Galaxy Zoo: Clump Scout: Surveying the Local Universe for Giant Star-forming Clumps

How to cite:
Adams, Dominic; Mehta, Vihang; Dickinson, Hugh; Scarlata, Claudia; Fortson, Lucy; Kruk, Sandor; Simmons, Brooke and Lintott, Chris (2022). Galaxy Zoo: Clump Scout: Surveying the Local Universe for Giant Star-forming Clumps. The Astrophysical Journal, 931(1)

For guidance on citations see FAQs.

© [not recorded]

https://creativecommons.org/licenses/by/4.0/

Version: Version of Record

Link(s) to article on publisher’s website:
http://dx.doi.org/doi:10.3847/1538-4357/ac6512

Copyright and Moral Rights for the articles on this site are retained by the individual authors and/or other copyright owners. For more information on Open Research Online’s data policy on reuse of materials please consult the policies page.
Galaxy Zoo: Clump Scout: Surveying the Local Universe for Giant Star-forming Clumps

Dominic Adams1, Vihang Mehta1, Hugh Dickinson2, Claudia Scarlata1, Lucy Fortson1, Sandor Kruk3,4, Brooke Simmons5, and ChrisLintott6

1 School of Physics and Astronomy, University of Minnesota, 116 Church Street SE, Minneapolis, MN 55455, USA
2 School of Physical Sciences, The Open University, Walton Hall, MK7 6AA Milton Keynes, UK
3 European Space Agency, ESTEC, Keplerlaan 1, NL-2201 AZ, Noordwijk, The Netherlands
4 Max-Planck-Institut für extraterrestrische Physik (MPE), Giessenbachstrasse 1, D-85748 Garching bei München, Germany
5 Physics Department, Lancaster University, Lancaster, LA1 4YB, UK
6 Oxford Astrophysics, Denys Wilkinson Building, Keble Road, Oxford, OX1 3RH, UK

Received 2021 July 18; revised 2022 March 2; accepted 2022 March 17; published 2022 May 19

Abstract

Massive, star-forming clumps are a common feature of high-redshift star-forming galaxies. How they formed, and why they are so rare at low redshift, remains unclear. In this paper we identify the largest sample yet of clumpy galaxies (7050) at low redshift using data from the citizen science project Galaxy Zoo: Clump Scout, in which volunteers classified 58,550 Sloan Digital Sky Survey (SDSS) galaxies spanning redshift 0.02 < z < 0.15. We apply a robust completeness correction by comparing with simulated clumps identified by the same method. Requiring that the ratio of clump to galaxy flux in the SDSS u band be greater than 8% (similar to clump definitions used by other works), we estimate the fraction of local star-forming galaxies hosting at least one clump (f\textsubscript{clumpy}) to be 3.22±0.34%. We also compute the same fraction with a less stringent relative flux cut of 3% (1.68±1.38%), as the higher number count and lower statistical noise of this fraction permit finer comparison with future low-redshift clump galaxy studies. Our results reveal a sharp decline in f\textsubscript{clumpy} over 0 < z < 0.5. The minor merger rate remains roughly constant over the same span, so we suggest that minor mergers are unlikely to be the primary driver of clump formation. Instead, the rate of galaxy turbulence is a better tracer for f\textsubscript{clumpy} over 0 < z < 1.5 for galaxies of all masses, which supports the idea that clump formation is primarily driven by violent disk instability for all galaxy populations during this period.

Unified Astronomy Thesaurus concepts: Starburst galaxies (1570); Galaxies (573); Galaxy evolution (594); Galaxy formation (595); Galaxy structure (622); Cosmological evolution (336)

Supporting material: machine-readable table

1. Introduction

The morphologies of low-redshift galaxies can generally be classified by their location on the Hubble sequence (Hubble 1926). However, in the past few decades, observations of star-forming galaxies at the peak of cosmic star formation (z ∼ 2) have revealed that these galaxies typically exhibit irregular, clumpy morphologies different from these classifications (Cowie et al. 1995; Elmegreen et al. 2004a, 2004b). Clumpy galaxies receive their name from the “giant star-forming clumps” that occupy them. Elmegreen (2007) estimated these clumps to be of characteristic mass between 10\textsuperscript{7} and 10\textsuperscript{9} M\textsubscript{☉} in galaxies in the Hubble Ultra Deep Field, much more massive than typical star-forming regions locally (though more recent papers have suggested lower characteristic masses; e.g., Fisher et al. 2017; Dessauges-Zavadsky & Adamo 2018). It has now been established that clumpy morphology in star-forming galaxies peaks near z ∼ 2, with >50% exhibiting clumpy behavior, before declining significantly with cosmic time (Murata et al. 2014; Guo et al. 2015; Shibuya et al. 2016; Guo et al. 2018).

It is still not clear why clumps are so dominant at z ∼ 2, nor why they are so much rarer in low-redshift galaxies. Two major modes of clump formation have been discussed at length in the literature. First, clumps may form “in situ” owing to “violent disk instability” (VDI) within their host galaxy, i.e., in disks where the Toomre stability parameter is ∼1 (Bournaud et al. 2014; Mandelker et al. 2014; Fisher et al. 2017). Such disks are called “marginally stable.” This state can occur in galaxies that are continuously fed by smooth cold accretion of intergalactic gas (Genzel et al. 2008; Dekel et al. 2009), though observational confirmation of this process has been limited (Scarlata et al. 2009; Bouché et al. 2013). Second, clumps may form owing to interactions, i.e., major or minor mergers. The merging galaxies can themselves become clumps, known as “ex situ” clumps, or they can generate local instabilities within the merging galaxies that collapse to form in situ clumps. It is expected that ex situ clumps are higher in mass and volume, older, and lower in star formation activity than their in situ counterparts; simulations and observations both suggest that a minority of high-redshift clumps are formed this way (Mandelker et al. 2014; Zanella et al. 2019). It is possible that different formation mechanisms are at work for different populations of clumpy galaxies. Guo et al. (2015) suggest that the clump formation mode depends strongly on galaxy mass, with high-mass galaxies (M > 10\textsuperscript{10.5} M\textsubscript{☉}) dominated by the VDI-driven “in situ” mode, while low-mass galaxies (M < 10\textsuperscript{10} M\textsubscript{☉}) are dominated by minor-merger-driven “ex situ” formation.

Clumps have been difficult to study largely because they are difficult to resolve with existing instruments. While clumps were originally thought to be kiloparsec-scale objects,
high-resolution observations of clumpy galaxies in the local universe (e.g., Overzier et al. 2009; Fisher et al. 2014, 2017; Messa et al. 2019) and in strongly lensed fields (e.g., Wuyts et al. 2014; Dessauges-Zavadsky et al. 2017; Dessauges-Zavadsky & Adamo 2018; Cava et al. 2018) have revealed that many clumps have a much smaller characteristic size, ranging from tens to hundreds of parsecs. It has therefore been theorized that high-redshift “kiloparsec-scale clumps” may be rare and that the size and mass of observed clumps are an order of magnitude lower than were originally estimated. Several papers point to “blending” of multiple small clumps at low resolution to explain high-redshift observations (Dessauges-Zavadsky et al. 2017; Fisher et al. 2017; Dessauges-Zavadsky & Adamo 2018). The characteristic size, mass, and properties of giant star-forming clumps are still a topic of debate.

To date, studies of local clumpy galaxies have focused on high-resolution imaging of small galaxy samples ($n < 50$). However, no studies have yet assembled a large catalog of local clumpy galaxies. Local clumpy galaxies, while rare, are observable in much greater detail and can act as analogs to their more distant counterparts. To this end, we launched the citizen science project Galaxy Zoo: Clump Scout (herein called Clump Scout). This project, which was active between 2019 and 2021, recruited volunteers to visually identify clumps in galaxies in the local universe. Each subject was examined by many volunteers, whose annotations were aggregated into consensus locations. The catalog of low-redshift clumps and clumpy galaxies from this project is the largest of its kind and can help to constrain models of clump formation and galaxy evolution by comparing it with high-redshift populations.

In this paper we will present the clump catalog assembled by Clump Scout and use it to estimate the fraction of clumpy galaxies ($f_{\text{clumps}}$) in the local universe. Guo et al. (2015), Shibuya et al. (2016), and others have used the evolution of $f_{\text{clumps}}$ with redshift to evaluate the likelihood of different clump formation mechanisms (i.e., by internal VDI or by galaxy interactions). We will extend their analysis using our own $f_{\text{clumps}}$ result.

This paper is structured as follows. Section 2 describes the galaxy sample, the Clump Scout citizen science project, and our methods for aggregating the annotations provided by Clump Scout volunteers. Section 3 describes our methods for recovering clump properties and correcting for incompleteness and describes the best practices for using the Clump Scout clump catalog. Section 4 details our method of computing the clumpy fraction and correcting it for incompleteness and compares this fraction to other results. In Section 5 we discuss the physical significance of our clumpy fraction result. Finally, Section 6 presents a summary and conclusions.

