Mon. Not. R. Astron. Soc. **000**, 000–000 (0000) Printed 15 March 2016 (MN LATEX style file v2.2)

Galaxy Zoo: Morphological Classifications for Galaxies in HST Legacy Imaging

Lead Author and other Galaxy Zoo science team members

*This publication has been made possible by the participation of more than 200,000 volunteers in the Galaxy Zoo project. Their contributions are individually acknowledged at http://authors.galaxyzoo.org/authors.html.

E-mail: lead.author@university.edu

15 March 2016

ABSTRACT

This will be the data release paper for GZ:Hubble. We present the classifications, the methodology for data reduction and corrections for redshift dependent biases in the observed morphologies.

1 INTRODUCTION

Usual due diligence for an intro science paper. Should cover:

(i) Motivation for studying morphology of galaxies
(ii) Particular scientific questions governed by galaxies at $z \sim 1$
(iii) Theoretical predictions for galaxy morphology evolution
(iv) Past imaging and morphology studies at $z \sim 1$
(v) Summary of GZ/citizen science work on galaxy morphology to date

2 SAMPLE AND DATA

2.1 Summary of HST Legacy Survey Imaging

• Hubble ACS imaging for the All-Wavelength Extended Groth Strip International Survey (AEGIS; Davis et al. 2007) covers a strip centered at $\alpha = 14^{\rm h}17^{\rm m}, \delta = +52°30'$. The strip was originally selected due to low extinction and Galactic/zodiacal emission, making it a prime target for multi-wavelength observations by space-based observatories. The ACS images covered 63 separate tiles over a total area of $\sim$ 710 arcmin^2. Images were in two bands, with exposure times of 2300 seconds in F606W (V_{606W}) and 2100 seconds in F814W (I_{814W}). The final mosaic images are dithered to a resolution of 0.03 ″/pixel. For extended objects, the limiting magnitudes of sources in AEGIS are 26.23 (AB) in V_{606W} and 25.61 (AB) in I_{814W}.

• The Great Observatories Origins Deep Survey (GOODS; Giavalisco et al. 2004) covers two well-studied fields in the northern and southern hemispheres: the Hubble Deep Field-North ($\alpha = 12^{\rm h}36^{\rm m}, \delta = +62°14'$) and the Chandra Deep Field-South ($\alpha = 03^{\rm h}32^{\rm m}, \delta = -27°48'$). Data including Hubble ACS images are referred to as GOODS-N and GOODS-S, respectively. ACS imaged the GOODS fields in 4 filters – F435W (B_{435W}), V_{606W}, F775W (I_{775W}), and F850LP (I_{850LP}). The mean exposure times for each epoch vary by band, from 1050–2100 seconds. The B_{435W} images were completed in a single epoch at the beginning of the survey, but the V_{606W}, I_{775W}, and I_{850LP} images were taken in five separate epochs separated by 40–50 days each. The ACS images are dithered to a pixel scale of 0.03 ″/pixel and covers a total area of $\sim$ 320 arcmin^2 (160 arcmin^2 per field). The 5σ limiting magnitudes for extended sources are 25.7 for V_{606W} and 25.0 for I_{775W}.

• The Cosmic Evolution Survey (COSMOS; Scoville et al. 2007) covers an area of $\sim$ 1.8 deg^2 centered at $\alpha = 10^{\rm h}00^{\rm m}28^{\rm s}, \delta = +02°12'21''$. Its location near the celestial equator was designed to enable coverage by ground-based telescopes in both the Northern and Southern Hemispheres, as well as the space-based observatories. The Hubble ACS data from COSMOS consists of 1 orbit with 2028 seconds per pointing in I_{814W}, consisting of 590 total pointings. The image resolution is dithered to 0.05 ″/pixel. The 50% completeness magnitude for a galaxy with a half-light radius of 0″.50 in I_{814W} is 24.7 mag.

• The Galaxy Evolution from Morphologies and SEDS (GEMS; Rix et al. 2004; Caldwell et al. 2008) survey is also centered on the Chandra Deep Field-South. The GEMS data covers $\sim$ 800 arcmin^2, and surrounds the area covered by GOODS-S. Images from ACS in GEMS have 1 orbit per pointing for a total of 63 pointings. The exposure times are 2160 and 2286 seconds in V_{606W} and I_{850LP}, respectively. The image resolution has a pixel scale of 0.03 ″/pixel. The 5σ limiting magnitudes for source detection are 25.7 AB in V_{606W} and 24.2 AB in I_{850LP}.

2.2 Image creation

The GOODS images in GZH use mosaics constructed from both 2-epoch and 5-epoch sets of data.

The filters that Griffith et al. (2012) uses for the colored

Table 1. Summary of Galaxy Zoo: Hubble imaging

Survey	$t_{\rm exp}$ [sec]	Filters	Resolution [″/pix]	Area [arcmin^2]	$N_{\rm galaxies}$
AEGIS	2100−2300	V_{606W} and I_{814W}	0.03	710	8157
COSMOS	2028	I_{814W}	0.05	6480	88530
GEMS	2160−2286	V_{606W} and I_{850LP}	0.03	800	9143
GOODS	1000−2100	B_{435W}, V_{606W}, I_{775W}, I_{850LP}	0.03	320	7336
GOODS-N	–	–	–	–	*2551*
GOODS-S	–	–	–	–	*4785*
total	–	–	–	8310	113166

Table 2. GZH redshifts by survey

	Griffith		3DHST			MUSYC		UltraVISTA	Total	
Survey	spec-z	photo-z	spec-z	grism-z	photo-z	spec-z	photo-z	photo-z	with redshift	in survey
AEGIS	3,656	2,941	12	515	249	0	0	0	7,373	8,507
COSMOS	7,201	77,435	35	358	26	0	0	2,665	85,020	92,808
GEMS	387	628	6	99	40	279	7,564	0	8,745	8,564
GOODS-N	1,947	37	418	1,545	1,381	0	0	0	5,398	12,930
GOODS-S	1,080	4	327	1,348	281	816	1,184	0	5,040	14,284
SDSS	0	0	0	0	0	0	0	0	37,545	51,861
Total	14,271	81,045	798	3,865	1,977	1,095	8,748	2,665	149,121	189,954

images were F606W and F775W for GOODS-N and F606W and F850LP for GOODS-S.

We use different filters for the north and south GOODS fields so that GEMS can be directly compared with GOODS-S (Figure 1).

Fake AGN

Stripe 82

Different treatment of colored noise in COSMOS; creating color gradients with Subaru data and using I_{814W} for illumination map.

FERENGI images.

2.3 Redshifts

We compiled redshifts from a variety of sources to include in the GZH catalog. For each galaxy, the redshift selected is in the `Z_Best` column of the data (see Table 5), its type (spectroscopic: `SPEC_Z`, photometric: `PHOTO_Z`, or grism: `GRISM_Z`) is listed in the column `Z_BEST_TYPE`, and the source catalog (`Griffith` (Griffith et al. 2012), `3DHST` (Momcheva et al. 2015), `MUSYC` (Cardamone et al. 2010), or `UltraVISTA` (Ilbert et al. 2013)) of the redshift is in column `Z_BEST_SOURCE`.

For galaxies which have redshifts from multiple sources, we use the following algorithm to select the `Z_BEST` redshift. We first prioritize spectroscopic redshifts; these are provided in the Griffith, 3DHST, and MUSYC catalogs. If a high quality spec-z exists in the Griffith catalog we use that, else 3DHST, else MUSYC. We show in Figure 2 that over 98% of the the spec-z's are consistent with each other, and therefore the priority order of selection makes no negligible difference. If no spectroscipic redshifts are available, we compare the 1-σ errors of the photometric (Griffith, 3DHST, MUSYC, UltraVISTA) and grism (UltraVISTA) redshifts, and use the redshift with the smallest error. Table 2.2 shows the results of this selection.

2.4 User weighting

The votes of individual users who classified galaxies in GZH are combined to make a vote fraction for each question on the classification tree. Users' votes are weighted slightly (in a method identical to that described in Willett et al. 2013) such that users who frequently disagree with all other users end up having very low weights. The majority of users have weights very close to $w = 1.0$ (**STEVEN: Is this true for GZH - do you have a plot of the distribution of user weights or consistencies we can include here?**.

3 GALAXY ZOO INTERFACE AND CLASSIFICATIONS

4 CORRECTING FOR REDSHIFT-DEPENDENT CLASSIFICATION BIAS

The previous versions of Galaxy Zoo morphology classifications (Lintott et al. 2008; Willett et al. 2013) were based on observations of galaxies in the Sloan Digital Sky Survey (SDSS) which are typically at $z < 0.1$. In these cases it was assumed that there was no cosmological evolution of the morphologies of galaxies and therefore any observed changes in the distribution of galaxies with different consensus morphologies was due to the effects of redshift on the image quality (*i.e.*. the reduction in physical resolution, surface brightness dimming, etc). For both previous releases of GZ morphologies, we provided a correction for redshift-dependent bias based on matching the classification fractions at the highest redshfts with those at the lowest redshift. See Bamford et al. (2009) and Willett et al. (2013) for the details.

In the GZH samples, the redshift range is large enough that we expect to measure cosmological evolution of the types and morphologies of galaxies in the sample. As a result, the previous methods of correcting for redshift dependent bias will not work. In addition, the effects of band shifting will change the images even more across these redshift ranges.

