de Astrofísica de Canarias (IAC), Universidad de La Laguna (ULL), E-38200 La Laguna, Spain*
[2]*LERMA, Observatoire de Paris, CNRS, PSL, Université Paris Diderot F-75013, France*
[3]*Instituto de Radioastronomía y Astrofísica, Universidad Nacional Autónoma de México, Apdo. Postal 72-3, 58089 Morelia, Mexico*
[4]*Max-Planck-Institut für Astrophysik, Karl-Schwarzschild-Str 1, D-85741 Garching, Germany*
[5]*Max-Planck-Institut für Astronomie, Königstuhl 17, D-69117 Heidelberg, Germany*
[6]*Department of Physics and Astronomy, University of Victoria, Victoria, British Columbia V8P 1A1, Canada*
[7]*Department of Physics and Astronomy, University of Pennsylvania, Philadelphia, PA 19104, USA*
[8]*Center for Computational Astrophysics, Flatiron Institute, 162 Fifth Avenue, New York, NY 10010, USA*
[9]*Columbia Astrophysics Laboratory, Columbia University, 550 West 120th Street, New York, NY 10027, USA*
[10]*Heidelberg Institute for Theoretical Studies, Schloss-Wolfsbrunnenweg 35, D-69118 Heidelberg, Germany*
[11]*Space Telescope Science Institute, 3700 San Martin Dr, Baltimore, MD 21218, USA*
[12]*Department of Physics, Kavli Institute for Astrophysics and Space Research, MIT, Cambridge, MA 02139, USA*

Accepted 2019 July 26. Received 2019 July 25; in original form 2019 March 18

ABSTRACT

We analyse the optical morphologies of galaxies in the IllustrisTNG simulation at $z \sim 0$ with a convolutional neural network trained on visual morphologies in the Sloan Digital Sky Survey. We generate mock SDSS images of a mass complete sample of $\sim 12\,000$ galaxies in the simulation using the radiative transfer code SKIRT and include PSF and noise to match the SDSS *r*-band properties. The images are then processed through the exact same neural network used to estimate SDSS morphologies to classify simulated galaxies in four morphological classes (E, S0/a, Sab, Scd). The CNN model classifies simulated galaxies in one of the four main classes with the same uncertainty as for observed galaxies. The mass–size relations of the simulated galaxies divided by morphological type also reproduce well the slope and the normalization of observed relations which confirms a reasonable diversity of optical morphologies in the TNG suite. However we find a weak correlation between optical morphology and Sersic index in the TNG suite as opposed to SDSS which might require further investigation. The stellar mass functions (SMFs) decomposed into different morphologies still show some discrepancies with observations especially at the high-mass end. We find an overabundance of late-type galaxies ($\sim$ 50 per cent versus $\sim$ 20 per cent) at the high-mass end [$\log(M_*/\mathrm{M}_\odot) > 11$] of the SMF as compared to observations according to the CNN classifications and a lack of S0 galaxies ($\sim$ 20 per cent versus $\sim$ 40 per cent) at intermediate masses. This work highlights the importance of detailed comparisons between observations and simulations in comparable conditions.

Key words: galaxies: abundances – galaxies: formation – galaxies: photometry.

1 INTRODUCTION

Understanding the physical processes that lead to the diversity of galaxy morphologies we see in today's Universe, i.e. the Hubble Sequence, is still a major goal in the field of galaxy evolution. Until recently, numerical simulations struggled to simulate galaxies with realistic morphologies. An improvement of spatial resolution together with more accurate numerical codes and treatments of physical processes has triggered the emergence of hydrodynamical cosmological simulations which produce galaxies with a variety of morphologies in the local universe (e.g. Dubois et al. 2015; Genel

★ E-mail: marc.huertas@obspm.fr

et al. 2015; Schaye et al. 2015). This allows us to move from a qualitative to a more quantitative approach in which the number densities as well as other scaling relations of different simulated morphologies can be compared to observations. It requires to consistently measure morphologies in the simulations as done in the observations. Radiative transfer codes enable one to forward model the simulation outputs and produce mock observations under some assumptions on the conversion from mass to light as well as on the absorption of light by dust.

However, there have been few works that precisely quantify the detailed morphologies of the simulated galaxies. This is partly due to the fact that quantifying galaxy morphology for large numbers of galaxies has also remained an elusive problem in the observations. An interesting attempt to process simulated galaxies with the same methodology was carried out by Dickinson et al. (2018) in the framework of the Galaxy Zoo project (Lintott et al. 2011). They classified a complete sample of simulated galaxies from the Illustris simulation using the citizen science approach developed by the Galaxy Zoo collaboration and showed that simulated galaxies still presented important differences with respect to observed ones They found in particular that simulated galaxies present more substructures than observed ones, especially at lower masses. One caveat of this approach is that it is very time consuming to consistently process different simulations with the same methodology. Other works have then followed a more automated approach. Bottrell et al. (2017a,b) performed bulge-disc decompositions of mock Illustris galaxies finding also significant differences with the observations. They find in particular a deficit of bulge-dominated galaxies which implies that the size–luminosity relations of Illustris galaxies present higher normalizations and smaller slopes than for real galaxies. More recently, Rodriguez-Gomez et al. (2019) used parametric and non-parametric morphological proxies of the new IllustrisTNG galaxies (see also Snyder et al. 2015 for a similar approach). They measured a significant improvement in terms of scaling relations compared to the original run but some discrepancies remain on the slopes and normalizations of the mass–size relations of early and late-type galaxies. Galaxies from the EAGLE simulation have also recently been processed through radiative transfer codes (Trayford et al. 2017) and used for example to identify barred systems (Elagali et al. 2018). However, there is no precise quantification of the photometric morphological mix. Most of the works focus on the morphologies traced by kinematics (e.g. Correa et al. 2017; Clauwens et al. 2018; Rosito et al. 2018; Thob et al. 2019; Trayford et al. 2019) which are more difficult to compare with large observational samples because of the lack of kinematic data on complete samples.

In recent years, machine learning and more precisely deep learning has emerged as an extremely efficient tool to estimate detailed visual like morphologies from images (e.g. Dieleman, Willett & Dambre 2015; Huertas-Company et al. 2015; Domínguez Sánchez et al. 2018). It offers an interesting and fast approach to consistently compare theory and observations since it becomes possible to efficiently apply the exact same methods to simulations and observations and obtain accurate detailed morphologies. The realism of the simulated galaxies can be quantified in detail.

The main purpose of this work is thus to quantify the detailed optical *visual* morphologies of the new TNG100 simulation of the IllustrisTNG suite (Nelson et al.) and compare with observations. We use to that purpose a convolutional neural network trained on the SDSS (Domínguez Sánchez et al. 2018) to classify a mass selected [$\log(M_*/\mathrm{M}_\odot) > 9.5$] sample of $\sim$ 12 000 simulated galaxies at $z = 0.05$ in four major morphological types (E, S0, Sab, Scd) in the SDSS *r* band. Bayesian neural networks are used to quantify the similarities between observations and simulations. We then analyse the stellar mass functions (SMFs) and mass–size relations of simulated galaxies divided by morphological type and compare with observations.

The paper proceeds as follows. Sections 3 and 2 describe the simulated and observational samples used in this work, respectively. We then describe in Section 4 the main methodology used to quantify galaxy morphologies with Convolutional Neural Networks (CNNs). In Section 5 we discuss the similarity between simulated and observed morphologies from the machine learning perspective. The main results regarding the scaling relations and SMFs of simulated and observed galaxies are presented in Sections 6 and 7, respectively.

2 OBSERVATIONS

2.1 SDSS parent sample: M15

The observational sample used in this work comes from the 670 722 galaxies selected by Meert, Vikram & Bernardi (2015) (hereafter the M15 sample) from the SDSS DR7 spectroscopic sample. We refer the reader to the aforementioned work for all the details regarding the selection. Very briefly, galaxies are selected from the SDSS DR7 database according to three main criteria: (1) the extinction-corrected *r*-band Petrosian magnitude is between 14 and 17.77. The limit at the bright end is to avoid large nearby galaxies which are typically split into multiple objects in the SDSS catalogue. The faint-end limit is the lower limit for completeness of the SDSS spectroscopic survey (Strauss et al. 2002); (2) the photometric pipeline classified the object as a galaxy; and (3) the spectrum was also identified as a galaxy. Some further cleaning of very nearby objects ($z < 0.0005$) and objects with catastrophic photometric redshifts results in sample of 670 722 galaxies. The median redshift of this sample is $\sim$0.09 and goes up to $z \sim 0.25$ which is slightly higher than that of the Illustris TNG $z = 0.05$ snapshot we consider. In order to reduce the impact of possible morphological evolution, we select for this work only objects with $z < 0.1$. This additional selection results in a final sample of 328 709 galaxies. The volume probed by the observational sample is roughly 40 times larger than the simulated TNG volume. The morphological mix, which is the main property we aim to measure in this work, might eventually change in small volumes, especially at the high-mass end. This is addressed in detail in Section 7.

2.2 Structural parameters and stellar masses

A large number of derived quantities exist for the dataset described above. In particular we use for this work the stellar masses computed for all galaxies (Bernardi et al.).[1] Stellar masses are derived using a Chabrier (2003) initial mass function (IMF) and the M/L ratio from Mendel et al. (2015). We do not include any variation in the IMF. We refer to Bernardi et al. (2018) for the implications of IMF gradients in the mass. The luminosity comes from the Sersic best-fitting models derived for all galaxies (see Meert et al. 2015 for more details). We also use the effective radii estimated through fitting Sersic models when comparing the scaling relations of observed and simulated galaxies. Domínguez Sánchez et al. (2018) also derived detailed visual morphologies for this sample with CNNs.

[1]Catalogue available at: http://alan-meert-website-aws.s3-website-us-east-1.amazonaws.com/fit_catalog/download/index.html

However, in this work we perform a new training to ensure that the neural networks are trained on data with similar properties to the simulations.

3 SIMULATIONS

3.1 The IllustrisTNG simulation: TNG

The IllustrisTNG Project (Marinacci et al. 2018; Naiman et al. 2018; Nelson et al. ; Pillepich et al. 2018a; Springel et al. 2018) is a suite of magneto-hydrodynamic cosmological simulations performed with the moving-mesh code AREPO (Springel 2010; Pakmor, Bauer & Springel 2011; Pakmor et al. 2016). See Weinberger et al. (2017) and Pillepich et al. (2018a) for a description of the TNG simulation model which is an improved version of the original Illustris simulation (Genel et al. 2014; Vogelsberger et al. 2014a,b; Sijacki et al. 2015). The IllustrisTNG model was especially designed to match some key observables: (i) the global star formation rate density at $z = 0-8$, (ii) the galaxy mass function at $z = 0$, (iii) the stellar-to-halo mass relation at $z = 0$, (iv) the black hole-to-stellar mass relation at $z = 0$, (v) the halo gas fraction at $z = 0$, and (vi) galaxy sizes at $z = 0$. In this work, we use the highest resolution version of TNG100, which follows the evolution of 2×1820^3 resolution elements within a periodic cube measuring $75\,h^{-1} \simeq 110.7$ Mpc. The main differences with respect to the first Illustris run consist of a new active galactic nucleus feedback model that operates at low accretion rates (Weinberger et al. 2017) and a reworking of the galactic winds (Pillepich et al. 2018a), and the inclusion of magnetic fields (Pakmor et al. 2011). The simulation output, along with ancillary data products, has been recently made publicly available (Nelson et al.).

3.2 Dataset and synthetic images

In this work, we consider a single simulation snapshot at $z = 0.0485$ (snapshot 95) as done in Rodriguez-Gomez et al. (2019). The redshift is consistent with the average redshift of the observational comparison sample ($z_{\rm med} = 0.06$). Within this snapshot, we consider all simulated galaxies with $\log(M_*/{\rm M}_\odot) > 9.5$ which is also roughly consistent with the stellar masses in the observational dataset detailed in the following section. The sample selected contains 12 468 galaxies (hereafter TNG sample). We do not require a perfect stellar mass and redshift match between observations and simulations since all properties will be explored at fixed stellar mass. It is important though that all types of galaxies are well represented in the training set used to train the CNNs. We will further explore this in Section 5.

From the parent sample we create synthetic images for all the galaxies in the snapshot using the radiative code SKIRT (Baes et al. 2011).[2] We refer the reader to Rodriguez-Gomez et al. (2019) for full details on how the images are created. In short, each galaxy is observed from a unique random viewing angle perpendicular to the *xy*-plane of the simulation volume. The field of view of each image is equal to 15 times the (3D) stellar half-mass radius of the corresponding galaxy. The number of pixels is tuned to match the SDSS pixel scale (0.396 arcsec, 0.38 kpc at $z = 0.05$). The stellar populations are modelled with the Bruzual & Charlot (2003) stellar population models for stars older than 10 Myr. Younger stars are considered starbursting regions and are modelled with the MAPPINGS-III photoionization code (Groves et al. 2008). All details of how parameters are set can be found in Rodriguez-Gomez et al. (2019). For computational reasons, dust is taken into account only if the fraction of star forming gas is above 1 per cent of the total baryonic mass. It is assumed for these objects that the dust content is traced by the star-forming gas. A constant dust-to-metal ratio of 0.3 is also assumed. The final output is a 3D data cube for each galaxy, consisting of a full rest-frame SED for each pixel. We then assume that the source is located at $z = 0.0485$ and generate the data cube that would be measured by a local observer, taking cosmological effects such as surface brightness dimming into account. Each SED is then multiplied by each of the SDSS filter curves (g, r, i, z) and integrated over the full wavelength range. We use only the *r* band in this work.

To include instrumental and observational effects, we insert the SKIRT synthetic images into real SDSS fields following the statistical observational realism approach of Bottrell et al. (2017a,b). In this approach, the insertion statistics are guided by a basis catalogue of real galaxies (with similar properties than the M15 dataset) such that the distributions of sky brightness, PSF resolution, and crowding by nearby sources for real galaxies are statistically reproduced in the synthetic images. Consequently, any biases that these or any other field-related properties may have on predicting morphology are equally likely to affect the synthetic data (TNG) and the real data (M15). The adapted Bottrell et al. (2017a) procedure is as follows for every synthetic image:

(i) A galaxy is randomly selected from the Simard et al. (2011) bulge + disc decomposition catalog of 1.12 million SDSS galaxies. The *r*-band field in which that galaxy resides is extracted and converted to electrons. A source mask is generated using SExtractor. An injection site is then selected randomly with the restriction that the *centre* of the injected image does not land on another object in the source mask.

(ii) An SDSS PSF corresponding to the injection site is reconstructed using the SDSS psField files and the dedicated read_PSF code. The SKIRT synthetic image (electrons $\rm s^{-1}$) is converted to electron counts using the SDSS exposure time of 53.9 s, the image is convolved with the SDSS PSF, and source Poisson noise is added.

(iii) The PSF-convolved and Poisson noise-added synthetic image is inserted into the SDSS Field at the selected location. A cut-out which now includes a real sky, real PSF, and real additional sources is then extracted corresponding to the desired FOV – which in our case is the size of each synthetic image (∼50 arcsec).

A detailed investigation showing the importance of observational realism for neural network analyses using synthetic images (specifically, with respect to galaxy merger-stage predictions) is carried out in Bottrell et al. (in preparation). The realism suite is also being made publicly available in tandem with their investigation.