2. Sample Selection and Preparation

2.1. The Galaxy Zoo: Clump Scout Project

Galaxy Zoo: Clump Scout was a citizen science project that was active from 2019 September 19 to 2021 February 11 on the Zooniverse platform. It presented volunteers with image cutouts of galaxies from the Sloan Digital Sky Survey (SDSS; York et al. 2000); for each galaxy image (referred to as a subject) the volunteer was asked to identify its central bulge, followed by all of its off-center clumps. The central bulge location was requested in order to discourage volunteers from marking the galaxy center as a clump, as well as to remove any identified “clumps” coinciding with galaxy center. To improve classification quality, new volunteers were presented with a classification tutorial; we also provided a field guide, as well as help pages for each task, which showcased many examples of quality clump annotations. In addition, the majority of volunteers were shown a small sample of expert-classified “training” images when they began classifying (≤ 10 randomly interspersed throughout their first 20 classifications). After classifying a training image, the volunteer was given feedback on how the classification compared to that of the experts.

For the clump-marking task, volunteers were provided with a “normal clump marker” and an “unusual clump marker” (see Figure 1). The “unusual” marker was intended to identify foreground star contaminants, i.e., Milky Way stars that overlap with the angular area of the target galaxy. Like clumps, foreground stars appear as point sources in SDSS images, and their colors in the $i$, $r$, and $g$ bands can be very similar to clumps; this makes them very difficult to distinguish from clumps by any simple rule. We therefore instructed volunteers to mark clumps as “unusual” if they were particularly bright, differently colored, or especially offset from their host galaxy. Examples of “unusual clumps” (which were generally foreground stars) were provided in the tutorial and field guide. A total of 11,089 distinct volunteers provided classifications while logged into the Zooniverse website, while volunteers who were not logged in undertook 10,200 distinct classification sessions. In total, we obtained 1,707,485 classifications on our galaxy sample.

2.2. Galaxy Selection Criteria

In this study, our goal was to identify as thorough a sample of clumpy galaxies in the local universe as possible, where we take “local” to refer to a region nearby enough that little to no cosmological evolution takes place (i.e., $z \lesssim 0.15$). To select for potentially clumpy galaxies, we relied on data from the Galaxy Zoo 2 (GZ2) citizen science project, which provides morphological classifications for over 300,000 local galaxies from the SDSS legacy survey. GZ2 itself consists of roughly 25% of galaxies identified in the SDSS main galaxy sample and comprises “the nearest, brightest, and largest systems for which fine morphological features can be resolved and classified” (Willett et al. 2013). From 243,500 GZ2 galaxies with spectroscopic redshifts, we selected for galaxies with detectable features by requiring a weighted vote fraction of >50% for the presence of “features or disk” (herein called $f_{\text{featured}}$). Major mergers were removed by requiring that the debiased vote fraction for the presence of “merger” (herein called $f_{\text{merger}}$) was <50%. (For an overview of GZ2 vote fractions, see Willet et al. 2013.) Galaxies with $z < 0.02$ were removed to ensure that the vast majority of clumps could not be resolved and would appear as point sources; this was required so that realistic-looking simulated clumps could be added to these...

---

7 https://zooniverse.org/ClumpScout

9 The fraction of volunteers who voted for a particular Galaxy Zoo 2 classification, weighted according to the estimated consistency of each volunteer.

10 Similar to the weighted vote fraction, but additionally corrected to remove systematic biases in galaxy properties due to their redshift.
galaxies for comparison. Finally, we limited our sample to galaxies whose masses had been estimated by the SDSS DR7 MPA-JHU value-added catalog\(^\text{11}\) (Kauffmann et al. 2003; Brinchmann et al. 2004). We did not make any cuts on galaxy size, as GZ2 already selects its galaxies to be resolvable (petroR90\(_r\) > 3 arcsec, where petroR90\(_r\) is the radius containing 90% of galaxy flux in the Petrosian aperture in the \(r\) band). The resulting sample contained 53,613 galaxies.

In addition to this main sample, we included a sample of 4937 additional galaxies from GZ2 with \(0.02 < z < 0.075\) and for which the GZ2-weighted vote fraction for “features or disk” was \(\leq 50\%\). The redshift limit of \(z < 0.075\) was applied to include a larger proportion of low-mass galaxies in our sample, since SDSS is dominated by high-mass galaxies at higher redshifts. In total, 168,603 galaxies in the GZ2 spectroscopic sample met the “features or disk” vote fraction cut \((f_{\text{featured}} > 0.5)\), and 74,519 of these additionally met the \(z < 0.075\) redshift cut. We selected only a subsample of these to study, as they were not classified as having features or a disk and therefore were unlikely candidates for hosting clumps. The 4937 galaxies chosen will be referred to as the “extra” sample, and we included the sample to permit us to extrapolate its clumpy statistics over the full population. With the “extra” sample included, a total of 58,550 galaxies were studied. Unless otherwise specified, photometry describing galaxies in our sample was obtained from the SDSS DR15 PhotoPrimary table (Aguado et al. 2019). Table 1 summarizes our galaxy selection criteria and number counts.

<table>
<thead>
<tr>
<th>Selection</th>
<th>Galaxy Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parent Sample</td>
<td></td>
</tr>
<tr>
<td>GZ2 with spec. redshift</td>
<td>243,500</td>
</tr>
<tr>
<td>With MPA-JHU mass estimates</td>
<td>239,950</td>
</tr>
<tr>
<td>With (z &gt; 0.02)</td>
<td>225,085</td>
</tr>
<tr>
<td>Regular Sample</td>
<td></td>
</tr>
<tr>
<td>With (f_{\text{merger}} &lt; 0.5)</td>
<td>211,236</td>
</tr>
<tr>
<td>With (f_{\text{featured}} &gt; 0.5)</td>
<td>53,613</td>
</tr>
<tr>
<td>Extra Sample</td>
<td></td>
</tr>
<tr>
<td>With (f_{\text{featured}} \leq 0.5)</td>
<td>168,603</td>
</tr>
<tr>
<td>With (z &lt; 0.075)</td>
<td>74,519</td>
</tr>
<tr>
<td>Randomly selected sample</td>
<td>4,937</td>
</tr>
</tbody>
</table>

Note. \(f_{\text{merger}}\) refers to the debiased GZ2 vote fraction for “merger,” while \(f_{\text{featured}}\) refers to the weighted GZ2 vote fraction for “features or disk.” The final sample consisted of all 53,613 galaxies with \(f_{\text{featured}} > 0.5\) and \(f_{\text{merger}} < 0.5\) (the “regular” sample), in addition to a sample of 4937 of the 168,603 galaxies with \(f_{\text{featured}} \leq 0.5\) (the “extra” sample).

2.3. Generating Images with Simulated Clumps

To estimate the completeness of our sample, we created additional cutouts of galaxies from the target sample with artificial clumps added. These will be called “simulated” subjects, whereas cutouts with no added clumps will be called “real.” The simulated sample consists of 84,565 simulated clumps in 26,736 galaxies, approximately half the galaxy count in the real sample. During Clump Scout, each volunteer was shown a random sequence of subjects drawn uniformly from the pool of all subjects, i.e., both real and simulated. A subject

---

\(^{11}\)http://www.sdss3.org/dr8/spectro/galspec.php
was retired (no longer shown to volunteers) after receiving 20 independent classifications.

Galaxies were selected for the simulated sample such that their mass distribution matched the mass distribution of clumpy galaxies in the Guo et al. (2018) sample (1132 galaxies spanning $0.5 < z < 3$). This lowered the characteristic mass of galaxies in the simulated sample: the galaxy sample from Guo et al. (2018) has a median mass of $10^{9.7} M_\odot$, compared to $10^{9.06} M_\odot$ for galaxies in the real sample. This was appropriate, as massive galaxies tend to be redder and smoother; visual inspection confirmed that a large fraction of star-forming clumps inserted into high-mass galaxies were easily distinguished as simulations. We allowed galaxies to appear in the simulated sample multiple times, applying a different image transformation (i.e., combination of rotations and reflections) to each cutout made of the galaxy to ensure that the final subject was unique.

To determine clump luminosity and color, we simulated clump spectra using spectral templates generated by the software Flexible Stellar Population Synthesis (Conroy et al. 2009; Conroy & Gunn 2010). We treated each clump as a single stellar population (with delta-function star formation history), assigning it an age and a V-band total extinction $A_V$. Clump ages spanned $10^{-2.5}$ Gyr (i.e., $\sim 3$ Myr) to $10^4$ Gyr. Dust extinction was given by a Calzetti et al. (2000) attenuation curve, and the clump’s $A_V$ value varied narrowly about the SDSS-estimated dust content for the host galaxy ($\sigma = 0.04$ in $\ln(A_V)$). Metallicity was fixed to the solar value. The resulting spectrum was redshifted to match the host galaxy and integrated to determine the clump’s broadband flux in each SDSS band.

Clump mass was selected uniformly from a distribution that depended on the clump age, shown in Figure 2. Qualitatively, we excluded low-mass, high-age clumps to avoid extremely faint (“known invisible”) objects. Similarly, we excluded the brightest clumps (high mass, low age), as these were “known visible” objects. The resulting sample probes the region where we are least certain of volunteers’ recovery capability. We discarded and regenerated a clump if its apparent magnitude was $>22.8$ in each of the $g$, $r$, $i$, and $z$ bands, as well as $>24.8$ in the $u$ band, since these clumps were considered “known invisible.” The resulting distribution in the age and mass of simulated clumps is shown in Figure 2.

