

**Integrating Human and Machine Intelligence in Galaxy
Morphology Classification Tasks:
Development and Use Cases**

A THESIS
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Melanie Renee Beck

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
Doctor of Philosophy

Advisor:
M. Claudia Scarlata

August, 2017

**© Melanie Renee Beck 2017
ALL RIGHTS RESERVED**

Acknowledgements

Getting to this point in one's academic career requires the help and support of dozens of people. I hope I can name you all. Without a doubt in my mind, I know I would not be where I am today without my advisor, Claudia Scarlata. She unwittingly saved me from transferring to the Statistics department by asking me to join her research group and telling me that female astronomers need to stick together. She's right. I will forever be grateful to her because I know staying in this program was the right choice. In addition to the myriad technical skills I gained from her, above all Claudia taught me how to be an assertive, confident researcher – no bullshit. She pushed me when I needed pushing and stood back when I needed space. She taught me the value in finishing a project even if I didn't love it – publish! She also taught me how to navigate the sometimes turbulent interpersonal relationship between graduate student and advisor. I like to think we turned out not just as colleagues, but also friends (even if she does still get a little motherly..).

I must also thank my unofficial co-advisors, Lucy Fortson and Chris Lintott. Chris taught me how to work independently, how to speak up when I need help, how to effectively argue my position, and how to correctly interpret the British (do not underestimate how important this skill can be!). His ability to fit the pieces into the Big Picture helped me stay grounded. Lucy provided constant support. Her ability to navigate and communicate sdfasdfsfsf

For the past three years I've had the delightful fortune to be part of the Galaxy Zoo science team. The galaxy zoo team including Brooke, Karen, Stephen, Becky The grad students: Micaela, Michael, Melanie, Tony, Brian

All the institutions that have contributed to funding me: Doctoral Dissertation Fellowship, Graduate School's Thesis Research Travel Grant, Oxford's Balzan Fellowship

My family for years and years of support.

Dedication

I dedicate this thesis to my brother, James Overton Beck IV, whose memory never ceases to inspire me to achieve.

Abstract

Quantifying galaxy morphology is a challenging yet scientifically rewarding task. As the scale of data continues to increase with upcoming surveys, traditional classification methods will struggle to handle the load. We present a solution through an integration of visual and automated classifications, preserving the best features of both human and machine. We demonstrate the effectiveness of such a system through a re-analysis of visual galaxy morphology classifications collected during the Galaxy Zoo 2 (GZ2) project. We reprocess the top-level question of the GZ2 decision tree with a Bayesian classification aggregation algorithm dubbed SWAP, originally developed for the Space Warps gravitational lens project. Through a simple binary classification scheme we increase the classification rate nearly 5-fold classifying 226,124 galaxies in 92 days of GZ2 project time while reproducing labels derived from GZ2 classification data with 95.7% accuracy.

We next combine this with a Random Forest machine learning algorithm that learns on a suite of non-parametric morphology indicators widely used for automated morphologies. We develop a decision engine that delegates tasks between human and machine and demonstrate that the combined system provides a factor of 11.4 increase in the classification rate, classifying 210,543 galaxies in just 32 days of GZ2 project time with 93.5% accuracy. As the Random Forest algorithm requires a minimal amount of computational cost, this result has important implications for galaxy morphology identification tasks in the era of *Euclid* and other large-scale surveys.

Contents

Acknowledgements	i
Dedication	iii
Abstract	iv
List of Tables	ix
List of Figures	x
1 Introduction	1
1.1 Galaxy formation	3
1.2 Standard morphology designations	4
1.3 Morphology as a tracer of galaxy evolution	5
1.3.1 Stellar populations and star-formation histories	6
1.3.2 Mass assembly and the emergence of the Hubble sequence	8
1.3.3 Morphology as a function of environment	10
1.3.4 Insights from rare morphologies	11
1.4 Obtaining morphologies	12
1.4.1 Visual classifications	12
1.4.2 Automated classifications	13
1.4.3 Machine learning	14
1.5 Overview of Galaxy Zoo: Express	15

2 Data: visual and automated morphologies	18
2.1 Galaxy Zoo	18
2.1.1 Galaxy sample selection	19
2.1.2 GZ2 decision tree and project history	20
2.1.3 Data reduction	22
2.2 Automatic morphology indicators	24
2.2.1 Imaging data	24
2.2.2 Image cleaning	25
2.2.3 Morphology indicators	26
2.2.4 Quality and consistency	33
2.3 Catalog of morphological indicators for 282350 SDSS galaxies	36
3 Intelligent management of visual classifications	40
3.1 Galaxy Zoo 2 Classification Data	40
3.2 Efficiency through intelligent human-vote aggregation	42
3.2.1 Gold-standard sample	43
3.2.2 Volunteer Bias	44
3.2.3 Fiducial SWAP simulation	48
3.2.4 Intelligent subject retirement	50
3.2.5 Reducing human effort	54
3.2.6 Disagreements between SWAP and GZ2	57
3.2.7 Summary	59
3.3 Exploring SWAP’s Parameter Space	60
3.3.1 Initial agent confusion matrix.	60
3.3.2 Subject prior probability, p_0	62
3.3.3 Retirement thresholds, t_F and t_N	63
4 Incorporating machine intelligence	66
4.1 Efficiency through incorporation of machine classifiers	66
4.1.1 Random Forests	66
4.1.2 Grid Search and Cross-validation	67
4.1.3 Feature Representation and Pre-Processing	67
4.1.4 Decision Engine	68

4.1.5	The Machine Shop	70
4.2	Results	70
4.2.1	Who retires what, when?	72
4.2.2	Machine performance	75
4.3	Looking Forward	78
4.3.1	Conclusions	79
4.4	Acknowledgements	80
5	A population of “clumpy” galaxies in local Universe	81
5.1	Introduction	81
5.2	Sample Selection & Data	83
5.3	Are these star-forming regions analogs of high-redshift clumps?	89
5.3.1	Clump galactic radial distance	89
5.3.2	Clump luminosity	89
5.3.3	Velocity dispersion and clump diameters	90
5.4	Summary	91
5.5	Clump Scout	92
5.5.1	Redshift evolution of f_{clumpy}	94
5.6	Summary and conclusions	94
6	Summary & Future Work	96
Appendix A. Spectro-polarimetry Confirms Central Powering in a Lyα Nebula at $z = 3.09$		110
A.1	Introduction	111
A.2	Ly α Polarization Basics	113
A.3	Observations, Reduction, and Calculations	115
A.3.1	Observations	115
A.3.2	Data Reduction	117
A.3.3	Polarization and Error Calculations	120
A.4	Polarization of Ly α in LAB1	122
A.4.1	Polarization integrated over the line profile	122
A.4.2	Polarization across the line profile	125

A.5	Discussion and Conclusions	129
A.6	The Future of Ly α Polarization	132

List of Tables

2.1	Summary of morphology measurements (UPDATE THIS)	33
2.2	Morphology Diagnostics for \sim 283K SDSS Galaxies in the GZ2 Sample	39
4.1	Simulation summary	72
A.1	Polarization signal-to-noise and fractional polarization for spatial binning of LAB1	124

List of Figures

1.1	Hubble tuning fork	4
1.2	Galaxy color-magnitude relation.	7
1.3	The history of cosmic star formation.	9
1.4	Schematic of the Galaxy Zoo: Express human+machine hybrid system.	16
2.1	Galaxy Zoo 2 decision tree	21
2.2	Example of the Galaxy Zoo 2 web interface	22
2.3	Example of Source Extractor segmentation maps.	25
2.4	Examples of image cleaning and morphology diagnostic measurements .	27
2.5	Examples of image cleaning and morphology diagnostic measurements .	28
2.6	Examples of image cleaning and morphology diagnostic measurements .	29
2.7	Comparison of Petrosian radius from this work to that computed in SDSS pipeline	34
2.8	Comparison of concentration index from this work to that computed from SDSS.	35
2.9	Automated morphologies as a function of Galaxy Zoo 2 f_{smooth}	37
2.10	Automated morphologies for the full GZ2 sample.	38
3.1	45
3.2	Volunteer confusion matrices achieved through SWAP reprocessing of GZ2 data.	46
3.3	Galaxy posterior probabilities realized through SWAP reprocessing of GZ2 data.	47
3.4	Confusion matrix between predictions and ground truth defines quality metrics.	49

3.5	Reprocessing GZ2 data with SWAP results in a factor of increase in the classification rate.	51
3.6	SWAP’s intelligent retirement mechanism 3.5 times fewer classifications than GZ2.	52
3.7	SWAP increase classification efficiency by retiring “easy” galaxies quickly.	55
3.8	SWAP’s volunteer-weighting mechanism provides a factor of three reduction in the required human effort for classification tasks.	56
3.9	SWAP’s prediction disagrees with the GZ2 label as a result of the confidence interval around the chosen threshold used to define the GZ2 label.	58
3.10	SWAP’s performance is robust to changes in initial volunteer confusion matrix and subject prior.	61
3.11	Receivor operating characteristic curve for the chosen SWAP retirement thresholds.	64
4.1	Random Forest learning curve	68
4.2	Performance of the human+machine combination – Galaxy Zoo: Express	71
4.3	Individual contributions of human and machine to galaxy classification	73
4.4	Random subsample of galaxy jpegs identified as false positives by the Random Forest	74
4.5	Distributions of measured galaxy morphology features.	75
4.6	Random Forest’s feature importances	77
5.1	Galaxy Zoo: Hubble decision tree.	84
5.2	Sample of “clumpy” galaxies determined by Galaxy Zoo: Hubble	86
5.3	Stellar mass, redshift, and Star-formation rate for a subset of our “clumpy” galaxies as measured by the GSWLC.	88
5.4	Properties of “clumps”: velocity dispersion, clump diameter, and clump galactic radial distance.	90
5.5	<i>Clump Scout</i> sample selection criteria from non-Stripe 82 GZ2 galaxy sample.	93
A.1	Science frames for Ly α polarimetry of LAB1	117
A.2	Slit position over LAB1 as compared with results from Hayes et. 2011	118
A.3	Polarization of spectrally integrated Ly α	123
A.4	2D Stokes parameters and polarization map of LAB1	125

Chapter 1

Introduction

A galaxy's morphology is the culmination of its formation, interactions, and evolution through environmental and internal processes. It is a snapshot into the current state of a galaxy's life as well as window to its past. The insights gleaned through the study of galaxy morphology have radically changed our view of universe since the time of Edwin Hubble.