In order to test and correct for the effects of redshift, we generated a set of calibration images. These images consist of the same galaxy as it would appear over a variety of redshifts. The input images are from the SDSS (York et al. 2000; Strauss et al. 2002) and are processed using the FERENGI code (Barden et al. 2008) to match the observational properties of the HST surveys out to $z = 1$. These images

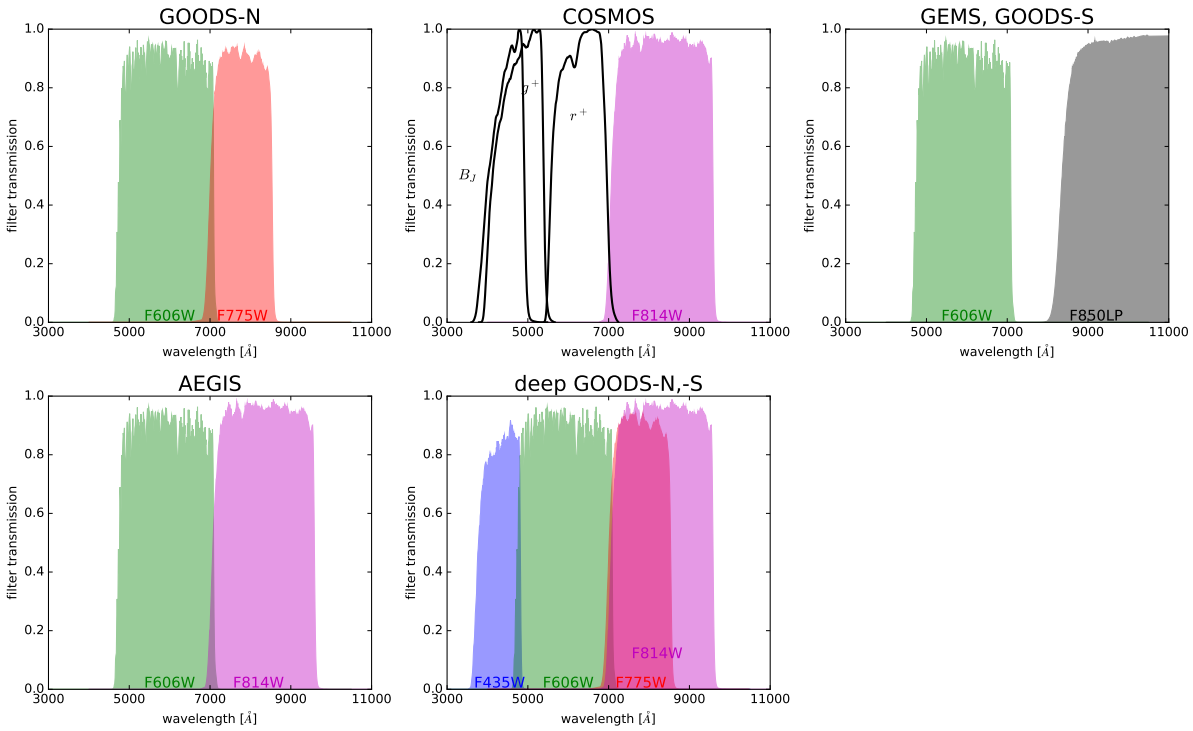


Figure 1. Transmission curves of the filters used by *HST* Advanced Camera for Surveys (ACS) in wide-field channel mode for the various surveys in Galaxy Zoo: Hubble. The unfilled black curves show the filters for the Suprime Camera on the *Subaru* telescope which were used to create color gradients in the composite images for COSMOS.

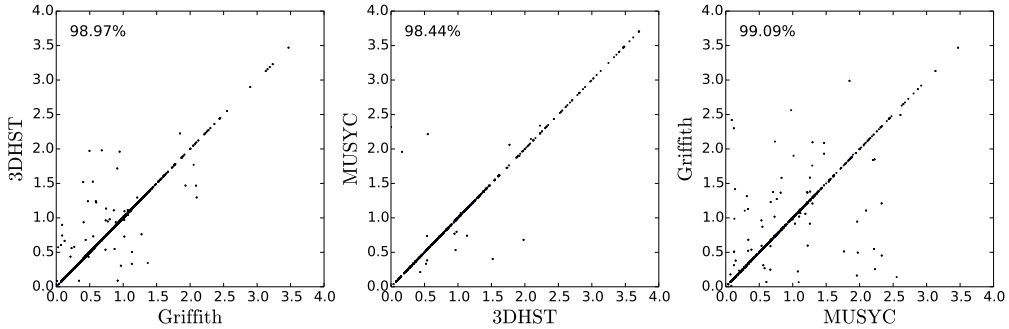


Figure 2. Spectroscopic redshifts from Griffith, 3DHST, and MUSYC catalogs. The number in the upper left of each plot is the percentage of redshifts which agree within $\Delta z < 0.05$ between the two catalogs being compared in each panel. Within this range there is over 98% agreement in redshifts between all three catalogs.

were classified in the Galaxy Zoo interface using the same classification scheme as the original HST images.

4.1 Selection of FERENGI input galaxies

We selected 288 unique galaxies from SDSS imaging to run through the FERENGI code. The selection spanned a variety of galaxy morphologies (as selected by GZ2 classifications) and r'-band surface brightnesses, and also spanned the redshift range of SDSS targets (in $N_z = 4$ bins) in order to be optimised for different target minimum redshifts in HST imaging.

The selection criteria for the different morphological categories is summarised in Table 3. The surface brightness selection ($N_\mu = 3$) was (1) low: $\mu > 21.5$ mag arcsec^{-2}; (2) mid: $20.5 < \mu < 21.5$ mag arcsec^{-2}; and (3) high: $\mu < 20.5$ mag arcsec^{-2}. For each of the four "target redshifts" ($z = 0.3, 0.5, 0.8$ and 1.0), the images were redshifted in $\Delta z = 0.1$ bins up to $z = 1.0$.

In addition to the physical parameters of the input images, the FERENGI output depends on assumptions of the global galaxy evolution model. This evolution is a crude mechanism that mimics the brightness increase of galaxies with increasing redshift (out to at least $z \sim 1-2$). The effect on the redshifted images is simply an empirical addition to the magnitude of a galaxy of the form $M' = e \times z + M$, where M' is the corrected magnitude, and e is the evolu-

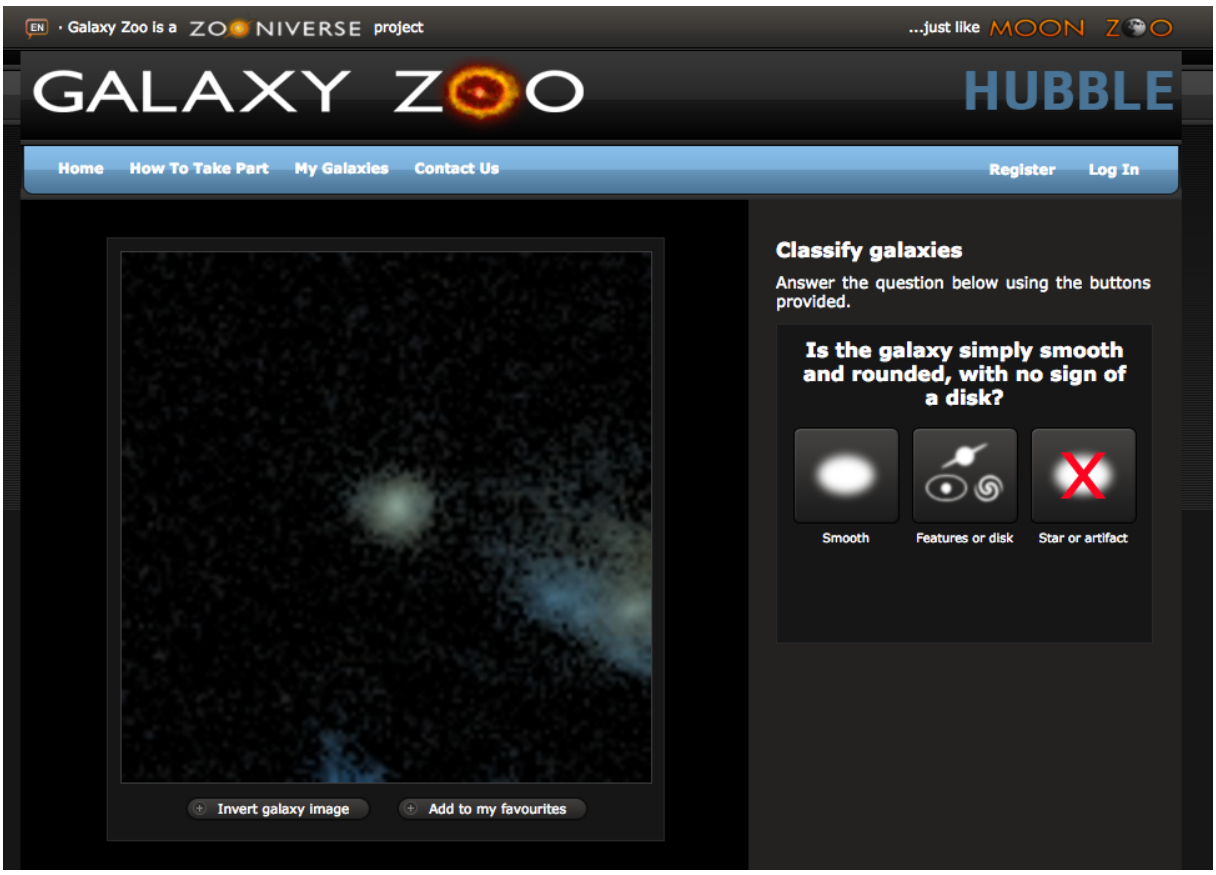


Figure 3. Screenshot of the Galaxy Zoo: Hubble interface showing an example COSMOS image at the first step in the decision tree.

Table 3. Summary of morphological categories selected for FERENGI sample.