Each galaxy in the simulated sample was assigned a number of clumps selected from a Poisson distribution with mean 3, where 0 results were rejected and resampled. (Since 0 results were rejected, the actual mean number of clumps per galaxy was 3.16.) This distribution was chosen to maximize the clump density per galaxy without appearing unrealistic to volunteers and is not meant to reflect the true distribution of clumps per galaxy. To assign locations to clumps, image segmentation was performed on the $r$-band imaging field containing the host galaxy using Source Extractor (Bertin & Arnouts 1996). For each field, a centered cutout was made, smoothed with a boxcar filter, and then segmented by Source Extractor to determine which pixels belonged to the host galaxy. A pixel was chosen uniformly at random from the host galaxy’s segment to be the clump’s location within the image. In order to probe the low surface brightness regions of a galaxy (without letting our sample be dominated by clumps in these regions), we randomly assigned 25% of clumps to a “wide” segment that encompassed a larger area than the standard galaxy segment.\footnote{For both the narrow and wide segments, the Source Extractor settings DETECT_THRESH = 16 and DEBLEND_MINCONT = 0.01 were used; the only parameter changed was the detection threshold DETECT_THRESH, which was set to 3 for the wide segment and 5 for the narrow. In addition, to generate the wide segment, the galaxy image was convolved with a boxcar filter of size 7 pixels, while the narrow image was convolved with a boxcar filter of size 3 pixels.}

Figure 2 provides an example of the simulation placement process. Each clump was then simulated as a point source with a $\gamma = 2.5$ Moffat profile; of the profiles we tried, it was found by visual inspection by the authors that a Moffat profile looked the most “real” to observers and was most difficult to distinguish from real clumps. For most clumps, the FWHM of this profile was equivalent to the FWHM of the $r$-band point-spread function (PSF) in the image. We also allowed a small number of clumps to be “extended” by assigning each clump an effective physical radius, selected from a uniform distribution over the range $[10$ pc, $500$ pc] (based on size limits observed by Messa et al. 2019). If this physical radius exceeded that of the PSF, this radius was used as the effective radius of the clump’s profile. Approximately 19% of simulated clumps (16,342 of 84,565) exceeded the seeing PSF size, by a median of 27%. The distribution of simulated clump sizes allows us to probe the effect of clump size on recovery statistics and is not intended to match the true distribution.

2.4. Aggregation Method

After subjects had been examined by Clump Scout volunteers, the locations of clumps needed to be determined from the collected annotations on each subject. To this end, we developed an aggregation algorithm with which all volunteer annotations on each subject were transformed into consensus clump locations. Broadly, our aggregation process consisted of two steps: first, clump candidates were identified via a clustering algorithm; second, clumps marked “unsual” by a sufficient fraction of volunteers ($>0.35$) were discarded, since these are likely to be foreground star contaminants.

The clustering algorithm we employed is adapted from Branson et al. (2017) and is specialized for citizen science applications. A complete description of the algorithm can be found in H. Dickinson et al. (2022, in preparation). Briefly, we assign each volunteer a “false-positive probability” (i.e., the chance that an annotation by this volunteer is not associated
The risk falls below a threshold value, clustering is no longer
volunteer-identi-
We additionally remove any clumps that coincide with the
clumps, while the majority of those below the threshold did.

improve the purity of our sample, we discard clumps with
false-positive probability

candidates; in turn, the identified clump candidates update the volunteer
statistics, and so on. To determine when each image has been
fully classi-
fed, a clump candidate is assigned a
false-positive probability

Figure 3. An illustration of the process of placing simulated clumps in an image. (a) The initial galaxy image, g band only. (b) The galaxy segmentation map. The dark-gray inner region is the “narrow” region where clumps are placed with 75% probability, while the dark- and light-gray regions together make up the “wide” region where clumps are placed with 25% probability. Selected locations for simulated clumps are plotted as crosses. (c) The final galaxy image with simulated clumps added (some are too faint to be detected). Dashed red lines are drawn around the simulations to highlight them. The final galaxy image shown to volunteers was rotated and/or reflected with respect to this one to distinguish its appearance from the original.

with any clump candidates), a “false-negative probability (i.e.,
the chance that the volunteer failed to mark a given clump
candidate), and a “scatter” value that estimates the distance
between an annotation and its intended target. These values
inform a clustering algorithm that identifies clump candidates;
in turn, the identified clump candidates update the volunteer
statistics, and so on. To determine when each image has been
fully classified, a “risk” value is computed for each one. Once
the risk falls below a threshold value, clustering is no longer
performed and its clump candidates are finalized. Each clump
candidate is assigned a “false-positive probability” based on the
number and properties of volunteers who marked it. To
improve the purity of our sample, we discard clumps with
false-positive probability >0.6. By comparison with a sample
of classifications by the authors, we found that the majority
of volunteer-identified clumps with false-positive probabilities
larger than 0.6 did not correspond to any expert-identified
clumps, while the majority of those below the threshold did.

We additionally remove any clumps that coincide with the
volunteer-identified central bulge of a galaxy. To locate a
galaxy’s central bulge, we take the median x and y coordinates
of central bulge annotations from all volunteers; we then prune
outlying annotations by removing any volunteer’s central bulge
annotation that falls more than 20 pixels from this location (in
the 400 × 400 cutout image), and we recalculate the central
bulge location without these. We then remove any clumps
within 1 PSF-FWHM of the central bulge.

In total (excluding unusual clumps), we identify 10,738
clumps over 7050 galaxies in our sample. An additional 3858
unusual clumps were found; although these are most likely
foreground contaminants, we include them in our public
catalog for transparency.

3. Clump Properties and Completeness

In this section, we will discuss our method for estimating the
flux and background of each clump identified by Clump Scout.

We estimated the flux of and galactic background for each
clump in each of the SDSS _ugriz_ bands using a method similar
to Guo et al. (2015). First, flux was measured in an aperture of
diameter 2.25 PSF-FWHM centered on the clump location.
Next, the background-per-pixel value (where “background”
refers to diffuse galaxy light) was estimated by taking the
median pixel value in an annulus spanning diameters 3–5 PSF-
FWHM and used to estimate the background flux in the central
aperture. Figure 4 provides a visual example of the aperture and
annulus sizes used. This background is subtracted to obtain a
clump flux estimate within the aperture. A random sample of
model PSFs from 1000 SDSS fields revealed that 84% ± 2% of
flux from a point source falls within an aperture of diameter
2.25 PSF-FWHM. We therefore multiplied the background-
substracted aperture flux by 1.191 to counteract this systematic.

There are a few known sources of systematic error in our
flux estimation process for simulated clumps. A small
percentage of flux (≤5%) is lost owing to a combination of
pixelation effects, contamination of the background region by
the clump, and offsets between the recovered and true locations
of clumps. In addition to these, the diffuse background flux is
slightly underestimated by the background annulus (median:
~85% of the true value). This results in overestimation of
clump fluxes, particularly for dimmer clumps. While these
systematics cannot fully be removed with the existing method,
we have minimized them so that they are smaller than the
scatter. It should be noted that these systematic effects may not
exactly match those for the real sample of clumps compared
with our simulations; as such, we simply minimize the
systematics and do not apply a correction to counteract them.
The bottom panel of Figure 5 visually quantifies the
effectiveness of this flux recovery method by displaying the
recovered versus input magnitudes of the simulated clump
sample and demonstrates that the systematic effects are less
than the scatter.

To estimate the error on clump fluxes, we first estimated the
per-pixel uncertainties in each SDSS field using the gain and
dark variance values provided for each CCD, as well as the
image calibration and sky image maps provided with each field.
We then fit an uncertainty model to all the pixels in each field
of the form $\sigma^2 = mf + b$, where $\sigma^2$ is the variance on a pixel’s
flux, $f$ is the pixel’s flux, and $m$ and $b$ are model parameters.
We sum the variances within a clump’s aperture to estimate the
aperture flux uncertainty, and we take the median pixel
variance in the background annulus and multiply this by the
aperture area to estimate the background uncertainty within the
aperture. Finally, to obtain the variance on the background-subtracted aperture flux, we sum the aperture and annulus variance estimates; we multiply this value by the same aperture correction (1.191) to estimate the uncertainty on the final, background-subtracted clump flux.

We also obtained a completeness estimate for each clump. For a given clump, its completeness estimate is the estimated fraction of clumps similar to it that Clump Scout recovered. To determine this, we relied on the recovery fractions of simulated clumps, where a simulated clump was considered “recovered” if Clump Scout volunteers located a clump within 0.75 PSF-FWHM of its location. We then examined the sample of simulated clumps with respect to three properties: a clump’s brightness (g-band magnitude), its color (g-minus-r magnitude), and its contrast against the diffuse background (clump-minus-background g-band magnitude). The simulated clumps were binned with respect to these properties, and the overall recovery fraction was calculated for each bin. We then

Figure 4. An illustration of the parameters used to estimate the flux from each clump. The left panel is a galaxy image shown to volunteers from the “real” sample (no simulated clumps), zoomed in to show detail. The right panel is an image that marks the locations of four recovered clumps. Solid outlines are drawn around the central apertures, and dashed lines are drawn around the background regions.

