DRAFT VERSION JANUARY 29, 2018
Typeset using LaTeX **twocolumn** style in AASTeX61

arXiv:1801.08541v1 [astro-ph.GA] 25 Jan 2018

GALAXY ZOO: MORPHOLOGICAL CLASSIFICATION OF GALAXY IMAGES FROM THE *ILLUSTRIS* SIMULATION

HUGH DICKINSON,[1] LUCY FORTSON,[1] CHRIS LINTOTT,[2] CLAUDIA SCARLATA,[1] KYLE WILLETT,[1] STEVEN BAMFORD,[3] MELANIE BECK,[1] CAROLIN CARDAMONE,[4] MELANIE GALLOWAY,[1] BROOKE SIMMONS,[5,*] WILLIAM KEEL,[6] SANDOR KRUK,[2] KAREN MASTERS,[7] MARK VOGELSBERGER,[8] PAUL TORREY,[8,†] AND GREGORY F. SNYDER[9]

[1]*School of Physics and Astronomy, University of Minnesota, 116 Church Street SE, Minneapolis, MN 55455, USA*
[2]*Oxford Astrophysics, University of Oxford, Denys Wilkinson Building, Keble Road, Oxford, OX1 3RH, UK*
[3]*School of Physics and Astronomy, The University of Nottingham, University Park, Nottingham, NG7 2RD, UK*
[4]*Department of Mathematics and Science, Wheelock College, Boston, MA 02215, USA*
[5]*Center for Astrophysics and Space Sciences, Department of Physics, University of California, San Diego, CA 92093, USA*
[6]*Department of Physics and Astronomy, University of Alabama, Box 870324, Tuscaloosa, AL 35487, USA*
[7]*Institute for Cosmology and Gravitation, University of Portsmouth, Dennis Sciama Building, Burnaby Road, Portsmouth, PO1 3FX, UK*
[8]*Department of Physics, Kavli Institute for Astrophysics and Space Research, Massachusetts Institute of Technology, Cambridge, MA 02139, USA*
[9]*Space Telescope Science Institute, 3700 San Martin Drive, Baltimore, MD 21218, USA*

ABSTRACT

Modern large-scale cosmological simulations model the universe with increasing sophistication and at higher spatial and temporal resolutions. These ongoing enhancements permit increasingly detailed comparisons between the simulation outputs and real observational data. Recent projects such as *Illustris* are capable of producing simulated images that are designed to be comparable to those obtained from local surveys. This paper tests the degree to which *Illustris* achieves this goal across a diverse population of galaxies using visual morphologies derived from Galaxy Zoo citizen scientists. Morphological classifications provided by these volunteers for simulated galaxies are compared with similar data for a compatible sample of images drawn from the Sloan Digital Sky Survey (SDSS) Legacy Survey. This paper investigates how simple morphological characterization by human volunteers asked to distinguish smooth from featured systems differs between simulated and real galaxy images. Significant differences are identified, which are most likely due to the limited resolution of the simulation, but which could be revealing real differences in the dynamical evolution of populations of galaxies in the real and model universes. Specifically, for stellar masses $M_\star \lesssim 10^{11} M_\odot$, a substantially larger proportion of *Illustris* galaxies that exhibit disk-like morphology or visible substructure, relative to their SDSS counterparts. Toward higher masses, the visual morphologies for simulated and observed galaxies converge and exhibit similar distributions. The stellar mass threshold indicated by this divergent behavior confirms recent works using parametric measures of morphology from *Illustris* simulated images. When $M_\star \gtrsim 10^{11} M_\odot$, the *Illustris* dataset contains substantially fewer galaxies that classifiers regard as unambiguously featured. In combination, these results suggest that comparison between the detailed properties of observed and simulated galaxies, even when limited to reasonably massive systems, may be misleading.

Corresponding author: Hugh Dickinson
* Einstein Fellow
† Hubble Fellow

1. INTRODUCTION

As large-scale simulations of the universe increase in size and in resolution, increasingly sophisticated comparisons with observations are becoming more feasible. While early work concentrated on matching features of the universe captured by simple parameterizations such as the mass function or scaling relations (e.g. Kauffmann et al. 1993; Cole et al. 1994), modern cosmological simulations produce galaxies with apparently realistic star formation histories, substructures, and colors (e.g. Genel et al. 2014; Crain et al. 2015; Kaviraj et al. 2017). The prospect of "observing" this simulated universe via the creation of artificial images offers the chance to test any such simulation's fidelity, and any discrepancies may provide new insights on the physics that drives galaxy formation and evolution.

The obvious comparison for simulations that model the present-day galaxy population is the Sloan Digital Sky Survey (SDSS; York et al. (2000); Strauss et al. (2002)), which has provided a wealth of information about a large number of local systems (see Strateva et al. 2001; Kauffmann et al. 2003; Tremonti et al. 2004; Brinchmann et al. 2004; Baldry et al. 2004, for just some of the most highly cited results). The SDSS augments its galaxy catalogs with a rich suite of spectral, photometric, and instrumental metadata. In particular, the availability of estimated galaxy redshifts and stellar masses is critical for our analysis.

Modern simulations such as *Illustris* (Vogelsberger et al. 2014a,b; Genel et al. 2014; Sijacki et al. 2015) have been used to construct simulated versions of the SDSS (Torrey et al. 2015), and comparisons between observed and simulated universes have utilized a large range of parameters derived from observations (Snyder et al. 2015; Bottrell et al. 2017a,b). However, much insight can still be gained by relying on morphological classification of galaxy images. Morphology is a sensitive probe of a galaxy's dynamical and star formation histories, and such classifications have been shown to reflect differences between systems that are often difficult to recover from purely parametric approaches (e.g. Bamford et al. 2009; Schawinski et al. 2009; Masters et al. 2010a), and have also helped to unveil previously unnoticed trends and behaviors (e.g. Schawinski et al. 2010; Masters et al. 2011; Simmons et al. 2013; Casteels et al. 2013; Galloway et al. 2015; Smethurst et al. 2016; Kaviraj 2014).

This paper uses visual morphological classifications as a metric for comparison between simulated and observed universes. Using calibrated citizen science data from the Galaxy Zoo project (Lintott et al. 2008; Willett et al. 2013), we provide non-parametric labels for a large number of simulated galaxies and compare these to SDSS galaxies labeled in the same way. In this manner, we aim to investigate the degree to which large cosmological simulations, and specifically *Illustris*, can claim to match the present-day galaxy population.

2. DATA

2.1. *The Illustris Sample*

Illustris is a suite of large volume, cosmological hydrodynamical simulations run with the moving-mesh code Arepo (Springel 2010; Genel et al. 2014). It includes a comprehensive set of physical models that are deemed critical for modeling the formation and evolution of galaxies across cosmic time. Galaxy formation processes in *Illustris* are simulated following the models described by Vogelsberger et al. (2013) and Torrey et al. (2014). Each of the *Illustris* simulations encompasses a volume of $106.5\ \mathrm{Mpc}^3$ and self-consistently evolves five different types of resolution element (dark matter particles, gas cells, passive gas tracers, particles that represent stars and their stellar winds, and supermassive black holes) from a starting redshift of $z = 127$ to the present day, $z = 0$. The *Illustris* simulation suite successfully reproduces a range of well established galaxy scaling relations. It implements a unique combination of high-resolution and total simulation volume, which provides an ideal test dataset for our purposes.

The *Illustris* image sample is generated using an ensemble of 6891 unique subhaloes that had assembled within the *Illustris* simulation volume by $z = 0$. Each subhalo is assumed to represent a single galaxy. These were chosen to have $M_\star \gtrsim 10^{10} M_\odot$[1], which corresponds to a typical number of stellar particles $\gtrsim 10^5$. Simulated galaxies comprised of fewer particles were deemed unlikely to accurately represent morphological features of interest (e.g. Torrey et al. 2015), and were therefore excluded from our sample.