Morphology	Label	Selection	$N_{\rm objects}$ $[N_z \times N_\mu]$
Features	Yes	$p_{\rm features} > 0.8$, $p_{\rm odd} < 0.1$	12
	Int.	$0.3 < p_{\rm smooth} < 0.6$, $p_{\rm odd} < 0.1$	12
	No	$p_{\rm smooth} > 0.8$, $p_{\rm odd} < 0.1$	12
Merger	No	$p_{\rm features} > 0.8$, $p_{\rm odd<0.1}$, $p_{\rm merger} < 0.1$	12
	Int.	$p_{\rm odd} > 0.5$, $0.1 < p_{\rm merger} < 0.4$	12
	Yes	$p_{\rm odd} > 0.5$, $p_{\rm merger} > 0.4$	12
Edge-on	Yes	$p_{\rm edgeon} > 0.8$, $p_{\rm features} > 0.5$	12
	Int.	$0.4 < p_{\rm edgeon} < 0.8$, $p_{\rm features} > 0.5$	12
	No	$p_{\rm edgeon} < 0.2$, $p_{\rm features} > 0.5$	12
Bar	No	$p_{\rm bar} < 0.1$, $p_{\rm features} > 0.5$, $p_{\rm edgeon} < 0.2$	24
	Int.	$0.2 < p_{\rm bar} < 0.4$, $p_{\rm features} > 0.5$, $p_{\rm edgeon} < 0.2$	24
	Yes	$p_{\rm bar} > 0.8$, $p_{\rm features} > 0.5$, $p_{\rm edgeon} < 0.2$	24
Visible spiral	No	$p_{\rm spiral} < 0.2$, $p_{\rm features} > 0.5$, $p_{\rm edgeon} < 0.2$, $p_{\rm bar} < 0.1$	12
	Int.	$0.2 < p_{\rm spiral} < 0.8$, $p_{\rm features} > 0.5$, $p_{\rm edgeon} < 0.2$, $p_{\rm bar} < 0.1$	12
	Yes	$p_{\rm spiral} > 0.8$, $p_{\rm features} > 0.5$, $p_{\rm edgeon} < 0.2$, $p_{\rm bar} < 0.1$	12
Oblique bulge size	No	$p_{\rm nobulge>0.6}$, $p_{\rm features} > 0.5$, $p_{\rm edgeon} < 0.5$, $p_{\rm bar} < 0.2$	12
	Int.	$p_{\rm justnoticeable} > 0.6$, $p_{\rm features} > 0.5$, $p_{\rm edgeon} < 0.5$, $p_{\rm bar} < 0.2$	12
	Yes	$p_{\rm obvious\|dominant} > 0.5$, $p_{\rm features} > 0.5$, $p_{\rm edgeon} < 0.5$, $p_{\rm bar} < 0.2$	12
Edge-on bulge shape	Round	$p_{\rm rounded} > 0.5$, $p_{\rm features} > 0.5$, $p_{\rm edgeon} > 0.5$	12
	Boxy	$p_{\rm boxy} > 0.4$, $p_{\rm features} > 0.5$, $p_{\rm edgeon} > 0.2$	12
	No bulge	$p_{\rm nobulge} > 0.5$, $p_{\rm features} > 0.5$, $p_{\rm edgeon} > 0.5$	12

Table 4. Summary of FERENGI artificial redshifting

$z_{\rm target}$	$N_{z\rm bins}$	$N_{\rm evolution}$	$e_{\rm max}$	$N_{\rm galaxies}$	$N_{\rm images}$
0.3	8	7	−3.0	72	4032
0.5	6	4	−1.5	72	1728
0.8	3	3	−1.0	72	648
1.0	1	3	−1.0	72	216

tionary correction in magnitudes (i.e., $e = -1$ essentially brightens the galaxy by 1 magnitude by $z = 1$). We ran FERENGI for values of e starting from $e = 0$ and decreasing to $e = -3.5$ in increments of $\Delta e = 0.5$. Figure 4 shows several examples of the effects of "losing" spiral/disc features with increasing redshift for two galaxies with $e = 0$.

The final number of FERENGI images produced for each galaxy is ultimately a function of galaxy's redshift, since the new images cannot be resampled at better angular resolution than the original SDSS data, as well as the number of e values selected. Table 4 summarizes the total sample of redshifted images produced for GZH.

4.2 Correcting GZH morphologies for classification bias

The approach used in GZH for correcting the weighted classifications for user bias rests on the assumption that the *amount* of bias is a function of the apparent size and brightness of the image as seen on screen. This is controlled by two types of parameters: **intrinsic** properties of the galaxy itself, such as its physical diameter and luminosity, and **extrinsic** properties, such as the distance (redshift) of the galaxy and its relative orientation. There are likely other parameters that affect user accuracy, such as the proximity of close companions ("distraction bias"; see Johnson et al. 2015) or bias as a function of the individual user. The combination of all such parameters forms a high-dimensional space, and we have insufficient data to measure their individual effects. Instead, we use just two parameters that are intended to capture the bulk of the change in bias (based on GZ1/GZ2): a galaxy's r'-band surface brightness (μ_r; intrinsic) and redshift (z; extrinsic).

The change in bias as a function of μ_r and z is measured using the FERENGI images over all the evolutionary correction factors. We assume that the "true" (ie, debiased) vote fraction $f_{\mu,z}$ for a galaxy can be expressed as:

$$f_{\mu,z} = (f_{\mu,z=0.3}) \times e^{\frac{z-z_0}{\zeta}}, \quad (1)$$

where $f_{\mu,z=0.3}$ is the "calibrated" vote fraction at the lowest redshift in the FERENGI bins ($z = 0.3$) and ζ is a positive parameter that controls the rate at which f decreases with increasing redshift. This formula fits the data relatively well (with almost no exceptions, the vote fractions for featured galaxies decrease monotonically with increasing redshift), and the exponential function bounds the observed vote fractions between $f_{\mu,z=0.3}$ and zero. Figure 5 show examples of the change in vote fraction and their fits to Equation 1 for a random selection of galaxies in the FERENGI images.

We use the values of ζ for *all* sets of artifically redshifted galaxies to fit the overall distribution as a function of surface brightness, since we expect the correction being applied to vary as a function of the intrinsic galaxy properties. We restrict the galaxies that can be used to measure the calibration to those with data at the pivot redshift of $z = 0.3$, non-zero $f_{\rm features}$ at $z = 0.3$, and with a reasonable fit to the exponential model ($\Delta\chi^2 > 3.0$).

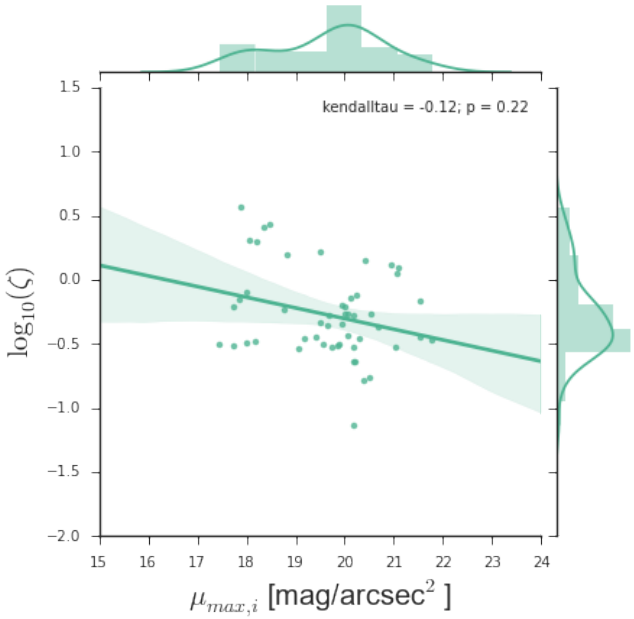


Figure 6. All fits for the vote fraction dropoff parameter ζ for $f_{\rm features}$ in the FERENGI galaxies as a function of surface brightness. This includes only the 37 galaxies with a reasonably bounded range on the dropoff ($-10 < \log(\zeta) < 10$) and sufficient points to fit the function.

Figure 6 shows the results of fitting the FERENGI images with Equation 1; the correction is a weak function of galaxy surface brightness. Higher-surface brightness galaxies have stronger average corrections, likely because these galaxies are more likely to have larger $f_{\rm features}$ values at high redshifts. Low surface brightness galaxies are more likely to begin low and remain low; the bounded nature of the dropoff (and Poissonian-like variance among the individual voters) means that the average magnitude of ζ will be less.

We fit the data in Figure 6 with a linear function such that:

$$\log_{10}(\hat{\zeta}) = \zeta_0 + \zeta_1 \times \mu, \quad (2)$$

where $\hat{\zeta}$ is the correction factor applied to each galaxy as a function of surface brightness. The best-fit parameters to the linear fit (from least-squares optimization) are $\zeta_0 = 0.1$, $\zeta_1 = 1.4$. To make the final debiased correction, we modify the simple exponential form of Equation 1 to bound the debiased vote fractions between f and 1:

$$f_{\rm features,debiased} = 1 - (1 - f)e^{\frac{z-z_0}{\hat{\zeta}}}. \quad (3)$$

4.3 Results of ζ approach

In Figure 8 we examine the change in $p_{\rm features}$ for the FERENGI galaxies relative to their lowest simulated redshift. In this analysis, only galaxies whose lowest simulated redshift

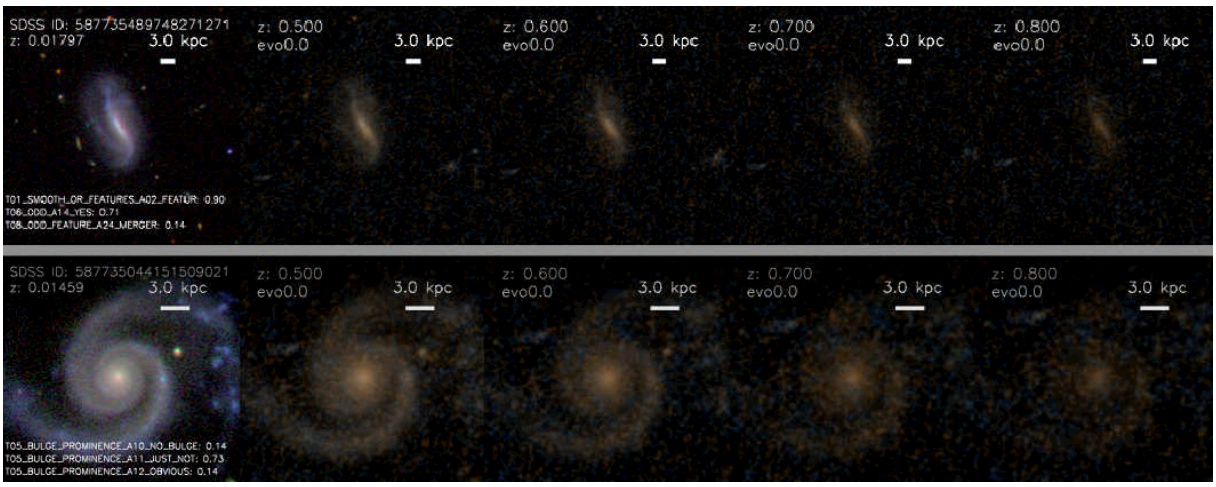


Figure 4. Examples of two galaxies which have been run through the FERENGI code to produce simulated HST images. The value of $p_{\rm features}$ for each panel is (1) Top row: $p_{\rm features}$ = 0.9, 0.625, 0.35, 0.35, 0.225 and (2) Bottom row: $p_{\rm features}$ = 1.00, 0.875, 0.875, 0.625, 0.375.