Figure 5. Here we illustrate the effectiveness of the aggregation and flux estimation process. For all simulated clumps that volunteers recovered, we compare the “input” properties of clumps (their original simulated values) with their “recovered” properties, following the aggregation process laid out in Section 2.4 and the flux estimation method in Section 3. Top: histogram of the offset between the recovered and true locations of simulated clumps recovered by volunteers. The median offset is 0.15 PSF-FWHM, as marked by the dashed line. Bottom: the flux recovery fraction (recovered flux/input flux) versus input magnitude for simulated clumps. Red points mark the binned medians in magnitude difference, while vertical bars span the 16th to 84th percentile flux recovery fractions. The median flux recovery fraction was 1.03 for clumps with simulated magnitude <22.5. There are a few sources of systematic error in the flux recovery process that could not be removed completely (described in Section 3).
calculated the same three properties for each real clump and compared with the recovery fractions of the simulated sample to obtain each clump’s completeness estimate. The three properties selected—brightness, color, and contrast—effectively predict the completeness of each clump, as we found that including additional properties, including clump galactocentric radius, galaxy redshift, galaxy size, or image resolution, had only a very small effect on our completeness estimates by comparison. The details of this process are described in Appendix B. Figure 6 shows the estimated recovery fraction statistics for real clumps in our sample.

We use these completeness estimates in this work to correct the fraction of clumpy galaxies \( f_{\text{clumpy}} \) for incompleteness (Section 4.3). Future studies relying on this clump catalog should generally incorporate these completeness estimates to accurately model the local population of clumps and should not rely solely on the observed number counts of clumps.

3.1. Catalog Release and Use

Along with the electronic release of this paper, we release the catalog of all of the clumps identified by the Clump Scout aggregator and their estimated properties. The columns are fully described by Table 3 at the end of this paper. Clumps with a high fraction (>35%) of “unusual” annotations are included but are marked by the flag `unusual_flag=1`.

To use this catalog for scientific purposes, we make a few “best practices” suggestions on how to filter this catalog:

1. Selecting a clean sample: We strongly recommend that clumps with `unusual_flag=1` should be rejected, as these are probable nonclump contaminants (i.e., foreground stars, background galaxies, or other point-like sources).

2. Mass completeness: This galaxy catalog is not mass complete, and the lower-limit mass on galaxies evolves significantly with redshift. We estimate that the catalog is mass complete down to \( 10^9 M_\odot \) for \( z < 0.035 \). If a mass-complete catalog is needed, care must be taken to limit the sample’s redshift.

4. The Local Clumpy Fraction of Galaxies

A particularly important observable in the study of clumpy galaxies is the fraction of star-forming galaxies with at least one clump, known as the “clumpy fraction” or \( f_{\text{clumpy}} \). The clumpy fraction is most simply defined by

\[
    f_{\text{clumpy}} = \frac{N(\text{SFGs with } \geq 1 \text{ clump})}{N(\text{SFGs})},
\]

Figure 6. The estimated recovery fraction for clumps as a function of the three clump properties \( m_g \text{clump} \), \( m_g \text{clump} - m_g \text{clump} \), and \( m_g \text{clump} - m_g \text{background} \). Plots at the top of each column show the median estimated recovery fraction for clumps as a function of each property (red line), with 68% scatter (red shaded region). Completeness estimates are not shown for bins with <50 clumps; the dashed black lines trace the number density of clumps as a function of each property. Plots below the diagonal show the 2D dependence of estimated recovery fraction vs. pairs of clump properties, where bins with <5 clumps are not shown.
where SFGs refer to star-forming galaxies (specific star formation rate (sSFR) > 0.1 Gyr⁻¹). This is an easily compared observable between different galaxy populations that can significantly constrain models of galaxy evolution. In this section, we establish a clump definition that makes estimating \( f_{\text{clumpy}} \) straightforward and precise, and then we present our \( f_{\text{clumpy}} \) estimate.

4.1. Selecting a Clump Definition for \( f_{\text{clumpy}} \)

A major difficulty in clumpy galaxy literature is that the definition of a “clump” is highly inconsistent. Past works have defined clumps as those objects identified by visual investigation (e.g., Elmegreen et al. 2007; Puech 2010; Overzier et al. 2009) or by applying detection algorithms that are robust to changes in resolution and depth, including the clumpfind algorithm from Williams et al. (1994) and others (e.g., Livermore et al. 2012; Guo et al. 2012; Tadaki et al. 2014; Zanella et al. 2019). However, comparisons between these different methods are not straightforward.

Guo et al. (2015) proposed an empirically motivated definition that the ratio of clump to galaxy UV luminosity (\( f_{\text{LUV}} \)) must exceed 8%. The 8% cutoff was chosen to select for star-forming regions at high redshift while excluding common star-forming regions locally; specifically, it includes many star-forming regions in HST-imaged galaxies spanning 0.5 < \( z < 3 \) but excludes >99% of star-forming regions identified in the galaxy M101 (blurred to match the resolution of the high-redshift sample). Local clumps exceeding the \( f_{\text{LUV}} > 8\% \) threshold are therefore expected to be rare, exceptional objects.

Several other recent works use this or a similar criterion (e.g., Shibuya et al. 2016; Mandelker et al. 2017; Fisher et al. 2017). However, it is not universally applicable: Dessauges-Zavadsky & Adamo (2018) rejected it to allow for study of the mass function of high-redshift clumps down to much lower masses, while Huertas-Company et al. (2020) used a clump mass cut of \( M_{\text{clump}} > 10^7 M_\odot \) instead to facilitate comparison with simulations.

In this paper, we calculate \( f_{\text{clumpy}} \) by specifying a relative flux cutoff in the SDSS \( u \) band, i.e., the ratio of clump to galaxy flux in the \( u \) band (\( f_{Lu} \)) is greater than some specified fraction. In particular, we use the relative flux cuts \( f_{Lu} > 8\% \) and \( f_{Lu} > 3\% \), and call the clumpy fractions under these criteria \( f_{\text{clumpy},8\%} \) and \( f_{\text{clumpy},3\%} \), respectively. The 8% fraction was selected to be comparable to existing works with a similar criterion, while the 3% fraction allows for larger number statistics and is easier and more accurate to estimate for the low-redshift universe, where clumps are less common.

Regarding the relationship between \( f_{Lu} \) and \( f_{\text{LUV}} \), it is worth noting that several past studies have used a relative flux threshold in the UV (\( \sim 2500 \, \AA \)) to define clumps (e.g., Guo et al. 2015; Shibuya et al. 2016). Because UV data are not available in SDSS, we instead use the lowest available wavelength band, the \( u \) band at \( \sim 3500 \, \AA \). Therefore, in order to compare our \( f_{\text{clumpy}} \) value with past studies, we must demonstrate that clumps selected under our \( f_{Lu} \) cut are comparable to those selected under the more commonly used \( f_{\text{LUV}} \) cut. To do so, we examined a highly complete sample of clumps from the Guo et al. (2018) sample, requiring \( f_{\text{LUV}} > 5\% \) to select for bright clumps and a highly complete sample. This sample is herein called the Guo+2018 sample. Guo+2018 was chosen because it includes a large sample of bright clumps (523 clumps meet the \( f_{\text{LUV}} > 5\% \) criterion) and because it spans a similar physical resolution to Clump Scout: a typical SDSS \( g \)-band PSF-FWHM at \( z \sim 0.05 \) is \( \sim 1.2 \, kpc \), compared with \( \sim 1.1–1.3 \, kpc \) for HST sources spanning \( 1 < z < 2.5 \) for similar wavelengths.

For each clump in the Guo+2018 sample, we examined its flux fraction in CANDELS filter bands that were analogous to near-UV and the SDSS \( u \) band in the rest frame. We find that there is a strong correlation between \( f_{Lu} \) and \( f_{\text{LUV}} \) using these filters, with the median clump having \( f_{Lu} = 0.86 f_{\text{LUV}} \). A total of 1170 clumps from the Guo+2018 sample meet the \( f_{\text{LUV}} > 8\% \) criterion, compared with 961, which meet \( f_{Lu} > 8\% \), a reduction of \( \sim 18\% \).

We performed a similar experiment on clumps in Clump Scout. Though we could not calculate \( f_{\text{LUV}} \) directly since SDSS does not provide near-UV data (~2500 Å), we obtained near-UV fluxes for local galaxies from the GALEX survey (Martin et al. 2005), using the cross-matched GALEX-SDSS catalog created by Bianchi & Shiao (2020). We then assumed that local clumps matched the SED distribution of clumps from the Guo+2018 sample and multiplied each clump’s \( u \)-band flux by the UV-to-\( u \) ratio of a randomly selected Guo+2018 clump. The resulting values for \( f_{Lu} \) and \( f_{\text{LUV}} \) are plotted alongside the Guo+2018 values in Figure 7. For both groups, \( f_{Lu} \) is a reasonably strong predictor of \( f_{\text{LUV}} \). Performing the same experiment using the \( g \) band rather than the \( u \) band revealed that \( f_{Lu} \) values are not as well correlated to \( f_{\text{LUV}} \) and are typically much smaller (the median clump had \( f_{Lu} = 0.73 f_{\text{LUV}} \)).

Based on these experiments, we conclude that \( f_{Lu} \) is the best available analog of \( f_{\text{LUV}} \) for SDSS data. We choose not to apply a conversion factor to convert \( f_{Lu} \) values to \( f_{\text{LUV}} \), since no statistically large sample of UV-measured local clumps exists that could verify such a conversion factor. Further, our results in this section suggest that \( f_{Lu} \) and \( f_{\text{LUV}} \)-defined values of \( f_{\text{clumpy}} \) are closely related and can be meaningfully compared, with the caveat that \( u \)-band-estimated values of \( f_{\text{clumpy}} \) are likely to be slightly smaller than UV-estimated values.