4 DEEP LEARNING R BAND *VISUAL* MORPHOLOGIES

4.1 Training set: N10

The training of the neural networks is performed using the visual morphologically classified sample of Nair & Abraham (2010) (hereafter N10 sample). The catalogue contains detailed morphologies of ∼ 14 000 galaxies performed by two professional astronomers. The authors associate to every galaxy in the sample a numeric value (T-Type) indicating the morphological type spanning from −5 (Elliptical) to 10 (Irr) – see table 3 in Nair & Abraham (2010). We used this dataset instead of the Galaxy Zoo catalogue even

[2] http://www.skirt.ugent.be/root/index.html

if it is smaller in size because the classification reflects well the standard Hubble Sequence which is not the case in the Galaxy Zoo classification tree. We notice that although the N10 dataset has been only classified by two astronomers, the classification is shown to be in very good agreement with other known classifications on common objects. It is therefore considered as a reference for detailed optical morphologies in the local Universe. The catalogue contains galaxies with $0.01 < z < 0.1$ which is compatible with the M15 and TNG datasets. However, the classification is only done for bright galaxies ($g < 16$) so the S/N is on average higher than for the M15 and TNG samples. Domínguez Sánchez et al. (2018) has shown that this S/N difference between training and test does not introduce significant biases in the final classification. Moreover and most importantly, any eventual bias will be present in both the TNG and M15 samples since they both have similar properties.

4.2 Network architecture and training

We use the same vanilla architecture as in Domínguez Sánchez et al. (2018) which has been shown to perform well on galaxy morphology. The network architecture is a standard CNN with four convolutional layers, each of them followed by a pooling layer of size (2×2). The number of filters in each layer is 32, 64, 128, and 128, respectively, and the kernel sizes are 6, 5, 2, and 3. The convolutional part is then followed by a fully connected network with two layers of sizes 64 and 1 (output layer). In this work, we performed a new training instead of directly using the published catalog because Domínguez Sánchez et al. (2018) used JPEG images from the SDSS to train the networks. Since we want to make sure in this work that the neural networks see exactly the same data in the observations and in the simulations, we use fits *r*-band images as input for the training. The input stamps are of fixed size 128×128 ($\sim$50 kpc $\times$ 50 kpc at $z = 0.05$) which is larger than two effective radii of ~ 90 per cent of galaxies in the sample. Before being fed into the CNN, images are background subtracted by removing the median of the pixels in an empty region and normalized to the maximum value so that all images span a similar range between 0 and 1. We also tried other non-linear normalizations such as hyperbolic sine to boost the signal in the outskirts of the galaxies. However this had no significant impact in the performance so we decided to keep a simpler linear normalization.

The training strategy is also a bit different than in Domínguez Sánchez et al. (2018). In that work, they performed a regression on the T-type of the galaxy. We simplified the problem here into a hierarchical binary classification problem as done in Huertas-Company et al. (2011) which is easier to evaluate and train (see Silla, Carlos & Alex 2011 for a review of hierarchical classifications). This is enough for our purposes since we will use only four main classes. We thus train three different binary classifiers with the same architecture. The first model (Model-1) is trained to separate early-type (Ttype ≤ 0) from late-type galaxies (Ttype ≥ 1). Model-1 delivers therefore a probability for a galaxy to be late-type: $P(\mathrm{Late}) = 1 - P(\mathrm{Early})$. A second model (Model-2) distinguishes Ellipticals (Ttype ≤ -3) from S0/a's ($-3 <$ Ttype < 1). Model-2 measures then the probability of being S0 with the prior that the galaxy is early-type: $P(\mathrm{S0/Early}) = 1 - P(\mathrm{E/Early})$. Finally a third model (Model-3) splits objects between Sab, Sb galaxies ($1 \leq$ Ttype < 4) and late-type spirals and irregulars (Ttype ≥ 4) with a training set made only of late-type galaxies. Model-3 estimates the probability for a galaxy to be a late-type spiral given that it is a late-type galaxy: $P(\mathrm{Scd/Late}) = 1 - P(\mathrm{Sab/Late})$. The sizes of the training set decreases as one goes deeper into the tree, but it is enough to avoid overfitting.

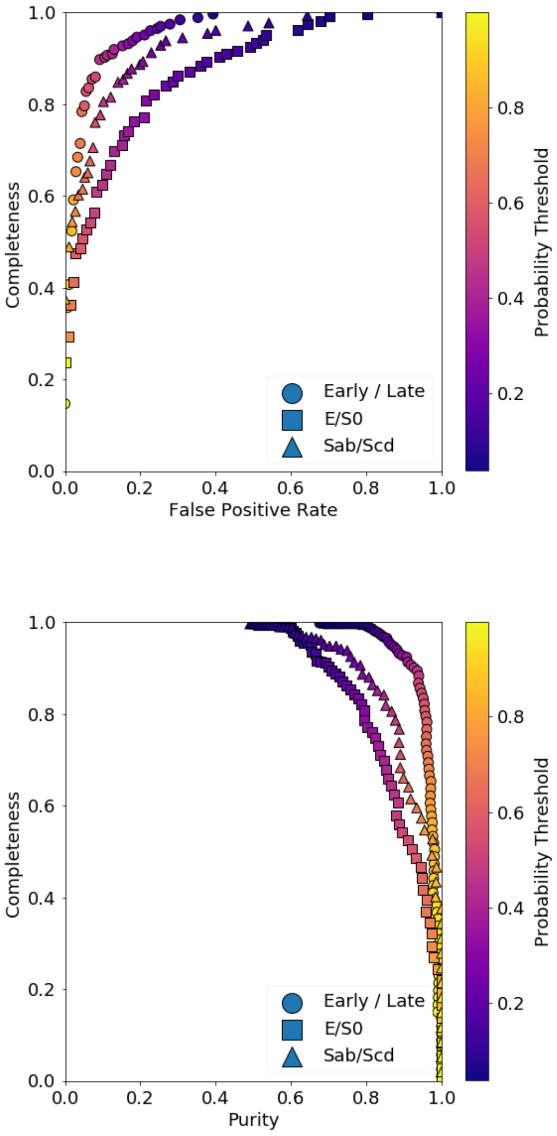

Figure 1. Performance of the three morphological classifiers used in this work (see text for details). The top panel shows the ROC curve and the bottom panel the precision–recall curve (purity–completeness). The circle, square, and triangle symbols indicate respectively the early/late, Sab/Scd, and E/S0 classifiers. The colour bar indicates the corresponding probability threshold. See text for details.

Fig. 1 shows the ROC (receiver operating characteristic) and precision–recall (or purity–completeness) curves for the three classifications computed on a test set never used for training. In a binary classification, the ROC curve shows the fraction of false positives (i.e. in our case the fraction of galaxies classified as late-type among all galaxies with an early-type label) versus the fraction of true positives (i.e. in our case, the fraction of galaxies classified as late-type among all galaxies with a late-type label). The true positive rate is typically called in astronomy completeness. Since the network outputs a probability and not a binary number, one can change the threshold to define positives (i.e. late-type galaxies). The smaller the threshold the larger the fraction of true positives but also the larger the fraction of false positives. This is shown in

the ROC curve in which every point shows the fractions of false positives and true positives for different thresholds. The closer the curve gets to the top left corner, the more accurate the classifier is. For comparison, a random classifier will always have an equal fraction of false and true positives. The precision–recall curve is another indicator in which, instead of plotting the fraction of false positives, plots the fraction of true positives among all positive examples classified by the network (i.e. fraction with a true late-type label among all objects classified as late-type). The latter is a proxy for purity. For the P-R curve, both quantities need therefore to be as close as possible to one.

As expected the best accuracy is achieved for the first classifier (early versus late) with a ~ 90 per cent purity and completeness consistent with previous works. The accuracy slightly decreases when more detailed morphologies are considered but still remains above 80 per cent in both purity and completeness. This is achieved for a typical probability threshold around 0.5 as expected for a well calibrated classifier. It is worth noticing that in building these ROC curves for E/S0s and Sab/Scds, we assume that the samples are free of contaminations. We also emphasize that the main purpose of this work is not to obtain the best possible match with a human-based classification but to apply exactly the same model to observations and simulations.

We then use the three models to classify both the TNG and M15 samples. Every galaxy in both samples has therefore three different probabilities. The output of Model-2 has however little meaning for galaxies classified as late-type by Model-1. The same is true for the output of Model-3 and galaxies classified as early-type by Model-1. Therefore we associate four probabilities to every galaxy using the Bayes theorem:

$$\begin{aligned} P(\mathrm{E}) &= P(\mathrm{Early}) \times P(\mathrm{E/Early}) \\ P(\mathrm{S0}) &= P(\mathrm{Early}) \times P(\mathrm{S0/Early}) \\ P(\mathrm{Sab}) &= P(\mathrm{Late}) \times P(\mathrm{Sab/Late}) \\ P(\mathrm{Scd}) &= P(\mathrm{Late}) \times P(\mathrm{Scd/Late}) \end{aligned} \quad . \tag{1}$$

We then simply put a galaxy in the class of maximum probability. In the following, early-type galaxies include ellipticals and S0/a's (also called lenticulars) and late-type galaxies include Sabs and Scds. We will also refer to Sab galaxies as early-type spirals, and to Scd objects as late-type spirals. Figs 2–5 show some example stamps of galaxies classified in the four types ordered by increasing stellar mass both in the TNG and in the M15 datasets. As can be appreciated, elliptical galaxies are mostly bulge-dominated systems. S0 or lenticular galaxies have a dominant bulge component but tend to have a disc with no marked features. Sab galaxies (early-type spirals) have smaller but still noticeable bulges and large discs with spiral arms and/or visible structure in the disc. Finally, Scd galaxies (late-type spirals) have a very small bulge or no bulge at all and a clumpy disc component or with irregular morphology.

The first thing to notice is that the CNN model trained on SDSS successfully identifies galaxies in the TNG simulation in the four morphological types and that simulated and observed galaxies in a given class share some obvious features. We emphasize that this does not mean that simulated and observed galaxies are not distinguishable. The network is forced to put galaxies in any of the four classes by construction. The fact that there are objects in the four classes, means only that some of the features learned by the networks to identify the different morphologies in the images are present both in the simulations and in the observations. As a matter of fact, Fig. 5 clearly reveals some discrepancies between simulated and observed Scd galaxies. Simulated objects appear systematically more extended and also generally more clumpy than observed Scds. The two edge-on systems also appear to be thicker than observed edge-on systems. However, the CNN still finds that the closest morphological type is a late-type spiral.

5 HOW REALISTIC ARE THE TNG MORPHOLOGIES FROM THE MACHINE LEARNING PERSPECTIVE?

The previous section has shown that the CNNs trained on the visual morphologies from the N10 samples find objects in all four classes also in the TNG simulation. One first interesting question is how *confident* the networks are about the classification in the simulation. Machine learning algorithms will always try to associate objects with the classes they were trained with because there is an implicit assumption that there is a perfect match between the training and test datasets. This is not necessarily the case in this work since we are training in the observational domain and inferring in the simulated one.

In the following we try to quantify the similarity between the simulated and observed morphologies from the neural network perspective. We adopt two different approaches. First we measure the network confidence by inferring the uncertainties through a bayesian approach. Secondly, we compare the features learned by the CNN in the simulated and observed samples. We stress that this exercise is not probing whether simulated and observed galaxies can be distinguished.

5.1 Bayesian neural networks

Dropout was first introduced as a method to reduce the risk of overfitting when training deep neural networks (Hinton et al. 2012). By randomly removing some neurons during the training phase, we do not allow neurons to become too specific and ease generalization. Gal & Ghahramani (2015) have shown that Monte Carlo dropout in the inference phase can be formally used to approximate the model uncertainty. We adopt this approach in this work to associate an uncertainty measurement to all classified galaxies. Every galaxy is classified 500 times dropping out a variable fraction of neurons ranging from 30 per cent to 50 per cent at each layer, including the convolutional layers. That way, instead of having a single softmax probability value, every galaxy has an associated probability distribution, arising from the 500 classifications. We repeat the dropout sampling for the three trained models and then compute the uncertainty for every galaxy by quadratic addition of the standard deviations of the different distributions:

$$\sigma_{\mathrm{tot}} = \sqrt{\sigma^2_{\mathrm{Model1}} + \sigma^2_{\mathrm{Model2}}} \text{ if Morphology} = \mathrm{E, S0}$$

$$\sigma_{\mathrm{tot}} = \sqrt{\sigma^2_{\mathrm{Model1}} + \sigma^2_{\mathrm{Model3}}} \text{ if Morphology} = \mathrm{Sab, Scd.}$$

The above equations assume that the probability distributions estimated through dropout are well approximated by Gaussian distributions. We inspected several of them and verified that they do not present complex shapes with double peaks. The Gaussian approximation is thus justified. We then compute the uncertainties for the M15 and TNG samples (see Section 2). The cumulative distributions of uncertainties for both datasets are shown in Fig. 6. We find that the distributions are very similar for both datasets indicating that the degree of uncertainty is comparable in the simulations and in the observations. This confirms the realism of