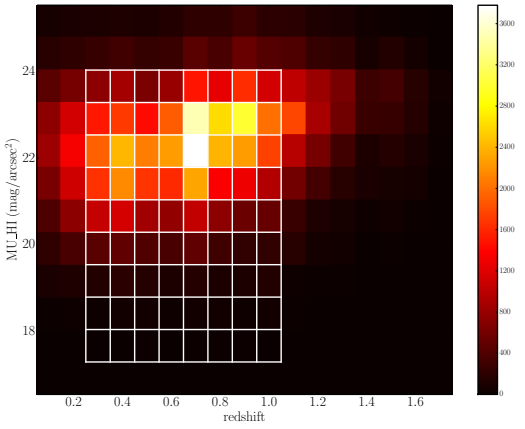


Figure 7. Surface brightness vs. redshift of 118,083 galaxies in the ACS sample. The white grid denotes the surface brightness and redshift range of the FERENGI images, subdivided in bins corresponding to fixed ranges used for analysis in Figure 8.

image was ($z_{\rm sim}$ = 0.3) were used (see Table 4), and only those which had detectable surface brightness measurements in SEXTRACTOR; this includes 3,950 of the total 6,466 images. For each simulated redshift value z, and at a fixed surface brightness μ, we plot $p_{\rm features,z}$, the value measured at that simulated redshift, vs $p_{\rm features,z=0.3}$, the value measured for the same galaxy imaged at $z = 0.3$.

Our objective is to use these data to predict, for a galaxy with a measured $p_{\rm features,z}$ value, what its $p_{\rm features}$ value *would have been* if it had been viewed at $z = 0.3$. This predicted value is defined as the debiased vote fraction $p_{\rm features,debiased}$, and is calculated by applying a correction to the measured value of $p_{\rm features}$, determined by the ζ function described in the previous section. A reliable predicted value can be obtained so long as the relationship between $p_{\rm features,z}$ and $p_{\rm features,z=0.3}$ is single-valued; that is, for a given $p_{\rm features,z}$, there is exactly one corresponding value of $p_{\rm features}$ at $z = 0.3$.

Figure 8 shows that the relationship between $p_{\rm features,z}$ and $p_{\rm features,z=0.3}$ is *not* always single valued; hence, it is not appropriate to correct galaxies that lie in certain regions of surface brightness/redshift/$p_{\rm features}$ space. These regions tend to have low $p_{\rm features}$ values at high redshift, but a wide range of values at $z = 0.3$. These regions contain two morphological types of galaxies: First are genuine ellipticals, which have low values of $p_{\rm features}$ at both high and low redshift. Second are disks whose features become washed out at high redshift; hence their $p_{\rm features}$ value at $z = 0.3$ may be quite high, while the value observed at high redshift is very low. This effect is strongest at high z and low μ, where features become nearly impossible to discern in the images.

Our criteria for determining whether a region of this space is single-valued, and therefore correctable, is as follows: In each surface brightness and redshift bin, we model the relationship between $p_{\rm features,z}$ and $p_{\rm features,z=0.3}$ by fitting the data with a polynomials of degrees 3, 2, and 1, and use the best fit out of the three. These fits are shown as the dashed black lines in Figure 8(a). Any flat regions of the polynomial fits are areas in which there is not a clear single-valued relationship between $p_{\rm features,z}$ and $p_{\rm features,z=0.3}$; we quantify this by setting a minimum slope cut of 0.4. Any data in which the polynomial fit has a slope less than this value is considered *not* one-to-one, and therefore "uncorrectable." These regions are highlighted in blue in figure 8(a). Uncolored (white) regions of the plot have sufficiently high slopes for us to consider the relationship to be single-valued; galaxies in these regions are considered "correctable", and only these are used in measuring the parameters for the ζ function (Section 4.2). Only surface brightness/redshift bins with at least 5 galaxies were considered; regions with fewer than 5 galaxies we consider to have "not enough information" to determine the $p_{\rm features,z}$ and $p_{\rm features,z=0.3}$ relationship, these are colored gray in Figure 8(a).

The unshaded regions in Figure 8(a) define discrete ranges of redshift, surface brightness, and $p_{\rm features}$ a galaxy

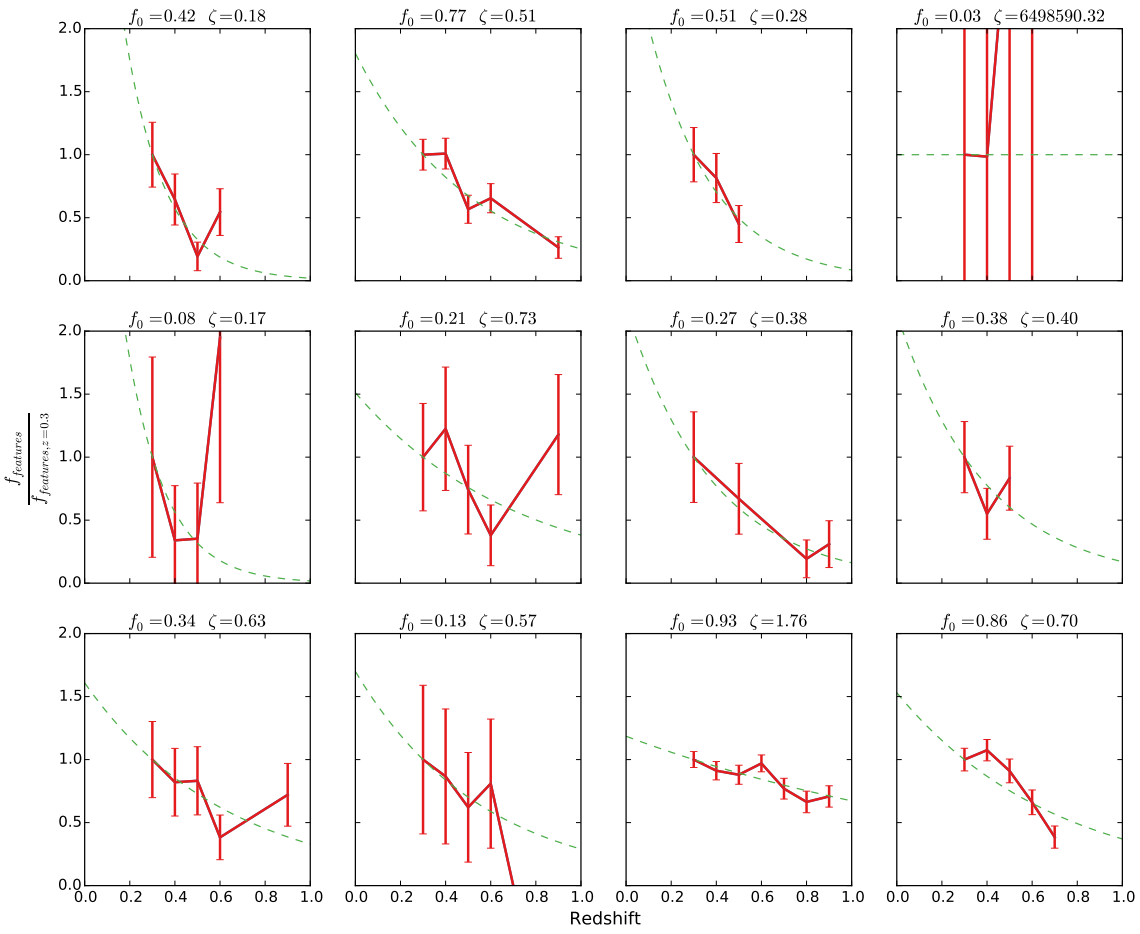
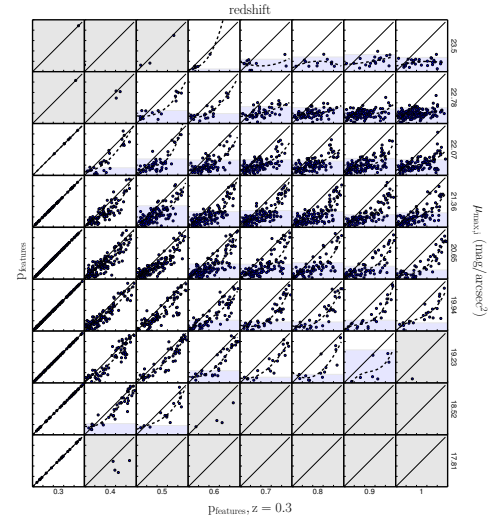


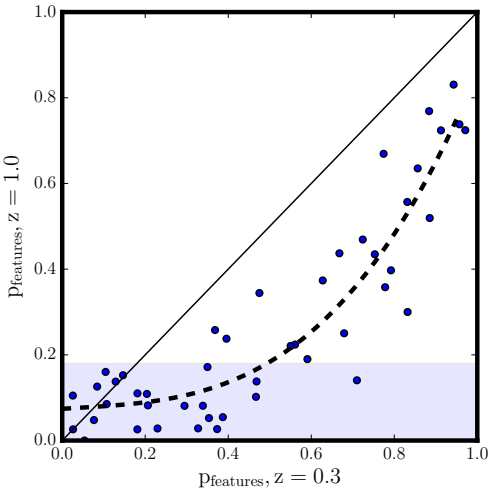
Figure 5. Behavior of the normalized, weighted vote fractions of features visible in a galaxy ($f_{\rm features}$) as a function of redshift in the artifical FERENGI images. Galaxies are a random selection of images with $e = 0$ and at least three detectable images in redshift bins of $z \geq 0.3$. The measured vote fractions (red points) are fit with an exponential function (Equation 1); the best-fit parameters are given above each plot. Error bars are Poissonian, assuming a median of 40 votes per galaxy.

must have in order for the ζ approach to be confidentely applied to a galaxy in the GZH sample. While the appropriate correctable regions were defined discretely, we assume the true correctable region is a smooth function of z, μ, and $p_{\rm features}$. To define this smooth space, we use a convex hull method to enclose the correctable and uncorrectable FERENGI galaxies is z-μ-$p_{\rm features}$ space. Due to scatter, the boundaries of the resulting hulls overlap. The boundaries are then adjusted until the contamination from both groups is minimized. We use the resulting hulls to define the correctable and uncorrectable regions for categorizing the Hubble galaxies. The results of this method and final categorization of the Hubble sample is displayed in Table 6. We find that of the galaxies at redshift higher than $z = z_0 = 0.3$, 17% of these are able to be debiased using the ζ method, 27% cannot be debiased, and 56% cannot be determined, due to a lack of redshift or information or due to a lack of FERENGI data corresponding to those galaxies' redshift/surface brightness values.