4.2. Calculating \( f_{\text{clumpy}} \)

The calculation of \( f_{\text{clumpy}} \) from the Clump Scout catalog has several steps. We calculate \( f_{\text{clumpy}} \) in the full sample, as well as in three distinct bins of galaxy mass, and the selection cuts and galaxy counts for each mass bin are detailed in Table 2. Here, we detail the steps for calculating \( f_{\text{clumpy}} \) in the broadest mass bin (\( M > 10^9 M_\odot \)), though the same process applies to every bin.

First, we isolate a star-forming, mass-complete sample of galaxies. SDSS is complete for galaxies down to \( 10^9 M_\odot \) at redshifts \( z < 0.035 \). Therefore, beginning with the Clump Scout parent sample defined in Section 2.2, we limit the sample to galaxies with sSFR > \( 10^{-1} \), \( M > 10^9 M_\odot \), and \( z < 0.035 \). In addition, because clumps may be more difficult to detect in edge-on galaxies than face-on galaxies, we remove all galaxies for which the ratio of the galaxy’s major to minor axis is less than 0.3, where this axis ratio is estimated by the SDSS exponential fit in the \( r \) band (\( e^{\expAB \cdot x} \) in the PhotoPrimary

---

14 The filters selected for UV were F435W for \( 0.5 < z < 1 \), F606W for \( 1 < z < 2 \), and F775W for \( 2 < z < 3 \), matching the scheme in Guo et al. (2015). We selected SDSS \( u \)-band analog filters to be closest to 3550 Å in the rest frame: F606W for \( 0.5 < z < 0.9 \), F775W for \( 0.9 < z < 1.2 \), F814W for \( 1.2 < z < 1.6 \), F105W for \( 1.6 < z < 2.2 \), and F125W for \( 2.2 < z < 3 \).
Our sample contains \( N_{\text{tot}} = 5843 \) galaxies passing these cuts.

We then combine the contributions from the “regular” Clump Scout sample of 53,613 galaxies and the “extra” sample of 4937 galaxies used to extrapolate over all galaxies in SDSS that Clump Scout did not directly examine. In total, \( N_{\text{reg}} = 2640 \) of 5843 galaxies passing all cuts for \( f_{\text{clumpy}} \) were examined in the regular sample.

Of the remaining 3203 galaxies, a sample of 230 were examined by volunteers as part of the “extra” sample. We use \( N_{\text{extra,samp}} \) to refer to the size of the sample of these galaxies that volunteers examined directly, i.e., \( N_{\text{extra,samp}} = 230 \), and \( N_{\text{extra,tot}} \) to refer to the size of the total population from which these galaxies were drawn, i.e., \( N_{\text{extra,tot}} = 3203 \).

For each group, the observed clumpy fraction \( f_{\text{clumpy},\text{obs}} \) is calculated and the completeness correction from Section 4.3 is applied. The corrected fraction \( f_{\text{clumpy},\text{corr}} \) over the “extra” sample is then extrapolated over all SDSS galaxies not examined by Clump Scout. This yields the total clumpy fraction:

\[
f_{\text{clumpy}} = \frac{N_{\text{reg}}}{N_{\text{tot}}} f_{\text{clumpy},\text{reg}} + \frac{N_{\text{extra,tot}}}{N_{\text{tot}}} f_{\text{clumpy},\text{extra}}.
\]

Table 2

<table>
<thead>
<tr>
<th>Parent sample</th>
<th>( z_{\text{max}} )</th>
<th>Regular</th>
<th>Extra (Examined)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>With sSFR &gt; 0.1 Gyr(^{-1})</td>
<td>53.613</td>
<td>168,603</td>
<td>49,373</td>
<td>222,216</td>
</tr>
<tr>
<td>With b/a ratio &gt; 0.3</td>
<td>12,671</td>
<td>20,969 (1,015)</td>
<td>33,640</td>
<td></td>
</tr>
<tr>
<td>All masses ((10^9 &lt; M/M_\odot))</td>
<td>0.035</td>
<td>2640</td>
<td>3203 (230)</td>
<td>5843</td>
</tr>
<tr>
<td>Low-mass galaxies ((10^8 &lt; M/M_\odot &lt; 10^{9.8}))</td>
<td>0.035</td>
<td>2042</td>
<td>2898 (118)</td>
<td>4940</td>
</tr>
<tr>
<td>Medium-mass galaxies ((10^9 &lt; M/M_\odot &lt; 10^{10}))</td>
<td>0.05</td>
<td>1741</td>
<td>1444 (149)</td>
<td>3185</td>
</tr>
<tr>
<td>High-mass galaxies ((10^{10} &lt; M/M_\odot &lt; 10^{11.4}))</td>
<td>0.09</td>
<td>601</td>
<td>720 (203)</td>
<td>1321</td>
</tr>
</tbody>
</table>

Note. “Regular” galaxies were fully examined by Clump Scout volunteers, while “extra” galaxies (with \( f_{\text{clumpy}} \leq 0.5 \)) were only partially examined (the number of examined galaxies is given in parentheses); the “total” column sums the regular and extra columns. Row 1 presents the parent sample of galaxies from Table 1. (The number of galaxies here is slightly smaller than that in Table 1, as galaxies with \( f_{\text{featured}} > 0.5 \) and \( f_{\text{featured}} > 0.5 \) were not examined and are left out of the total.) Rows 2 and 3 enumerate the sample removing quiescent galaxies and edge-on galaxies, respectively. The final four rows detail the four mass-binned samples used for calculating \( f_{\text{clumpy}} \).

The sampling error is estimated separately on the “regular” and “extra” samples using the standard error formula on a proportion, taking the sample size \( N \) to be the number of examined galaxies in the “regular” or “extra” group:

\[
\sigma_{\text{clumpy,reg}} = \sqrt{\frac{f_{\text{clumpy,reg}}(1 - f_{\text{clumpy,reg}})}{N_{\text{reg}}}}
\]

\[
\sigma_{\text{clumpy,extra}} = \sqrt{\frac{f_{\text{clumpy,extra}}(1 - f_{\text{clumpy,extra}})}{N_{\text{extra,samp}}}}.
\]

We then scale these by their contribution to the total value of \( f_{\text{clumpy}} \) and add them in quadrature to obtain

\[
\sigma_{\text{clumpy}}^2 = \left( \frac{N_{\text{reg}}}{N_{\text{tot}}} \right)^2 \sigma_{\text{clumpy,reg}}^2 + \left( \frac{N_{\text{extra,tot}}}{N_{\text{tot}}} \right)^2 \sigma_{\text{clumpy,extra}}^2.
\]

To estimate the total error on \( f_{\text{clumpy}} \), we use a Monte Carlo method to include contributions from the uncertainty on clump fluxes and clump incompleteness, as well as from sampling error. Over 100 trials, clump fluxes are allowed to randomly vary within a normal distribution defined by their estimated error values (see Section 3), while the clump completeness map is recalculated on each trial by the method described in Appendix B. On each Monte Carlo trial, we include sampling error by calculating an initial value of \( f_{\text{clumpy}} \) and then reassigning it to a random value selected from \( N(f_{\text{clumpy}}, \sigma_{\text{clumpy}}) \). Of these sources of error, sampling error is by far the most significant: in a trial run where error contributions from clump flux error and the completeness correction were ignored, the error bars were typically within 10% of their original values.
4.3. Correcting $f_{\text{clumpy}}$ for Incompleteness

Our estimate of $f_{\text{clumpy}}$ must also take into account the incompleteness of our clump catalog. We therefore use the following method for completeness-correcting the clumpy fraction of galaxies.

The end goal of the $f_{\text{clumpy}}$ completeness correction is to calculate $P_{\text{FN}}$, the probability that a galaxy is a “false negative” for clumps—in other words, the probability that the galaxy contains one or more clumps, but that none of its clumps were detected. To calculate $P_{\text{FN}}$, we work in steps from the completeness estimates on individual clumps in the sample. We define $P_{\text{rec},i}$ to be the recovery fraction of clump $i$. (Refer to Appendix B for a full overview of how $P_{\text{rec},i}$ is determined for each clump.)

We begin by calculating the recovery probability of a randomly selected clump within the sample, $P_{\text{rec}}$:

$$P_{\text{rec}} = \frac{n_{\text{obs}}}{n_{\text{true}}} \approx \frac{n_{\text{obs}}}{\sum_{i=1}^{n_{\text{obs}}} 1/P_{\text{rec},i}},$$  

where for each clump $i$, $1/P_{\text{rec},i}$ approximates the true number of clumps with similar properties.