We use images from Torrey et al. (2015), which have been processed as described in Snyder et al. (2015) to produce 'observationally realistic' images. This process produces synthetic *Illustris* images that are square arrays with side length 424 pixels, with a typical angular pixel scale $0\farcs05 - 0\farcs10$ per pixel. For each image, the precise pixel scaling is adjusted to ensure that the central $\frac{2}{3}$ of each subject image corresponds to twice the simulated galaxy's projected Petrosian radius. This scaling emulates the approach used to generate the original Galaxy Zoo 2 subject images. Each image is convolved with a nominal PSF with Full Width at Half Maximum (FWHM) $\sim 1\farcs0$, which is similar to the $\sim 1\farcs4$ average seeing for the SDSS DR7; the two sets of images should be broadly comparable. It should be noted that these images represent a simulation of galaxies that have evolved until redshift zero, but projected as if they lie at $z = 0.05$. We expect little evolution in the galaxy population between $z = 0.05$ and the present, and so this displacement should not significantly affect the comparison we wish to make. Observational evidence also indicates that galaxy populations in the real universe exhibit little evolution in this redshift interval (e.g. Rudnick et al. 2003; Blanton et al. 2003).

[1] *Illustris* generates several definitions of the stellar mass for each simulated galaxy. Throughout this paper, we use the *total* stellar mass, labeled as `mass_stars` in the *Illustris* catalog.

Images of each galaxy were generated for four orientations that model observation from the separate vertices of a tetrahedron with the subhalo at the center (the tetrahedron is oriented with respect to the simulation and so randomly relative to the galaxy). Backgrounds are randomly selected from real SDSS images. The 'target' galaxy is assumed to be in the foreground and in rare cases may be superimposed over systems that are actually closer than the projected distance of the simulated galaxy ($z = 0.05$). Four separate backgrounds for each galaxy were used to mitigate this and other systematic effects. The final sample that is potentially available for classification therefore comprises a total of 16 images per subhalo, making a total of 110,256 distinct subjects.

2.2. *The SDSS Sample*

To provide a valid comparison for the *Illustris* sample, described in §2.1, we begin by selecting SDSS galaxies with $M_\star > 10^{10}\ M_\odot$ and with redshifts between $z = 0.045$ and $z = 0.055$.

The left-hand panel of Figure 1 shows the stellar mass[2] distributions of the raw, redshift-selected SDSS and *Illustris* datasets. The distributions are obviously mismatched due to a combination of the a-priori galaxy mass selection applied to the *Illustris* sample and incomplete sampling of faint, low-mass galaxies in the SDSS.

Within the narrow redshift range spanned by our SDSS sample, the inferred stellar mass provides a good proxy for galactic size and luminosity, which are both likely to influence the observability of morphological features. We therefore use bootstrap resampling to construct a final SDSS sample with a mass distribution that matches the *Illustris* sample that was ultimately classified (see §3). The SDSS sample is drawn from 100 bins, equally separated in log-mass space. The right-hand panel of Figure 1 illustrates the resulting distribution in $M_\star$ of our bootstrap-resampled SDSS dataset. This dataset contains 7159 entries, of which 5556 are unique. Among those remaining images that are sampled repeatedly, the vast majority are pairs; very few images appear more than twice.

For reference, Figure 2 compares mass-matched, but otherwise randomly selected images from the *Illustris* and SDSS subject sets.

2.3. *Predictable differences between the Illustris and SDSS images*

Several assumptions and simplifications were adopted when generating synthetic galaxy images based on the *Illustris* simulation data. Accordingly, some predictable differences between simulated and real images are inevitable, and we outline the most significant of these here. Intrinsic dust reddening was not considered when generating synthetic images based upon the simulated *Illustris* galaxy structures. Dust formation occurs in dense molecular clouds, which are not fully resolved at the ~ 1 kpc spatial resolution that *Illustris* achieves, so modeling of the dust within simulated galaxies requires augmentation of the simulation output with a number of *ad-hoc* assumptions[3]. In contrast, the three-dimensional positions of the *Illustris* galaxies' stellar populations are directly resolved by the simulation. Accordingly, synthetic images that omit dust modeling provide a faithful representation of the raw simulation output, which ultimately simplifies inference of the performance of *Illustris* using visual classification data. Nonetheless, dust obscuration is known to be significant for some local galaxies (e.g. Masters et al. 2010b), and this omission is manifested in Figure 3 as clear mismatches between the distributions of absolute magnitude for the five SDSS filters (u, g, r, i, z) between the *Illustris* and *resampled* SDSS samples that worsens for increasingly blue filters.

In addition, Snyder et al. (2015) note that the sizes of simulated and real galaxies (measured by the half-mass radius for *Illustris* and Petrosian 50% radius for SDSS) are comparable at masses of 10^{11} and above, but at lower $M_\star$ the *Illustris* galaxies are comparatively more extended. The discrepancy amounts to a factor of two at a mass of $10^{10} M_\odot$.

3. GALAXY ZOO CLASSIFICATION INFRASTRUCTURE

Galaxy Zoo is a set of citizen science projects that have collectively engaged hundreds of thousands of volunteers in the classification of galaxy images drawn from large ground-based surveys and from those conducted by the *Hubble Space Telescope* (Lintott et al. 2008; Fortson et al. 2012). Such classifications have been shown to be a good match to expert classifications (Lintott et al. 2008; Willett et al. 2013; Simmons et al. 2017; Willett et al. 2017). Moreover, the degree of consistency between the classifications provided by multiple volunteers for the same galaxy image provides a measure of the precision of their aggregate classification.

Classification of a galaxy image in Galaxy Zoo entails answering a series of questions, each evaluating a particular aspect of a galaxy's morphological appearance. The earliest questions segregate the subject set into broad morphological categories before subsequent questions investigate increasingly intricate aspects of a galaxy's appearance. The full question set is subjected to hierarchical filtering such that questions are only asked if they remain pertinent following earlier responses. Accordingly, sampling becomes increasingly sparse for questions that appear later in the classification hierarchy and the degree of statistical uncertainty associated with each subject's consensus response increases. For this project, *Illustris* images were classified via a decision tree emulating the tree used for the Galaxy Zoo 2 project and described in Willett et al. (2013).

[2] Stellar masses for the SDSS galaxies were derived from the `P97P5` column of the MPA-JHU catalog (Brinchmann et al. 2004).

[3] Trayford et al. (2015) showed how different modeling assumptions pertaining to dust obscuration affect the inferred observational colours of simulated galaxies in the EAGLE simulation.