Astronomers have made use of visual galaxy morphologies to understand the dynamical structure of these systems for nearly ninety years (e.g., Hubble, 1936; ?; Sandage, 1961; van den Bergh, 1976; Nair & Abraham, 2010; Baillard et al., 2011). The division between early-type and late-type systems corresponds, for example, to a wide range of parameters from mass and luminosity, to environment, color, and star formation history (e.g., Kormendy, 1977; Dressler, 1980; Strateva et al., 2001; Blanton et al., 2003a; Kauffmann et al., 2003; Nakamura et al., 2003; Shen et al., 2003; Peng et al., 2010); while detailed observations of morphological features such as bars and bulges provide information about the history of their host systems (e.g., Kormendy & Kennicutt, 2004; Elmegreen et al., 2008; Sheth et al., 2008; Masters et al., 2011; Simmons et al., 2014). Modern studies of morphology divide systems into broad classes (e.g., Conselice, 2006; Lintott et al., 2008; Kartaltepe et al., 2015; Peth et al., 2016), but a wealth of information can be gained from identifying new and often rare classes, such as low redshift clumpy galaxies (e.g., Elmegreen et al., 2013), polar-ring galaxies (e.g., Whitmore et al., 1990), and the green peas (Cardamone et al., 2009).

Obtaining these morphologies has traditionally been a time-consuming visual endeavor and only in the past twenty years have automated morphological assignment been possible. [WEIRD TRANSITION automated CAN do everything] Even with the varied automated approaches currently being exploited, an era of Even Bigger Data looms for the field of astronomy. The next decade will herald the first light of more powerful ground- and spaced-based telescopes such as the Large Synoptic Survey Telescope (LSST), Euclid, and WFIRST. The surveys planned for these instruments promise to revolutionize the field of astrophysics providing several orders of magnitude more data than is currently available. Traditional techniques will not be sufficient for extracting galaxy morphologies on a pertinent timescale.

This thesis details a solution to the scalability of galaxy morphology designations by examining classifications obtained as part of the Galaxy Zoo project, a crowd-sourcing initiative that has obtained morphological classifications from thousands of volunteers for over a million galaxies from several astrophysical surveys. Though innovative, even crowd-sourcing will be unable to sustain the classification load for future surveys. Instead, these classifications are combined with supervised machine learning algorithms that train on automated measurements of galactic morphological structural indicators. This thesis begins with a detailed account of the data utilized in this work as well as the methodology used to obtain these morphology diagnostics (Chapter 2). Chapter 3 demonstrates how crowd-sourcing techniques can be optimized by applying a Bayesian approach to aggregation of Galaxy Zoo classifications, while Chapter 4 explores the combination of these visual classifications with machine learning algorithms. Also included is a preliminary analysis of a rare sample of “clumpy” galaxies in the local universe discovered during the course of the analysis of Galaxy Zoo classifications (Chapter 5). While common in the distant universe, these potential local counterparts could shed light on star formation as it pertains to the formation of galactic bulges. This Introduction provides a brief overview of galaxy formation, the science achieved through the study of morphology, and galaxy classification techniques, as well as an overview of Galaxy Zoo: Express (GZX), an integrated framework of human and machine galaxy classifiers.

1.1 Galaxy formation

According to the Λ CDM model, the first galaxies are theorized to have formed from primordial density fluctuations shortly after the Big Bang. These fluctuations are characterized as local overdensities. These regions experience gravitational instability which causes them to grow with time according to a power law. When the overdensity is in the linear regime ($\delta\rho/\rho \ll 1$) the size of the regions expand with the expansion of the universe. As they continue to grow they eventually reach a turning point ($\delta\rho/\rho \sim 1$), break away from the expansion, and start to collapse, eventually forming self-gravitating structures.

These collapsing structures contain both baryonic and dark matter. The latter relaxes violently to create a dark matter halo, while the baryonic matter shocks to the virial temperature. In order for stars to form this gas must cool, and various mechanisms are possible depending on the temperature: bremsstrahlung emission in the hottest regimes, and various methods of excitation and de-excitation in cooler regimes. As the gas cools it collapses further and likely fragments into small, highly dense cores that eventually form the first stars.

Population III stars were formed entirely of primordial gas consisting solely of hydrogen and helium, as heavier elements were not yet synthesized. These stars were likely the most massive stars ever created and thus had short life spans before dying brilliantly in the first supernovae. As these stars formed at $z \sim 11$ no confirmed observations have been recorded and thus the details are obscure. However it is well established that their demise distributed the first heavier elements from which subsequent stars formed. Today, Population I stars are the most metal-rich.

If the collapse of the gas in the dark matter halo retained angular momentum it would be conserved through the formation of a disk. However, the disks of galaxies in the distant universe may not be like those of today as discussed later in this chapter. Feedback processes such as the blast waves and energy released during supernovae, and the active accretion of supermassive black holes known as active galactic nuclei (AGN) can affect the course of a galaxy's development.

The current paradigm also supports that these systems grow hierarchically in that larger dark matter halos are built from the coalescence of smaller systems. If these halos

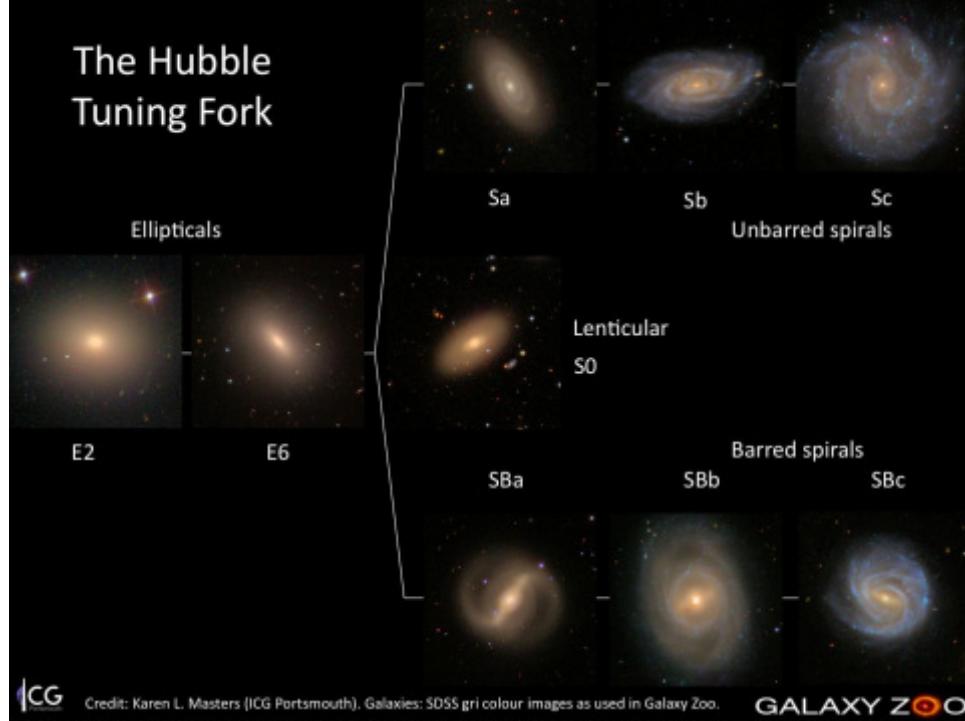


Figure 1.1 The Hubble “tuning fork” with example images for each galaxy type created from *gri*-composite SDSS imaging. Credit: Karen L. Masters and the Sloan Digital Sky Survey (SDSS) Collaboration.

contain galaxies, those too can merge which results in identifiably distinct morphologies as will be discussed below.