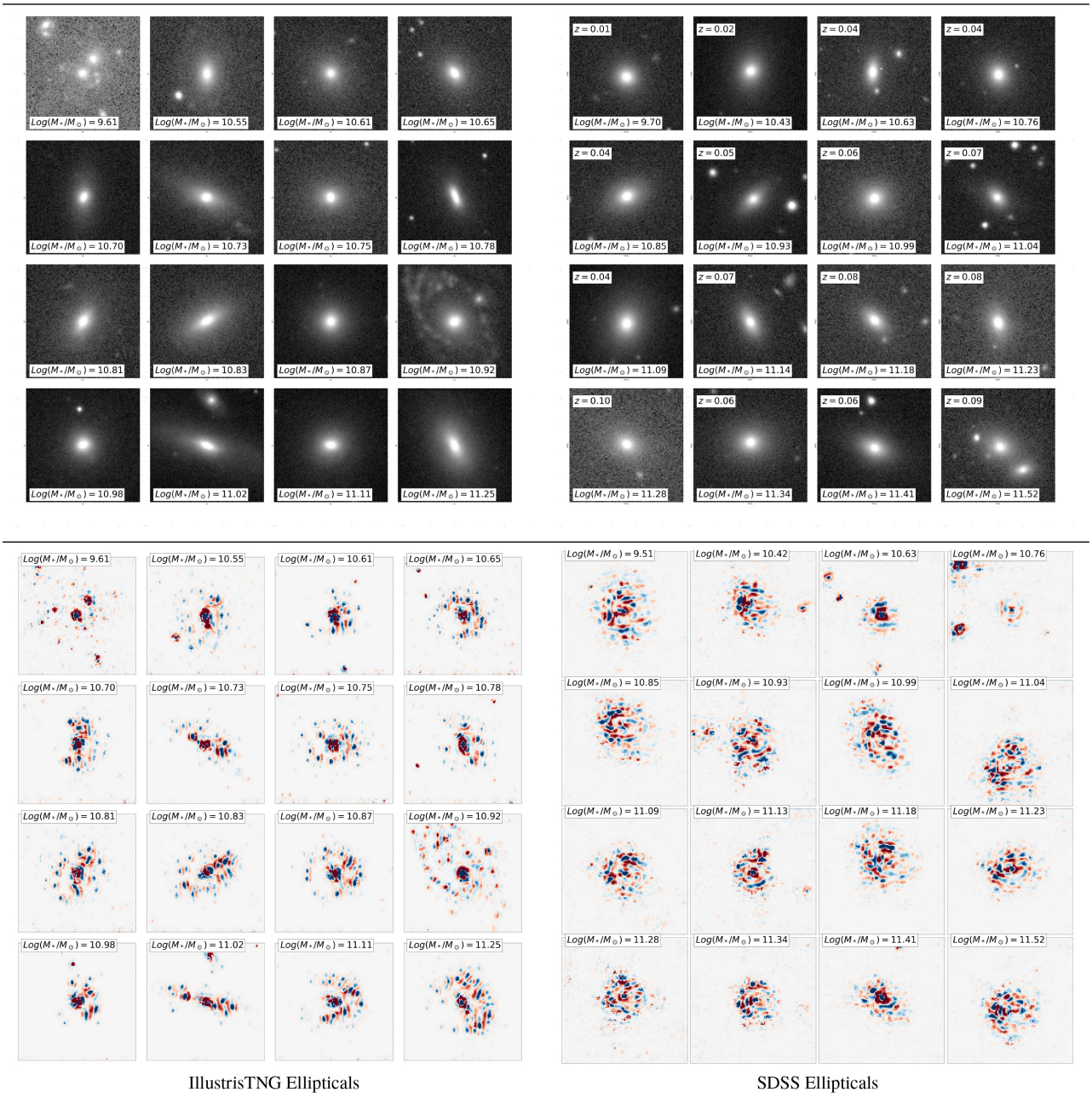

Figure 2. Example stamps of galaxies classified as Es (ellipticals) by the CNN sorted by increasing stellar mass. The left-hand panel shows simulated galaxies and the right-hand panel observed ones. The stamps are all 128 × 128 pixels (50 arcsec FoV). For visualization purposes, images have been normalized and converted to jpg with a non-linear hyperbolic sine normalization to better appreciate the outskirts. The grey scale is arbitrary. The bottom panels show the attribution maps of the same images computed through integrated gradients. Blue and red colours indicate negative and positive values, respectively. No response (0 values) are represented in white. The maps are normalized between the maximum and minimum values so the units are arbitrary and they only reflect variations from a blank image.

galaxy morphologies in the TNG run or at least that the features learned by the CNNs to identify the different morphologies are found in both the simulations and the observations. One interesting question that arises is *what are these features*. We try to address this in the following section by exploring the attribution maps.

The fraction of outliers can also be used to quantify differences between real and synthetic data. We use the observational dataset as a reference to define objects with large uncertainties in the TNG dataset. To that purpose we compute the median error value and the standard deviation of the distribution of uncertainties in the M15 dataset and define as outliers objects with a measured uncertainty larger than three times the standard deviation (dashed vertical line in Fig. 6). The number of outliers defined that way is only ~ 1 per cent in TNG. From this we confirm that simulated galaxies do not show more uncertain classifications than observations. In Fig. 7, we investigate the nature of outliers in both datasets. The top row of the figure shows images of the 16 galaxies in TNG and M15 with the largest uncertainties. In the majority of the cases, the large

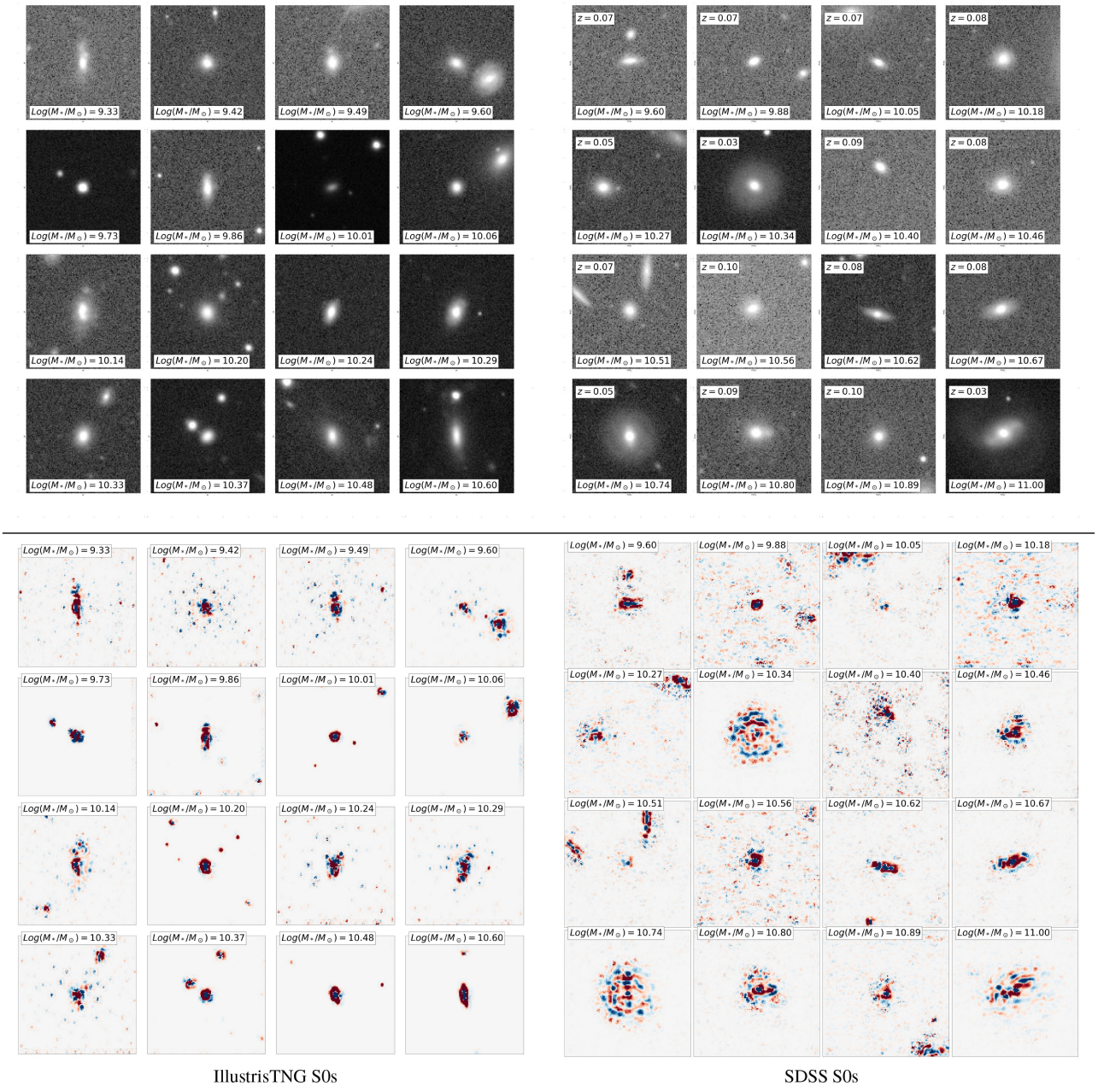

Figure 3. Example stamps of galaxies classified as S0s (lenticulars) sorted by increasing stellar mass. The left-hand panel shows simulated galaxies and the right-hand panel observed ones. The stamps are all 128 × 128 pixels (50 arcsec FoV). Images have been converted to jpg with a non-linear hyperbolic sine normalization to better appreciate the outskirts. The grey scale is arbitrary. The bottom panels show the attribution maps of the same images computed through integrated gradients. Blue and red colours indicate negative and positive values, respectively. No response (0 values) are represented in white. The maps are normalized between the maximum and minimum values so the units are arbitrary and they only reflect variations from a blank image.

uncertainties are due to the presence of very bright companions (stars or galaxies) or too large galaxies that do not fit in the stamps. This confirms that our error measurement is sensitive to outliers and also that the observational realism included in the TNG dataset is satisfactory given that the fractions of such cases are comparable (except for the centring problems which are not included in TNG).

Finally, the measurement of uncertainties also allows us to quantify how stable are the morphological classes in the simulated sample as compared to the observed one. To that purpose, we perform 500 different classifications for all galaxies by randomly changing probabilities with a Gaussian random number with a standard deviation equal to the dropout uncertainty. We then compute the typical scatter in the final morphological type. A value lower than 1 means that the galaxy does not change morphological class. Results are shown in the right-hand panel of Fig. 6. The fraction of objects with a morphological standard deviation larger than 1 is very small indicating that for most of the galaxies the morphological type is well constrained. Also the distributions are almost identical for the TNG and M15 datasets, which is an indication that the probability distributions are very similar for both populations. This

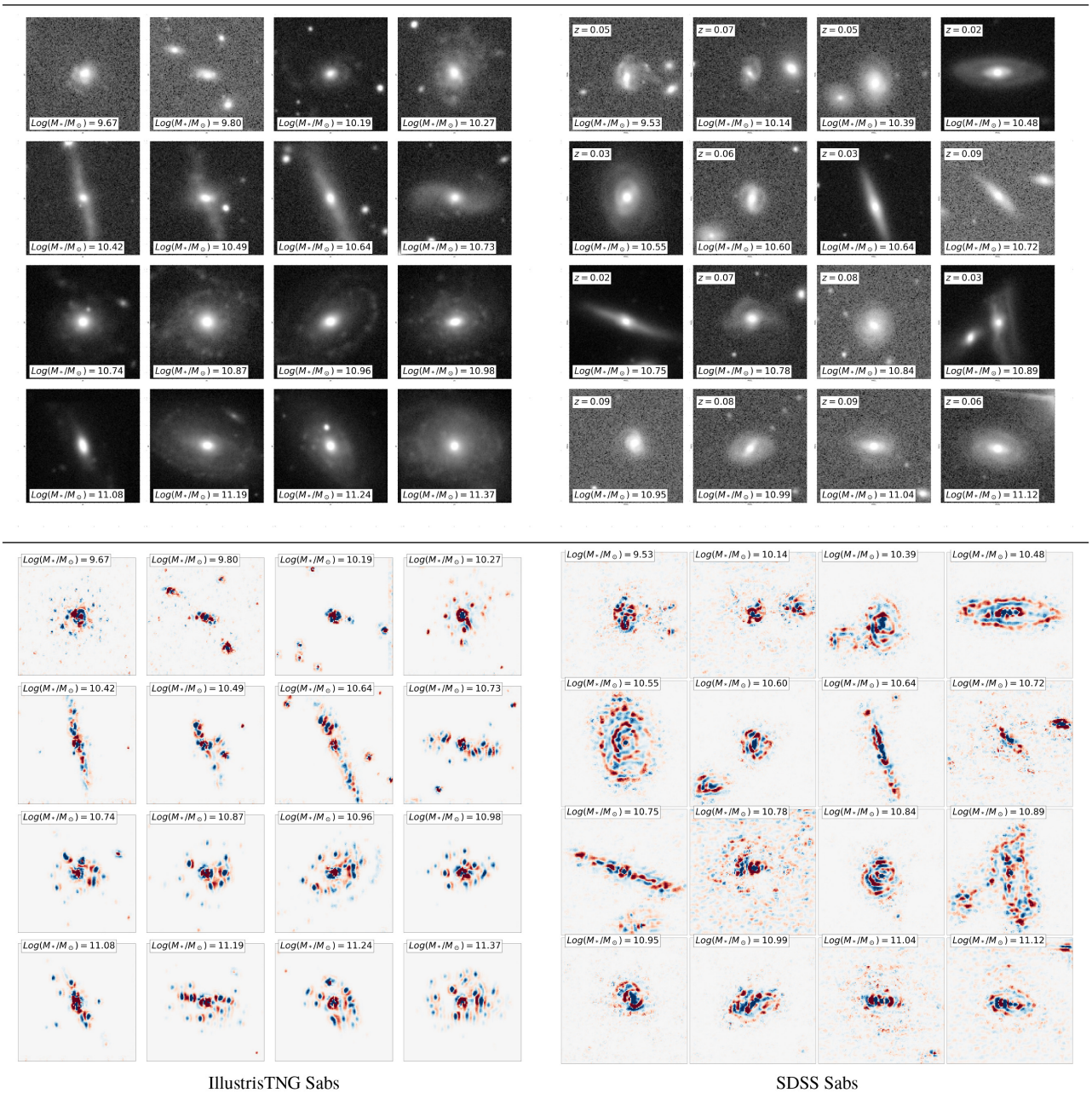

Figure 4. Example stamps of galaxies classified as Sabs (early-type spirals) sorted by increasing stellar mass. The left-hand panel shows simulated galaxies and the right-hand panel observed ones. The stamps are all 128 × 128 pixels (50 arcsec FoV). Images have been converted to jpg with a non-linear hyperbolic sine normalization to better appreciate the outskirts. The grey scale is arbitrary. The bottom panels show the attribution maps of the same images computed through integrated gradients. Blue and red colours indicate negative and positive values, respectively. No response (0 values) is represented in white. The maps are normalized between the maximum and minimum values so the units are arbitrary and they only reflect variations from a blank image.

is somehow in contrast with the main findings of Dickinson et al. (2018). It points to an improvement of TNG with respect to the first Illustris (Rodriguez-Gomez et al. 2019) but also might indicate that the CNNs are sensitive to different features than human classifiers. In the bottom row of Fig. 7 we show some random examples of galaxies with large morphological uncertainty [$\sigma(T_{\rm type}) > 1$]. Those are typically small objects sometimes in crowded regions. This is expected since the N10 sample does contain few examples of small galaxies and also because the resolution prevents a proper morphological classification of these objects. Also the fraction of crowded fields in the N10 sample is limited which explains the large uncertainty. Additionally and more interestingly, there are several cases in TNG of bulge-dominated galaxies presenting a ring of clumpy star formation around which do not seem to exist in the observations. A deeper exploration of the formation history of these systems might be interesting.