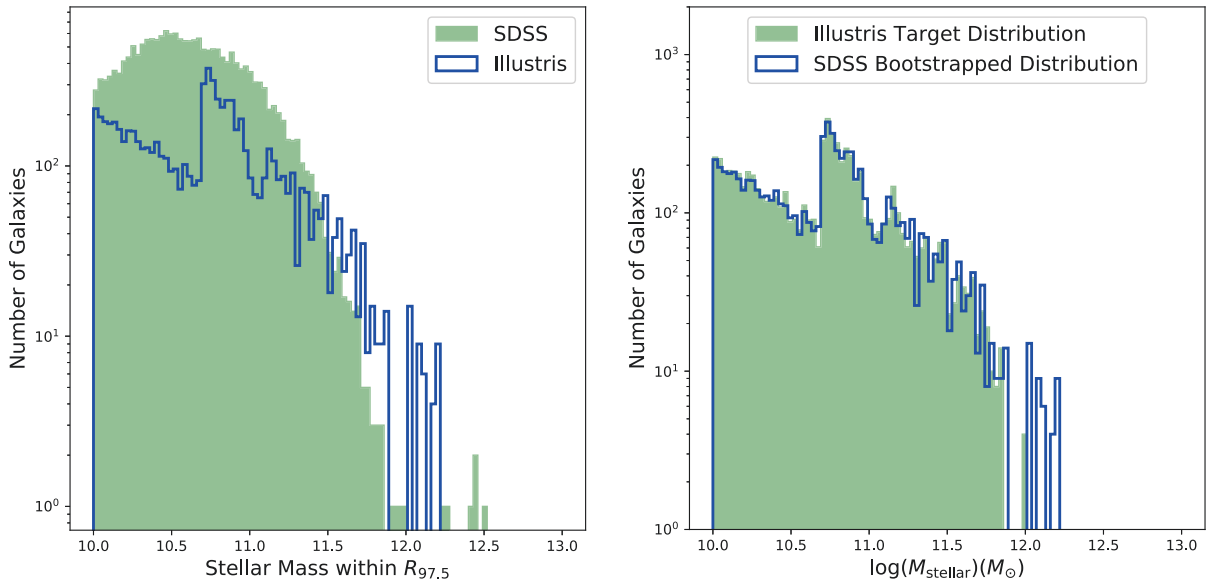

Figure 1. Raw (left) and resampled (right) stellar mass distributions for the *Illustris* (blue hollow) and SDSS (green filled) datasets. Distributions are shown for the inferred stellar mass within 97.5% (left) of the galaxy's Petrosian radius.

In Galaxy Zoo, each galaxy image is classified by at least forty[4] nominally independent volunteers. The individual responses to each question are then aggregated to yield an overall consensus classification. For questions that require a binary response, the availability of multiple independent responses permits the aggregate classification to be encapsulated as a real-valued vote fraction, which is evaluated as the ratio of the number of positive (or negative) responses to the total number of responses.

The *Illustris* classifications used for this study were accumulated via the Galaxy Zoo web-based interface between 2015 September and 2017 August. During this interval, 164,627 volunteers contributed 814,283 morphological assessments for 20248 distinct galaxy images. Classification began with an initial subject set comprising 17046 images for simulated galaxies with stellar masses $10 \leq \log_{10}(M_\star/M_\odot) \leq 13$ The initial sample was designed to facilitate the assessment of potential systematic biases that were anticipated but were not ultimately evident during analysis. To isolate the effect of background and viewing angle on morphological classification, a subset of 10832 images were derived from 677 distinct subhaloes that were selected by uniform random sampling from within two narrow ranges of total **halo** mass $10.5 \leq \log_{10}(M_{\rm halo}/M_\odot) \leq 11$, $12.5 \leq \log_{10}(M_{\rm halo}/M_\odot) \leq 13$. Each subhalo was imaged from the four directions corresponding with the vertices of a regular tetrahedron and superimposed over four randomly selected background images per vertex, as described in §2.1. The remaining 6214 images sample the complementary ranges of halo mass, facilitating mass-independent morphological comparison with observed SDSS galaxies. Each synthetic image in this subset corresponds to a distinct subhalo, viewed from a single, randomly selected viewing angle and superimposed over a single randomly selected background.. To enhance the sample of classifications for the most massive *Illustris* galaxies, the initial set was subsequently augmented with 3202 additional images for which the corresponding stellar masses exceeded $10^{10.5} M_\odot$.

For our SDSS sample, we use data from Galaxy Zoo 2 (Willett et al. 2013), which provides detailed morphological classifications of nearly 250,000 galaxies drawn from the 7th SDSS data release (Abazajian et al. 2009). The subset of the SDSS used for Galaxy Zoo 2 is described by Willett et al. (2013) and was further subsampled to provide a comparison dataset for the *Illustris* images and their corresponding morphologies.

4. RESULTS

We identify discrepancies between the Galaxy Zoo classifications that were obtained for the *Illustris* dataset and those obtained for a redshift- and mass-matched sample of SDSS galaxies by comparing the distributions of vote fractions obtained for each sample. For this investigation, we concentrate on the first, most fundamental question in the Galaxy Zoo 2 decision tree, which distinguished galaxies with features - predominately disk-dominated systems - from those where no such features are apparent. Even this crude distinction reflects significant differences in the underlying dynamical and star formation history of a galaxy, which dictate its visual morphology. Accordingly, it is an excellent test of the realism of the images produced by the *Illustris* simulation.

[4] The mean number of classifications per subject is 40.2.

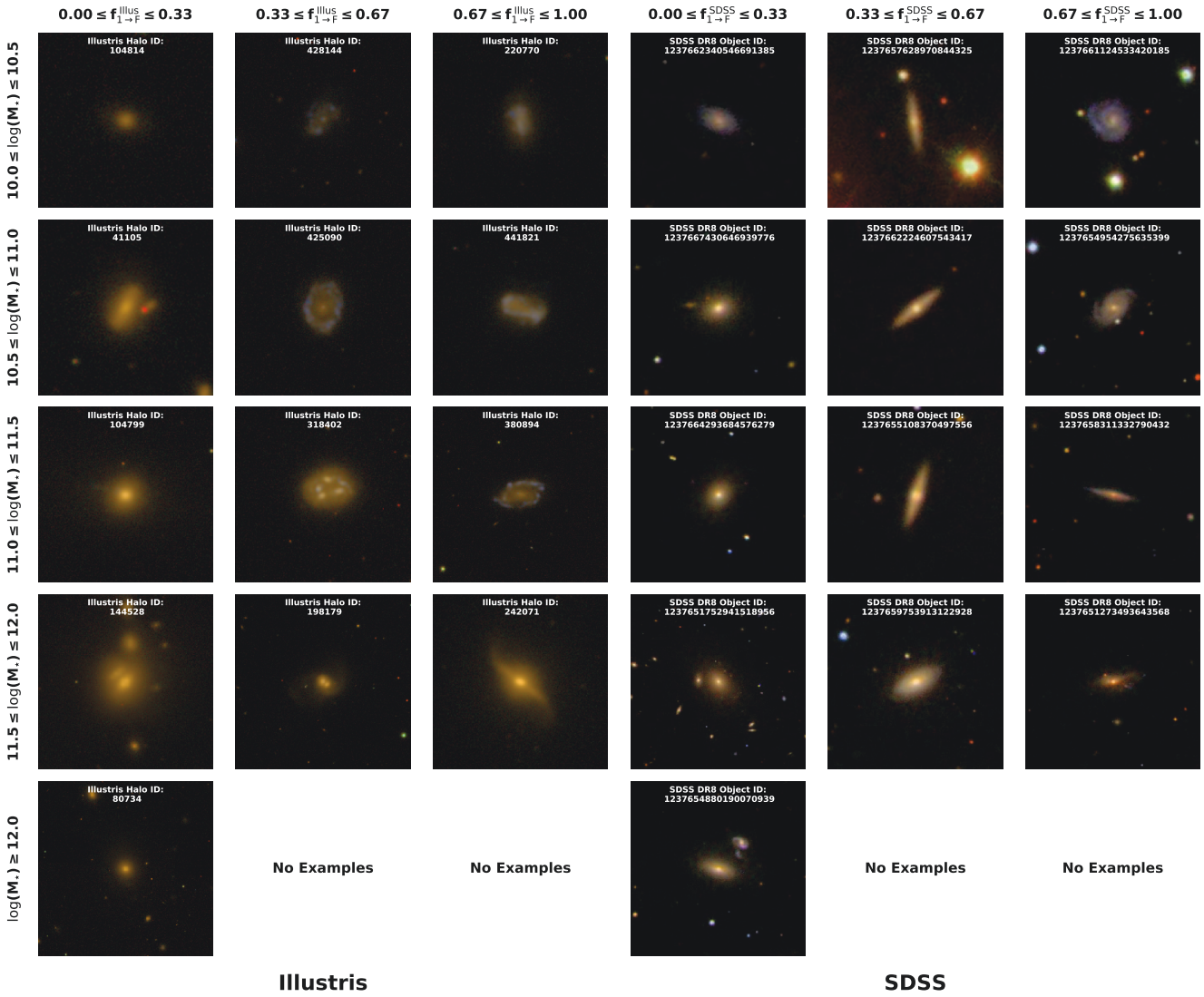

Figure 2. Comparison between mass-matched *Illustris* (three left-hand columns) and SDSS (three right-hand columns) subject images. Each row shows a triplet of galaxies drawn from broad mass bins for each survey. Listing from the top row to the bottom, the chosen mass bins correspond to $10 \leq \log(M_\star/M_\odot) \leq 10.5$, $10.5 \leq \log(M_\star/M_\odot) \leq 11.0$, $11.0 \leq \log(M_\star/M_\odot) \leq 11.5$, $11.5 \leq \log(M_\star/M_\odot) \leq 12$, and $12 \leq \log(M_\star/M_\odot) \leq 50$