1.2 Standard morphology designations

The oldest and arguably simplest system for categorizing galaxy morphology dates back to the 1930s and Edwin Hubble’s famous “tuning fork” (Hubble, 1926, 1936). Based on a small sample, Hubble classified galaxies into two major groups, elliptical and spiral. Visually, elliptical galaxies possess a smooth light distribution, while spirals are characterized by a well-defined disk structure often with spiral arms. Hubble assigned a number to elliptical galaxies denoting the degree of their ellipticity, where 0 corresponds to a nearly perfectly round galaxy and 7 being highly elongated. Spirals were given additional designation in the form of letters ‘a’ through ‘c’, characterizing the compactness

of their spiral arms. For example, “Sa” galaxies are tightly wound, whereas “Sc” spirals are looser. The spiral category was then further subdivided by galaxies that exhibited a central bar. There are also indications that the bulges inherent to many spiral galaxies share a close connection to elliptical galaxies, thus the transitional “S0” category: galaxies with disks that are dominated by the bulge component.

It became common to call galaxies on the left side of the diagram “early-types,” and those on the right “late-types”. Contrary to this author’s belief, Hubble never intended these designations to imply galactic evolution. Instead, this terminology was borrowed from stars, where massive O and B stars were referred to as “early-type,” while older stars were known as “late-type”(Buta, 2011). It has subsequently become clear that, much to the contrary, ellipticals are dominated by late-type stars, while disk galaxies are typically composed of young, early-type stars. Unfortunately, the misnomer has stuck.

This method of galaxy classification was based on an extremely small sample of which only a few percent did not conform to the basic designations originally posited by Hubble. These leftover galaxies were dubbed “irregulars” or “peculiars”. It wasn’t until much later that it was discovered this galaxy type was far more prevalent than Hubble originally thought, especially in the more distant universe. Since this early attempt at classification, several other systems have been put forward but most share the same basic categories (e.g., de Vaucouleurs, 1959; Conselice, 2006). Indeed, even this simplistic approach has yielded nearly a hundred years of science that has advanced our understanding of galaxy formation, structure, and evolution.

1.3 Morphology as a tracer of galaxy evolution

That galaxies exhibit different features is obvious, but what, if anything, do those features tell us? Because we cannot observe the entire lifespan of a single galaxy it is fair to question whether or not morphology is primarily an indicator of age, with galaxies marching starkly through the Hubble sequence, or whether dynamical processes shape that morphology. Or both. In this section we discuss some of the major roles that galaxy morphology has played in understanding the evolution and formation of these systems.

1.3.1 Stellar populations and star-formation histories

At its heart, morphology simply traces an integrated 2D projection of a galaxy’s light distribution. As such, it encodes information on the distribution of a galaxy’s stellar, gas, and dust content. However, these components are best traced through different wavelengths of light.

Consider a coevolving stellar population with a mass distribution according to your favorite initial mass function. Stars on the Main Sequence (MS) radiate in the blue and ultraviolet (UV) end of the spectrum due to their high effective temperatures. The most massive quickly evolve off the MS and become red supergiants causing a decrease in the UV flux and an increase in the near infrared (NIR). As the low mass stars in this stellar population continue to evolve off the MS, the UV flux steadily decreases until eventually the red giant branch becomes the dominant source of flux, radiating in the IR. This so-called *passive* evolution indicates that as a galaxy ages it becomes redder. A galaxy’s morphology is tightly correlated with its color: massive elliptical galaxies are typically referred to as “red and dead,” possessing old stellar populations, while disk galaxies are generally still undergoing star formation and thus possess young, bluer stellar populations (e.g., Strateva et al., 2001; Baldry et al., 2004a; Cirasuolo et al., 2007; Lee et al., 2013; Taylor et al., 2015). This dichotomy is so prevalent that color has been used as a proxy for morphology when acquiring the latter was impractical (e.g., Shen et al., 2003; Blanton et al., 2003b).

Furthermore, this color bimodality correlates with luminosity resulting in the color-magnitude relation (CMR) (Baldry et al., 2004b; Bell et al., 2004). Now ubiquitous, this relation visualizes the separation of galaxy colors as a function of luminosity resulting in three main categories: the blue cloud, the red sequence and, more recently recognized, the green valley. Studies have shown that the red sequence is dominated by early-type galaxies like elliptical/S0, while disk galaxies reside in the blue cloud. There is a distinct gap between these two galaxy populations but recent studies have shown that, though sparse, galaxies residing in this region could be in the middle of active evolution sparking morphological change (Schawinski et al., 2007). The top panel of Figure 1.2 shows an example of the color-magnitude relation for a sample of SDSS galaxies, while the bottom panel depicts a schematic of the associated morphologies (Kormendy & Bender, 2012).

The color-magnitude relation is tighter for early-type galaxies providing a tool

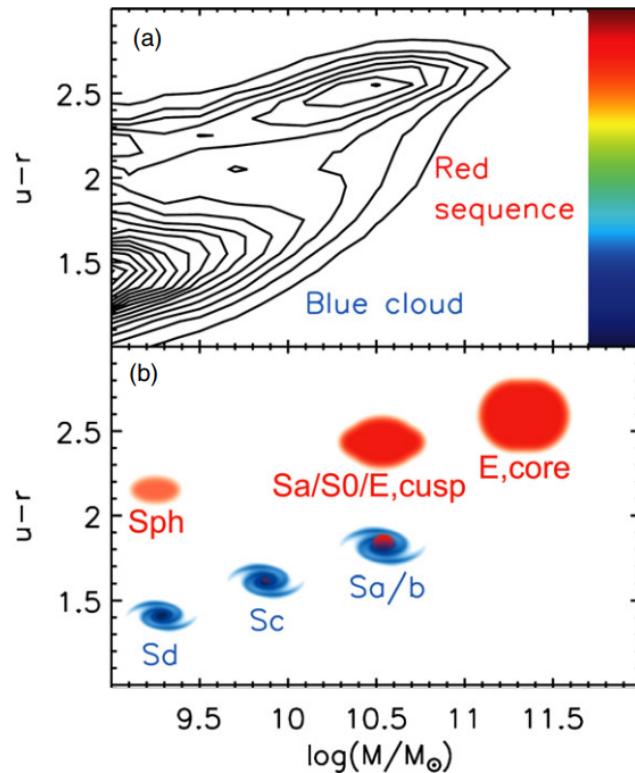


Figure 1.2 Color-magnitude relation adapted from Kormendy & Bender (2012). Though the x-axis is in units of log stellar mass, mass correlates strongly with a galaxy's luminosity. The top panel shows the contours of galaxy number density (Baldry et al., 2004a) with the rainbow bar denoting the associated $u - r$ color. The bottom panel depicts the dominant galaxy morphologies associated with each location.

to constrain the star-formation histories of this galaxy population (Sandage & Visvanathan, 1978; Tully et al., 1982). There exists a degeneracy between the age and metallicity of stellar populations: though older stars tend to be redder, this effect can also be achieved through an increase in stellar metallicity. It is now believed that the slope of the CMR is due primarily to metallicity effects (Bower et al., 1992; Kodama & Arimoto, 1997), which then allows for age estimates to be placed on the last episode of significant star formation in early-type galaxies (e.g., López-Cruz et al., 2004). The small scatter in this relation reflects that these galaxies have old, passively evolving stellar populations.

put this somewhere (Mei et al., 2009)

1.3.2 Mass assembly and the emergence of the Hubble sequence

The most fundamental characteristic of a galaxy is its mass. Though not as easy to measure as luminosity, it is now believed that several processes within a galaxy are directly tied to stellar mass. Stellar mass functions (SMFs) are thus a key first-order observable that statistically traces the formation of stars in the universe. Another key property is a galaxy's star formation rate (SFR). A cottage industry for decades, constructing mass and star-formation density functions has provided insights into the build-up of baryonic mass over cosmic time.

Understanding the rate of growth and distribution of baryonic mass in galaxies provides constraints on the mechanisms necessary to produce such observations. It is well known that star formation density of the universe increased from the earliest epochs up until $z \sim 2$ and has substantially declined since then, as shown in the Lilly-Madau plot in Figure 1.3 (Madau & Dickinson, 2014, and references therein). However this rapid build up of stellar mass was not distributed equally among all galaxies. Furthermore, the subsequent evolution of these systems provide clues to the emergence of the galaxy morphologies we see today.