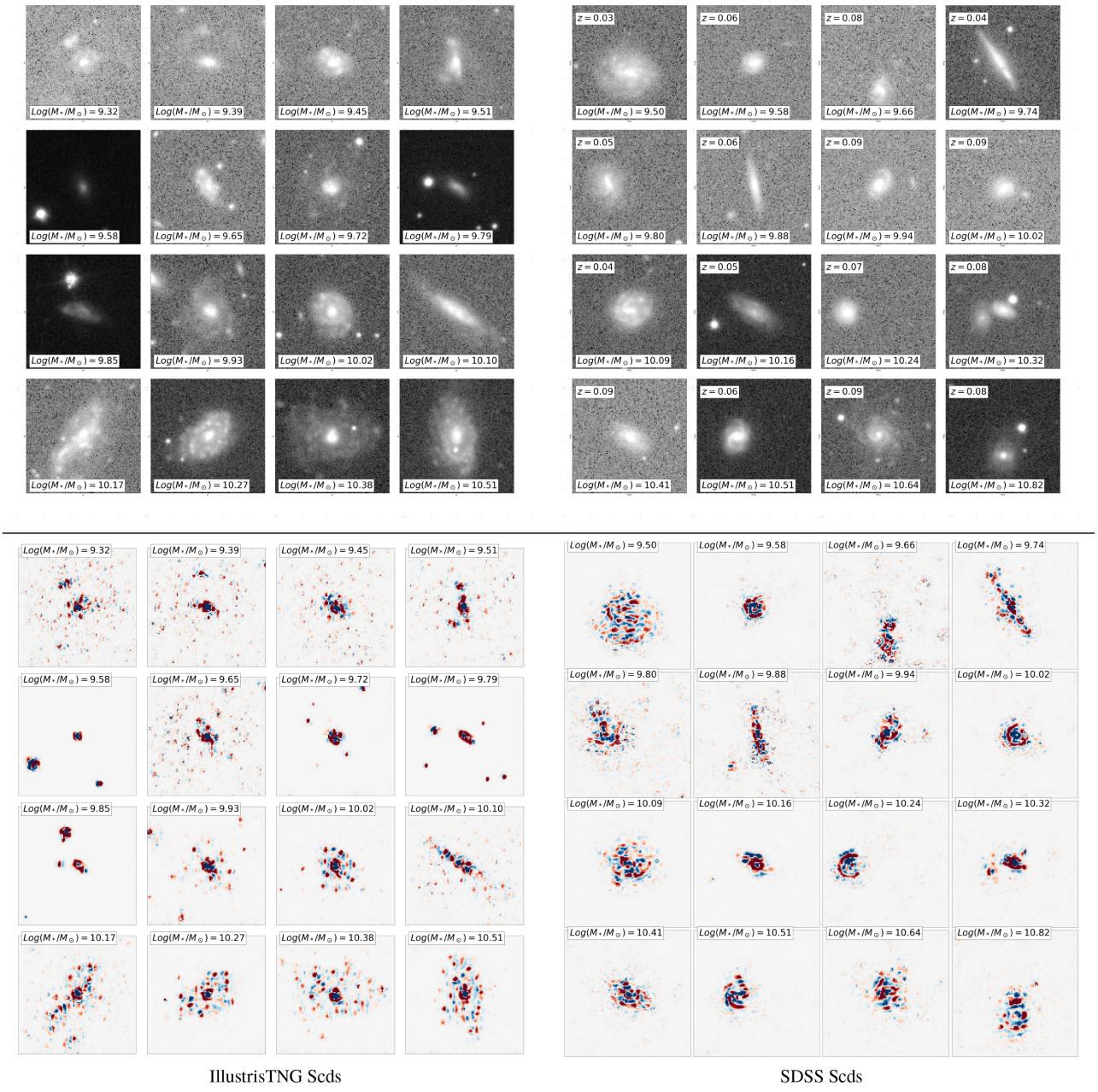

Figure 5. Example stamps of galaxies classified as Scds (late-type spirals) sorted by increasing stellar mass. The left-hand panel shows simulated galaxies and the right-hand panel observed ones. The stamps are all 128 × 128 pixels (50 arcsec FoV). Images have been converted to jpg with a non-linear hyperbolic sine normalization to better appreciate the outskirts. The grey scale is arbitrary. The bottom panels show the attribution maps of the same images computed through integrated gradients. Blue and red colours indicate negative and positive values, respectively. No response (0 values) is represented in white. The maps are normalized between the maximum and minimum values so the units are arbitrary and they only reflect variations from a blank image.

5.2 Comparison of attribution maps and features learned

Deep neural networks have been proved to produce very high classification accuracies. The price to pay is less control on the features used by the networks to perform the classification. In this section we try to explore the similarity between the features extracted by the network in the simulations and in the observations as a way to quantify how close are the morphologies between TNG and SDSS. Figs 2–5 show the attribution maps for the same galaxies represented in the top panels. The maps indicate the pixels that contributed most to the network decision for a given image. They are computed here using Integrated Gradients (Sundararajan, Taly & Yan 2017), but other attribution techniques we tested gave similar results. The maps are directly not translatable into a *physical* set of features but can help localizing where in the galaxies is the information used for classification. We do see that the attribution maps generally trace the pixels belonging to the galaxy, confirming that the network is ignoring the noise when classifying galaxies. This might appear like a trivial statement, but it is not given that there might be some S/N trends in the training set. For example, Scd galaxies tend to be fainter than Ellipticals so the networks might

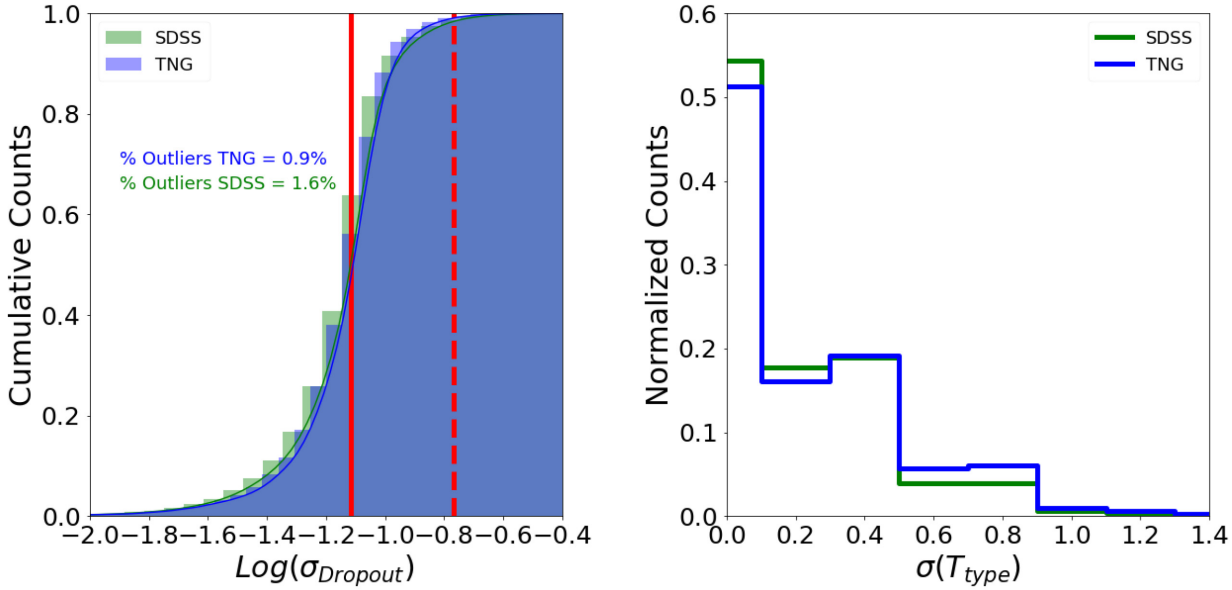

Figure 6. Left-hand panel: Cumulative distribution of the logarithm of uncertainties of the morphological classification estimated through Monte Carlo dropout. The blue histogram corresponds to the TNG sample and the red one is for the M15 dataset. The fraction of outliers defined as the fraction of objects with uncertainties larger than three times the standard deviation of the M15 distribution is also indicated. The red solid line shows the median value and the dashed line is the 3σ limit used to define outliers. Right-hand panel: Standard deviation distribution of the morphological type for the M15 (green) and TNG (blue) datasets.

have learned that a lower S/N is correlated with the morphological type. The attribution maps show it is not the case. We do also observe that for elliptical galaxies the important pixels seem to be more concentrated towards the central regions than for later types which indicates that the network is focusing on the bulge component for these systems as one would expect. The attribution maps also reveal that in cases where there are several galaxies of comparable brightness in the stamps, they both contribute to the morphological classification. As discussed in the previous section, this might be a consequence of the lack of such objects in the training set. It also explains the larger uncertainty measured in these systems (see Section 5.1). We emphasize that this bias is the same for synthetic and real data.

The comparison between the maps in TNG and SDSS qualitatively reveals that similar features are found in both datasets at fixed morphology. For Sab galaxies for example (Fig. 4) we appreciate how the activation pixels similarly trace the disc region and even a kind of spiral structure when it is visible. In summary the maps provide limited information (or at least not easily interpretable in terms of physical quantities measured in images) but allow one to confirm that no major differences are found between the simulated and observed samples.

In order to better understand if the CNN appreciates noticeable differences between the TNG and SDSS datasets we perform an exploration of the features learned by the network of Model1. We extract the features after the last convolutional layer before the dense part of the network and compare them. To that purpose we create a feature vector for a subsample of 500 galaxies from the N10 sample used for training, 500 additional galaxies from the M15 dataset, and 500 galaxies from TNG. Since the feature space is highly dimensional ($\sim 100\,000$) we project it into a two-dimensional space for visualization purposes using the dimensionality reduction algorithm tSNE (t-distributed stochastic neighbour embedding, van der Maaten & Hinton 2008). We use a learning rate of 900 and a perplexity value of 30. However changes in these parameters do not change the main trends. The result of this exercise is shown in Fig. 8. Recall that the axes have arbitrary units and do not encapsulate any physical meaning. The TNG (blue squares) and the TNG (green circles) samples form a unique cluster in the space defined by the two features extracted with tSNE. It confirms that similar features are found by the CNN in both datasets, which trigger comparable responses of the neurons. We also observe that data points from the M15 and TNG samples are slightly off centred with respect to the N10 sample (red crosses). The reason seems to be that the M15 and TNG samples contain fainter galaxies. This is in some sort a limitation of the approach followed in this work since we are using a sample of bright galaxies to train while we are inferring on significantly fainter objects. As explained in previous sections, this is not critical in this work since the same biases are propagated to both TNG and M15 samples. It confirms however that the visualization of the feature space is sensitive to differences in the galaxy properties. The main conclusion from this plot is therefore that simulated galaxies do not show significant differences in the feature space.

From these different tests, we can conclude that, at least from the neural network perspective, comparable features are found in the simulations and in the observations. Training on N10 and inferring in TNG seems justified. We can be confident that the morphologies estimated in the TNG sample are reliable. In the next sections, we analyse their physical properties and abundances as compared to observations.

6 STRUCTURAL PROPERTIES

We now explore the similarity between simulated and observed galaxies from a more physical perspective, i.e. by looking at the structural properties and scaling relations.

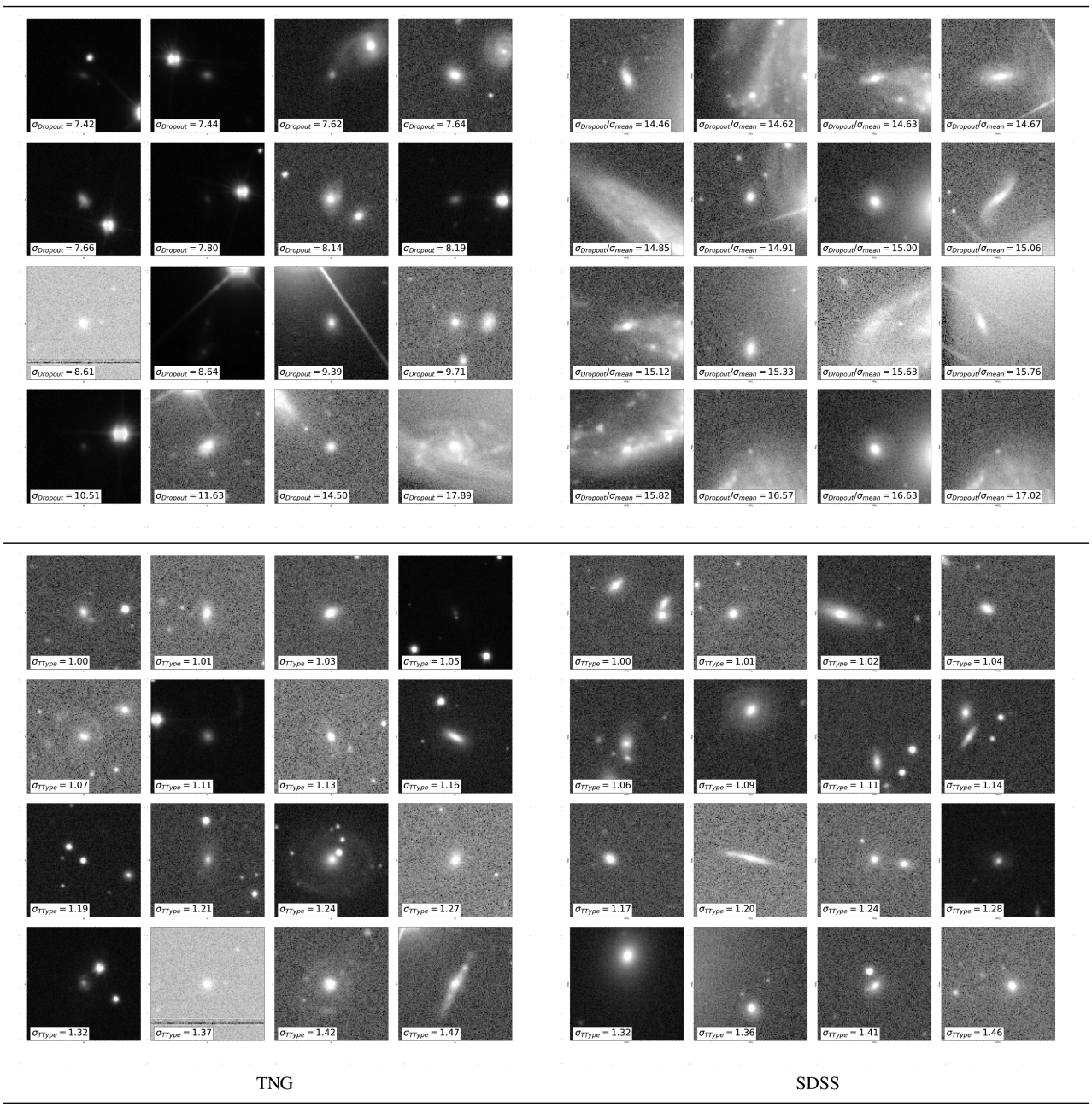

Figure 7. Top row: Objects with the largest uncertainties as inferred from the Monte Carlo dropout technique used in this work. Bottom row: Galaxies for which the morphological uncertainty [$\sigma(T_{\rm type})$] is larger than one type. See text for details. The left-hand (right-hand) panels show the TNG sample and the right-hand panels the M15 one.