Figure 4 illustrates the unweighted[5] vote fraction distributions for the response "disk or features" to the question "Is the galaxy simply smooth and rounded, with no sign of a disk?" (hereafter $f_{1\rightarrow \mathrm{F}}$) for the *Illustris* and SDSS samples[6]. Consequently, a high value of $f_{1\rightarrow \mathrm{F}}$ implies that the imaged galaxy probably has features, while $f_{1\rightarrow \mathrm{F}} \rightarrow 0$ implies the converse. A surprisingly marked disparity is evident. The SDSS galaxies show a broadly bimodal distribution, with many (visibly featureless) systems clustered around low featured vote fractions, and a smaller number of systems that have high vote fractions. The SDSS distribution arises primarily from genuine morphological separation between elliptical and spiral systems but is augmented at low $f_{1\rightarrow \mathrm{F}}^{\mathrm{SDSS}}$ by

[5] Previous analysis of Galaxy Zoo 2 has used a weighting system, which downweights highly inconsistent classifications; as the population of classifiers has changed between the original GZ2 run and classifications of *Illustris* simulated galaxies, introducing such a weighting here would introduce a new systematic difference between the samples. For most systems, the weighting makes little difference in practice. Therefore, we choose to use unweighted vote fractions to avoid even the possibility of introducing a systematic difference between the samples.

[6] In addition to the nominal positive and negative responses, a third option, which labels the putative galaxy as an "artifact" is also possible. All votes for "artifact" were discarded when computing the vote fractions we present in this paper. We verified that omitting artifact votes from our analysis does not qualitatively affect our results.

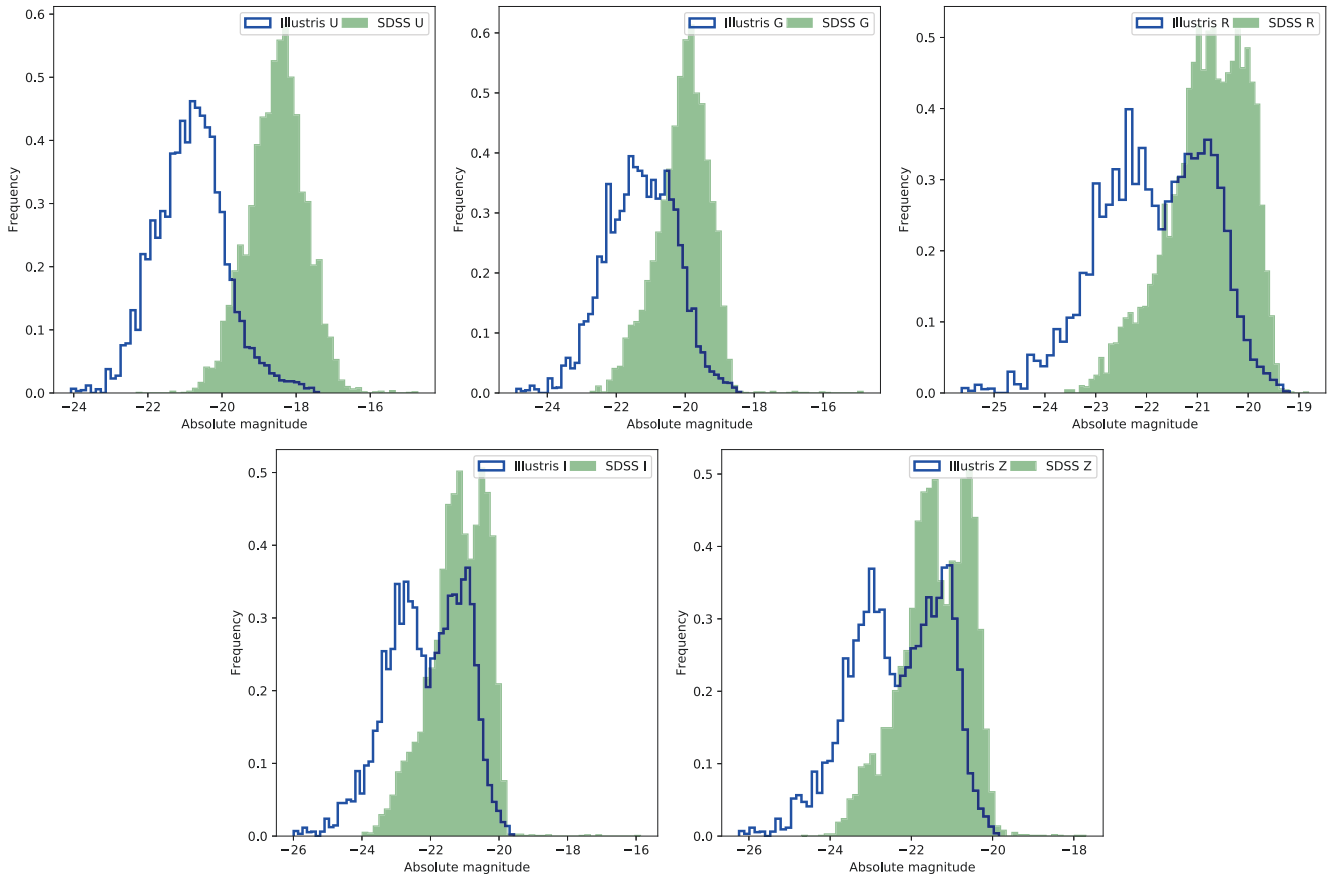

Figure 3. Illustration of the mismatch between the distributions of absolute magnitude for the 5 SDSS filters (u, g, r, i, z) for the *Illustris* and *resampled* SDSS datasets.

galaxies that would exhibit features but are too faint for any intrinsic substructure to be visible in subject images.

The *Illustris* sample, by contrast, is characterized by a prevalence of galaxies with visible substructure, which is evident in Figure 4 as a dominant peak around a modal vote fraction of around 0.6. It is clear from even this simple comparison that there are significant differences between the two samples.

In Figure 5 we subdivide the *Illustris* and SDSS samples into disjoint subsamples according to galaxy stellar mass, $M_\star$. For $10 \leq \log(M_\star/M_\odot) \leq 10.5$, the mismatch between the distributions of $f_{1\rightarrow \mathrm{F}}$ that was evident for the full range of galaxy masses is qualitatively reproduced. For subsamples that correspond to higher stellar masses, the $f_{1\rightarrow \mathrm{F}}$ distributions become increasingly similar, and for $M_\star \gtrsim 10^{11}\mathrm{M}_\odot$, we see a significant fraction of galaxies in the *Illustris* sample with low vote fractions as expected from SDSS observations.

We verified that the observed overabundance of featured galaxies in *Illustris* is not an artifact of viewing angle by individually analyzing four subsets of images corresponding to the distinct vertices of the tetrahedral imaging structure described in §2.1 and verifying that qualitatively similar vote fraction distributions are obtained. We also verified that the observed dependence on $M_\star$ is preserved for each subset of the data.