At $z > 2$, most of the stellar mass density resides in irregular galaxies, and the existence of disk galaxies is questionable (Dickinson, 2000; Papovich et al., 2005; Cameron et al., 2011; Conselice et al., 2005, 2011; Buitrago et al., 2013). However, mass functions constructed according to morphological type as a function of redshift show that the stellar mass density gradually begins to reside more in disk systems at $1 < z < 2$, and in

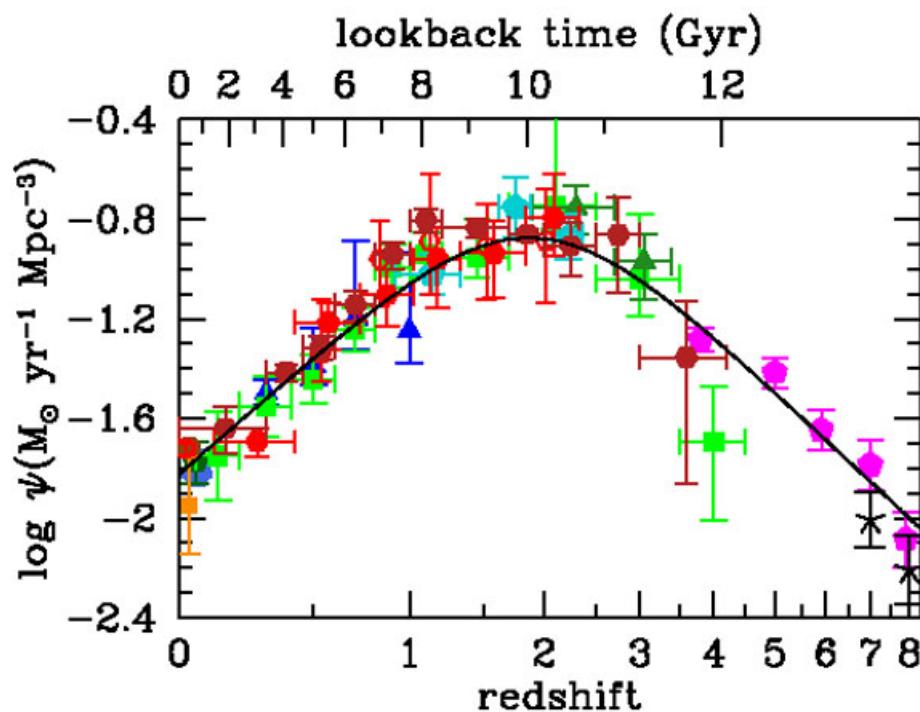


Figure 1.3 The history of cosmic star formation as determined by FUV+IR rest-frame measurements from several galaxy samples (credit: Madau & Dickinson, 2014, and references therein). This shows that cosmic star formation experienced peaks around $z \sim 2$ when the universe was approximately 3.5Gyr old followed by a gradual decline until present day.

the local universe, is dominated by early-type galaxies though statistics differ (Brinchmann & Ellis, 2000; Bell et al., 2003; Bundy et al., 2005; Mortlock et al., 2013; Kelvin et al., 2014; Huertas-Company et al., 2016; Thanjavur et al., 2016). Thus considerable evolution has occurred such that high-redshift systems settle down into the galaxies we see in the local universe.

Heirarchical galaxy formation theory predicts the build up of smaller systems before the creation of larger galaxies by way of violent mergers, simultaneously transforming disk morphologies into spheroidal systems (Driver et al., 2013; Bluck et al., 2009; Man et al., 2012). Indeed, this is likely the case in the high-redshift universe, but after $z < 2$, the dominant mechanism for galaxy evolution switches to cold gas accretion and with it, the formation of disk galaxies from irregular systems (Conselice et al., 2013). Both mergers and gas accretion processes are observed to be mass-dependent with the most massive galaxies settling into the familiar Hubble sequence more quickly than low-mass galaxies (Buitrago et al., 2013; Conselice et al., 2011; Mortlock et al., 2013). This so-called “downsizing” process, though seemingly contradictory to the heirarchical theory, is nevertheless favored observationally (Cowie et al., 1996; Brinchmann & Ellis, 2000; Bundy et al., 2005).

1.3.3 Morphology as a function of environment

A galaxy’s environment also has a direct relationship with its morphology. First quantified by Dressler (1980), the morphology-density relation is an emperical observation that elliptical galaxies tend to reside in the densest environments, i.e., rich groups and clusters, at all stages of cosmic time. in the richest and densest clusters of galaxies, the dominant morphology is elliptical, while for field galaxies, the dominant morphology is spiral. Furthermore, it’s been determined that this relation evolves over time, at least for the densest environments. All of this points to scenarios in which the environment of a galaxy influences its morphology (e.g., Fasano et al., 2000; Shen et al., 2003; Smith et al., 2005; Peng et al., 2010)

Much study has been done to separate the effects of environment from mass, i.e., secular, causes of morphological features. Peng, Shen,,etc. Cluster environments are, compared to the field, extremely hot and dense. Galaxies residing in these environments have faster velocity dispersion, on average. It’s estimated that field galaxies are, on

average, 1 per Mpc (?). The merging rate depends on the density of galaxies. In the past galaxies were closer together so merging was a bigger deal. Logically then, it would seem that merging happens all the time in clusters and that this subsequently causes a morphological change from disk to elliptical. However, galaxies are moving so briskly in clusters that it becomes less likely that they collide! Instead, mergers seems to be a predominant morphological mechanism in groups rather than clusters. In clusters, instead, other physical processes shape the way a galaxy evolves and looks over time.

Spirals convert to S0/elliptical galaxies in clusters due to several processes: they should have lost their cold atomic and molecular gas via ram pressure stripping (but see Tonnesen & Bryan 2009, for the effect of ram pressure on molecular clouds), harassment (Moore et al. 1996), strangulation (Kawata & Mulchaey 2008), etc.

1.3.4 Insights from rare morphologies

While great insight into the formation and evolution of galaxies has been gleaned through examination of broad morphological categories, more can be learned by digging into the details. Any theory of galaxy evolution will have to account for the dizzying array of galactic forms including the development of bars, bulges, and rings, as well as rare morphologies such as the “green peas,” or giant clumps of star formation in galaxies in the nearby universe.

Identifying finer details of galactic structure has become easier with the advent of cameras aboard the Hubble Space Telescope. Features like bars and bulges directly relate to a galaxy’s history. However, it has only been since the development of large scale surveys like SDSS, CANDELS, DeCALS(?) that astronomers have become aware of rare populations of galaxies. While small, these populations provide a means to constrain formation and evolution theories.

First recognized as an individual class of galaxies by Galaxy Zoo volunteers in 2007, the “green peas” are a type of luminous blue compact galaxy whose name reflects the hue of these galaxies in the false color SDSS images presented during the Galaxy Zoo project. These oxygen-rich objects are observed to have some of the largest star formation rates with some of the smallest masses. It is surmised that these galaxies were commonplace in the early universe and likely played a large role in the reionization of the universe. That such galaxies exist in the local universe provides a way to probe the cosmic past.

Another class of potential local analogs of high-redshift counterparts are low redshift “clumpy” galaxies. As previously discussed, galaxies of the past were largely irregular with star formation rates much higher than today. Much of the peculiar shapes of these galaxies are due to massive knots of star formation. These galaxies underwent processes that transformed them into disk galaxies with the result being that this particular morphology is rather rare in the local universe. These star-forming clumps are thought to form via gravitational disk instability (Toomre, 1964). If these features are long-lived compared to their host galaxy, it is possible they contribute to the growth of the galactic bulge (Conselice, 2014). Chapter 5 presents a more in depth discussion and preliminary analysis of a sample of such galaxies discovered through the Galaxy Zoo project.

1.4 Obtaining morphologies

A galaxy’s morphology is thus an integral component for understanding the nature of these systems as well as deriving a fuller comprehension of their formation and evolution. However, obtaining such morphological information poses several challenges. This section will discuss the methods in which galaxy morphologies are collected and quantified.

1.4.1 Visual classifications

For most of the past century, galaxy morphologies were determined by a small number of expert astronomers beginning with Hubble. Visual classification, though highly accurate due to the human mind’s unique pattern recognition capability is, however, incredibly time consuming. For decades assignment of morphological type to galaxies resulted in small samples with varying degrees of descriptive complexity and often lacking in statistical significance (Hubble, 1936; Sandage, 1961; Sandage & Tammann, 1981; de Vaucouleurs, 1963; de Vaucouleurs et al., 1991). With surveys like the Sloan Digital Sky Survey (SDSS, Abazajian et al., 2003) and the CFHTLS-Deep Survey, coupled with cartels of graduate student classifiers, samples approached the tens of thousands (Fukugita et al., 2007; Schawinski et al., 2007; Nair & Abraham, 2010).

Unfortunately, this approach still cannot take full advantage of the depth and scope of such large scale surveys. This necessitated the birth of the Galaxy Zoo (GZ) project

(Lintott et al., 2008, 2011; Willett et al., 2013, 2017; Simmons et al., 2017), the first effort to crowd-source the task of galaxy morphology to the general public. With the efforts of thousands of citizen scientists, GZ has released visual morphologies for over one million galaxies, providing a solution that scales visual classification for current surveys and producing a prolific amount of scientific output (e.g., Land et al., 2008; Bamford et al., 2009; Darg et al., 2010; Schawinski et al., 2014; Galloway et al., 2015; Smethurst et al., 2016). A hybrid approach is the system developed by the Cosmic Assembly Near-infrared Dark Energy Legacy Survey (CANDELS; Grogin et al., 2011; Koekemoer et al., 2011) team, imaging produced by the Hubble Space Telescope. This scheme (Kartaltepe et al., 2015) crowd-sources not to the general public, but to dozens of expert astronomers, collecting visual classifications for over 50,000 galaxies in the CANDELS fields.

1.4.2 Automated classifications

Another approach has been the automated extraction of morphologies with the development of parametric (Sersic, 1968; Odewahn et al., 2002; Peng et al., 2002), and non-parametric (Abraham et al., 1994; Conselice, 2003; Abraham et al., 2003; Lotz et al., 2004; Freeman et al., 2013) structural indicators. While these scale well to large samples (e.g., Simard et al., 2011; Griffith et al., 2012; Casteels et al., 2014; Holwerda et al., 2014; Meert et al., 2016), they often fail to capture detailed structure and can provide only statistical morphologies with large uncertainties (e.g., Abraham et al., 1996; Bershady et al., 2000). We briefly highlight a few of these indicators here while a more detailed discussion can be found in Chapter 2.