6.1 Sersic index distributions and kinematic morphology

The Sersic index (n) is commonly used as a proxy for morphology. It is therefore interesting to see how well the relation between n and morphological type is reproduced in the simulations. For the simulations, Rodriguez-Gomez et al. (2019) performed also Sersic fits on the projected 2D light maps that we use here. Although both studies did not use strictly speaking the same method we assume that the one-component Sersic fits are stable enough so that no major systematics are introduced. Fig. 9 shows the distribution of Sersic indices for the four morphological types considered in this work. The different morphological types show clearly different distributions in the M15 dataset which confirms that the CNN classification is identifying galaxies with different structural properties. Galaxies of type Scd have a Sersic index distribution clearly peaking at 1 which is indicative of a pure exponential profile with no bulge. For elliptical galaxies the distribution is skewed towards values larger than three typical of pure bulge-dominated systems. In between the Sab galaxies have low Sersic index values but larger than Scds which is indicative of the presence of a bulge. Finally, S0s present the broadest distribution which is also something expected. We notice that this also shows that S0s and Elliptical galaxies have different structural properties. Simulated galaxies qualitatively follow similar trends but the differences in the distributions are less clear. In particular, we notice the distribution

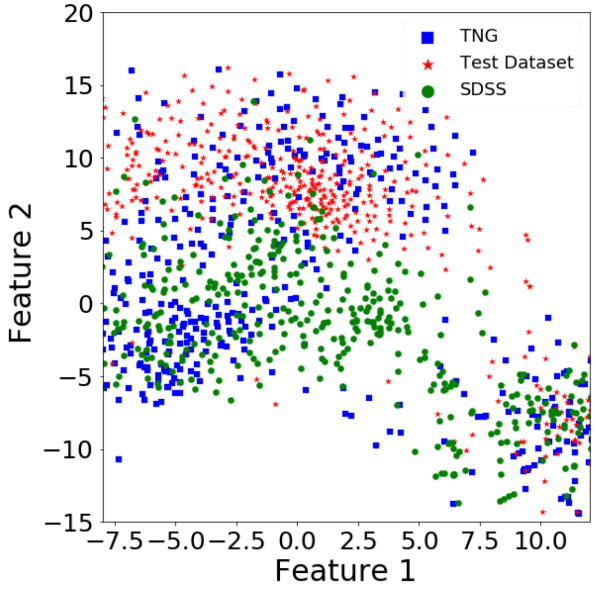

Figure 8. Visualization of the features learned by the CNN in a two-dimensional space obtained with tSNE (see text for details). The red points show the N10 sample, the green points are galaxies in the M15 dataset, and the blue points are simulated galaxies from TNG.

for elliptical peaks at $n \sim 2$ instead of $n \sim 4$ in SDSS. The difference between S0 and Sab galaxies is also less apparent in the simulations. A word of caution should be raised when interpreting these differences since the methods used to compute the Sersic index in simulations and observations are different. However, if these differences are confirmed, it would indicate that the surface brightness profile of these simulated galaxies (especially ellipticals) differs from observations. This does not necessarily prevent the CNN to establish a morphological class for TNG galaxies with high confidence as detailed in the previous section. It suggests that other *global* morphological features are used to classify galaxies.

For simulated galaxies, we have also access to kinematic morphology measurements which is an independent indicator. We plot, in Fig. 10, the distribution of the $\kappa_{\rm rot}$ parameter for the four morphological types. $\kappa_{\rm rot}$ measures the fraction of the kinetic energy that is invested in ordered rotational motion (e.g. Sales et al. 2012; Rodriguez-Gomez et al. 2017). It is defined as the fraction of the total kinetic energy contributed by the azimuthal component of the stellar velocities, where the z-axis coincides with the total angular momentum of the galactic stellar component (see Sales et al. 2012; Rodriguez-Gomez et al. 2017 for details):

$$\kappa_{\rm rot} = \frac{1}{K}\sum_i \frac{1}{2} m_i \left(\frac{j_{z,i}}{R_i}\right)^2,$$

where K is the total kinetic energy of the stellar component, m_i represents the mass of the particle, $j_{z,i}$ is the z-component of the specific angular momentum, and R_i is the projected radius. The figure shows that, as expected, ellipticals are the ones with the lower average value of rotational support ($\kappa_{\rm rot} \sim 0.4$) and Scd galaxies present the larger value on average ($\sim$0.55). S0s and Sabs are in between. Globally, $\sim$ 60 per cent of galaxies with $\kappa_{\rm rot} < 0.5$ are early-type and $\sim$ 75 per cent of objects with $\kappa_{\rm rot} > 0.5$ are late-type. This confirms that the CNN-based optical morphologies do correlate with stellar kinematics as one would expect. However, the distributions are quite broad showing that a selection based on kinematics does not perfectly match an optical-based selection (e.g. Emsellem et al. 2007; Bernardi et al. 2019). We emphasize however that the purpose of this work is to compare the morphological properties of simulations and observations in comparable conditions, not to find the optimal definition of morphology. This said, comparing the relation between optical and kinematic morphology in observations and simulations might provide additional constraints. It could be done by simulating for example Manga (Bundy et al. 2015) cubes of TNG galaxies (Pérez-Montaño et al., in preparation).

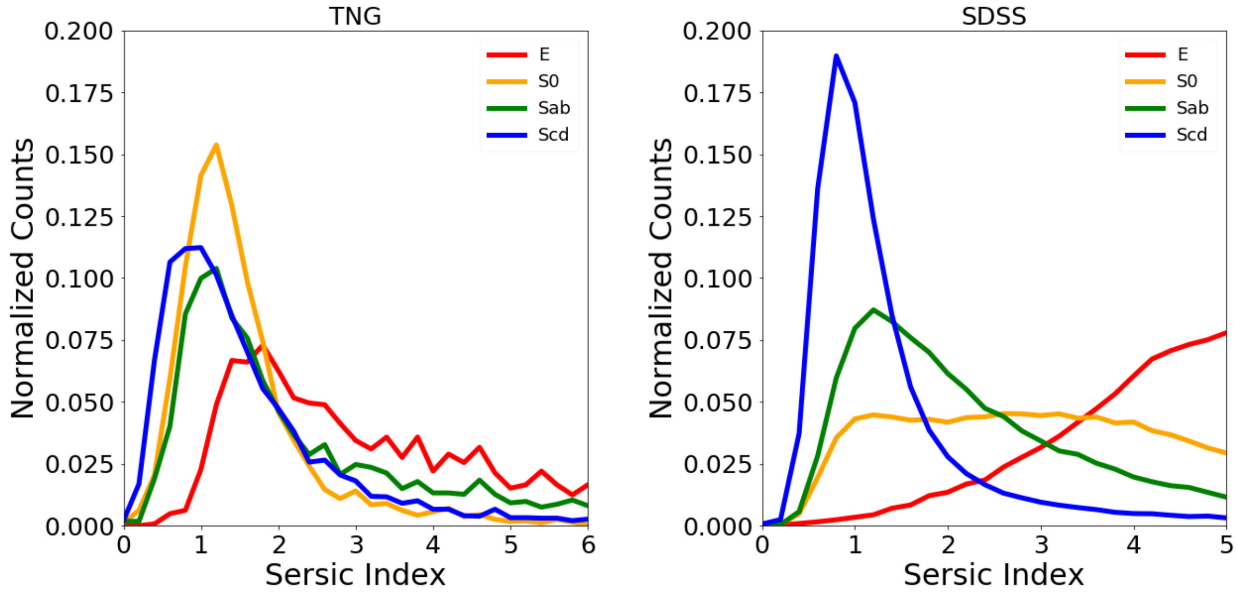

Figure 9. Sersic index distributions for simulated galaxies (left-hand panel) and observed (right-hand panel) for the four morphological types considered in this work.

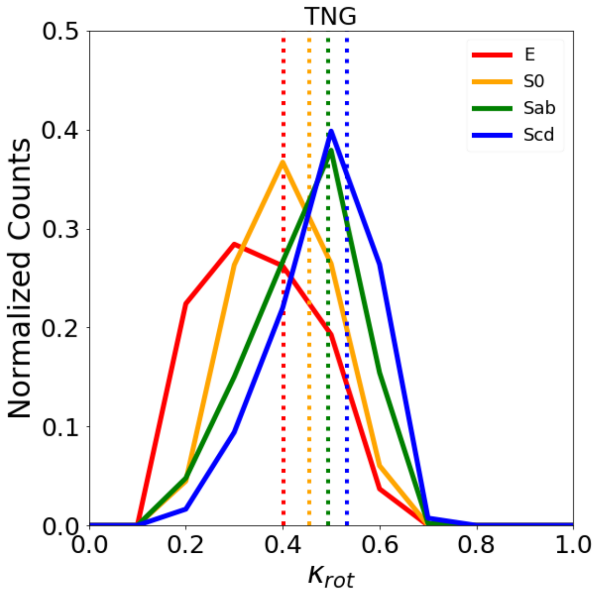

Figure 10. Distribution of $\kappa_{\rm rot}$ for the four different morphological types in TNG. $\kappa_{\rm rot}$ measures the fraction of the kinetic energy that is invested in ordered rotational motion (see text for details). The vertical dashed lines show the median values.

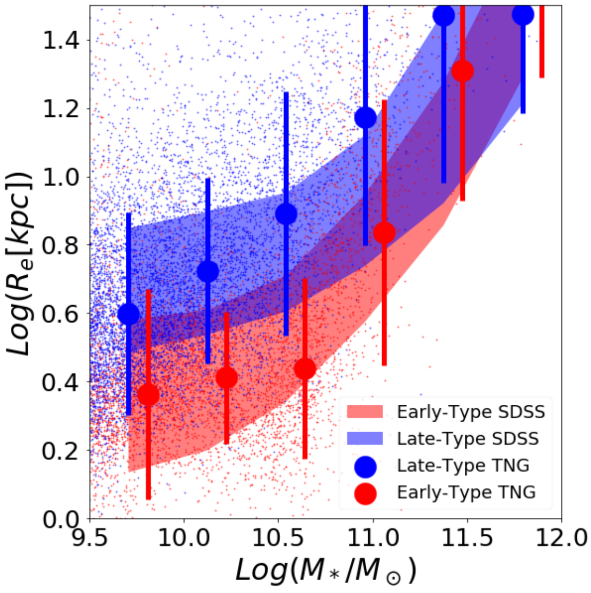

Figure 11. Stellar mass–size relation of early- and late-type galaxies in IllustrisTNG and in SDSS. The shaded red (blue) regions show the observed median mass–size relations along with the $1-\sigma$ scatter for early- (late-) type galaxies. The red (blue) points show the distribution of individual simulated early- (late-) type galaxies. The large dots with error bars indicate the medians and scatters in bins of stellar mass, respectively.

6.2 Stellar mass–size relation

We now focus on the stellar mass–size relation. We use the semimajor axis of the best Sersic model as a size estimator for galaxies both in the simulations and in the observations. As detailed in Section 2, the fitting approach in the M15 sample is fully described in Meert et al. (2015). For the stellar mass in the SDSS, we use, for consistency, the stellar mass estimated using the luminosity from the best single Sersic model. In the simulations we use an aperture of 30 kpc. This aperture has been shown to provide stellar mass estimates in good agreement to those within Petrosian radii in observations (Schaye et al. 2015) and a reasonable compromise for comparison with observations also towards the highest-mass end (Pillepich et al. 2018a). As will be shown in Section 7, the 30 kpc aperture provides an excellent match to the observed SMF.

Fig. 11 shows the mass–size relations for early- and late-type galaxies. We observe a reasonably good match between observed and simulated galaxies. Namely the simulations reproduce well the largely reported trend in the observations that late-type galaxies are larger than early-type galaxies at fixed stellar mass (e.g. Bernardi et al. 2014). The slopes of both relations are also well captured in the simulations. Notice that using the same synthetic images, Rodriguez-Gomez et al. (2019) found no significant differences in the sizes of early- and late-type galaxies in TNG. This might be due to the fact that they used only the Sersic index to define the two morphological classes while here we are using a definition based on the global appearance of the galaxies. It emphasizes the importance of using accurate global descriptors of morphology. This is a remarkable improvement as compared to the first Illustris run which showed significant discrepancies in the scaling relations (i.e. Bottrell et al. 2017a), namely shallower slopes and higher normalizations than in the observations. Fig. 11 shows instead that both the slope and the normalization match reasonably well the observations. The simulations present still a slightly larger scatter in size at fixed stellar mass. Note that Genel et al. (2018) also measured a size difference between quenched and star-forming galaxies in the TNG simulation (with no morphological selection), also in reasonable agreement with observations not only at $z = 0$ but also to higher redshifts. One possible explanation of this improvement could also be that the mass–size relations match by construction. The neural networks could indeed use the size as a parameter to estimate the morphology. There are several reasons why this is unlikely. First of all not all galaxies of a given morphology have the same size. The effective radius depends both on redshift and mass as shown in Fig. 11. The effective radii of elliptical galaxies for example changes by a factor of $\sim$10 from low to high mass. This would imply that the CNN has learned both the redshift and mass dependence which is even more unlikely given that the training set (N10) is not complete (it contains brighter galaxies). Finally, the fact that the scatter in size at fixed mass is significantly larger in the simulations suggests that size is not a primary estimator used by the network.

In Fig. 12 we now explore the mass–size relations divided in finer morphological classes. We also do find a remarkable agreement between observed and simulated galaxies for all the morphological types. The median sizes in TNG generally within the 1σ confidence interval of the observations.

7 STELLAR MASS FUNCTIONS OF DIFFERENT HUBBLE TYPES

We explore in this section the SMFs of the different morphologies in the observations and in the simulations. Fig. 13 shows first the total SMF as well as the SMF of early- and late-type galaxies. The bottom panels of Fig. 13 indicate the fraction of early- and late-type galaxies as a function of stellar mass as well as the ratio between simulated and observed galaxies in bins of stellar mass. The total SMF is in excellent agreement with the results of Bernardi et al. (), i.e. to the 20–40 per cent level. This is not too surprising given that

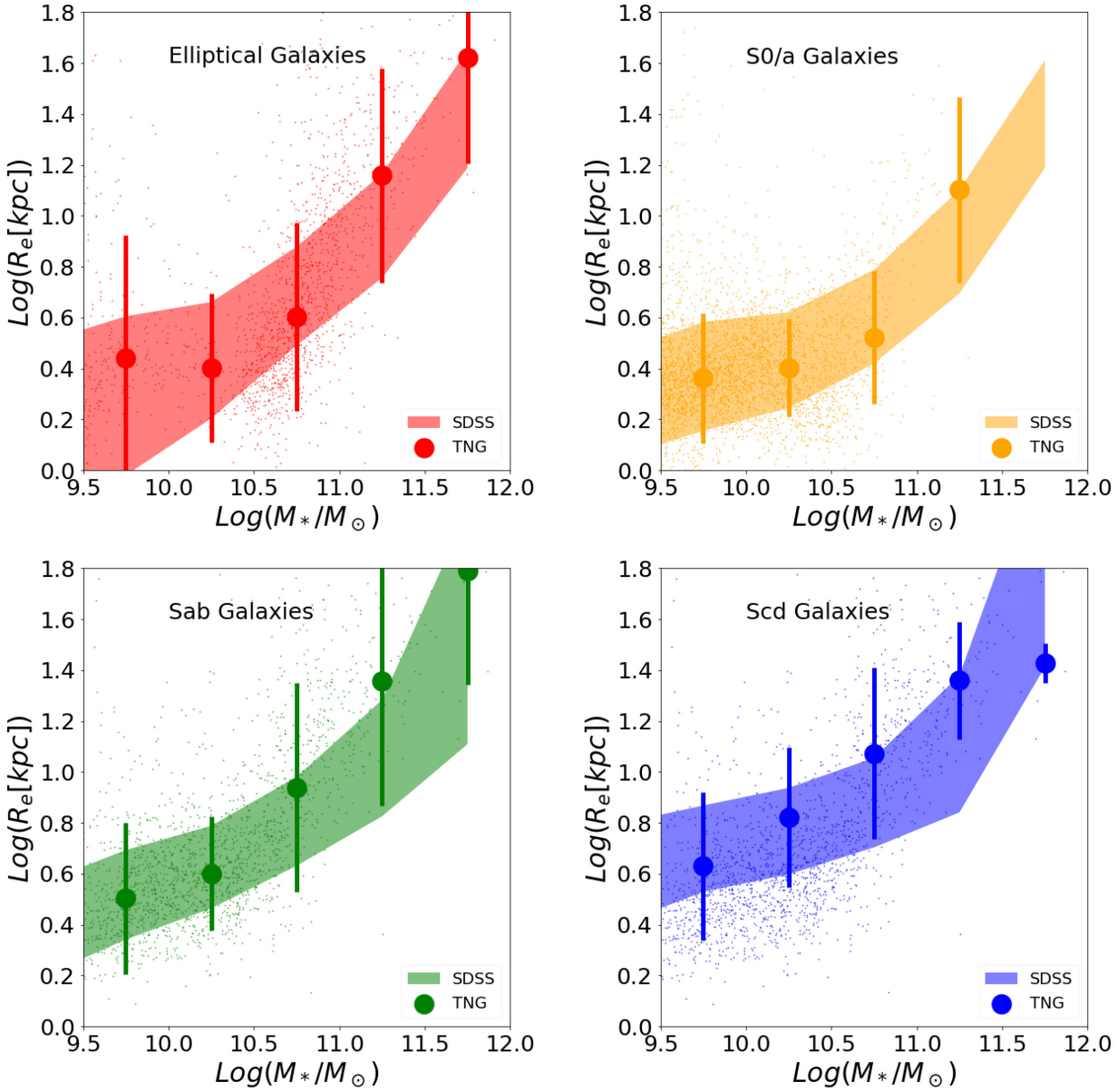

Figure 12. Stellar mass–size relation of observed and simulated galaxies divided in four morphological types as labelled. The shaded regions indicate the observed relations. Small dots indicate individual TNG galaxies and the large dots with error bars are median values and scatter.

the TNG model was designed to improve upon the original Illustris in matching the SMF at $z = 0$ (Pillepich et al. 2018a).