The other notable difference between the two samples is manifested for $M_\star \geq 10^{10.5}\mathrm{M}_\odot$ as a significant subset of SDSS galaxies with very high featured vote fractions ($f_{1\rightarrow \mathrm{F}} \gtrsim 0.85$). A population of galaxies that almost all classifiers identify as spiral in the SDSS is either missing in the simulated universe or classified differently in the *Illustris* sample. Figure 6 shows representative samples of galaxy images drawn from the mismatching region of $(M_\star - f_{1\rightarrow \mathrm{F}})$ parameter space for the *Illustris* (left-hand columns) and SDSS (right-hand columns) datasets. While *Illustris* does produce a population of featured galaxies with $M_\star \geq 10^{10.5}\mathrm{M}_\odot$, the SDSS image sample appears to include a larger fraction of nearby grand design spirals that the majority of volunteers would classify as obviously featured. In contrast, the *Illustris* galaxy images appear slightly more ambiguous, with less prominent disks, and it seems plausible that the apparent deficiency of galaxies that are unanimously perceived as featured reflects this ambiguity.

The intentional omission of dust modeling when generating the synthetic *Illustris* images (see §2.3) is another factor that likely contributes to the mismatched visual classi-

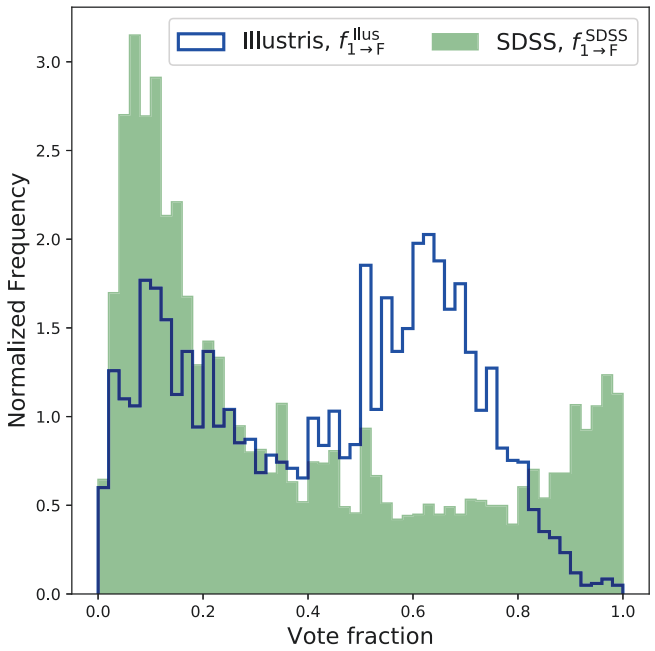

Figure 4. Comparison between the normalized distributions for $f_{1\to F}^{Illus}$ and $f_{1\to F}^{SDSS}$ corresponding to the full *Illustris* and SDSS samples, respectively. A high value of $f_{1\to F}$ implies that the majority of volunteers discerned discrete substructure in the galaxy image, while $f_{1\to F} \to 0$ implies the converse. While the SDSS distribution is dominated by systems with low $f_{1\to F}^{SDSS}$, the Illustris sample apparently contains many more galaxies that exhibit visible substructure and yield more intermediate vote fractions.

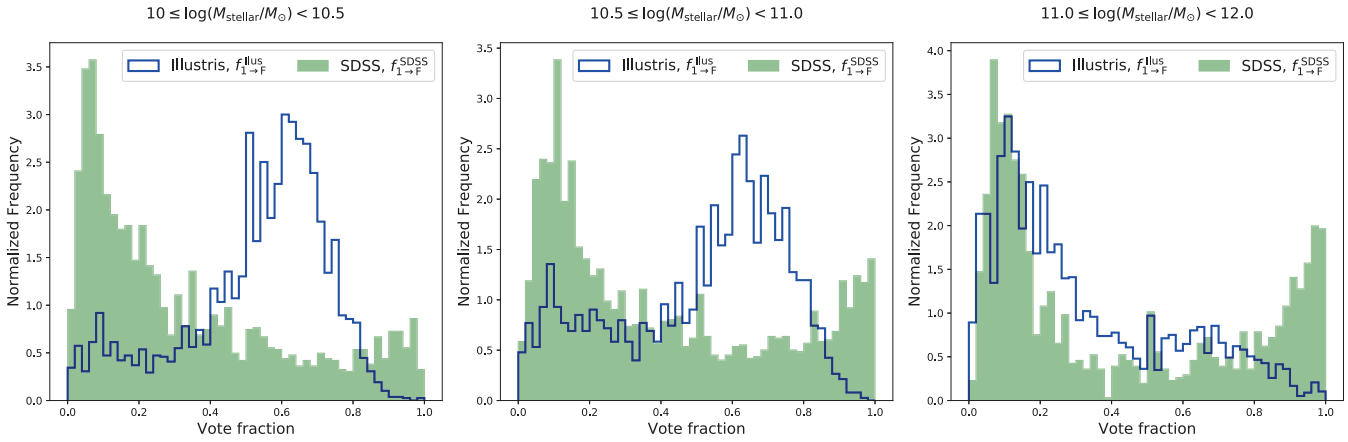

Figure 5. The $f_{1\to F}$ vote fractions in intervals of $\log(M_\star/M_\odot)$. Proper interpretation of $f_{1\to F}$ is explained in the main text as well as in the caption of Figure 4. Below $\log(M_\star/M_\odot) \sim 11$, the SDSS and *Illustris* $f_{1\to F}$ distributions match very poorly. At higher masses, overall agreement between the distributions is substantially improved, albeit with a residual discrepancy between the numbers of obviously featured galaxies.

fications. To illustrate how intrinsic dust extinction affects the classifications that are gathered for *real* galaxy images, Figure 7 plots featured vote fraction distributions for disjoint subsets of the SDSS sample that were segregated based upon the observed axial ratio $(B/A)_{SDSS}$ between the projected semi-minor (B) and semi-major (A) axes of each galaxy[7]. Remarkable differences between the four distributions are

[7] The values for A and B correspond to those listed in the SDSS DR7 calatog for the *exponential* or de Vaucouleurs profile model that provided the best fit to each galaxy's light distribution.