One of the first and most popular parametric approaches for modelling a galaxy’s light distribution is the Sérsic profile:

$$I(R) = I_e \exp \left\{ -b_n \left[\left(\frac{R}{R_e} \right)^{1/n} - 1 \right] \right\} \quad (1.1)$$

where I_e is the intensity at the “effective” radius R_e that encloses half of the total light from the model, n is the Sérsic index that essentially describes the concentration of the light profile, and b_n is a term that depends on n . The de Vaucouleurs law is produced when $n = 4$, which well describes the light profile of elliptical galaxies (?). On the other hand, a Sersic index of $n = 1$ reduces the equation to an exponential which is a good

description for disk galaxies. This index has been used for decades as a broad method to classify galaxies into early- and late-type categories.

A drawback to the parametric approach is the need to assume the underlying distribution and while this works technique works well for galaxies that are obviously elliptical or spiral, it produces mixed results for other morphological types, i.e., irregulars or peculiars, which have low central concentration resulting in a low Sérsic index, but which do not have disks or spiral arms. Additionally, fitting this function to thousands or even millions of galaxies is time consuming. Non-parametric structural indicators require fewer assumptions and are derived empirically.

Closely related to the Sérsic index is the non-parametric diagnostic of concentration. Originally conceived by Abraham et al. (1996), it has several definitions but each measures the ratio of the aggregated light within two concentric apertures about the galaxy's center: one close to the galaxy's center and the other further out. Typically, these apertures contain 50% and 90% of the galaxy's total light. Measuring a galaxy's concentration can be much faster than fitting it with a Sérsic profile and less prone to fitting failures.

Once these parametric and non-parametric diagnostics have been computed for a sample of galaxies it is then common practice to place a cut on one or more of these parameters to separate the sample into early- and late-types (Shen et al., 2003). More sophisticated approaches involve measuring several automated diagnostics and separating galaxies in a two dimensional plane, a technique that has been highly successful at identifying not only distinctions between spheroidal and disk-like galaxies but also merging and interacting systems (Lotz et al., 2004; Conselice et al., 2000; Conselice, 2003; Freeman et al., 2013).

1.4.3 Machine learning

Machine learning techniques are becoming increasingly popular for classification and image processing tasks. Another automated approach, these generally work by defining a set of features that describe the morphology in an N -dimensional space. These features can be anything: color, mass, spectral index, velocity dispersion, and of course the parametric and non-parametric indicators discussed above. Choosing which features are most appropriate will depend on the classification task at hand, the particular machine

learning algorithm chosen, and the strength of the correlation between a given feature and the galaxy’s intended class.

The location in this N -dimensional morphology space defines a morphological type for each galaxy. Learning the morphology space can be achieved through algorithms such as Support Vector Machines (Huertas-Company et al., 2008) or Principal Component Analysis (Watanabe et al., 1985; Conselice, 2006; Scarlata et al., 2007; Peth et al., 2016). Another approach is through deep learning, a machine learning technique that attempts to model high level abstractions. Algorithms like convolutional and artificial neural networks (CNNs, ANNs) have been used for galaxy morphology classification with impressive accuracy (Ball et al., 2004; Banerji et al., 2010; Dieleman et al., 2015; Huertas-Company et al., 2015).

A drawback to all machine learning classification techniques is the need for standardized training data, with more complex algorithms requiring more data. Furthermore, these data are best when consistent for each survey: differences in resolution and depth can be implicitly learned by the algorithm making their application to disparate surveys challenging.

1.5 Overview of Galaxy Zoo: Express

In this work we present a system that preserves the best features of both visual and automatic classifications, developing for the first time a framework that brings both human and machine intelligence to the task of galaxy morphology to handle the scale and scope of next generation data. We demonstrate the effectiveness of such a system through a re-analysis of visual galaxy morphology classifications collected during the Galaxy Zoo 2 project, and combine these with a Random Forest machine learning algorithm that trains on a suite of non-parametric morphology indicators widely used for automated morphologies. In this paper we focus on the first question of the Galaxy Zoo decision tree. We demonstrate that our method provides a factor of 11.4 increase in the rate of galaxy morphology classification while maintaining at least 93.5% classification accuracy as compared to Galaxy Zoo 2 published data. We first present an overview of our framework, which also serves as a blueprint for this paper.

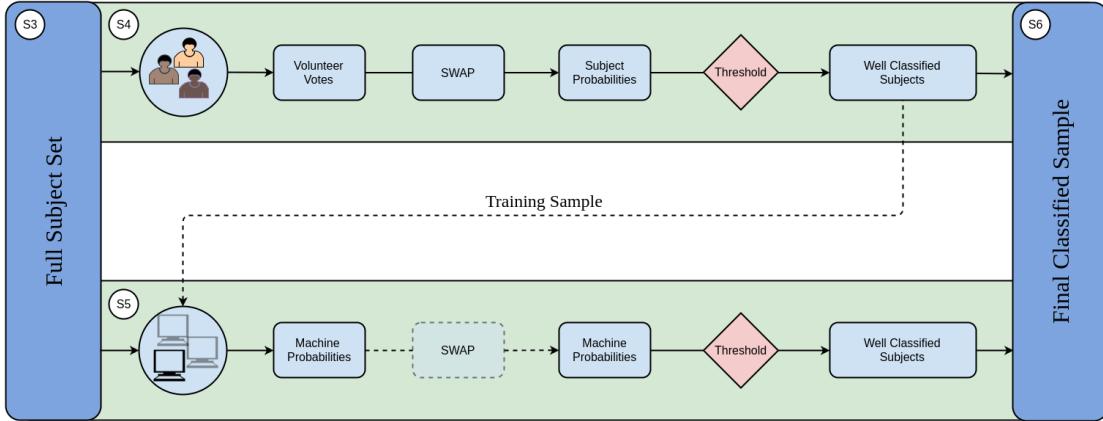


Figure 1.4 Schematic of our hybrid system. Humans provide classifications of galaxy images via a web interface. We simulate this with the Galaxy Zoo 2 classification data described in Chapter 2. Human classifications are processed with an algorithm described in Chapter 3. Subjects that pass a set of thresholds are considered human-retired (fully classified) and provide the training sample for the machine classifier as described in Chapter 4. The trained machine is applied to all subjects not yet retired. Those that pass an analogous set of machine-specific thresholds are considered machine-retired. The rest remain in the system to be classified by either human or machine. This procedure is repeated nightly.

The Galaxy Zoo Express (GZX) framework combines human and machine to increase morphological classification efficiency, both in terms of the classification rate and required human effort. Figure 1.4 presents a schematic of GZX including section numbers as a shortcut for the reader. We note that transparent portions of the schematic represent areas of future work which we explore in Chapter 6. Any system combining human and machine classifications will have a set of generic features: a group of human classifiers, at least one machine classifier, and a decision engine which determines how these classifications should be combined.

In this work we demonstrate our system through a re-analysis of Galaxy Zoo 2 (GZ2) classifications. This allows us to create simulations of human classifiers. These classifications are used most effectively when processed with SWAP, a Bayesian code described in Chapter 3, first developed for the Space Warps gravitational lens discovery project (Marshall et al., 2016). These subjects provide the machine's training sample.

In Chapter 4, we incorporate a machine classifier. We have developed a Random

Forest algorithm that trains on measured morphology indicators such as Concentration, Asymmetry, Gini coefficient and M_{20} well-suited for the top-level question of the GZ2 decision tree, discussed in Chapter 2. After a sufficient number of subjects have been classified by humans, the machine is trained and its performance assessed through cross-validation. This procedure is repeated nightly and the machine's performance increases with size of the training sample, albeit with a performance limit. Once the machine reaches an acceptable level of performance it is applied to the remaining galaxy sample.

Even with this simple description, one can see that the classification process will progress in three phases. First, the machine will not yet have reached an acceptable level of performance; only humans contribute to subject classification. Second, the machine's performance will improve; both humans and machine will be responsible for classification. Finally, machine performance will slow; remaining images will likely need to be classified by humans. This blueprint allows even modest machine learning routines to make significant contributions alongside human classifiers and removes the need for ever-increasing performance in machine classification.

Chapter 2

Data: visual and automated morphologies

This chapter presents all data used in this research. It begins with an in depth overview of the Galaxy Zoo 2 project, including the galaxy sample and procurement of visual galaxy morphology classifications. It then covers in considerable detail the methodology for obtaining the morphological structural indicators measured for the Galaxy Zoo 2 sample.

2.1 Galaxy Zoo

Founded by co-creators Chris Lintott and Kevin Schawinski, Galaxy Zoo is a crowd-sourced initiative to visually classify large numbers of galaxies by enlisting members of the general public. It began in 2007 with the classification of 893,212 galaxy images from the Sloan Digital Sky Survey (SDSS) Data Release 6 with $r < 17.77$ Petrosian AB magnitudes (Strauss et al., 2002; Adelman-McCarthy et al., 2008). The first iteration of the project was simplistic, inviting volunteers to determine whether a galaxy was elliptical, spiral, or a star / artifact. The project (hereafter, GZ1) was an immediate success both in terms of the public interest and the resulting science: following its completion (Lintott et al., 2008), over a dozen peer-reviewed articles utilizing GZ1 classifications¹ were published. In addition to explorations of galaxy morphology and its dependence

¹ <https://www.zooniverse.org/about/publications>

on color and environment (Skibba et al., 2009; Bamford et al., 2009), significant results also included substantial populations of red disks (Masters et al., 2010), blue ellipticals (Schawinski et al., 2009), and discoveries of rare objects such as the “green peas” (Cardamone et al., 2009) and Hanny’s Voorwerp (Lintott et al., 2009), the first obvservation of a AGN ionization echo.