Surprisingly, when considering the SMF divided by two broad morphological types, we observe that relative fractions are well reproduced until a stellar mass of $\sim 10^{11}$ solar masses where TNG presents an excess of late-type galaxies as compared to SDSS. In the SDSS, the high-mass end is clearly dominated by early-type galaxies ($\sim$ 80 per cent) as reported by many previous works (e.g. Bernardi et al. 2013). However, in TNG the high-mass end of the SMF appears to have around 50 per cent of late-type galaxies. As described in Section 2, the volume probed by the simulations is $\sim$40 times smaller than in the observations. The morphological mix at the high-mass end could be affected by small statistics. In order to evaluate the impact of this, we recompute the SMF in the SDSS in 40 smaller volumes. The result is shown with dashed lines in Fig. 13. The difference measured between observations and simulations is larger than the variations caused by measuring abundances in smaller volumes. Therefore, even if TNG is able to produce realistic morphologies, this result suggests that the abundances might require some additional tuning.

We explore further the origins of this discrepancy by dividing the sample in finer morphological classes as described in Section 4. The results are shown in Fig. 14. It confirms that most of the discrepancies come from the early-type population.

At intermediate masses, the number densities are a factor of $\sim$8 smaller in TNG for S0/a galaxies. This effect is the opposite in the elliptical population of similar mass. The S0 class is traditionally the more challenging one since it is the most difficult to define. S0 and elliptical galaxies differ on the disc component. However, the disc is not always obvious, especially when it is seen face-on. As detailed in the previous sections, the difference between S0s and Sabs classes resides essentially in the size and structure

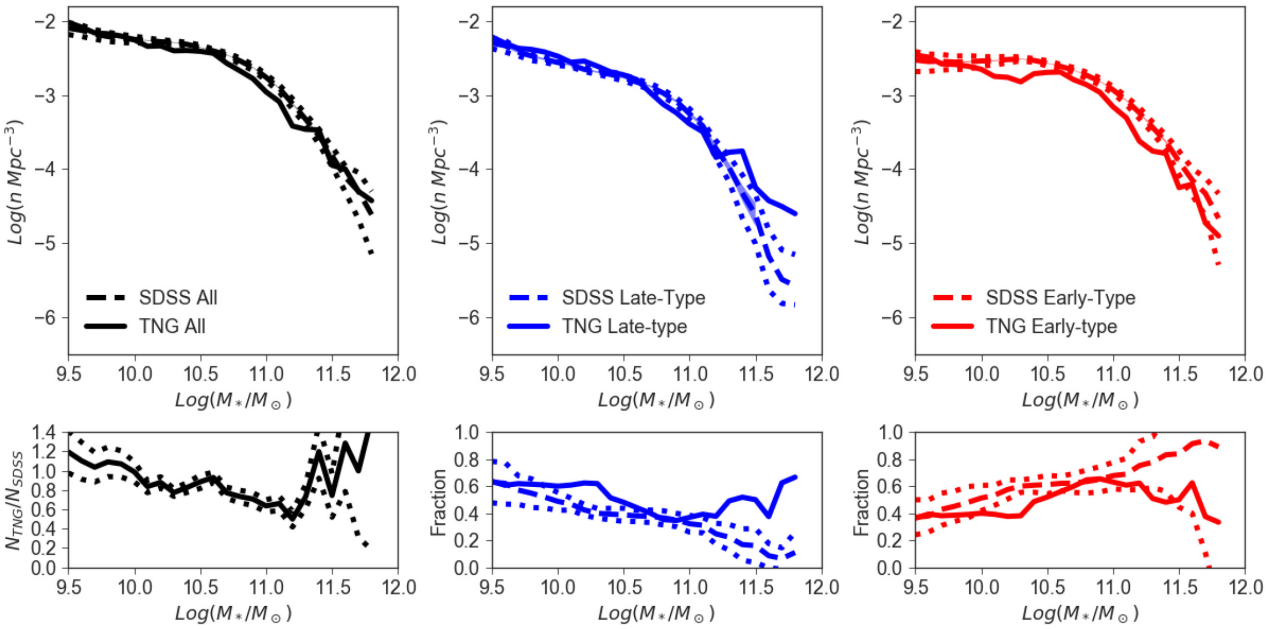

Figure 13. SMFs of all (top left panel), early- (top right panel) and late-type (top middle panel) galaxies. The solid lines show the IllustrisTNG measurements and the dashed lines are in SDSS. The dotted lines show the variations due to volume and the shaded regions are Poisson errors. The bottom panels show the fraction of early- and late-type galaxies as a function of stellar mass in the observations (dashed lines) and in the simulations (solid lines). The dotted lines show the maximum fluctuation in the relative abundances due to volume.

of the disc component. Sa galaxies are expected to have more features in the disc such as spiral arms. This is not always easy to appreciate with limited spatial resolution. For example, using a different classification method based on support vector machines and colour information, Bernardi et al. (2013) finds more Sa galaxies at the low-mass end (in detriment of S0s) than our measurements here. However since the same CNN model was used to classify both the simulations and the observations, there is an internal consistency which allows us to safely argue that the discrepancy is real. The TNG simulations presents a lack of S0 galaxies at intermediate masses. Also notice that, as shown in the previous section, the scaling relations of S0, Sab, and elliptical galaxies are different and well reproduced by the simulations. If this was a problem of classification errors in the simulations we would have measured some deviations in the mass–size relations in Fig. 12.

At the very high-mass end, in which observations are completely dominated by elliptical galaxies, the simulations present an over-abundance of late-type systems (Sabs). Although, TNG presents limited statistics at these stellar masses, the difference seems a bit too large in order to be fully explained by volume issues. In Fig. 15, we show some examples of massive late-type galaxies [$\log(M_*/\mathrm{M}_\odot) > 11$] in IllustrisTNG. Although the galaxies have a prominent bulge component, and are on average rounder than typical discs, they also present a clear extended featured structure which is most probably the feature used by the network to classify the galaxy as a late-type system. These galaxies are in particular different from the typical elliptical galaxies of similar mass shown in the right-hand panel of Fig. 15 and closer to Sab galaxies (Fig. 4) in the sense that the extended low surface brightness component presents more structure, probably due to on-going star formation. This probably causes the network to interpret the structure as a disc. We notice that the difference between these two populations of massive galaxies is not measurable using the Sersic index as a proxy. Fig. 16 shows the Sersic index distribution of massive galaxies [$\log(M_*/\mathrm{M}_\odot) > 11$] classified as late and early type. Both distributions look very similar indicating that the central bulge dominates the surface brightness distribution and is also very similar in both populations as can also be appreciated in Fig. 15. The difference in classification is certainly driven by the diffuse disc component. As a matter of fact, the right-hand panel of Fig. 15 shows that the distribution of κ for the disky population is more skewed towards larger values, suggesting that the different CNN classifications are justified.

At the low mass end, the simulated and observed SMFs match reasonably well. The galaxy population below 10^{10} solar masses is essentially dominated by Scd galaxies in both datasets.

As a final note, one could argue that these discrepancies might be partially caused classification errors in the observations since the training set used lacks faint galaxies (see Section 5). We have checked that this is not the case by computing the abundances of the different morphological types only in the N10 sample. We measure very similar trends as in the whole M15 sample although with more noise given the incompleteness and low statistics. Another potential source of discrepancy could arise from the way images of simulated TNG galaxies are generated. Discreteness effects from the finite particle resolution and smoothing procedure applied to the stars (e.g. as discussed/explored in Torrey et al. 2015) influence the presence of feature/structure that can be interpreted by the CNNs as discs (Bottrell et al. 2017a). In future work, we plan to quantify the impact of this by experimenting with the smoothing prescription in low stellar density regimes.

8 DISCUSSION: THE ASSEMBLY HISTORIES OF DIFFERENT MORPHOLOGIES

The previous sections have shown that the visual morphologies of galaxies in the TNG simulation reproduce fairly well the observed morphologies both in terms of global, visual morphology and scaling relations. However, there are still some discrepancies in the

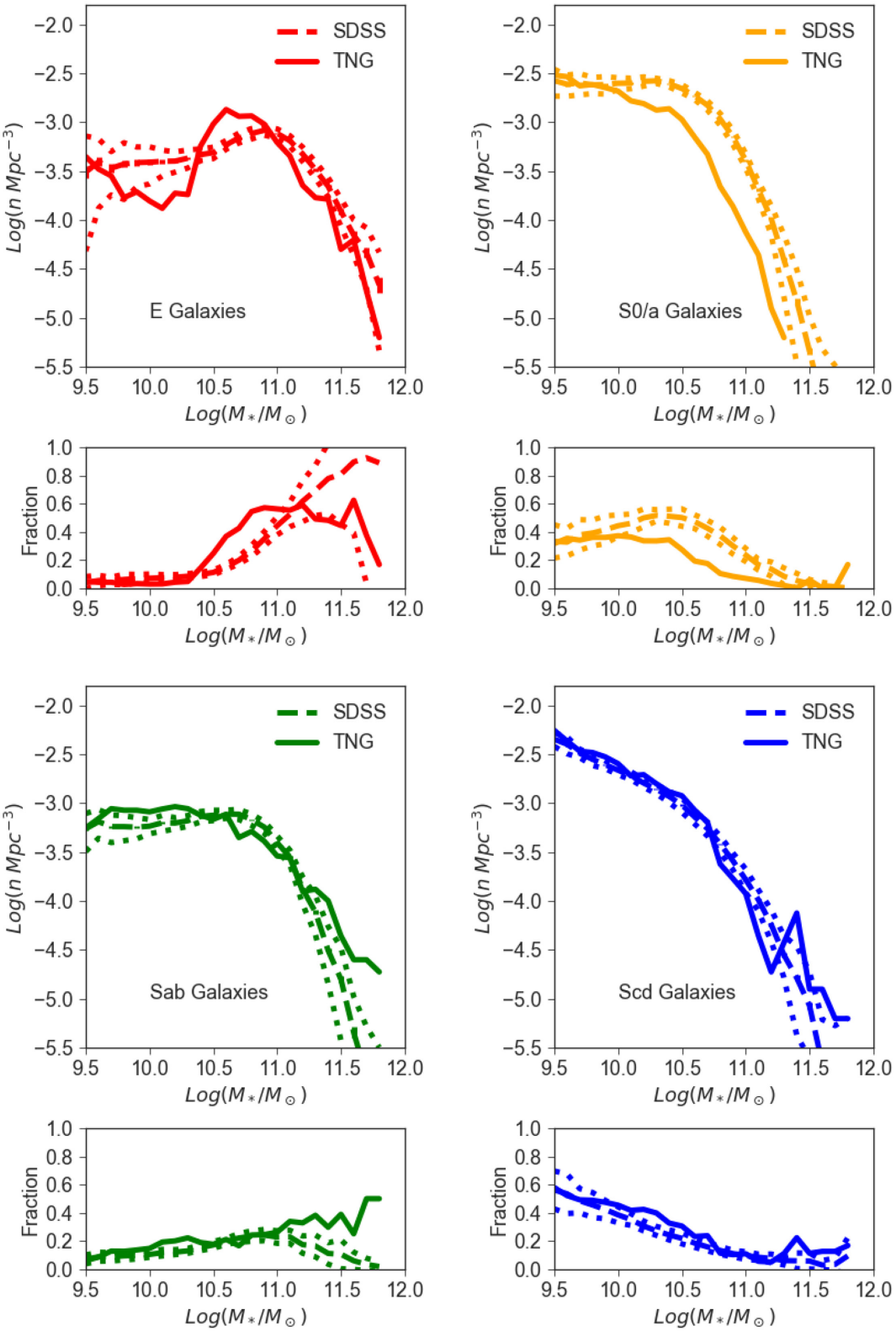

Figure 14. SMFs of different morphological types as labelled. The solid lines indicate the TNG simulations and the dashed lines show the measurements in the SDSS. The small panels show the fractions with respect to the total as function of stellar mass.