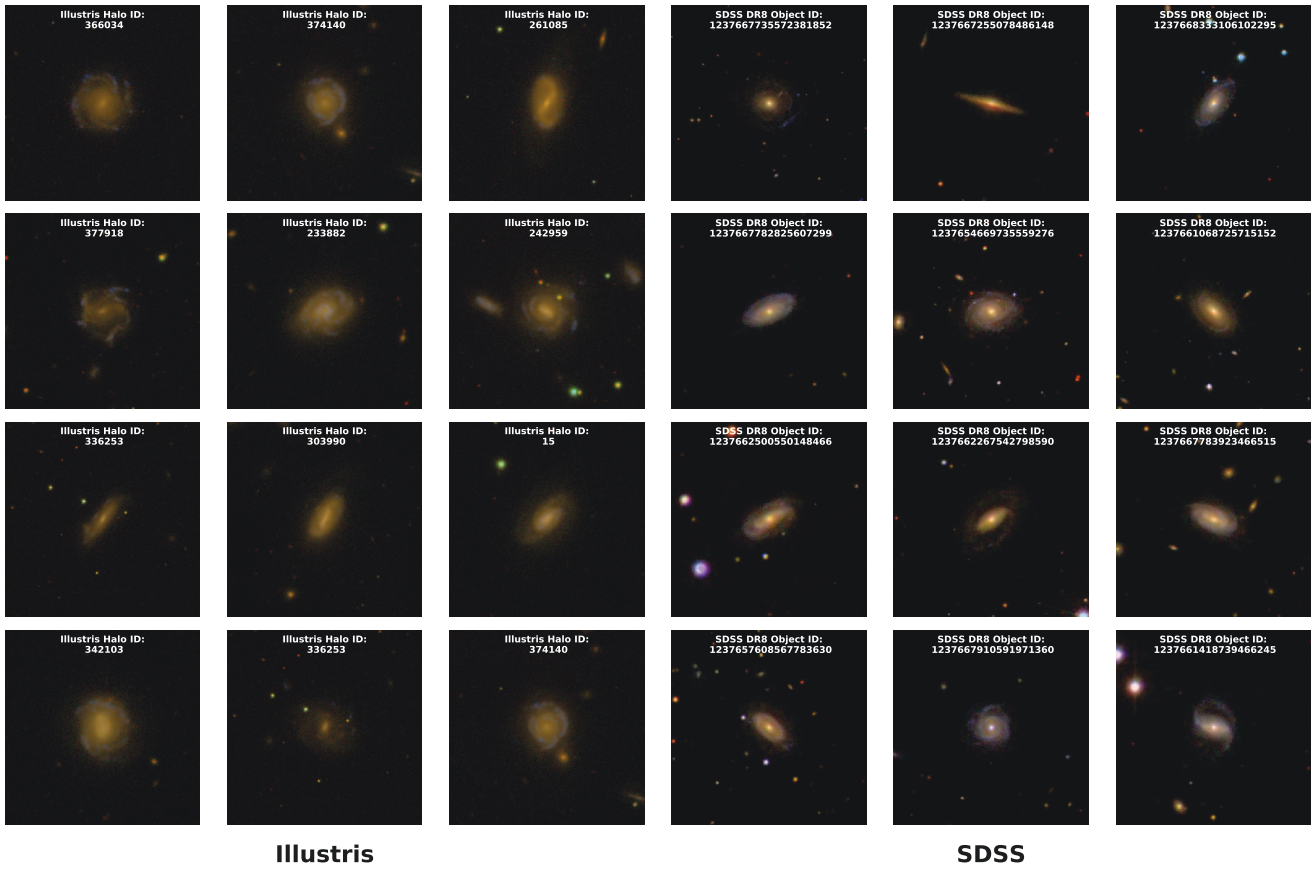

Figure 6. Example images of *Illustris* (three left-hand columns) and SDSS (three right-hand columns) galaxies with $f_{1\rightarrow \mathrm{F}} > 0.85$ and $\log(M_\star/M_\odot) > 11$

evident with volunteers labeling many more featured galaxies as the typical axial ratio for each subset increases from zero (edge-on) to unity (face-on).

This phenomenon is likely dual in origin. Intrinsic dust extinction within the target galaxy may obscure discernible features while superimposed substructures along the line of sight may lead them to appear as a single luminous mass. Focusing on structurally disk-like galaxies, small values of $(B/A)_{\mathrm{SDSS}}$ suggest that the target was observed with an edge-on orientation. This configuration increases the probability of discrete substructures occupying nearby sightlines and becoming visually indistinguishable. Moreover, escaping starlight that would reveal such features must traverse a much larger column of dust on average without being absorbed in order to reach the observer. Conversely, as $(B/A)_{\mathrm{SDSS}} \rightarrow 1$, galaxies with face-on orientations predominate and discrete substructures become more visible.

The procedure used to generate the *Illustris* subject images did not model dust extinction, and we show the normalized featured vote fraction distribution for the full *Illustris* sample in all four panels of Figure 7. The *Illustris* and SDSS distributions do not coincide well for *any* of the $(B/A)_{\mathrm{SDSS}}$ ranges considered. For $(B/A)_{\mathrm{SDSS}} \gtrsim 0.25$, the disparity is clearly manifested as an excess of apparently featured galaxies among the *Illustris* sample. It is plausible that the galaxies contributing to this excess would shift to lower $f_{1\rightarrow \mathrm{F}}$ if dust attenuation were properly simulated when preparing the *Illustris* subject images. Such migration might dilute or even eliminate the apparent morphological disparities between the two samples.

5. SUMMARY AND CONCLUSIONS

We have used visual classifications from Galaxy Zoo to compare the coarse morphological appearance of simulated galaxies from the *Illustris* cosmological simulation with those of a population drawn from the Sloan Digital Sky Survey, matched in mass and redshift. This set of visual classifications allows a direct comparison to be made with observations, with any differences indicating potentially missing physics in the simulation, the inevitably limited resolution of such simulations, or the choices made in producing 'observationally realistic' images. In any case, understanding how selection by morphology might influence comparisons between simulation and observation is essential.

Figure 4 reveals two marked disparities between the two samples. The fraction ($f_{1\rightarrow \mathrm{F}}^{\mathrm{Illus}}$) of classifiers who report noticeable features in *Illustris* galaxy images exceeds that for the equivalent quantity ($f_{1\rightarrow \mathrm{F}}^{\mathrm{SDSS}}$) for classifications of SDSS