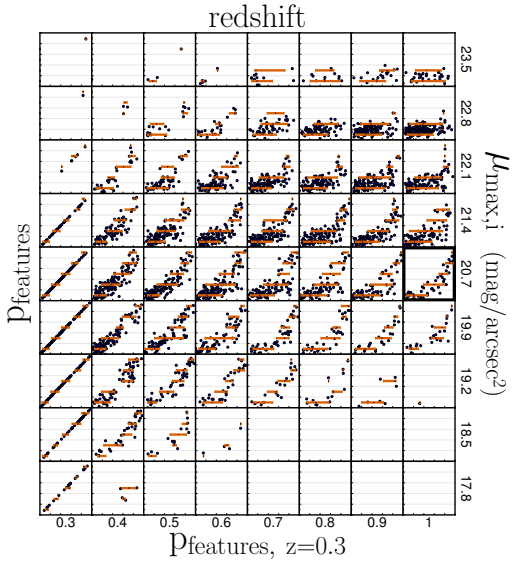
For the "uncorrectable" galaxies, those for which we cannot confidentely assign a single debiased $p_{\rm features}$ value, we instead determine a likely *range* of debiased values, using a method visualized in Figure 8(c). Here we again use the FERENGI simulated data to analyze the range of intrinsic $p_{\rm features,z=0.3}$ values for any given observed $p_{\rm features}$ value, again as a function of surface brightness and redshift. In each z,μ bin, we examine the spread of intrinsic values of $p_{\rm features,z=0.3}$ for 4 ranges of observed $p_{\rm features}$. We quantify the range of intrinsic values as the inner 80% of the data; this range is represented by the orange bars in Figure 8(c). For any galaxy which can't be directly debiased by the ζ method, then, we use these ranges to denote the upper and lower limits on what we expect $p_{features,z=0.3}$ to be for any observed value of $p_{\rm features}$.



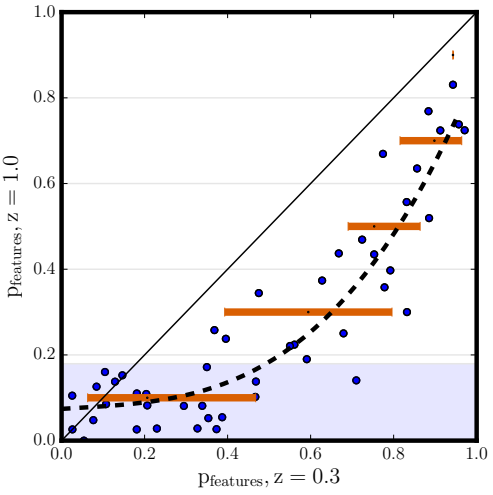
[6a]



[6b]



[6c]



[6d]

Figure 8. Effects of redshift bias in 3,950 images in the FERENGI sample. [6a]: Each point in a given redshift and surface brightness bin represents a unique galaxy. On the y-axis in each bin is the $p_{\rm features}$ value of the image of that galaxy redshifted to the value corresponding to that redshift bin. On the x-axis is the $p_{\rm features}$ value of the image of the same galaxy redshifted to $z = 0.3$. The dashed black lines represent the best-fit polynomials to the data in each square. The solid black line represents $p_{\rm features,z} = p_{\rm features,z=0.3}$. Regions in which there is a single-valued relationship between $p_{\rm features}$ at high redshift and at $z = 0.3$ are white; those in which there is not are blue, and those with not enough data ($N < 5$) are gray. [6b]: A larger version of the dark-outlined square in [6a], containing FERENGI galaxies that have been artificially redshifted to $z = 1.0$ and have surface brightnesses between $20.3 < \mu < 21.0$ (mag/arcsec^2). [6c]: The same data as [6a] is shown. Each z, μ bin is divided into 4 sub-bins to determine the range of intrinsic $p_{\rm features,z=0.3}$ for a given range of observed $p_{\rm features,z}$ values. In each sub-bin, the orange bars represent the inner 80th percentiles of the data, the boundaries of which are the lower and upper limits of the debiased values. [6d]: The same data as [6b], but highlighting the upper and lower limit regions.

4.4 Challenges of debiasing questions beyond "smooth or features"

Each FERENGI image does not have the same number of users answering each question, due to the structure of the decision tree. Every user answers the first question, "Is the galaxy smooth and rounded, with no sign of a disk?"; as such the vote fractions $\rm p_{smooth}$, $\rm p_{features}$, and $\rm p_{artifact}$ are all computed with the minimum statistical error for any question, with roughly 40 total answers (see Section 3). The number of users to answer any subsequent question, however, is always equal to or less than the number to answer the preceding question. For this reason, some galaxies may have very few (or even zero) answers to a question further down the tree (see Figure make-figure-of-count-distribution-for-each-

Table 5. Distribution of FERENGI images analysed in Figure 8. Correctable images had a single-valued relationship between their measured $p_{\rm features}$ values at high and low redshifts (white regions in Figure 8). Uncorrectable images had a non single-valued relationship (blue regions). NEI images had undetermined relationships due to a lack of data ($N < 5$) in their corresponding z-μ bins (gray regions).

	N	%
Correctable	1,884	48%
Uncorrectable	1,986	50%
NEI	80	2%
Total	3,950	100%

question). To minimize statistical error in computing vote fractions, a cut on the number of answers to a given question is always implemented.

In the FERENGI data, we find that this places large limitations on the amount of information we can extract for the higher order questions. We require that at least 5 users answer each question for a galaxy image at $z = 0.3$ *and* its image at higher z. This requirement placed on both images is not met by a significant number of galaxies for questions beyond question 1. Without sufficient galaxies in each surface brightness/redshift bin, we cannot accurately measure a relationship between vote fractions and redshift; for this reason we only offer debiased vote fractions for question 1. **perhaps compute number of galaxies that can be fit to zeta for each question, show a table? overkill?** In Section A2 we show results of an attempt to measure ζ for $\rm p_{bar}$.

- talk about where the Hubble sample falls in this space, reference table 6
- justify $N > 5$ and spread < 0.2 (or find a better way to choose criteria)
- check out corrections for correctable and NEI, show some sample images of corrected galaxies
- show some data for $p_{\rm bar}$, determine or justify why we won't debiase them

4.4.1 TODO LIST

We need to do:

- Calculate the magnitudes, surface brightnesses and sizes of the galaxies in the FERENGI images....
- Plot of magnitude distribution of galaxies in each of the four GZH subsamples with the magnitudes of our fake galaxies over plotted.
- Instructions of how to link the $z = 0$ p_X values for galaxies with a given size, magnitude (surface brightness) in the GZH images.

4.5 Morphological measurements in GZH beyond Task 1 - effects of debiasing?

4.6 Duplicate images

4.7 Effect of changing depth for GOODS

4.8 SDSS Stripe 82 images

4.9 Fake AGN

5 THE CATALOG

The data release for GZH includes morphological data for 181,101 galaxies. The full table can be accessed at some website. We also include a secondary metadata table, which contains data from a variety of sources explained in Section 2.

For each galaxy we list its unique objid, as well as the source's RA, DEC, and survey (AEGIS, COSMOS, GEMS, GOODS North (full and shallow depth), GOODS South (full and shallow depth), SDSS. For each of the 40 (?) questions in the GZH decision tree, the following classification data is provided: For each question, $\rm N_{votes}$ is the number of users to answer that question. For each unique answer, `fraction` is the fraction of users to select that answer ($\rm N_{answer}/N_{votes}$, and `weighted` is the weighted fraction, which takes into account user consistency (Section 2.4).

The GZH vote fractions can be largely dependent on the resolution of the image. Two otherwise morphologically identical galaxies which differ significantly in redshift, brightness, or size may result in very different vote fractions for any given question, given that many features of a galaxy are difficult to discern in less-resolved images (bars, spiral arms, disk structure, etc). For this reason, it is necessary to take caution utilizing vote fractions as cut-offs to determine morphological structure; we offer guidelines for careful classification in Section 6.

We corrected for the biases described for the first question of the GZH decision tree, which asks "Is the galaxy smooth and round, with no sign of a disk?" The method is described in Section 4. For this question, we provide the additional parameters `debiased`, `lower limit`, `upper limit`, and `best` vote fractions. The `best` fraction for $p_{features}$is chosen based on the categorization of the galaxy: if it is "correctable", `best` = `debiased`, if "uncorrectable", `best` = `lower limit`, and if neither, `best` = `weighted`. The debiased vote fractions for $\rm p_{smooth}$ were calculated on the critera that vote fractions for all answers must sum to unity. Explicitely: $\rm p_{smooth} = 1 - p_{features} - p_{artifact}$.

To include: detailed description of each column in the machine-readable tables.

Needs some test cases for extracting a given set of objects (eg, clump galaxies in a particular redshift range) and evaluation of the results. Possibly include suggested thresholds, á la GZ2.

6 USING THE CATALOG

Include cookbook for selecting morphologies.