Next, we estimate the true distribution of clumps per clumpy galaxy, $P_{\text{count}}(n_e)$. To do so, we begin with a proposed distribution $P_{\text{count}}(n_e)$ and then simulate 10,000 clumpy galaxies with this distribution. We then remove a fraction $P_{\text{rec}}$ of clumps at random to simulate the observed distribution, discarding galaxies with no observed clumps. The proposed distribution is adjusted until the mean number of clumps per galaxy in the simulated distribution closely matches the observed mean. For mathematical expediency, we used an exponential distribution to model $P_{\text{count}}(n_e)$ with coefficient $\lambda$, i.e., $P_{\text{count}}(n_e) \sim e^{-\lambda n_e}$. (It should be noted, though, that the observed distributions are also well fit by Poisson distributions, and the exponential model does not necessarily have physical significance.) In Figure 8, we plot the observed distributions of $f_{\text{Lu}} > 3\%$ clumps per galaxy for each galaxy mass bin we examine, along with the best-fit exponential model.
Given $P_{\text{rec}}$ and $P_{\text{count}}(n_c)$, $P_{\text{FN}}$ is given by

$$P_{\text{FN}} = \sum_{n_c=1}^{\infty} P_{\text{count}}(n_c) (1 - P_{\text{rec}})^{n_c}. \quad (5)$$

The $P_{\text{FN}}$ sum is dominated by the first few terms. For $f_{Lu} > 8\%$ clumps in our broadest galaxy mass bin ($M > 10^9 M_\odot$ with $z < 0.035$), we estimate $P_{\text{rec}} \approx 49.3\%$ and $\lambda \approx 0.85$; given these values, the first three terms of the $P_{\text{FN}}$ sum account for approximately 78\%, 92\%, and 99\% of missed clumpy galaxies, respectively.

Given the galaxy false-negative probability $P_{\text{FN}}$, the completeness of $f_{\text{clump}}$ is given by $1 - P_{\text{FN}}$. It is then straightforward to correct the clumpy fraction:

$$f_{\text{clumpy}}^{\text{corr}} = (1 - P_{\text{FN}})^{-1} f_{\text{clumpy}}^{\text{obs}}. \quad (6)$$

Parameter $f_{\text{clumpy}}^{\text{corr}}$ should be taken as our estimate of the “true” value, as it accounts for galaxies whose clumps were not detected; however, we present both the observed and corrected values in our results.

### 4.4. Results

Here we present our results for $f_{\text{clumpy}, 8\%}$ and $f_{\text{clumpy}, 3\%}$ (the clumpy fraction using the thresholds $f_{Lu} > 8\%$ and $f_{Lu} > 3\%$, respectively), both overall and within several different galaxy mass bins. All of these numbers are collected in Table 4.

Within the regular Clump Scout sample ($N_{\text{reg}} = 2640$), we detected 104 galaxies with clumps passing $f_{Lu} > 8\%$; corrected for incompleteness, we estimate the true number to be $\sim 156$. Within the extra sample ($N_{\text{extra}} = 230$), we observed just two galaxies with clumps passing $f_{Lu} > 8\%$, and we estimate the true number to be $\sim 3$ correcting for incompleteness. In total, we estimate that 188 of $N_{lu} = 5843$ galaxies have clumps (corrected for incompleteness), leading to an estimate $f_{\text{clumpy}}^{\text{corr}} = 3.22^{+0.34}_{-0.34}\%$.

We apply the same procedure to galaxies using the $f_{Lu} > 3\%$ cut and observe 410 clumpy galaxies in the regular sample ($\sim 641$ corrected for incompleteness) and 5 clumpy galaxies in the extra sample ($\sim 9$ corrected for incompleteness). This yields a total estimate $f_{\text{clumpy}}^{\text{corr}} = 12.68^{+1.38}_{-0.88}\%$.

To characterize the local distribution of clumps, we also examine $f_{\text{clumpy}}$ in three different bins of galaxy mass: $9 < \log_{10}(M/M_\odot) < 9.8$, $9.8 < \log_{10}(M/M_\odot) < 10.6$, and $10.6 < \log_{10}(M/M_\odot) < 11.4$; we refer to the galaxies in these bins as low-mass, medium-mass, and high-mass galaxies, respectively. These match the mass bins over which Guo et al. (2015) estimated the clumpy fraction. To obtain a complete galaxy sample, different redshift limits were applied to each bin: $z < 0.035$ for low-mass galaxies, $z < 0.05$ for medium-mass galaxies, and $z < 0.09$ for high-mass galaxies (see Figure 9).

Following the same procedure as for the overall clumpy fraction, we estimate that $f_{\text{clumpy}}^{\text{corr}}$ is $3.12^{+0.26}_{-0.26}\%$ for low-mass galaxies, $2.48^{+0.30}_{-0.33}\%$ for medium-mass galaxies, and $2.57^{+0.56}_{-0.57}\%$ for high-mass galaxies. A more complete list of statistics can be found in Table 4; they are also plotted in Figure 10. Example images of galaxies in each of the mass bins are shown in Figure 11.

### 4.5. Comparisons to Other Studies

To place our estimates of $f_{\text{clumpy}}^{\text{corr}}$ at $z < 0.1$ in context, we compare them with high-redshift ($z > 0.5$) results from other works, in particular Shibuya et al. (2016) for galaxies of all masses $M > 10^9 M_\odot$ and Guo et al. (2015) for galaxies in bins of low, medium, and high mass (matching the mass bins used in this paper). We plot these values in Figure 12.
Shibuya et al. (2016) found that the clumpy fraction peaks between redshifts 1 and 2 at a value of >50%, before declining over $z < 1$. To model this trend, they use a fit function with the same form as is commonly used to model the trend in the cosmic SFR density with redshift (Madau et al. 1996; Lilly et al. 1996). To compare with their results, we take their best-fit model for $f_{\text{clumpy}}$ and extend it beyond their data to $z \sim 0$; this is plotted in Figure 12. Their model predicts a value of $f_{\text{clumpy}} \approx 4\%$ at $z \sim 0$, which is in line with our result of $f_{\text{clumpy}} \approx 3.32\%$.

Guo et al. (2015) observed a nearly constant $f_{\text{clumpy}} \approx 50\%$ for low-mass galaxies over $0.5 < z < 3$, while our results indicate a value of $\sim 3\%$ at $z \sim 0$ for galaxies of the same mass.

As explained in Section 4.1, these results can be meaningfully compared because the $f_{\text{Lu}} > 8\%$ cut used in this work is similar to the $f_{\text{Lu}} > 8\%$ cut used by Guo et al. (2015). We note that the differences between our $z < 0.1$ results for $f_{\text{clumpy}}$ and the $z > 0.5$ results from Guo et al. (2015) are an order of magnitude or more, which is much larger than the difference we expect between $f_{\text{clumpy}}$ computed with a UV versus $\alpha$-band clump definition (as discussed in Section 4.1). These two works also probe a similar physical resolution; the physical resolution of CANDELS images is $\sim 1$ kpc at all redshifts, compared to $\sim 0.5$–$1.7$ kpc over the range $0.02 < z < 0.09$ for SDSS. We therefore expect that this drop in the clumpy fraction between $0 < z < 0.5$ is real and cannot be the result of different identification methods.

There are few other studies that estimate $f_{\text{clumpy}}$ in the local universe. Murata et al. (2014) studied clumpy galaxies selected from HST/ACS F814W imaging from the COSMOS field spanning redshifts $0.2 < z < 1$. The F814W filter approximately corresponds to the SDSS $r$ and $g$ bands over this redshift range, and the nearest galaxies in this sample ($z \sim 0.2$) are likely similar to Clump Scout galaxies ($z < 0.1$). The fraction of optically bright galaxies with multiple star-forming clumps was found to decrease from 0.35 at $z \sim 1$ to 0.05 at $z \sim 0.2$. While this decrease is qualitatively in line with our results, the definition of $f_{\text{clumpy}}$ used by Murata et al. (2014) is...
significantly different from that used here: rather than apply a relative flux criterion, clumpy galaxies were selected to have multiple star-forming clumps of comparable brightness. The low completeness of our sample prevents us from applying the Murata et al. (2014) condition, as it requires the detection of at least three clumps per clumpy galaxy. Overzier et al. (2009) also studied clumpy galaxies at $z<0.3$ but only studied a sample of 30 “Lyman break analog” galaxies with extremely

Figure 11. A sample of galaxies, each containing at least one clump with $f_{Lu} > 8\%$. Ten galaxies are presented from each of the low-, medium-, and high-mass bins; the SDSS specobjid of each galaxy is provided at the top of each image, while the R.A. and decl. values are provided in parentheses at the bottom. Each identified clump is circled in yellow, while the central bulge locations identified by volunteers are identified by red crosses. Dotted circles indicate clumps with $f_{Lu} < 3\%$, dashed circles indicate those with $3\% < f_{Lu} < 8\%$, and solid circles those with $f_{Lu} > 8\%$. 

high UV fluxes; the clumpy fraction obtained from this sample is not comparable to that of our broader sample.

5. Discussion

5.1. Interpretation of $f_{\text{clumpy}}$

Given this paper’s focus on the fraction of clumpy galaxies, it is worth discussing exactly how this quantity is defined and how it should be used. Parameter $f_{\text{clumpy}}$ is a particularly good probe of trends in clumpiness across cosmic time. By controlling for galaxy mass and SFR, we ensure that $f_{\text{clumpy}}$ is computed between groups of similar galaxies even at different redshifts. However, using a relative flux criterion ($f_{\text{LUV}} > 8\%$) to define clumps may select for very different sets of physical objects depending on galaxy mass. Naïvely, assuming a linear relation between mass and $u$-band luminosity in our galaxy sample, our lowest galaxy mass bin ($10^9 - 10^{9.8} M_\odot$) includes clumps that are $\sim 2$ dex less massive than our highest-mass bin ($10^{10.6} - 10^{11.4} M_\odot$). It is therefore not straightforward to compare $f_{\text{clumpy}}$ between bins of different galaxy mass. The validity of this comparison depends on the clump luminosity function: for example, if the clump luminosity function experiences an exponential cutoff (e.g., as proposed by Livermore et al. 2012) over three galaxy mass bins. We find that the turbulent fraction qualitatively matches the patterns observed in $f_{\text{clumpy}}$, i.e., that it declines significantly over $0 < z < 1.5$ and that high-mass galaxies begin this decline the soonest. By comparison, we find that the minor merger fraction remains approximately constant over the same time period and is a poor tracer of $f_{\text{clumpy}}$ for any mass bin when including our results at $z \sim 0$.