The early sucess of GZ1 led to several subsequent and progressively more complex projects. To date, Galaxy Zoo has provided morphologies for over a million galaxies from multiple imaging surveys of various wavebands and redshifts using classifications provided from over a million volunteers. The research presented in this thesis utilizes data from the Galaxy Zoo 2 (GZ2) project (Willett et al., 2013), the immediate successor of GZ1. The following sections provide an overview of the GZ2 project including the galaxy sample, the decision tree structure, and a brief description of how volunteer votes are converted into descriptive classifications.

2.1.1 Galaxy sample selection

The original GZ1 project sought classifications for nearly one million galaxies in SDSS. Due to the staggering galaxy sample size, the morphologies collected were broad, seeking to determine between early-type, late-type and mergers. However, much can be gained by probing detailed morpholgies such as the existance of bars, bulges, dust lanes, rings, etc. Galaxy Zoo 2 thus selected the nearest and brightest 25% of galaxies from the original GZ1 sample, galaxies for which fine morphological structure could be resolved and classified. Pulled from the Data Release 7 Legacy catalog (Abazajian et al., 2009) which imaged the the North Galactic Cap, this galaxy sample required the Petrosian half-light magnitude be brighter than 17.0 in the r -band, along with a size limit such that $\text{petror90_r} > 3''$, where petror90_r is the radius containing 90% of the r -band Petrosian aperture flux. Spectroscopic redshifts were pulled from the SDSS Main Galaxy Sample (Strauss et al., 2002) and galaxies outside of $0.0005 < z < 0.25$ were removed, though objects without reported redshifts remained in the sample. This resulted in a sample of 273,783 galaxies.

In addition to the DR7 Legacy catalog, galaxies were included from Stripe 82, a multiply-imaged strip along the celestial equator in the Southern Galactic Cap. Galaxies in this region were selected to have $m_r \leq 17.7$. GZ2 included multiple samples from

this region: a set of 21,522 single-exposure images (though only about half conformed to the shallower Legacy magnitude cut specified above), and two sets of $\sim 30K$ co-added images from multiple exposures resulting in an object detection limit approximately two magnitudes deeper than the normal depth imaging. The research presented in this thesis utilizes the final GZ2 single-depth sample consisting of 295,305 galaxies of which 11,334 have the deeper magnitude limit.

2.1.2 GZ2 decision tree and project history

GZ2 was the first Galaxy Zoo project to utilize a decision tree wherein, with the exception of the first question, subsequent tasks depended on the response to the current question. The full decision tree is shown in Figure 2.1. Volunteers are allowed to select a single option for each question and are immediately taken to the next task after responding. Using GZ2 nomenclature, a *classification* is the total amount of information about a subject obtained by completing all tasks in the decision tree. Each step in the tree is a *task* consisting of a *question* and a set of *responses*. A volunteer’s response is referred to as a *vote*. The first question in the tree is a modification of the GZ1 project, asking volunteers to identify whether a galaxy is ‘smooth’, has ‘features or a disk’, or is a ‘star or artifact’.

For the single-depth sample, volunteers were shown color images generated from the SDSS ImgCutout web service. Each image is a *gri* color composite scaled to $0.02 \times \text{petror90_r}$. Throughout the life of GZ2 these images were randomly served to a web interface similar to that shown in Figure 2.2. Towards the end of the project galaxy images with few responses were shown more frequently in order to ensure that each galaxy had a sufficient number of classifications to adequately characterize the likelihood of the classification distribution. This resulted in a median of 44 classifications per galaxy with a wide spread. The full project spanned just over 14 months with the final dataset consisting of over 16 million classifications from over 80 thousand volunteers.

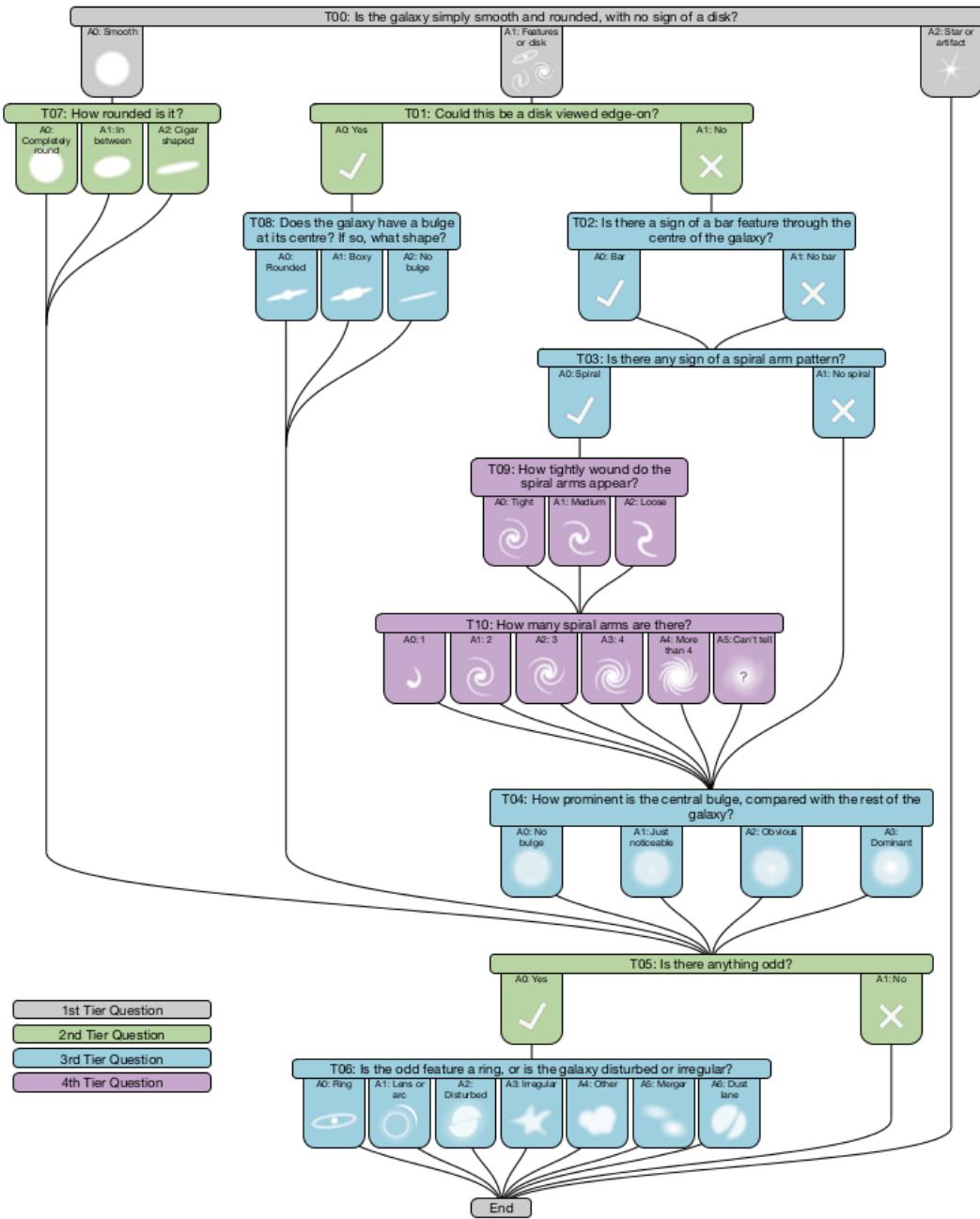


Figure 2.1 Galaxy Zoo 2 decision tree structure.

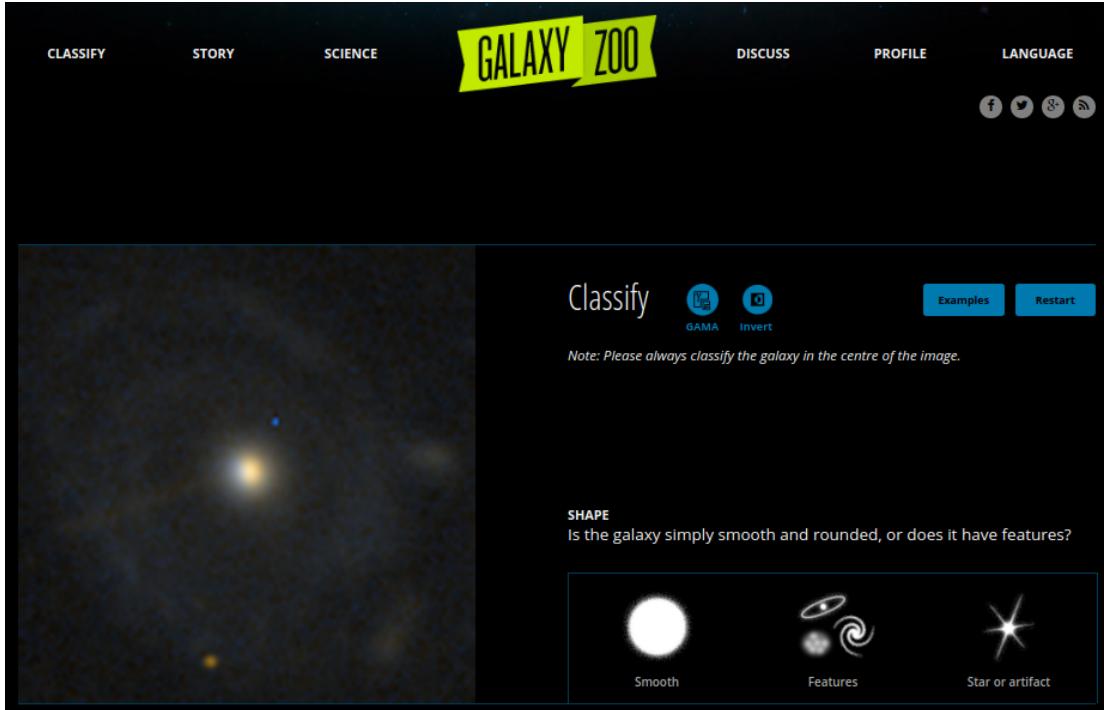


Figure 2.2 Example of the Galaxy Zoo 2 web interface showing the first question in the decision tree.