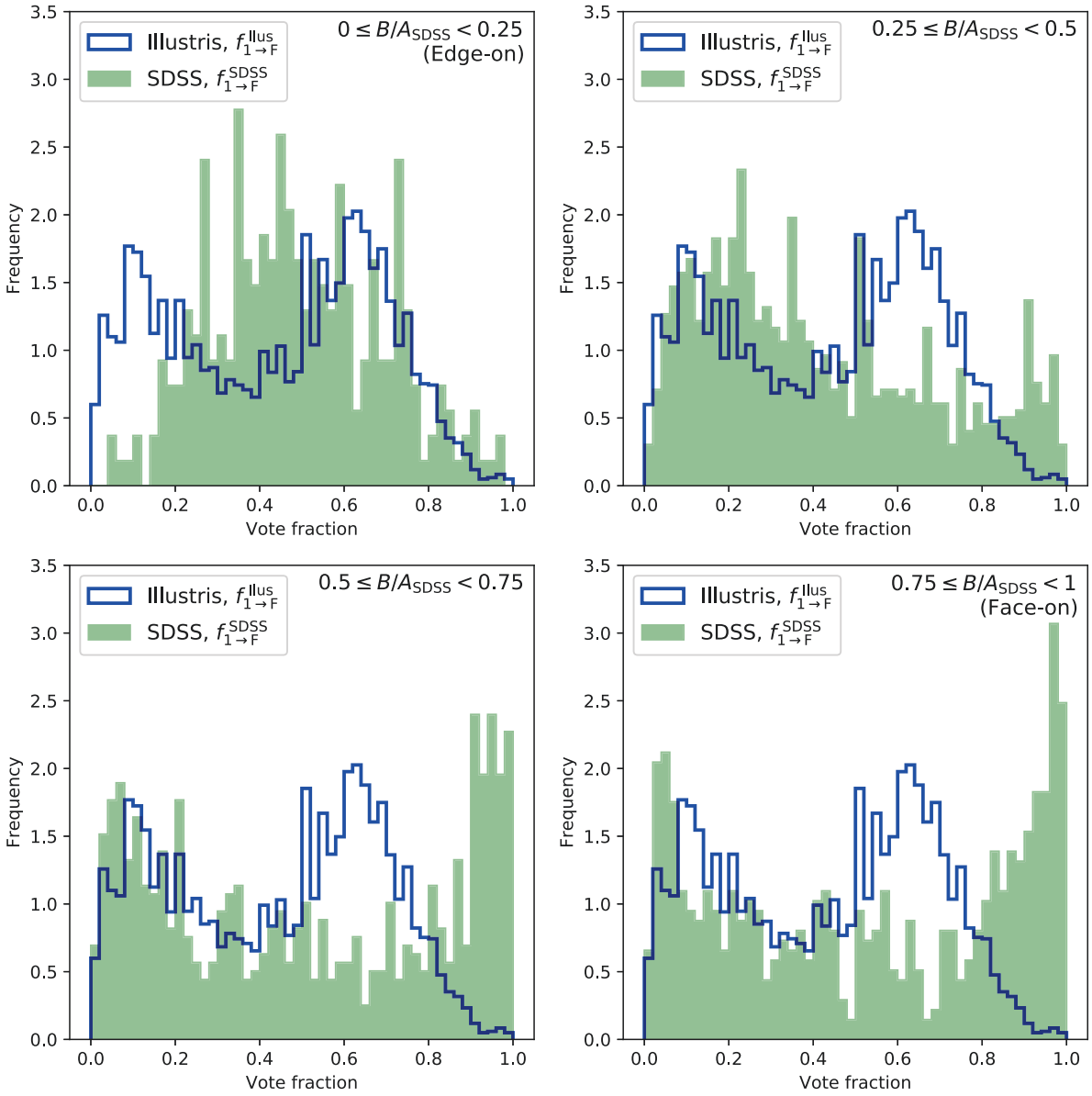

Figure 7. Distributions of $f_{1\to F}^{SDSS}$ for disjoint subsets of the SDSS sample (*green*) that were segregated based upon the observed axial ratio $(B/A)_{SDSS}$ between the projected semiminor (B) and semimajor (A) axes of each galaxy. Proper interpretation of $f_{1\to F}$ is explained in the main text as well as in the caption of Figure 4. Many more galaxies exhibit visible features (attain high $f_{1\to F}^{SDSS}$) as $(B/A)_{SDSS}$ increases from zero to unity. For comparison, the distribution of $f_{1\to F}^{Illus}$ (blue line) for the entire *Illustris* sample is shown in all panels.). For the SDSS distribution shown in the upper-left panel (smallest $(B/A)_{SDSS}$, most edge-on), no galaxies were unambiguously classified as featured or smooth. This indicates that edge-on galaxies may be particularly difficult to separate base on visual inspection.

subjects. Indeed, Figure 5 illustrates that for $\log(M_\star/M_\odot) < 10.5$ the distributions of $f_{1\to F}^{Illus}$ and $f_{1\to F}^{SDSS}$ are almost mirror images of each other. While volunteer classifiers clearly discern features in a large majority of the *Illustris* sample, a far smaller proportion report them for the SDSS galaxy images. There is also a small set of galaxies with high featured vote fractions in SDSS but this population is absent in *Illustris*. While the *Illustris* images are simulated to an observational resolution of $1''$ compared to an achieved average seeing of $1.4''$ for the SDSS, this small difference is unlikely to be responsible for such a large observed difference.

The absence of moderate and high-mass, unambiguously featured galaxies in the *Illustris* sample that was noted in §4 is perhaps the most surprising result. It may represent the response of volunteer classifiers to simulated objects, which, despite the care taken in preparing the images, are often eas-

ily distinguished from their SDSS counterparts. Features such as bright knots, over-prominent arms, and so on are seen in many *Illustris* images. These artifacts are the result of insufficient particle resolution and may confuse classifiers, reducing the consensus on features. Alternatively, it may be that the simulation is failing to producing realistic grand design spirals.

We also see a failure to produce the correct fraction of smooth galaxies. The importance of this mismatch between the *Illustris* and SDSS samples appears to depend strongly on the stellar mass range of the galaxies under consideration. Figure 5 plots analogues of Figure 4 for mass-selected subsets of the *Illustris* and SDSS. It is apparent that the distributions of $f_{1\to\mathrm{F}}^{\mathrm{Illus}}$ and $f_{1\to\mathrm{F}}^{\mathrm{SDSS}}$ become markedly less disparate for **stellar** masses $M_\star > 10^{11} M_\odot$. However, correspondence between the two datasets remains imperfect, and a population of highly featured galaxies that are present in the real universe, but absent in *Illustris* becomes apparent above $M_\star > 10^{10.5} M_\odot$.

The underproduction of unambiguously featured galaxies with large $M_\star$ that we identify in *Illustris* may indicate that accumulation of stellar mass involves simulated processes that also disrupt or destroy spatially discrete substructures. The most massive galaxies in *Illustris* are predominantly formed by the hierarchical assembly of smaller systems (Rodriguez-Gomez et al. 2016). Repeated interactions between simulated galaxies provide a plausible mechanism for suppression of visible features. To investigate this possibility, we searched for indications that the time since the most recent major merging event in a simulated galaxy's history predicts its morphological classification for galaxies with $M_\star > 10^{11} M_\odot$. No compelling correlations were observed. The two-sample Kolmogorov–Smirnoff test yields a p-value of 0.104 when comparing the distributions of the time since the most recent major merging event for subsamples of visually smooth ($f_{1\to\mathrm{F}}^{\mathrm{Illus}} < 0.3$) and featured ($f_{1\to\mathrm{F}}^{\mathrm{Illus}} > 0.85$) galaxies. This is consistent with both subsamples being drawn from the same parent distribution. We also checked for a significant correlation between the fraction of galactic stellar mass that was formed in-situ and the visibility of features in the *Illustris* galaxy images. In this case, the two-sample Kolmogorov–Smirnoff test yields a p-value of 4.1×10^{-7} when comparing the samples of smooth and featured galaxies. This result indicates that $f_{1\to\mathrm{F}}^{\mathrm{Illus}} \geq 0.85$ comprise a larger proportion of stars that were formed in-situ, which is broadly supportive of the hypothesis that visually featured galaxies experienced comparatively fewer interactions during their formation. A more rigorous verification that accumulation of ex-situ stellar mass is indeed responsible for the disruption of visually apparent substructures would require detailed examination of each galaxy's assembly history, which is beyond the scope of this paper.

Given that the ability of a simulation to represent a galaxy depends coarsely on the number of particles used to model it, some mass dependence should be expected; indeed, this is why galaxies with stellar masses less than $10^{10} M_\odot$ were excluded from the study. Such differences have been seen before, in particular by Bottrell et al. (2017a) who showed that a threshold at $M_\star > 10^{11} M_\odot$ also emerges when attempting morphological classification using parametric fits to the galaxy's light profile. Below this critical mass, the simulation produces a large proportion of disk-dominated galaxies; we confirm this result and show that it has a significant effect not only on the parametric measurements but on the overall visual morphology of the system being studied. In some cases, non-parametric morphological metrics for *Illustris* galaxies also appear to differ from those of their physical counterparts when $M_\star \lesssim 10^{11} M_\odot$. For example, Bignone et al. (2017) show that the measured asymmetry of merging *Illustris* galaxies appears artificially large in comparison with mass-matched observational samples. In the same mass range, Snyder et al. (2015) identify a peculiar population of galaxies that exhibit distinctive ring-like structures of enhanced star formation, resulting in unexpectedly extended morphologies (examples of several such systems are included in Figure 2). Snyder et al. (2015) suggest that these ring-like structures may reflect an imperfect model for coupling between feedback mechanisms and the interstellar medium (ISM) in *Illustris* galaxies. Alternatively, the rings of star formation may be an inherent manifestation of the ISM equation of state that is assumed for the *Illustris* simulation. Earlier studies (e.g. Hambleton et al. 2011) compared the properties of simulated galaxy samples with those of locally observed systems using non-parametric morphological estimators. Similar discrepancies pertaining to excessive asymmetry and clumpy substructure were identified.