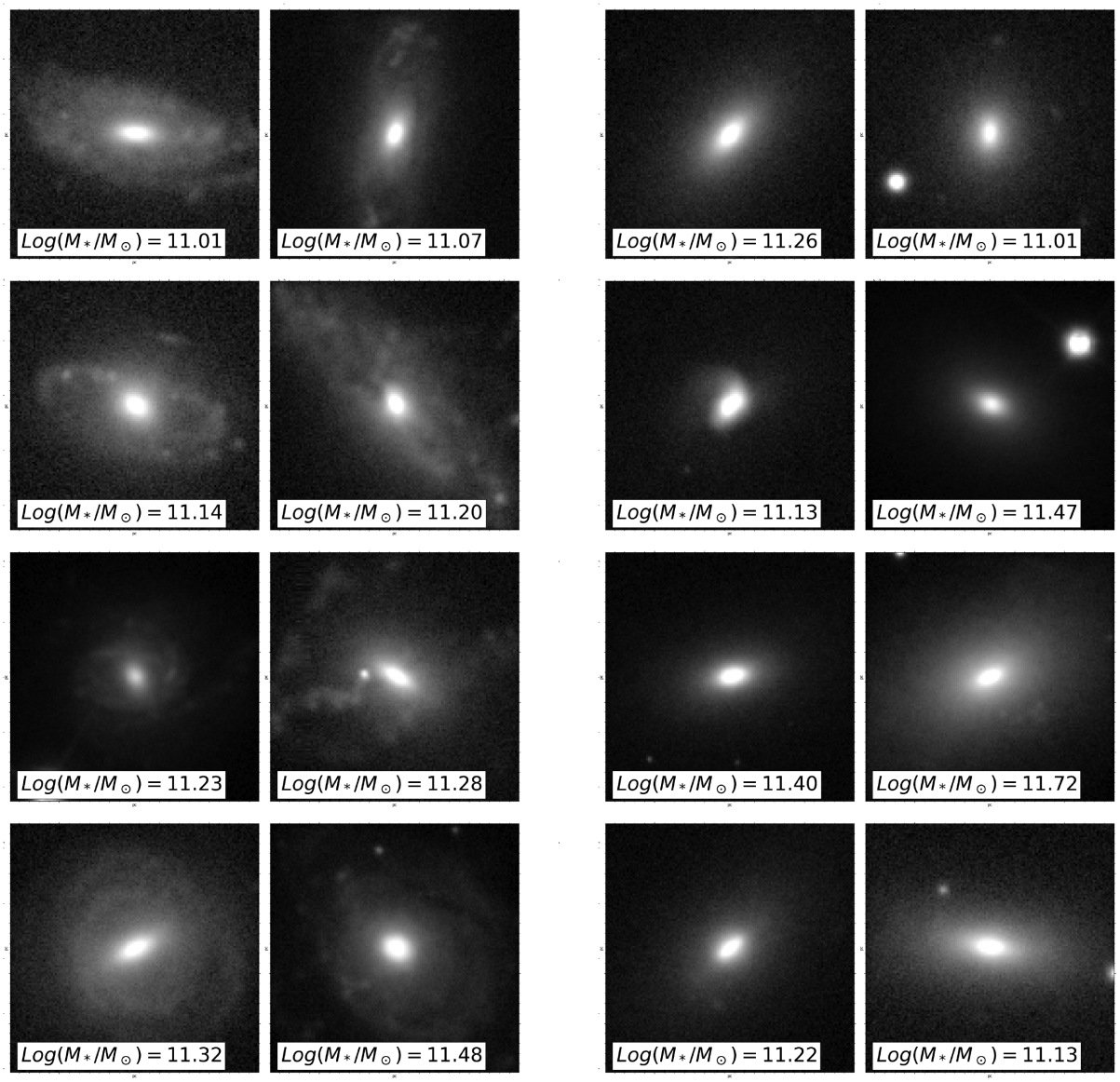

Figure 15. Example of stamps of massive galaxies ($M_*/M_\odot > 10^{11}$) classified as late-type galaxies (left) and early-type by the CNN in the IllustrisTNG simulation.

abundances of early- and late-type galaxies. It suggests that while the assembly channels of the different morphological types produce a realistic morphological distribution, the relative importance of the different mechanisms does not seem to be fully correct so the number densities are not always well reproduced.

In order to better understand the origin of the morphological classes, as well as possibly of these discrepancies especially at the high-mass end, we explore in Fig. 17 some tracers of the assembly histories. In particular, we first look at the contribution of mergers in the stellar mass assembly. The left-hand panel of Fig. 17 shows the fraction of *ex situ* stellar mass as a function of stellar mass (*Ex situ* Stellar Mass). By *ex situ* we mean stars that have formed not *in situ*, i.e. from gas condensing within the innermost regions of the observed galaxy (or its main progenitors) but in other galaxies that have been accreted, stripped and that have possibly merged with a galaxy prior to the time of observation (see Rodriguez-Gomez et al. 2016 and Pillepich et al. 2018b for operational definitions and basic results from Illustris and the IllustrisTNG simulations). The *ex situ* stellar mass fraction should be a proxy of the importance of mergers in the assembly histories.

First, as previously shown (e.g. Rodriguez-Gomez et al. 2016, Pillepich et al. 2018b and reference therein), the *ex situ* fraction is a very strong function of galaxy mass. Below $\sim 10^{10.5}$ solar masses, the amount of accreted stellar mass is negligible (<15–20 per cent) for all morphologies. Fig. 14 shows that $\sim$ 40 per cent low-mass galaxies in TNG have a bulge component (typically S0 or Sab galaxies). Depending on the bulge-to-total mass fraction, it could be that bulges in these systems might have grown through (also or exclusively) internal processes. Above $\sim 10^{10.5}$ solar masses, the amount of accreted stellar mass starts to be significant and reaches almost 80 per cent at $10^{11.5}$ solar masses, with large galaxy-to-galaxy variations. The scatter in *ex situ* fraction at fixed galaxy mass for different morphological types is also large. Yet, we do observe that massive ellipticals have on average larger *ex situ* fractions than later types, at least around the 10^{11} $M_\odot$ scale: e.g. 65 per cent versus 45 per cent at $10^{11.25}$ $M_\odot$. This finding

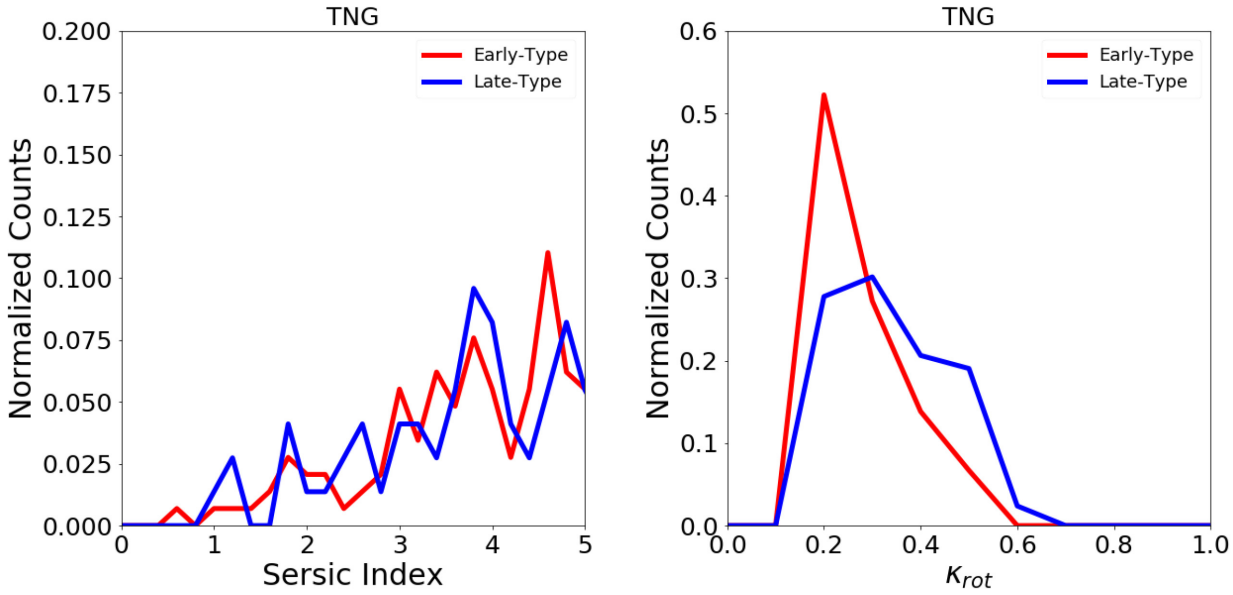

Figure 16. Left-hand panel: Sersic index distribution of massive galaxies ($M_*/M_\odot > 10^{11}$) classified as late-type (blue) or early-type (red). Right-hand panel: $\kappa_{\rm rot}$ distribution for the same massive galaxies.

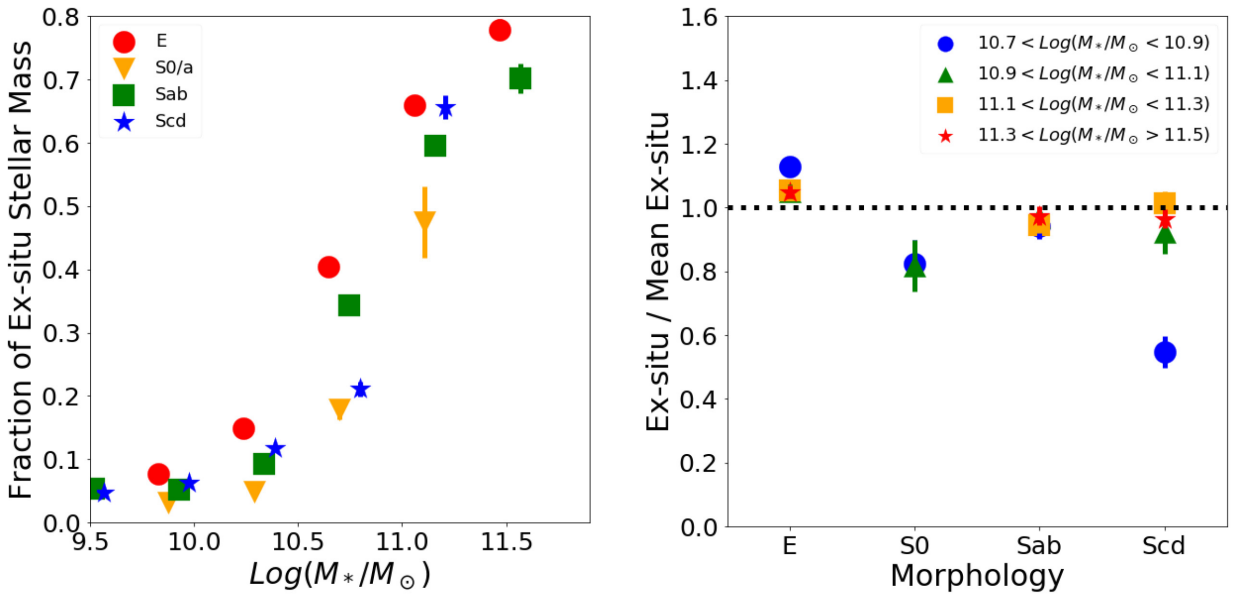

Figure 17. Fraction of the stellar mass formed outside the galaxies as a function of stellar mass and morphology. The left-hand panel shows the fraction as function of stellar mass at fixed morphological type. The right-hand panel shows the relative fraction of *ex situ* mass for a fixed morphological type as compared to the mean *ex situ* mass in a given stellar mass bin. The dotted line indicates a relative fraction of 1 for reference. Error bars are errors on the median values. Only points with more than five galaxies are plotted.

is consistent with the results of Rodriguez-Gomez et al. (2017). However, S0 galaxies seem to exhibit lower *ex situ* fraction than all other types at all masses. Although this might be a consequence of low statistics, it appears to be a systematic trend at all masses.

We expand on these trends in the right-hand panel. *Ex situ* fractions are plotted at fixed stellar mass for different morphologies, by focusing on the mass regime where the *ex situ* contribution is non-negligible. The panel shows the relative excess of *ex situ* mass fraction of a given morphological type as compared to the average *ex situ* mass in a given stellar mass bin. The curves show a weak trend of average *ex situ* fraction with morphology in bins of stellar mass, with a strong under abundance of *ex situ* mass for massive S0 galaxies (again possibly due to low statistics, as revealed by the SMFs of Section 7). At the highest mass end ($>10^{11.3}$), the differences in *ex situ* mass fractions across morphological types are very small (curves are essentially flat). This result does not mean that massive elliptical galaxies are not formed through mergers. Indeed the high mass end of the SMF in TNG is populated by a significant fraction of ellipticals which are likely formed through mergers given the large fraction of *ex situ* mass. However, it looks like a significant fraction of mergers do not produce early-type galaxies. Instead, the final

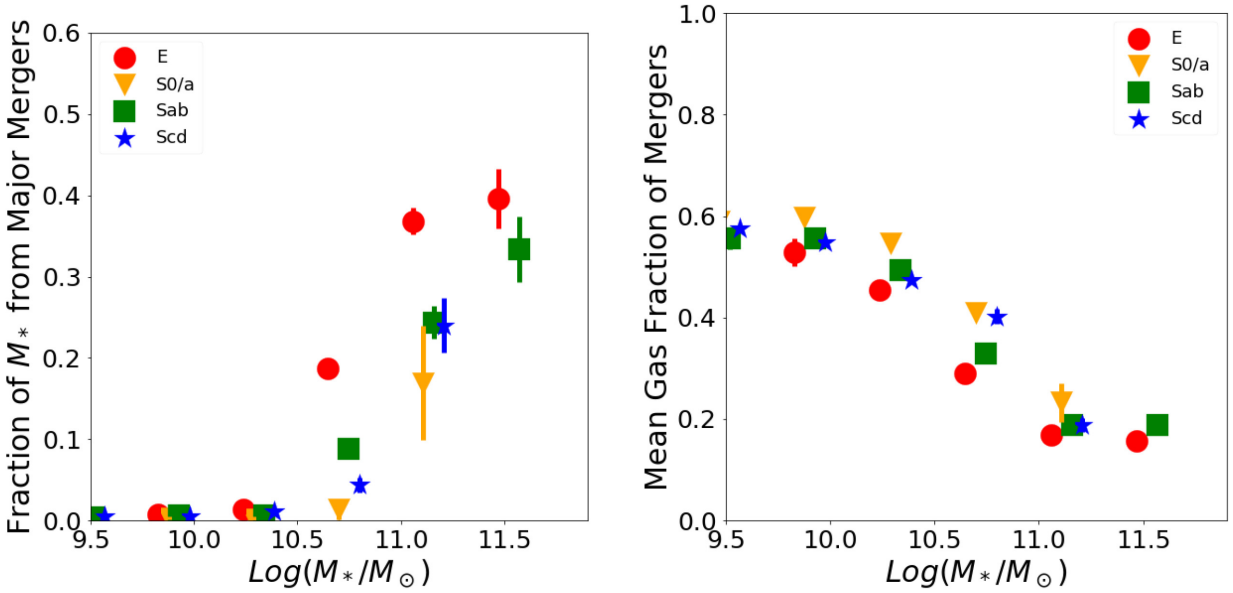

Figure 18. Left-hand panel: Fraction of the stellar mass formed through major mergers as a function of stellar mass and morphology. Right-hand panel: Mean gas fraction integrated over all the merger events as a function of morphology and stellar mass. Error bars are errors on the median values. Only points with more than five galaxies are plotted.

morphologies are also Sab galaxies which therefore dominate the high mass end of the SMF. The figures show that there are no strong differences in the assembly histories of different morphologies, suggesting that subtle differences in assembly histories may be responsible for changes in the morphological type. If the merger history is similar, what determines that a massive galaxy will end up as Sab or Elliptical? The answer might be in the properties of the mergers and accretion events.

In Fig. 18 we plot the fraction of mass coming from major mergers only (left-hand panel) as well as the median gas fraction involved in the mergers (right-hand panel). Consistently with the *ex situ* mass fraction findings, below $10^{10.5}$ $M_\odot$ galaxies mostly accrete stars via minor mergers. The relative contributions of major merges is again a strong function of galaxy mass above $\sim 10^{11}$ solar masses. At the high mass end, the figure suggests that the merger mass ratio might be a relevant factor in determining the final morphology in the simulation. Elliptical galaxies tend to have a larger fraction of stellar mass coming from major mergers than early-type spirals (0.4 versus 0.2). Surprisingly there is little difference in terms of gas fraction, although elliptical galaxies tend to be formed in slightly dryer mergers, whereas S0 in more gas-rich events.

The tentative picture that seems to emerge is that at the low-mass end ($< 10^{10.5}$ $M_\odot$) the assembly history has very little to null impact in setting galaxy morphologies. At the massive end ($> 10^{10.5}$ $M_\odot$), if the accretion and merger histories contribute to determine galaxy morphologies, their manifestations are subtle. The amount and types of mergers do manifest differently for different morphological types, but this is the case only for those galaxy masses where there is significant *ex situ* contribution and with relatively weak trends. Furthermore, in our model, a larger fraction of major mergers will tend to form an elliptical galaxy. In at least half the cases however (recall that the massive end of the SMF is populated by a significant amount of Sab galaxies in TNG) the galaxies will end up with a disky morphology even if ∼50 per cent of their stellar mass is accreted. It is unclear if this is because the mergers are not big enough to destroy the disc or because there is still a fairly large amount of available gas that is re-accreted. In fact, it is likely that morphological transformations are also associated with the nature and strength of the feedback mechanisms in place, particularly the feedback from the central supermassive black holes, which may act in conjunction with galaxy mergers to set galaxy morphologies. We postpone to future work the investigation of the relation between morphological types and feedback history.