2.1.3 Data reduction

The GZ2 catalog provides several types of morphologies computed from volunteer classifications consisting of a vote fraction for each response to every task, denoted f_{response} . The most basic of these is computed simply as $f_r = n_r/n_t$, where n_r is the number of votes of response r , and n_t is the total number of votes for task t . In future chapters, this type of vote fraction is referred to as the *raw* vote fraction as no post-processing has been performed.

All GZ projects perform a weighting scheme that evaluates the consistency of individual volunteers by assessing how their votes deviate from the majority for each task in the decision tree. This process effectively downweights volunteers whose responses are consistent with a random classifier. A volunteer's consistency, κ , for a given task is

defined as

$$\kappa = \frac{1}{N_r} \sum_{i=1}^{N_r} \kappa_i \quad (2.1)$$

where N_r is the total number of responses to a task, and κ_i is f_r if the volunteer's vote corresponds to response i , otherwise $\kappa = (1 - f_r)$. Each volunteer is then assigned a mean consistency, $\bar{\kappa}$, which is the average consistency over all tasks. A weighting function is then applied according to

$$w = \min(1.0, (\bar{\kappa}/0.6)^{8.5}). \quad (2.2)$$

All vote fractions are then recomputed using the volunteer weights and the process is repeated three times to assure convergence. The resulting vote fractions are dubbed *weighted*. For GZ2, $w = 1$ for $\sim 95\%$ of volunteers and thus the majority are treated equally. It's important to note that there is no up-weighting of exceptionally consistent volunteers.

Finally, when considering morphologies, be they visual or automatic, for a sample of galaxies in a small redshift range it is unlikely that galaxy evolution plays a major role in morphology variations. Thus, the presumed culprit is instead classification bias in that galaxies in the more distant universe tend to be, on average, smaller and dimmer. This obscures identification of features such as spiral arms, bars, and more. GZ2 corrects for this effect, briefly described below, producing *debiased* vote fractions.

The general approach is such that, for a galaxy of a given size and brightness, a sample of other galaxies with similar characteristics will statistically share the same mix of morphologies. Thus the GZ2 main galaxy sample is binned by Petrosian absolute magnitude (M_r) and the Petrosian half-light radius, R_{50} , as well as by redshift. A baseline morphology ratio is computed in the lowest redshift bin for those galaxies in the same with confirmed spectroscopic redshifts and with sufficient numbers of votes to yield statistically reliable classifications. This baseline is then used to correct more distant redshift bins. A more detailed account can be found in Willett et al. (2013).

2.2 Automatic morphology indicators

This thesis seeks to draw on the best qualities of all galaxy morphology classification methods including using automated morphologies and machine learning algorithms which provide speed and brute force. Any machine classifier must have a set of features from which to learn to differentiate between classes. These features can be anything that correlates or distinguishes among class labels. In the case of galaxy morphology possible features could include pixel values, spectral indices, magnitude, color, etc. Choosing which features are most appropriate for each task is a difficult task as there are no clear cut rules for feature selection. Too few features and a machine learning algorithm will be unable to learn the parameter space; too many features can result in the Curse of Dimensionality: as the dimensionality of the parameter space increases linearly, the number of samples required to learn that space increases exponentially!

In this work we draw on the Zurich Estimator of Structural Types (ZEST, Scarlata et al., 2007). ZEST utilized five features measured from the light profile of galaxy images combined with a PCA analysis to determine morphologies for 120K COSMOS galaxies. These features are well known to correlate strongly with the distinction between early- and late-type galaxies. In this section we discuss how these values are measured for the GZ2 SDSS galaxy sample.

2.2.1 Imaging data

The Galaxy Zoo 2 main galaxy sample contained 295,305 galaxies though 11,334 are single-epoch imaging from the Stripe 82 region with $m_r > 17.0$, fainter than the rest of the galaxy sample. These galaxies are excluded from the main GZ2 classification catalog though classifications for these galaxies exist in Stripe 82-exclusive catalogs. However, because we utilize the raw volunteer classifications from the original GZ2 project, we include all single-epoch galaxy imaging in our current sample, regardless of magnitude limit.

We obtain *i*-band imaging (with central wavelength 7480Å) from SDSS Data Release 12 for 290,059 galaxies in the GZ2 project. Image identifier such as CAMCOL, RUN, and FIELD are used to select over 151,987 SDSS fields. Because the original GZ2 project obtained imaging from DR7, we surmise that the route to some galaxies in DR12 have

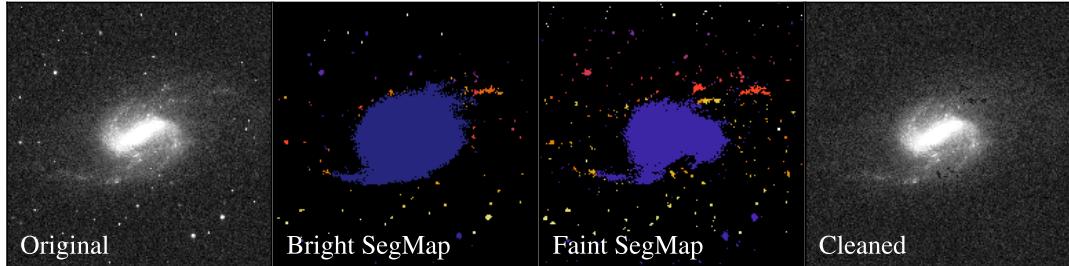


Figure 2.3 Example of SExtractor segmentation maps generated during the postage stamp cleaning process. The left panel shows the original *i*-band postage stamp, the middle two panels show the bright and faint segmentation maps where individual objects detected by SExtractor are shown on an arbitrary rainbow scale, and the right panel shows the resulting cleaned postage stamp.

switched locations or identifiers thus explaining the loss of 5246 galaxies. Because this represents only 1.8% of the total population, these galaxies were not tracked down at this time. Postage stamps of each galaxy are cut from these fields where the dimensions of each cutout are $4 \times$ Petrosian radius as measured by the SDSS pipeline. Galaxies located within 4 Petrosian radii of the edge of a field were excluded as image mosaicing was not performed. This removed 7962 galaxies resulting in a final sample of 282,350 GZ2 galaxy postage stamps or 95.6% of the original sample.

2.2.2 Image cleaning

These postage stamps next undergo a cleaning process in order to remove the light from nearby sources so as not to contaminate the light profile of the galaxy of interest. Each stamp is processed through Source Extractor (ver. 2.8.6; Bertin & Arnouts, 1996), a software that automatically detects sources in CCD imaging based on a set of input parameters that control its sensitivity. Two sets of parameters are used: the first is designed to identify bright sources, while the second is better optimized to detect fainter objects. This software produces segmentation maps that identify the boundaries of each detected object in an image. By design, the galaxy of interest is located at the center of the cutout. Extraneous sources are then identified from both the bright and faint segmentation maps and the pixels corresponding to these sources are replaced with a random value consistent with the background in that postage stamp. An example of the

segmentation maps created during this process is shown in Figure 2.3. Additionally, the first two columns in Figures 2.4 to 2.6 depict random samples of original and cleaned cutouts for a variety of “difficulties,” where the difficulty of successfully cleaning an image of all stray light from other nearby sources is dependent upon how many other sources exist in the postage stamp and how close those sources are to the galaxy of interest.

2.2.3 Morphology indicators

Defining the flux associated with a galaxy is a challenging task: galaxies do not have constant radial surface brightness, hard edges, or uniform shapes. One way to measure the flux is to consider the light within some isophote but the fraction of light included within will be dependent on the amplitude of the galaxy’s surface brightness which diminishes due to cosmological redshift and Galactic extinction. To combat these issues, it is common practice to use the Petrosian (1976) system which computes a radius defined by the empirical shape of the galaxy’s light profile. Specifically, the Petrosian radius, R_p , is such that the ratio of the surface brightness at R_p to the mean surface brightness within R_p is equal to a fixed value η , i.e.,

$$\eta = \frac{\mu(r_p)}{\bar{\mu}(r < r_p)} \quad (2.3)$$

where η is traditionally set to 0.2. Because this is a ratio of surface brightnesses, cosmological factors are mitigated.