As in *Illustris* a galaxy's stellar mass broadly maps to the number of stellar particles comprising the simulated galaxy, we conclude that below $10^{11} M_\odot$, the number of stellar particles comprising a galaxy is apparently insufficient to represent the simulated physics reliably, and observed structures are often likely to result from resolution-induced artifacts. The effects are subtle, and the images produced by the simulation are clearly perceived as realistic, but as a population there remain differences between simulated and observed galaxies. These differences complicate more detailed comparisons between the *Illustris* and SDSS galaxy morphologies. Below $M_\star \sim 10^{11} M_\odot$, the coarse morphological differences between observed and simulated galaxies could artificially distort the later stages of classification, because early volunteer responses restrict the set of questions that are subsequently posed. For the most massive galaxies, a limited number of subject images results in excessively sparse sampling of the Galaxy Zoo classification hierarchy that prevents reliable inference of morphological characteristics. Future studies that match SDSS and *Illustris* samples should be aware of the $10^{11} M_\odot$ threshold we have identified and its effects on the comparison being made. We have also shown that insight can be derived from visual analysis of large samples of images derived from simulations and recommend this procedure for future data products.

Acknowledgements: The data in this paper are the result of the efforts of the Galaxy Zoo volunteers, without whom none of this work would be possible. Their efforts are individually acknowledged at authors.galaxyzoo.org. Please contact the author(s) to request access to research materials discussed in this paper.

H.D., L.F., C.S., M.B., and M.G. gratefully acknowledge support from the US National Science Foundation grant AST1716602 (H.D., L.F., C.S. also supported by NSF grant AST1716602).

C.J.L. was supported by STFC under grant ST/N003179/1.

B.D.S. acknowledges support from the National Aeronautics and Space Administration (NASA) through Einstein Postdoctoral Fellowship Award Number PF5-160143 issued by the Chandra X-ray Observatory Center, which is operated by the Smithsonian Astrophysical Observatory for and on behalf of NASA under contract NAS8-03060.

P.T. acknowledges support from NASA through Hubble Fellowship grants HST-HF2-51384.001-A awarded by the STScI, which is operated by the Association of Universities for Research in Astronomy, Inc., for NASA, under contract NAS5-26555.

Funding for the SDSS and SDSS-II has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Science Foundation, the U.S. Department of Energy, the National Aeronautics and Space Administration, the Japanese Monbukagakusho, the Max Planck Society, and the Higher Education Funding Council for England. The SDSS website is http://www.sdss.org/. The SDSS is managed by the Astrophysical Research Consortium for the Participating Institutions. The Participating Institutions are the American Museum of Natural History, Astrophysical Institute Potsdam, University of Basel, University of Cambridge, Case Western Reserve University, University of Chicago, Drexel University, Fermilab, the Institute for Advanced Study, the Japan Participation Group, Johns Hopkins University, the Joint Institute for Nuclear Astrophysics, the Kavli Institute for Particle Astrophysics and Cosmology, the Korean Scientist Group, the Chinese Academy of Sciences (LAMOST), Los Alamos National Laboratory, the Max-Planck-Institute for Astronomy (MPIA), the Max-Planck-Institute for Astrophysics (MPA), New Mexico State University, Ohio State University, University of Pittsburgh, University of Portsmouth, Princeton University, the United States Naval Observatory, and the University of Washington.

REFERENCES

Abazajian, K. N., Adelman-McCarthy, J. K., Agüeros, M. A., et al. 2009, ApJS, 182, 543

Baldry, I. K., Glazebrook, K., Brinkmann, J., et al. 2004, ApJ, 600, 681

Bamford, S. P., Nichol, R. C., Baldry, I. K., et al. 2009, MNRAS, 393, 1324

Bignone, L. A., Tissera, P. B., Sillero, E., et al. 2017, MNRAS, 465, 1106

Blanton, M. R., Hogg, D. W., Bahcall, N. A., et al. 2003, ApJ, 594, 186

Bottrell, C., Torrey, P., Simard, L., & Ellison, S. L. 2017a, MNRAS, arXiv:1701.01451

—. 2017b, MNRAS, 467, 2879

Brinchmann, J., Charlot, S., White, S. D. M., et al. 2004, MNRAS, 351, 1151

Casteels, K. R. V., Bamford, S. P., Skibba, R. A., et al. 2013, MNRAS, 429, 1051

Cole, S., Aragon-Salamanca, A., Frenk, C. S., Navarro, J. F., & Zepf, S. E. 1994, MNRAS, 271, 781

Crain, R. A., Schaye, J., Bower, R. G., et al. 2015, MNRAS, 450, 1937

Fortson, L., Masters, K., Nichol, R., et al. 2012, Galaxy Zoo: Morphological Classification and Citizen Science, ed. M. J. Way, J. D. Scargle, K. M. Ali, & A. N. Srivastava (CRC Press, Taylor & Francis Group), 213–236

Galloway, M. A., Willett, K. W., Fortson, L. F., et al. 2015, MNRAS, 448, 3442

Genel, S., Vogelsberger, M., Springel, V., et al. 2014, MNRAS, 445, 175

Hambleton, K. M., Gibson, B. K., Brook, C. B., et al. 2011, MNRAS, 418, 801

Kauffmann, G., White, S. D. M., & Guiderdoni, B. 1993, MNRAS, 264, 201

Kauffmann, G., Heckman, T. M., Tremonti, C., et al. 2003, MNRAS, 346, 1055

Kaviraj, S. 2014, MNRAS, 440, 2944

Kaviraj, S., Laigle, C., Kimm, T., et al. 2017, MNRAS, 467, 4739

Lintott, C. J., Schawinski, K., Slosar, A., et al. 2008, MNRAS, 389, 1179

Masters, K. L., Mosleh, M., Romer, A. K., et al. 2010a, MNRAS, 405, 783

Masters, K. L., Nichol, R., Bamford, S., et al. 2010b, MNRAS, 404, 792

Masters, K. L., Nichol, R. C., Hoyle, B., et al. 2011, MNRAS, 411, 2026

Rodriguez-Gomez, V., Pillepich, A., Sales, L. V., et al. 2016, MNRAS, 458, 2371

Rudnick, G., Rix, H.-W., Franx, M., et al. 2003, ApJ, 599, 847

Schawinski, K., Lintott, C., Thomas, D., et al. 2009, MNRAS, 396, 818

Schawinski, K., Urry, C. M., Virani, S., et al. 2010, ApJ, 711, 284

Sijacki, D., Vogelsberger, M., Genel, S., et al. 2015, MNRAS, 452, 575

Simmons, B. D., Lintott, C., Schawinski, K., et al. 2013, MNRAS, 429, 2199

Simmons, B. D., Lintott, C., Willett, K. W., et al. 2017, MNRAS, 464, 4420

Smethurst, R. J., Lintott, C. J., Simmons, B. D., et al. 2016, MNRAS, 463, 2986

Snyder, G. F., Torrey, P., Lotz, J. M., et al. 2015, MNRAS, 454, 1886

Springel, V. 2010, MNRAS, 401, 791

Strateva, I., Ivezić, Ž., Knapp, G. R., et al. 2001, AJ, 122, 1861

Strauss, M. A., Weinberg, D. H., Lupton, R. H., et al. 2002, AJ, 124, 1810

Torrey, P., Vogelsberger, M., Genel, S., et al. 2014, MNRAS, 438, 1985

Torrey, P., Snyder, G. F., Vogelsberger, M., et al. 2015, MNRAS, 447, 2753

Trayford, J. W., Theuns, T., Bower, R. G., et al. 2015, MNRAS, 452, 2879

Tremonti, C. A., Heckman, T. M., Kauffmann, G., et al. 2004, ApJ, 613, 898

Vogelsberger, M., Genel, S., Sijacki, D., et al. 2013, MNRAS, 436, 3031

Vogelsberger, M., Genel, S., Springel, V., et al. 2014a, Nature, 509, 177

—. 2014b, MNRAS, 444, 1518

Willett, K. W., Lintott, C. J., Bamford, S. P., et al. 2013, MNRAS, 435, 2835

Willett, K. W., Galloway, M. A., Bamford, S. P., et al. 2017, MNRAS, 464, 4176

York, D. G., Adelman, J., Anderson, Jr., J. E., et al. 2000, AJ, 120, 1579