Figure 12. The value of $f_{\text{clumpy,8\%}}$ vs. redshift, as estimated by our study at $z \sim 0$ and others at $z > 0.5$. Top: $f_{\text{clumpy}}$ for galaxies of all masses ($> 10^9 M_\odot$). Our result is in line with the model by Shibuya et al. (2016) that was originally fit to their high-redshift $f_{\text{clumpy}}$ results. Bottom: $f_{\text{clumpy}}$ divided into mass bins. The mass bins and methods used in this work match closely with those used by Guo et al. (2015) to estimate $f_{\text{clumpy}}$ at $z > 0.5$, but our clumpy fractions at $z \sim 0$ are significantly lower than those at higher redshift. For comparison, we have plotted estimates of the minor merger fraction and observations of the “turbulent fraction” of galaxies between $0 < z < 1.5$. The minor merger fraction is modeled by Lotz et al. (2011) and plotted here with observability timescales of 0.5, 1.25, and 2 Gyr (with gray error regions containing the best-fit range). The turbulent fraction comes from kinematic observations by Kassin et al. (2012) over three galaxy mass bins. We find that the turbulent fraction qualitatively matches the patterns observed in $f_{\text{clumpy}}$, i.e., that it declines significantly over $0 < z < 1.5$ and that high-mass galaxies begin this decline the soonest. By comparison, we find that the minor merger fraction remains approximately constant over the same time period and is a poor tracer of $f_{\text{clumpy}}$ for any mass bin when including our results at $z \sim 0$.

5.2. Physical Implications of $f_{\text{clumpy}}$

A major motivation for determining $f_{\text{clumpy}}$ over large redshift ranges is to distinguish between different proposed
modes of clump formation. There are two primary modes by which clumps are thought to form. In the in situ mode, clumps form owing to gas collapse within the host galaxy due to turbulent disk dynamics (i.e., VDI). VDI is expected in galaxies that are actively accreting gas via “cold-mode” accretion, in which gas flows into the galaxy via smooth, cold streams. This accretion process adds kinetic energy to the disk and can drive the Toomre parameter below unity, making gas unstable to collapse (Dekel et al. 2009). Alternatively, in the ex situ mode of formation, clumps originate as minor mergers; the clump forms as a satellite galaxy with its own dark matter component, only later merging with its host. It should be noted that clumps are short-lived structures on a cosmological scale: simulations find that massive clumps in disk galaxies have a maximum lifetime of \( \lesssim 500 \, \text{Myr} \), by which time they are slowed owing to dynamical friction and have merged with their host’s central bulge (Bournaud et al. 2014; Mandelker et al. 2014). Therefore, the presence of clumps indicates that the clump formation process is ongoing or recent, and \( f_{\text{clumpy}} \) can effectively act as a tracer of galaxy behavior. It remains unclear which is the dominant formation process, as different processes may dominate different galaxy populations.

To determine the primary formation process of clumps (i.e., via in situ or ex situ formation), we can examine trends in the rate of VDI and the minor merger rate over cosmic time and compare these with trends in \( f_{\text{clumpy}} \). In Figure 12, we have plotted our estimate of \( f_{\text{clumpy}, 85\%} \), along with comparable estimates at higher redshift (Guo et al. 2015; Shibuya et al. 2016) and estimates of the fraction of galaxies experiencing VDI and with observable signatures of minor mergers.

We use the minor merger rate estimate from Lotz et al. (2011), which was obtained by subtracting the number of galaxies with close pairs (major mergers) from the number with disturbed, uneven morphologies (major and minor mergers). The best-fit model to the minor merger fraction takes the form \( f_{\text{merg, minor}} \propto T_{\text{obs}}(1+z)^{\alpha} \), with best-fit exponent \( \alpha = 0.2 \pm 0.2 \). The “observability timescale” refers to the time during which the host galaxy’s morphology is measurably disturbed, which is dependent on the detection method (and distinct from the lifetime of an ex situ clump formed during a merger). To represent the uncertainties in these parameters, we plot the best-fit model for \( T_{\text{obs}} \) values of 0.5, 1.25, and 2 Gyr over the fit range \( 0 < z < 1.5 \). In all cases, the minor merger rate rises or remains constant over the full redshift range owing to the fit parameter \( \alpha = -0.1 \pm 0.1 \).

To determine the fraction of galaxies experiencing turbulence, we use measurements of galaxy kinematics from Kassin et al. (2012). These measurements reveal that galaxies of a wide range of masses \( (10^{9} - 10^{10.7}) M_\odot \) tend to “settle” and become rotationally dominated over the period \( 0 < z < 1.2 \). Moreover, they find that the highest-mass galaxies have the lowest fraction of turbulence at any epoch. The “turbulent fraction,” defined as the fraction of galaxies for which \( V_{\text{circ}}/\sigma_{\text{gas}} < 3 \) (i.e., the fraction of galaxies experiencing VDI), is plotted in Figure 12 for three of the mass bins examined by Kassin et al. (2012), spanning \( 10^{9} - 10^{10.7} M_\odot \). All turbulent fractions decline over \( 0 < z < 1.2 \), with higher-mass galaxies declining more quickly.

We then turn to trends in \( f_{\text{clumpy}, 85\%} \) and compare them to the trends in the two clump formation mechanisms (VDI and minor mergers) described above. Ignoring our low-redshift data for a moment, the data from Guo et al. (2015) suggested that two different clump formation mechanisms may be dominant in high-mass galaxies and low-mass galaxies. The clumpy fraction for high-mass galaxies declines significantly with time over the span \( 0.5 < z < 3 \) from \( \sim 55\% \) to \( \sim 15\% \), while for low-mass galaxies it remains constant at \( f_{\text{clumpy}} \sim 60\% \) over the same time span. To explain this difference, it was suggested that the primary formation mechanism for clumps in high-mass galaxies may be VDI (in situ) and trace the turbulent fraction over this time span, while those in low-mass galaxies form owing to minor mergers (ex situ) that are roughly stable over the same time span.

However, our low-redshift estimates of \( f_{\text{clumpy}} \) challenge this two-mechanism formation model. For galaxies of all masses, we now observe a significant decline in \( f_{\text{clumpy}, 85\%} \) to \( \lesssim 5\% \) at \( z \sim 0 \). Even assuming that our \( z \sim 0 \) estimates of \( f_{\text{clumpy}, 85\%} \) are too small by a factor of several, the observed fraction would still be far lower at \( z \sim 0 \) than at \( z > 0.5 \) in every mass bin. This result matches the conventional wisdom about clumps, i.e., that giant star-forming clumps are common at high redshift and rare locally. However, the roughly constant minor merger rate over the period \( 0 < z < 0.5 \) is inconsistent with the significant low-redshift decline that we observe.

Instead, we suggest that in situ, VDI-driven formation is the primary mode of clump formation in galaxies of all masses, at least over the redshift range \( 0 < z < 1.5 \). The trends in galaxy turbulence over this time span match closely with the trends in \( f_{\text{clumpy}} \); all galaxies show evidence of a decline in turbulence, with low-mass galaxies remaining turbulent the longest. The VDI-driven formation model provides a natural mechanism for the decline in \( f_{\text{clumpy}} \), which is the decline in the cosmological rate of gas accretion by galaxies: as the availability of intergalactic gas decreases, so too do the rates of star formation, turbulent dynamics, and clump formation (Dekel et al. 2009).

Adding to this picture, smaller case studies have already provided limited evidence to link VDI to clumpiness directly. Studies of the kinematics of high-redshift clumpy galaxies find that they have turbulent morphologies (Elmegreen et al. 2009; Genzel et al. 2011), with Genzel et al. (2011) in particular noting that clumps appear in regions of the galaxy where the Toomre instability parameter is subunity. However, these studies examined spirals with stellar masses \( \gtrsim 10^{10.6} M_\odot \), corresponding to the highest-mass bin in our work; similar kinematic examination of galaxies with lower masses remains to be done. In total, the current body of evidence points to a picture of clump formation that is dominated by in situ formation due to turbulent disk dynamics, though more concrete evidence is needed to confirm this.

6. Summary and Conclusions

In this work we present the largest catalog yet of local star-forming clumps \( (z \lesssim 0.1) \), consisting of 10,738 clumps in 7050 galaxies. The clumps were identified via the citizen science project Galaxy Zoo: Clump Scout, which asked volunteers to identify star-forming clumps in a sample of 58,550 galaxies selected from the parent sample Galaxy Zoo 2. Consensus locations for these clumps are determined via an aggregation technique adapted from Branson et al. (2017). We estimate the completeness of our clump sample by comparing with a sample of simulated clumps identified via the same process.