Table 6. Breakdown of what we can correct out of the GZH data, by sample. *updated from 3-8-16: Switching to full depth for all GOODS data. Shallow depth information in appendix.*

	AEGIS	COSMOS	GEMS	GOODS-N	GOODS-S	SDSS	Total
Correctable	1,654	15,170	1,837	993	835	0	20,489
Uncorrectable	1,917	26,113	2,423	1,385	1,282	0	33,120
No Correction Needed ($z \leq 0.3$)	955	11,926	1,175	415	400	37,545	52,416
NEI	2,847	34,511	3,308	2,535	2,523	0	45,724
No Redshift Information	1,134	5,088	561	687	102	14,316	21,888
Total	8,507	92,808	9,304	6,015	5,142	51,861	173,637

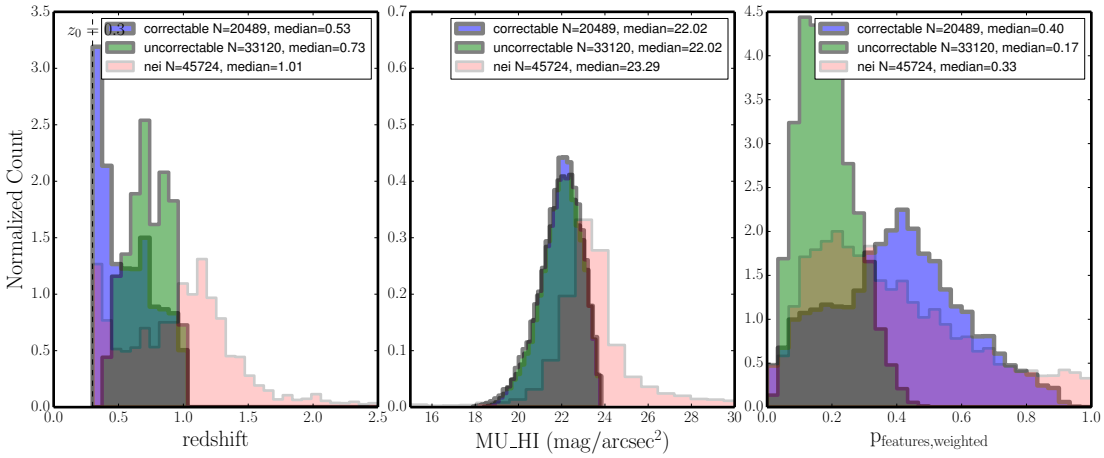


Figure 9. Distributions of redshift, surface brightness, and $p_{features}$ for correctable (purple), uncorrectable (green), and NEI (pink) galaxies in the full GZH sample. The uncorrectable galaxies tend towards higher redshift, slightly lower in surface brightness, and lower values of $p_{features}$ than the correctable galaxies. The long tail of NEI galaxies in redshift and surface brightness demonstrates the limits of the FERENGI sample, for which there is no data at $z > 1$ or $\mu > 24$.

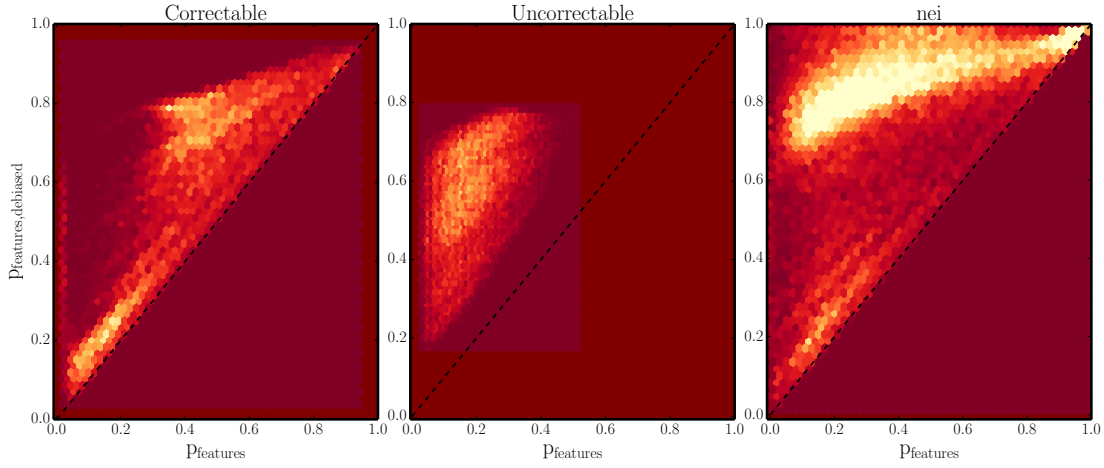


Figure 10. Debiased $p_{features}$ corrected to $z = 0.3$ vs weighted $p_{features}$ for the correctable (left), uncorrectable (middle), and NEI (right) galaxies in the GZH sample.

Table 7. GZH morphological classifications

				t01_smooth_or_features_			t01_smooth_or_features_a01_smooth_					...
OBJNO	RA	DEC	Imaging	Correctable Category	N_{votes}	fraction	weighted	debiased	lower limit	upper limit	best	
13035128	214.99	53.02	AEGIS	correctable	121	0.46	0.50	0.12	0.14	0.73	0.12	
13035131	215.01	53.02	AEGIS	nei	134	0.61	0.65	0.20			0.65	
...	...	...	...	...	...	...	...	...	...	...	...	
20055697	150.52	2.09	COSMOS	correctable	52	0.52	0.55	0.38	0.55	0.81	0.38	
20055698	150.50	2.09	COSMOS	uncorrectable	31	0.61	0.70	0.47	0.47	0.80	0.70	
...	...	...	...	...	...	...	...	...	...	...	...	
90021901	53.29	-27.85	GEMS	correctable	131	0.21	0.21	0.15	0.14	0.35	0.15	
90021903	53.34	-27.83	GEMS	$z < 0.3$	142	0.78	0.80	1.04			0.80	
...	...	...	...	...	...	...	...	...	...	...	...	
GDS_N_2333	189.02	62.12	GOODS-N	nei_needs_redshift	40	0.33	0.33	0.87			0.33	
GDS_S_3132	53.04	-27.71	GOODS-S	uncorrectable	40	0.83	0.83	0.45	0.38	0.85	0.83	
...	...	...	...	...	...	...	...	...	...	...	...	
587731173843665742	315.58	0.25	SDSS	$z < 0.3$	48	0.79	0.89	1.23			0.89	
587731173843730515	315.61	0.30	SDSS	$z < 0.3$	53	0.68	0.75	1.06			0.75	

Note. — Full version is online at some website, here are 10 rows out of 181,101 in the full version.

7 ANALYSIS

7.1 Demographics of morphology

Summarize the broad trends that are seen regarding the fraction of galaxies with various morphologies, how that relates to color, size, etc. Briefly discuss results as compared with literature and theory.

7.2 Comparison to other catalogs

Compare GZH data to:

- Scarlata et al. (ZEST; 2007) (COSMOS)
- Tasca (COSMOS)
- Cassata (COSMOS)
- Zajmoski (COSMOS)
- GEMS morphologies?
- AEGIS morphologies?
- GOODS N/S morphologies?
- expert visual inspection?

Address trends seen in broad morphological classes, possible reasons for difference. Also should attempt to map between the GZH vote fractions and whatever classification systems are used in the above systems.

8 SUMMARY

Now people go and do science with these awesome GZH classifications.

ACKNOWLEDGEMENTS. This publication has been made possible by the participation of more than 200,000 volunteers in the Galaxy Zoo project. Their contributions are individually acknowledged at `http://www.galaxyzoo.org/volunteers`.

We thank Meg Schwamb and the ASIAA for hosting the "Citizen Science in Astronomy" workshop, 3-7 Mar 2014 in Taipei, Taiwan, at which some of this analysis was done.

This project made heavy use of the Astropy packages in Python (Astropy Collaboration et al. 2013), the `seaborn` plotting package (Waskom et al. 2015), and astroML (Vanderplas et al. 2012).

HST acknowledgements.

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

Astropy Collaboration et al., 2013, A&A, 558, A33
Bamford S. P. et al., 2009, MNRAS, 393, 1324
Barden M., Jahnke K., Häußler B., 2008, ApJS, 175, 105
Caldwell J. A. R. et al., 2008, ApJS, 174, 136
Cardamone C. et al., 2010, ApJS, 189, 270
Davis M. et al., 2007, ApJ, 660, L1
Giavalisco M. et al., 2004, ApJ, 600, L93
Griffith R. L. et al., 2012, ApJS, 200, 9
Ilbert O. et al., 2013, A&A, 556, 55
Johnson L. C. et al., 2015, ApJ, 802, 127
Lintott C. J. et al., 2008, MNRAS, 389, 1179
Momcheva I. G. et al., 2015
Rix H.-W. et al., 2004, ApJS, 152, 163
Scarlata C. et al., 2007, ApJS, 172, 406
Scoville N. et al., 2007, ApJS, 172, 1
Strauss M. A. et al., 2002, AJ, 124, 1810
Vanderplas J., Connolly A., Ivezić Ž., Gray A., 2012, in Conference on Intelligent Data Understanding (CIDU), pp. 47 –54
Waskom M. et al., 2015, seaborn: v0.6.0 (june 2015)
Willett K. W. et al., 2013, MNRAS, 435, 2835
York D. G. et al., 2000, AJ, 120, 1579

Table A1. Breakdown of what we can correct out of the GOODS shallow depth data.

	GOODS-N	GOODS-S	Total
Correctable	748	514	1,262
Uncorrectable	526	1,143	1,669
No Correction Needed ($z \leq 0.3$)	267	267	534
NEI	851	2,670	3,521
No Redshift Information	159	319	478
Total	2,551	4,913	7,464

APPENDIX A: GOODS SHALLOW DEPTH DATA

GZH used both 5-epoch and 2-epoch sets of data to construct the GOODS set of images. The 11,157 full depth 5-epoch images are used in the main catalog; the classifications for the 7,464 shallow depth 2-epoch images are offered as a supplementary table. Here we briefly analyze the effect of image depth on the ability of the GZ users to identify features or disk structure in the images.