A possible resulting effect of what is seen thus far seems to be that the TNG simulations lack of an efficient way to form lenticular galaxies. It seems that either the disc is fully destroyed or it remains too important and featured to be considered an S0. A possible explanation could be that S0s are preferentially formed in high-density environments such as clusters. Some observational works have shown that there is a larger fraction of S0 galaxies in clusters as compared to the field (e.g. Mei et al. 2009; Huertas-Company et al. 2013). Since the TNG volume is relatively small ($\sim 100\,\mathrm{Mpc}^3$), the number of haloes at the cluster scale is small (about 10 haloes more massive than 10^{14} $M_\odot$ in total mass) and so the environmental effects may not be well represented in comparison to the Universe's demographics. This could be investigated with the larger TNG300 volume, however at the expenses of resolution. Here we point out that a quick inspection of the fraction of different morphological types as a function of halo mass in the SDSS sample shows that S0 and Sab galaxies live in very similar environments. Elliptical galaxies do tend to live in denser environments but their number densities are better reproduced. It is therefore unlikely that the discrepancy we measure can be fully explained by environmental considerations.

In order to better understand the origin of these discrepancies especially at the high mass end we explore in Fig. 17 some tracers of the assembly histories of the different morphologies in the simulation. In particular, we first look at the contribution of mergers in the assembly. The left-hand panel of Fig. 17 shows the fraction of stellar mass formed outside the galaxies as a function of stellar mass (*ex situ* stellar mass). This should be a proxy of the importance of mergers in the assembly histories. Interestingly we see very mild

dependence with the morphological type. The trend seems to be essentially driven by stellar mass. Below $\sim 10^{10.5}$ solar masses, the amount of accreted stellar mass is negligible for all morphologies. Fig. 14 shows that the majority of low mass galaxies in TNG have a bulge component (*S0* and *Sab* galaxies account for $\sim$ 80 per cent of the number densities). The bulges in these systems must have grown through internal processes. Above $\sim 10^{10.5}$ solar masses, the amount of accreted stellar mass starts to be significant and reaches almost 80 per cent at $10^{11.5}$ solar masses. However this trend seems to be still pretty independent of the morphological type. We do observe a slight tendency for a larger *ex situ* fraction for elliptical galaxies as one would expect (60 per cent versus 40 per cent), but all morphologies remain consistent at the 1σ level. This is roughly consistent with the results of Rodriguez-Gomez et al. (2017). The right-hand panel confirms this trend. *Ex situ* fractions are plotted at fixed stellar mass for different morphologies. The curves are mostly flat. There seems to be a decrease for massive S0 galaxies but this is likely due to low statistics as revealed by the SMFs of Section 7. Notice that this result does not mean that massive elliptical galaxies are not formed through mergers. Indeed the high mass end of the SMF in TNG is populated by a significant fraction of ellipticals which are likely formed through mergers given the large fraction of *ex situ* mass. However, it looks like mergers are also efficient in producing Sab galaxies. The figures show that there are no strong obvious differences in the assembly histories of different morphologies, suggesting that subtle differences only change the morphological type. If the merger history is similar, what determines that a massive galaxy will end up as Sab or Elliptical? The answer might be in the properties of the merger. In Fig. 18 we plot the fraction of mass coming from major mergers (mass ratio larger than 0.25) only (left-hand panel) as well as the median gas fraction involved in the mergers (right-hand panel). Although the trends are still very mild, the figure suggests that the merger mass ratio might be a relevant factor in determining the final morphology in the simulation. Elliptical galaxies tend to have a larger fraction of stellar mass coming from major mergers than early-type spirals (0.4 versus 0.2). Surprisingly there is little difference in terms of gas fraction although elliptical galaxies tend to be formed in slightly dryer mergers. The tentative picture that seems to emerge is that at the massive end, all galaxies seem to have a comparable contribution of merger to their assembly history. However, a larger fraction of major mergers will tend to form an elliptical galaxy. In the majority of the cases though (recall that the massive end of the SMF is dominated by Sab galaxies in TNG) the galaxies will end up with a disky morphology even if $\sim$ 50 per cent of their stellar mass is accreted. It is unclear if this is because the mergers are not big enough to destroy the disc or because there is still a fairly large amount of available gas that is re-accreted. The resulting effect seems to be that the simulations lack of an efficient way to form lenticular galaxies. It seems that either the disc is fully destroyed or it remains too important and thin to be considered an S0. A possible explanation could be that S0s are preferentially formed in high-density environments such as clusters. Some observational works have shown that there is a larger fraction of S0 galaxies in clusters as compared to the field (e.g. Mei et al. 2009; Huertas-Company et al. 2013). Since the TNG volume is relatively small ($\sim 100\,\mathrm{Mpc}^3$), the number of haloes at the cluster scale is small and so the environmental effects are not well incorporated. A quick inspection of the fraction of different morphological types as a function of halo mass in the SDSS shows that S0 and Sab galaxies live in very similar environments. Elliptical galaxies do tend to live in denser environments but their number densities are well reproduced. It is therefore unlikely that the discrepancy we measure can be fully explained by environmental considerations.

9 SUMMARY AND CONCLUSIONS

We have analysed the visual morphologies of galaxies at $z \sim 0$ in the IllustrisTNG simulation. We have trained a CNN on detailed visual morphologies estimated on 14 000 galaxies in the SDSS and applied the same network to classify a complete sample of 12 000 galaxies in TNG with stellar mass larger than $10^{9.5}$ solar masses. In order to produce images with similar properties than in the observations, the output of the simulations was post-processed with a radiative transfer code to create realistic mock observations with realistic instrumental effects. Our morphological classes include early-type galaxies, in turn divided in ellipticals E and lenticulars S0/a, and late-type galaxies, in turn classified as early-type spirals (Sab) and late-type spirals (Scd), the latter including irregulars.

Our main results are as follows:

(i) The TNG simulation reproduces well the diversity of morphologies in the local universe. A CNN trained on the SDSS is able to find galaxies in different morphological types in the simulation with comparable uncertainty. Even if some differences might exist, it means that the main features learned by the networks at the SDSS resolution are present both in the simulations and in the observations. An analysis of these features shows indeed that they cluster similarly. However, the TNG suite shows a weak correlation between optical morphology and Sersic index as opposed to observed galaxies. This discrepancy should be further investigated using exactly the same fitting methods on both datasets.

(ii) The mass–size relations of simulated galaxies reproduce well the slope and the normalization of the observed relations for all morphological types. This includes the global trends for the early- and late-type populations, but also when galaxies are divided in finer morphological classes. The scatter at fixed stellar mass remains slightly larger in the simulations. This is a significant improvement as compared with the original Illustris run.

(iii) We measure some discrepancies in the SMFs divided by morphological type especially at the high-mass end. The high-mass end of the SMF in the simulation presents an overabundance of late-type systems as compared to SDSS ($\sim$ 50 per cent of galaxies more massive than 10^{11} solar masses are found to be late-type in TNG as opposed to $\sim$ 20 per cent in SDSS). We show that this is due to a lack of lenticular galaxies in the TNG simulation at intermediate masses and ellipticals at the high-mass end, probably because there is still too much available gas to build discs.

(iv) At the low-mass end ($< 10^{10.5}\,\mathrm{M}_\odot$) the merger histories of galaxies have no manifest impact in setting morphologies There is a small (but statistically significant) difference between the merger histories of massive galaxies with different morphologies. At $\log(M_*/\mathrm{M}_\odot) \sim 11$, the fraction of accreted mass through mergers in elliptical galaxies is $\sim$ 10 per cent larger than in spiral galaxies. The nature of the mergers is also different. Elliptical galaxies have on average $\sim$ 20 per centBerna more stellar mass coming from major mergers. However, the influence of assembly history in setting the other morphological types remains unclear.

This work has shown that there is potentially interesting information to be learned by consistently comparing simulations and observations in the same observational frame. In particular, detailed morphologies which can be now estimated with high accuracy can

provide additional constraints on the physical processes driving galaxy assembly and help improving the next generation of simulations. In future work we will extend this analysis to high redshift using the high-resolution TNG50 simulations (Nelson et al. 2019; Pillepich et al. 2019) and *Hubble Space Telescope* imaging. We also plan to explore generative models as a more general way to confront models of galaxy formation and observations.

ACKNOWLEDGEMENTS

The authors thank the anonymous referee for his/her very constructive report which helped improving the work. MHC acknowledges support from the Google Faculty Award delivered to Prof. Joel Primack which enabled the first discussions about this project at the University of California Santa Cruz.

REFERENCES

Baes M., Verstappen J., De Looze I., Fritz J., Saftly W., Vidal Pérez E., Stalevski M., Valcke S., 2011, ApJS, 196, 22
Bernardi M., Meert A., Sheth R. K., Vikram V., Huertas-Company M., Mei S., Shankar F., 2013, MNRAS, 436, 697
Bernardi M., Meert A., Vikram V., Huertas-Company M., Mei S., Shankar F., Sheth R. K., 2014, MNRAS, 443, 874
Bernardi M. et al., 2018, MNRAS, 475, 757
Bernardi M., Domínguez Sánchez H., Brownstein J. R., Drory N., Sheth R. K., 2019, preprint (arXiv:1904.11996)
Bottrell C., Torrey P., Simard L., Ellison S. L., 2017a, MNRAS, 467, 1033
Bottrell C., Torrey P., Simard L., Ellison S. L., 2017b, MNRAS, 467, 2879
Bruzual G., Charlot S., 2003, MNRAS, 344, 1000
Bundy K. et al., 2015, ApJ, 798, 7
Chabrier G., 2003, PASP, 115, 763
Clauwens B., Schaye J., Franx M., Bower R. G., 2018, MNRAS, 478, 3994
Correa C. A., Schaye J., Clauwens B., Bower R. G., Crain R. A., Schaller M., Theuns T., Thob A. C. R., 2017, MNRAS, 472, L45
Dickinson H. et al., 2018, ApJ, 853, 194
Dieleman S., Willett K. W., Dambre J., 2015, MNRAS, 450, 1441
Domínguez Sánchez H., Huertas-Company M., Bernardi M., Tuccillo D., Fischer J. L., 2018, MNRAS, 476, 3661
Dubois Y., Volonteri M., Silk J., Devriendt J., Slyz A., Teyssier R., 2015, MNRAS, 452, 1502
Elagali A., Lagos C. D. P., Wong O. I., Staveley-Smith L., Trayford J. W., Schaller M., Yuan T., Abadi M. G., 2018, MNRAS, 481, 2951
Emsellem E. et al., 2007, MNRAS, 379, 401
Gal Y., Ghahramani Z., 2015, preprint (arXiv:1506.02142)
Genel S. et al., 2014, MNRAS, 445, 175
Genel S., Fall S. M., Hernquist L., Vogelsberger M., Snyder G. F., Rodriguez-Gomez V., Sijacki D., Springel V., 2015, ApJ, 804, L40
Genel S. et al., 2018, MNRAS, 474, 3976
Groves B., Dopita M. A., Sutherland R. S., Kewley L. J., Fischera J., Leitherer C., Brandl B., van Breugel W., 2008, ApJS, 176, 438
Hinton G. E., Srivastava N., Krizhevsky A., Sutskever I., Salakhutdinov R. R., 2012, preprint (arXiv:1207.0580)
Huertas-Company M., Aguerri J. A. L., Bernardi M., Mei S., Sánchez Almeida J., 2011, A&A, 525, A157
Huertas-Company M. et al., 2013, MNRAS, 428, 1715
Huertas-Company M. et al., 2015, ApJS, 221, 8
Lintott C. et al., 2011, MNRAS, 410, 166
Marinacci F. et al., 2018, MNRAS, 480, 5113
Meert A., Vikram V., Bernardi M., 2015, MNRAS, 446, 3943
Mei S. et al., 2009, ApJ, 690, 42
Mendel J. T. et al., 2015, ApJ, 804, L4
Naiman J. P. et al., 2018, MNRAS, 477, 1206
Nair P. B., Abraham R. G., 2010, ApJS, 186, 427
Nelson D. et al., 2018, MNRAS, 475, 624
Nelson D . et al., 2019, preprint (arXiv:1902.05554)
Pakmor R., Bauer A., Springel V., 2011, MNRAS, 418, 1392
Pakmor R., Springel V., Bauer A., Mocz P., Munoz D. J., Ohlmann S. T., Schaal K., Zhu C., 2016, MNRAS, 455, 1134
Pillepich A . et al., 2018a, MNRAS, 473, 4077
Pillepich A . et al., 2018b, MNRAS, 475, 648
Pillepich A . et al., 2019, preprint (arXiv:1902.05553)
Rodriguez-Gomez V. et al., 2016, MNRAS, 458, 2371
Rodriguez-Gomez V. et al., 2017, MNRAS, 467, 3083
Rodriguez-Gomez V. et al., 2019, MNRAS, 483, 4140
Rosito M. S., Tissera P. B., Pedrosa S. E., Rosas-Guevara Y., 2018, preprint (arXiv:1811.11062)
Sales L. V., Navarro J. F., Theuns T., Schaye J., White S. D. M., Frenk C. S., Crain R. A., Dalla Vecchia C., 2012, MNRAS, 423, 1544
Schaye J. et al., 2015, MNRAS, 446, 521
Sijacki D., Vogelsberger M., Genel S., Springel V., Torrey P., Snyder G. F., Nelson D., Hernquist L., 2015, MNRAS, 452, 575
Silla C., Freitas A., 2011, Data Min. Knowl. Discov., 22, 31
Simard L., Mendel J. T., Patton D. R., Ellison S. L., McConnachie A. W., 2011, ApJS, 196, 11
Snyder G. F. et al., 2015, MNRAS, 454, 1886
Springel V., 2010, ARA&A, 48, 391
Springel V. et al., 2018, MNRAS, 475, 676
Strauss M. A. et al., 2002, AJ, 124, 1810
Sundararajan M., Taly A., Yan Q., 2017, preprint (arXiv:1703.01365)
Thob A. C. R. et al., 2019, MNRAS, 485, 972
Torrey P. et al., 2015, MNRAS, 447, 2753
Trayford J. W. et al., 2017, MNRAS, 470, 771
Trayford J. W., Frenk C. S., Theuns T., Schaye J., Correa C., 2019, MNRAS, 483, 744
van der Maaten L. J. P., Hinton G. E., 2008, J. Mach. Learn. Res., 9, 2579
Vogelsberger M . et al., 2014, Nature, 509, 177
Vogelsberger M. et al., 2014, MNRAS, 444, 1518
Weinberger R. et al., 2017, MNRAS, 465, 3291

SUPPORTING INFORMATION

Supplementary data are available at *MNRAS* online.

TNG_morph_MHC19_MNRAS

Please note: Oxford University Press is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

This paper has been typeset from a TEX/LATEX file prepared by the author.