The clump catalog generated by this work is versatile and can be used for many purposes. While this paper focused on estimating the clumpy fraction of galaxies, the catalog can also...
be used to answer other questions about clumps. In follow-up work, we intend to investigate the mass and age functions of these clumps using photometric SED fitting. Obtaining these statistics will permit comparison between the properties of clumps at low and high redshift, and they can be also used to directly test theories of clump formation and evolution that make predictions on the mass or age distribution of clumps.

We define two different measures of the clump fraction of galaxies, \( f_{\text{clumpy,8}} \) and \( f_{\text{clumpy,3}} \), which measure the fraction of galaxies with at least one clump emitting at least 8% and 3% of the galaxy flux in the filter (\( f_{\text{Lu}} \)) and (\( f_{\text{Lu}} \)), respectively. Parameter \( f_{\text{Lu}} \) is found to be the closest analog to \( f_{\text{Lu}} \) (the fraction of galaxy flux emitted in the near-UV) available in SDSS data. Parameter \( f_{\text{clumpy,8}} \) is presented because it has been used in the past for high-redshift studies of clumpy galaxies, while \( f_{\text{clumpy,3}} \) is presented for comparison with future local studies. Both fractions are corrected for incompleteness. We find \( f_{\text{clumpy,8}} = 3.22^{+0.38}_{-0.34} \% \) and \( f_{\text{clumpy,3}} \sim 12.68^{+1.38}_{-0.88} \%, \) with considerable variation over different mass bins.

Our low value of \( f_{\text{clumpy,8}} \) is qualitatively in line with other low-redshift surveys (e.g., Murata et al. 2014), though few are available. It is, however, much lower than the values of \( f_{\text{clumpy,8}} \) estimated at high redshift (Guo et al. 2015; Shibuya et al. 2016). We suggest that the extreme decrease in clumpy morphology is not in line with minor-merger-driven clump formation (as suggested by Guo et al. 2015 for low-mass galaxies) because the minor merger rate does not show similar change over this period (Lotz et al. 2011). Instead, we suggest that a better tracer of \( f_{\text{clumpy,8}} \) is the turbulent fraction of galaxies. Kassin et al. (2012) observed a decline in turbulence for galaxies of all masses \((10^8 - 10^{10.7} \, M_\odot)\), but in particular they noted that larger galaxies settle quickly after \( z \sim 1.2 \) while less massive galaxies remain turbulent for a longer time, mimicking the trends in \( f_{\text{clumpy}} \). In total, the current body of evidence supports a picture where clumps primarily form in situ owing to disk instability, though more observations are needed.

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 http://authors.galaxyzoo.org. We would like to thank the anonymous referee, whose suggestions were insightful and led to substantial improvements to this paper.

This research is partially supported by the National Science Foundation under grant AST 1716602.

This material is based on work supported by the National Aeronautics and Space Administration (NASA) under grant No. HST-AR-15792.002-A.

This research made use of Montage. It is funded by the National Science Foundation under grant No. ACI-1440620 and was previously funded by the National Aeronautics and Space Administration’s Earth Science Technology Office, Computation Technologies Project, under Cooperative Agreement No. NCCS-626 between NASA and the California Institute of Technology.

This publication uses data generated via the Zooniverse.org platform, development of which is funded by generous supporters, including a Global Impact Award from Google, and by a grant from the Alfred P. Sloan Foundation.

### Appendix A

#### Galaxy Cutout Creation

Here we detail the process by which we created galaxy image cutouts. For each galaxy, we began by downloading the \( i, r, \) and \( g \)-band fields listed in the SDSS PhotoPrimary table for the galaxy. From each field, we selected a cutout centered on the target galaxy and scaled to 6 times the \( r \)-band 90% Petrosian radius on each side. Finally, we projected images in our three image bands to a uniform coordinate system using the Montage library (Jacob et al. 2010) to avoid subpixel offsets in the RGB composite image. This coordinate system had a scale of 0.396 pixel \(^{-1} \) to match SDSS’s native resolution. To create a color-composite image, the \( i, r, \) and \( g \)-band cutouts were mapped to red, green, and blue, respectively, and color scaling was performed via the “Lupton scaling” technique described in Lupton et al. (2004). The particular scaling performed obeyed the equation

\[
x = \frac{\text{asinh}(Q/(\text{band} \_\text{scaling} - \text{minimum}))}{\text{stretch}},
\]

where \( I \) is the input intensity in a given image band (in counts) and \( x \) the output intensity for display (ranging from 0 to 1). “Q,” “minimum,” and “stretch” were tunable parameters set to 7, 0, and 0.2, respectively, while “\text{band} \_\text{scaling}” is a per-band parameter set to 1.818, 1.176, and 0.7 for the \( i, r, \) and \( g \) bands, respectively. The result resembles SDSS’s standard color balance with a significant emphasis on the \( g \) band to emphasize star-forming regions. Finally, images were resized to a standard 400 \( \times \) 400 pixel scale.

#### Appendix B

##### Calculating the Completeness Estimate on Clumps

In this appendix we describe the mathematical process used to calculate the completeness estimate of each clump identified by Clump Scout volunteers.

We estimate the completeness of each recovered clump as a function of three measured properties: the clump’s brightness (traced by \( m_g \)), the clump’s color (traced by \( m_r - m_g \)), and the clump’s brightness relative to the diffuse background (traced by \( m_g - m_g \)). The background magnitude \( m_g \) measures the estimated background in a circular aperture with diameter 1 PSF-FWHM. The \( g \) band was selected as the primary band because it traces the bluest wavelength of the filters used in our image cutouts (\( i, r, \) and \( g \)) and correlates most strongly with star formation. Other properties, such as galactocentric distance, galaxy mass, redshift, or image resolution, were found to have minimal impact, possibly because they are correlated with the three properties already used.

We then partitioned real clumps into bins based on their features \( x = \{x_1, x_2, x_3\} \), where

\[
x_1 = m_g, \quad x_2 = m_r - m_g, \quad x_3 = m_g - m_g.
\]

Partition edges were selected such that each feature \( i \) was divided into partitions \( k(x_i) \) containing equal number counts of real clumps. Partitions over all three features were then given...
by \( k(x) = \{ k_1(x_i), k_2(x_j), k_3(x_k) \} \). (Note that since clump properties are not independent, the three-dimensional partitions \( k \) do not generally contain the same number counts of clumps.)

Next, we used simulated clumps to estimate the recovery fraction \( f_{\text{recov}}(k) \) in each partition. A naive estimate would be to equate \( f_{\text{recov}}(k) \) to the fraction of simulated clumps in bin \( k \) recovered by volunteers. However, this method is imprecise for bins with few recovered clumps and provides no simple method for estimating error. Instead, we used a Bayesian method: we assumed that the likelihood distribution of \( f_{\text{recov}} \) obeys a beta distribution defined by the number of recovered and missed clumps within \( k \). That is,

\[
P(f_{\text{recov}}(k)) = B(n_{\text{recov}}(k) + 1, n_{\text{missed}}(k) + 1),
\]

where \( n_{\text{recov}} \) is the number of clumps in bin \( k \) that volunteers recovered and \( n_{\text{missed}} \) is the number that they missed. We then took \( f_{\text{recov}}(k) \) to be the median of the \( P(f_{\text{recov}}(k)) \) distribution. (This method is described in Cameron 2011 and provides robust uncertainties for estimated fractions even when number count is low.) Figure 13 displays the number of (total and recovered) simulated clumps in each bin, as well as the uncorrected distribution of recovered real clumps over these bins.

Finally, we interpolated the discrete map \( f_{\text{recov}}(k) \) to a continuous map over all clump properties, \( f_{\text{recov}}(x) \). We began with a grid of values for \( f_{\text{recov}}(x) \) by assuming that \( f_{\text{recov}}(k) = f_{\text{recov}}(x_{50}(k)) \), where \( x_{50}(k) \) is the median value of \( x \) for simulated clumps in partition \( k \). We then linearly interpolated over this grid to obtain the continuous map. It should be noted that, because the interpolation grid was defined over the median clump locations in each bin \( (x_{50}) \), some clumps in the outermost bins fell outside of the interpolation grid; these clumps were assigned \( f_{\text{recov}}(x_{50}(k)) \), i.e., the value for their bin. A total of 0.4% of real clumps fell in bins with fewer than five simulated clumps, and these clumps were discarded for further analysis.

Once the map \( f_{\text{recov}}(x) \) is established, we can estimate the specific completeness estimate for each clump in our sample: for clump \( i \) with estimated properties \( x(i) \), we define its estimated completeness \( P_{\text{rec,i}} \) as \( P_{\text{rec,i}} = f_{\text{recov}}(x(i)) \). These \( P_{\text{rec,i}} \) values can then be used to correct \( f_{\text{clumpy}} \) for incompleteness, as explained in Section 4.3.

This method of estimating completeness also allows us to easily estimate the uncertainty on the estimate using a Monte Carlo method. Over 100 trials, we allowed the fraction of recovered clumps in each bin \( f_{\text{recov}}(k) \) to vary randomly over its estimated distribution \( P(f_{\text{recov}}(k)) \), rather than taking the 50th percentile value. Repeated trials yielded an approximate distribution on \( f_{\text{recov}(x)} \).