A1 Comparing shallow and full depth morphologies

. Of the 11,157 galaxies in the GOODS-N and GOODS-S full depth sample, 4,461 of these are in the shallow-depth sample. In Figure A1 we find a strong correlation between $p_{features}$ for both sets of images. The mean change in $p_{features}$ from the shallow to full depth images $p_{features,full} - p_{features,shallow} = \Delta p = 0.00$, with a standard deviation of $\sigma = 0.17$. While there is some variance in Δp in the whole sample, the change is usually small and not often significant enough to change a morphological classification. Defining a clean sample of disk galaxies as those with $p_{features,best} > 0.8$, elliptical galaxies as those with $p_{smooth,best} < 0.2$, and intermediate as those in between, we find that 75% of the sample would not change morphology. Of the remaining 25% that would change morphology, only 0.3% (representing 10 galaxies total) drastically change morphology from smooth to featured or visa versa, while the rest would transition to or from the "intermediate" morphology. Details can be seen in Table A2 and examples of images representing the 9 possible changes (or lack of) in morphology are shown in Figures A2,A3, and A4.

A2 FERENGI analysis of pbar

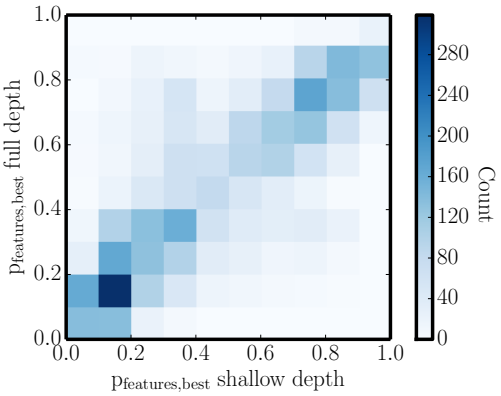


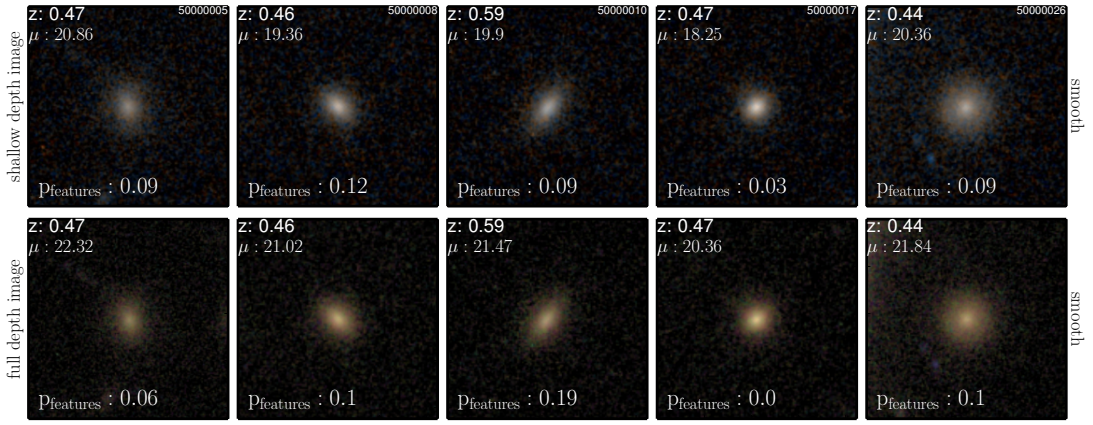
Figure A1. shallowfull

Table A2. Properties of galaxies whose morphologies changed or stayed the same in the shallow vs full images. Featured here is defined as $p_{features,best} > 0.8$, intermediate $= 0.2 < p_{features,best} < 0.8$, smooth $= p_{smooth,best} < 0.2$.

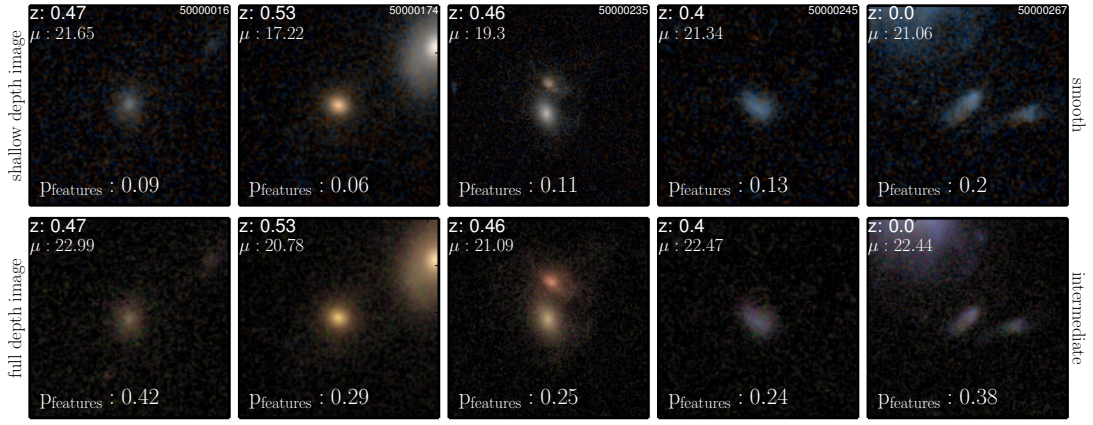
shallow to full morphology	N	%	$< \Delta p >$	$< z >$
smooth to smooth	758	17.0	-0.00	0.69
smooth to intermediate	367	8.2	0.18	0.69
smooth to featured	7	0.2	0.76	0.57
intermediate to smooth	214	4.8	-0.18	0.65
intermediate to intermediate	2,303	51.6	0.01	0.78
intermediate to featured	168	3.8	0.19	0.83
featured to smooth	3	0.1	-0.74	0.71
featured to intermediate	337	7.6	-0.18	0.68
featured to featured	301	6.8	-0.05	0.71
Total	4,461	100		

Table A3. Distribution of FERENGI images analysed in Figure A5. Correctable images had a single-valued relationship between their measured $p_{\rm bar}$ values at high and low redshifts (white regions in Figure A5). Uncorrectable images had a non single-valued relationship (blue regions). NEI images had undetermined relationships due to a lack of data ($N < 5$) in their corresponding z-μ bins (gray regions). Only 17% (maximum) of FERENGI galaxies in the sample were considered "correctable", which is not sufficient to compute a ζ function applicable to the Hubble data.

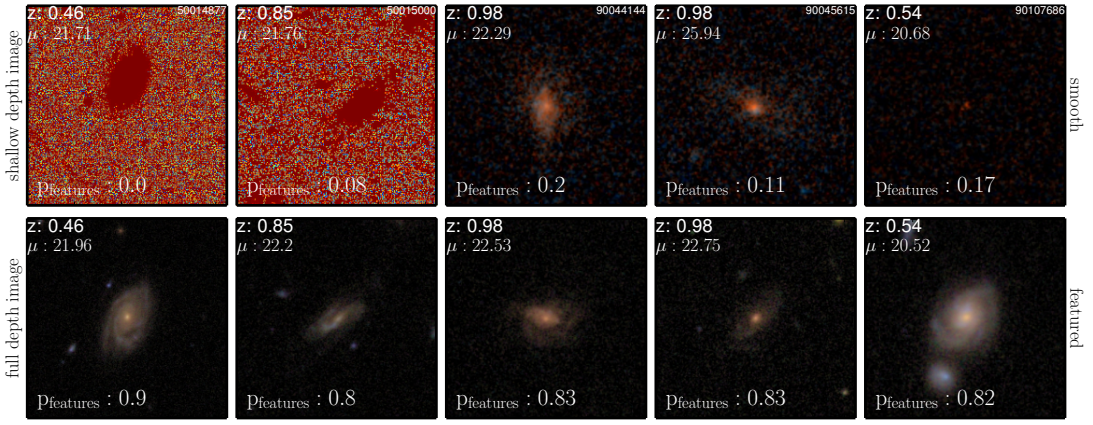
	N	%
Correctable	664	17%
Uncorrectable	483	12%
NEI	2,803	71%
Total	3,950	100%



[a]

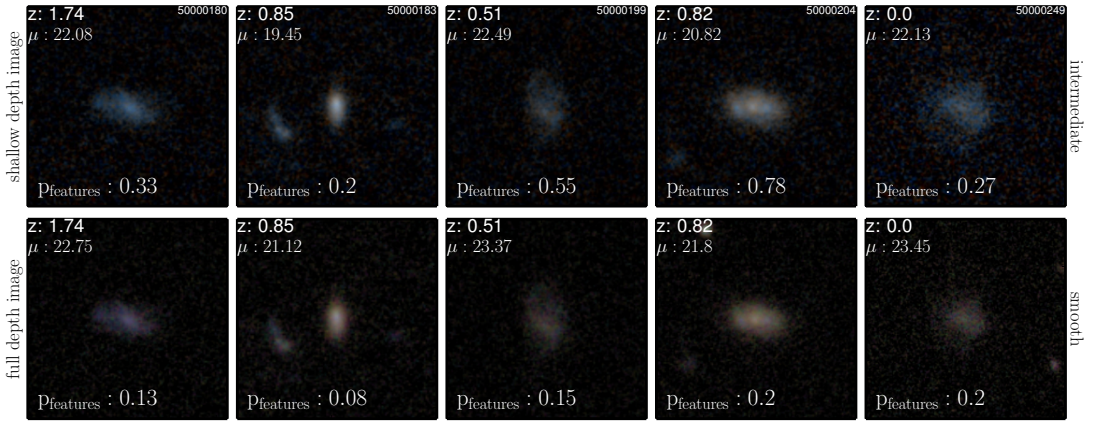


[b]

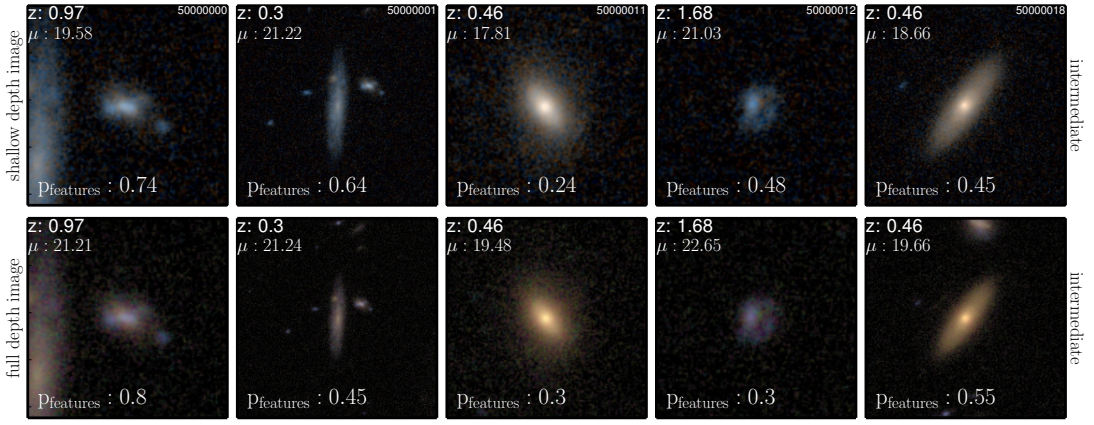


[c]

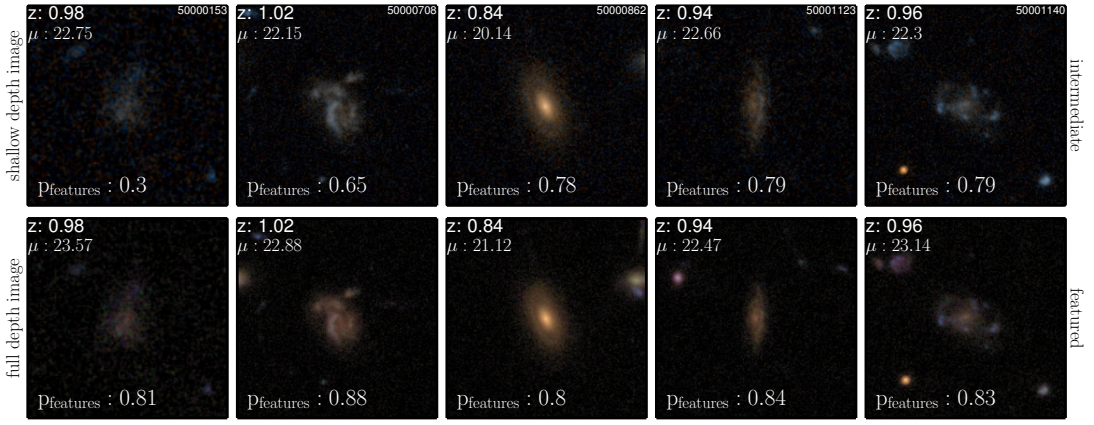
Figure A2. Galaxies whose shallow images were classified as smooth and full depth images were classified as smooth, intermediate, or featured.



[b]

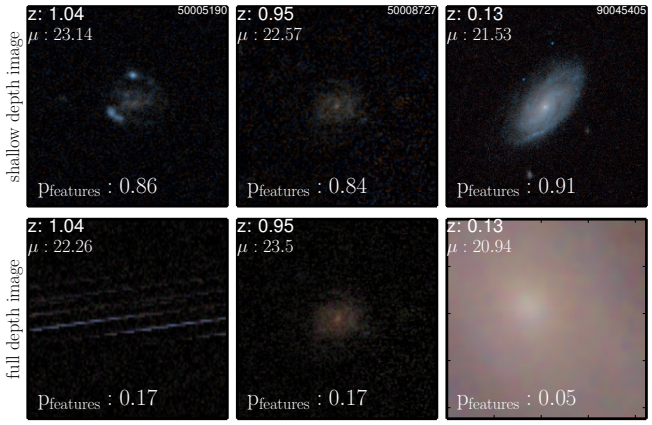


[b]

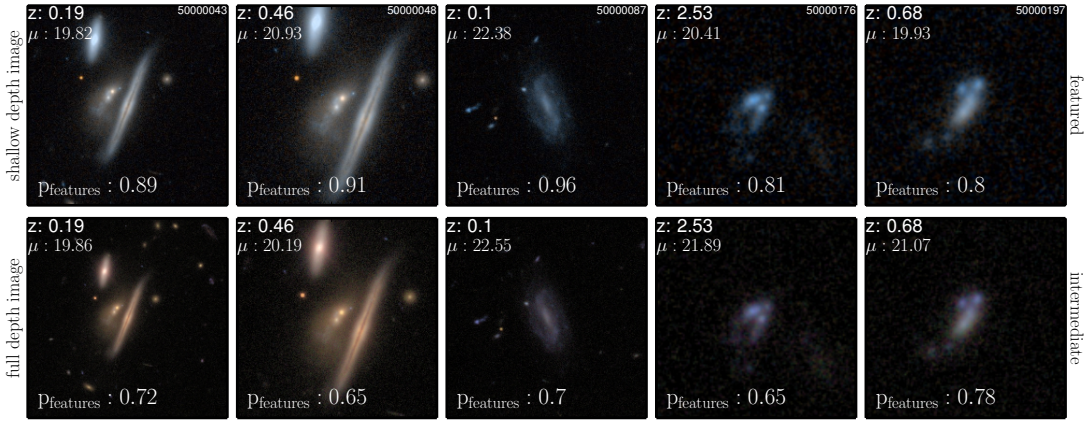


[b]

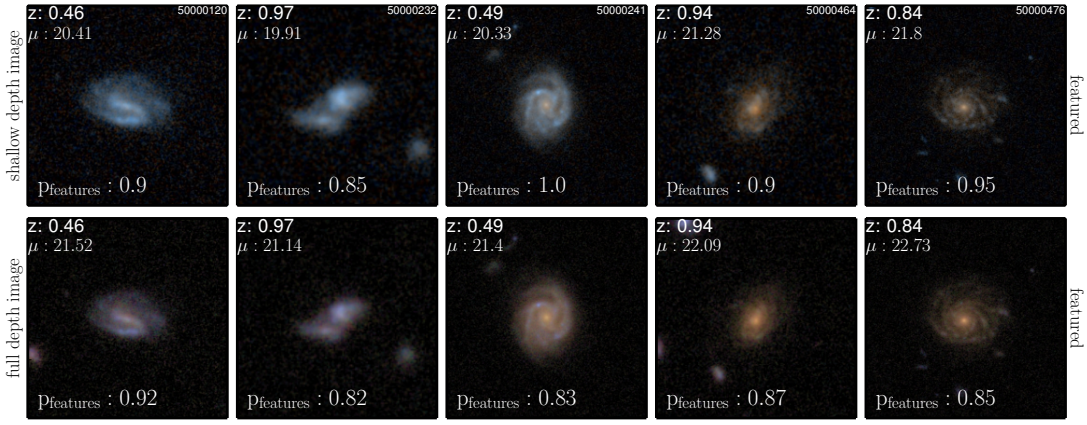
Figure A3. Galaxies whose shallow images were classified as intermediate and full depth images were classified as smooth, intermediate, or featured.



[b]

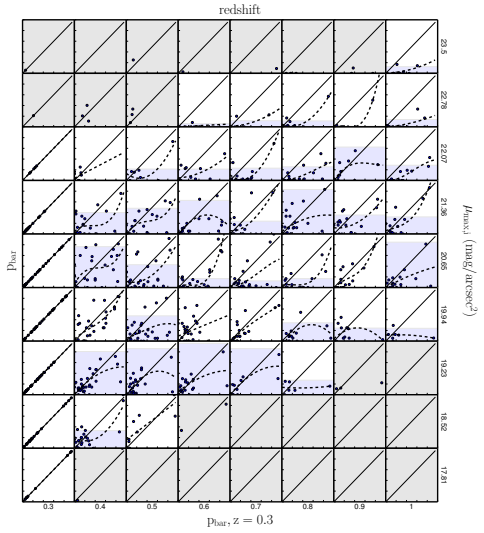


[b]

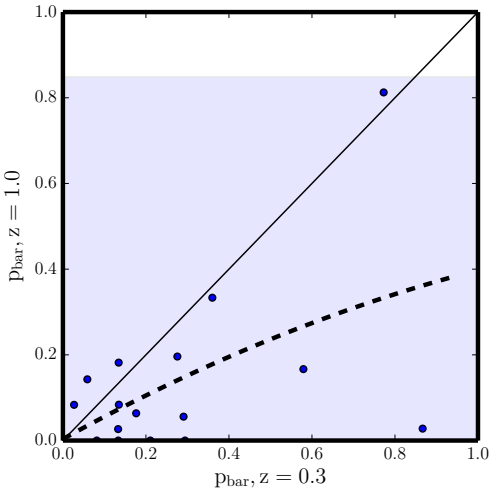


[b]

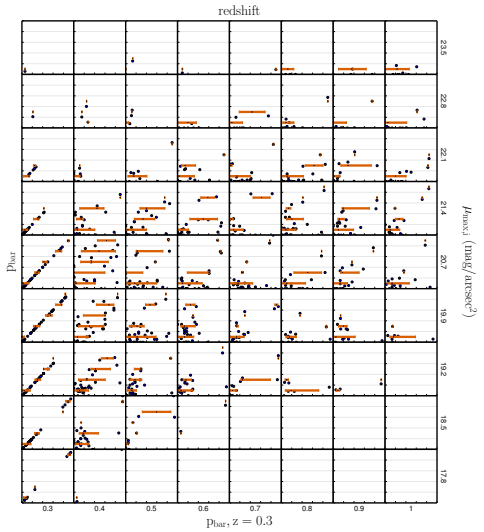
Figure A4. Galaxies whose shallow images were classified as featured and full depth images were classified as smooth, intermediate, or featured.



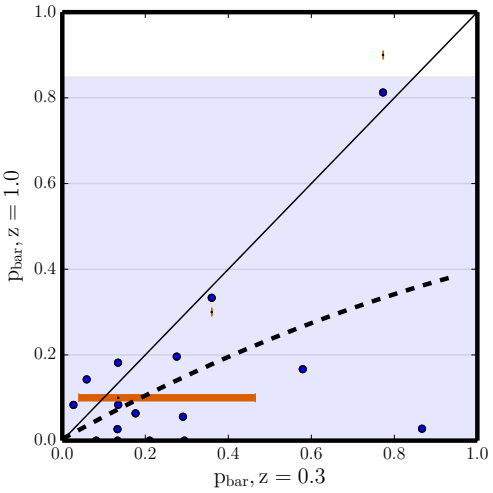
[a]



[b]



[c]



[d]

Figure A5. Same as Figure 8, but with the bar question.