We compute R_p by first generating a set of elliptical annuli centered on the galaxy and equidistant in logspace. At least 20 annuli are used for each postage stamp. Elliptical apertures minimize the contribution from noisy background pixels. The position angle and galaxy center are taken from the SExtractor catalogs generated earlier. For each annulus, $\mu(r_i)$ is computed as the flux within the annulus divided by the annulus area, while $\mu(r < r_i)$ is computed as the total flux integrated to the center of that annulus divided by πr_i^2 . This provides a crude estimate of η which is then interpolated onto a finer set of radial values. R_p is that radius for which η intersects 0.2. Examples of surface brightness profile are shown in Figure XXX. XX demonstrates a case in which the cleaning algorithm did not fully remove nearby contaminating light from

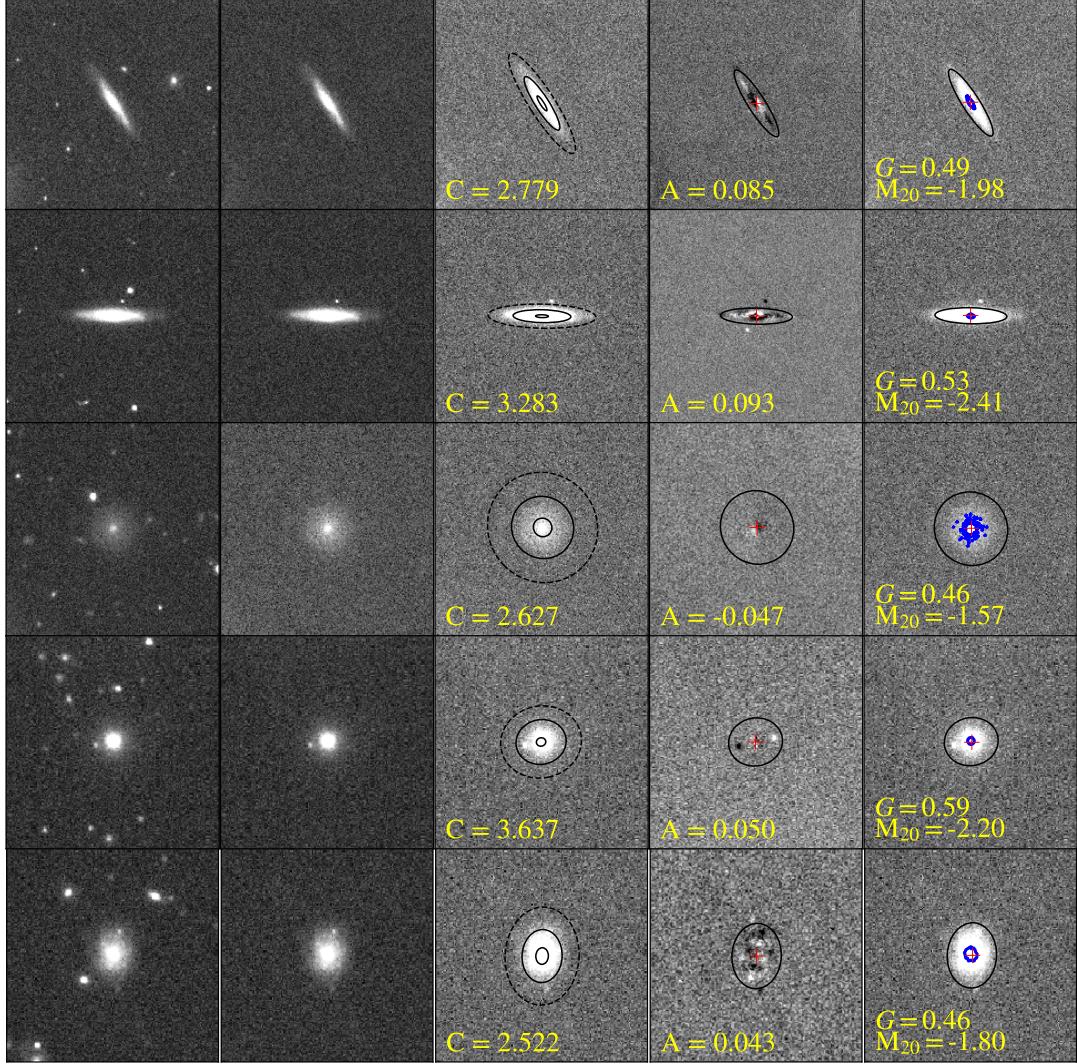


Figure 2.4 Examples of the postage stamp cleaning and morphology diagnostic measuring processes. The first and second columns show the original and cleaned *i*-band postage stamps. The third column shows the apertures used to calculate the concentration index where the two solid ellipses represent the apertures enclosing 20% and 80% of the galaxy total light which is defined as $1.5 \times R_p$ and shown as a dashed ellipse. The fourth column shows the residual asymmetry image generated according to Equation 2.5 where the red cross denotes the galaxy's asymmetry center. The final column shows the Gini and M_{20} values where the red cross denotes the galaxy's M_{20} center and the blue contours trace the brightest 20% of galaxy pixels. In the two rightmost columns, the solid ellipse represents $1R_p$ within which all morphology diagnostics are computed.

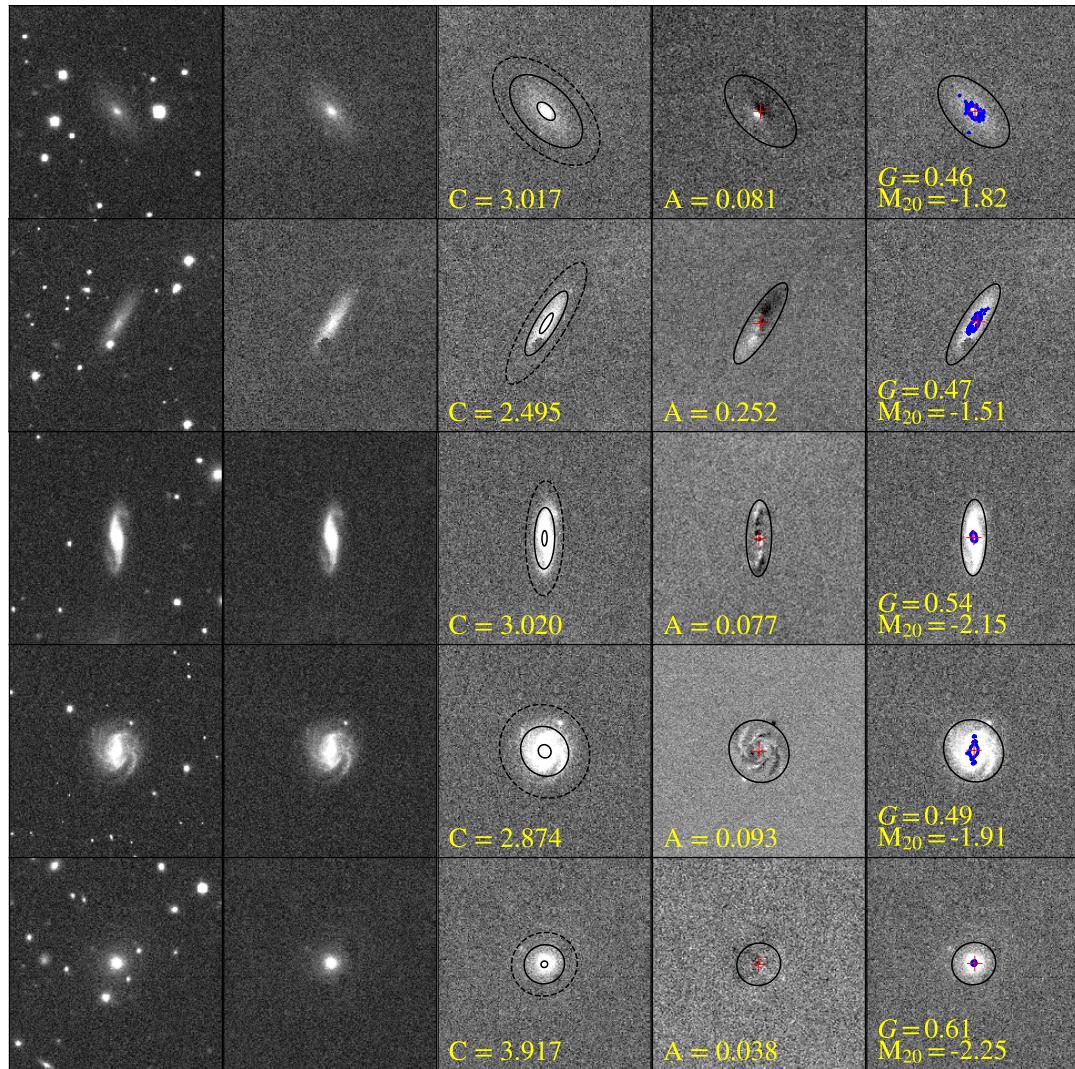


Figure 2.5 Additional examples of the postage stamp cleaning and morphology diagnostic measuring processes. See caption from Figure 2.4.

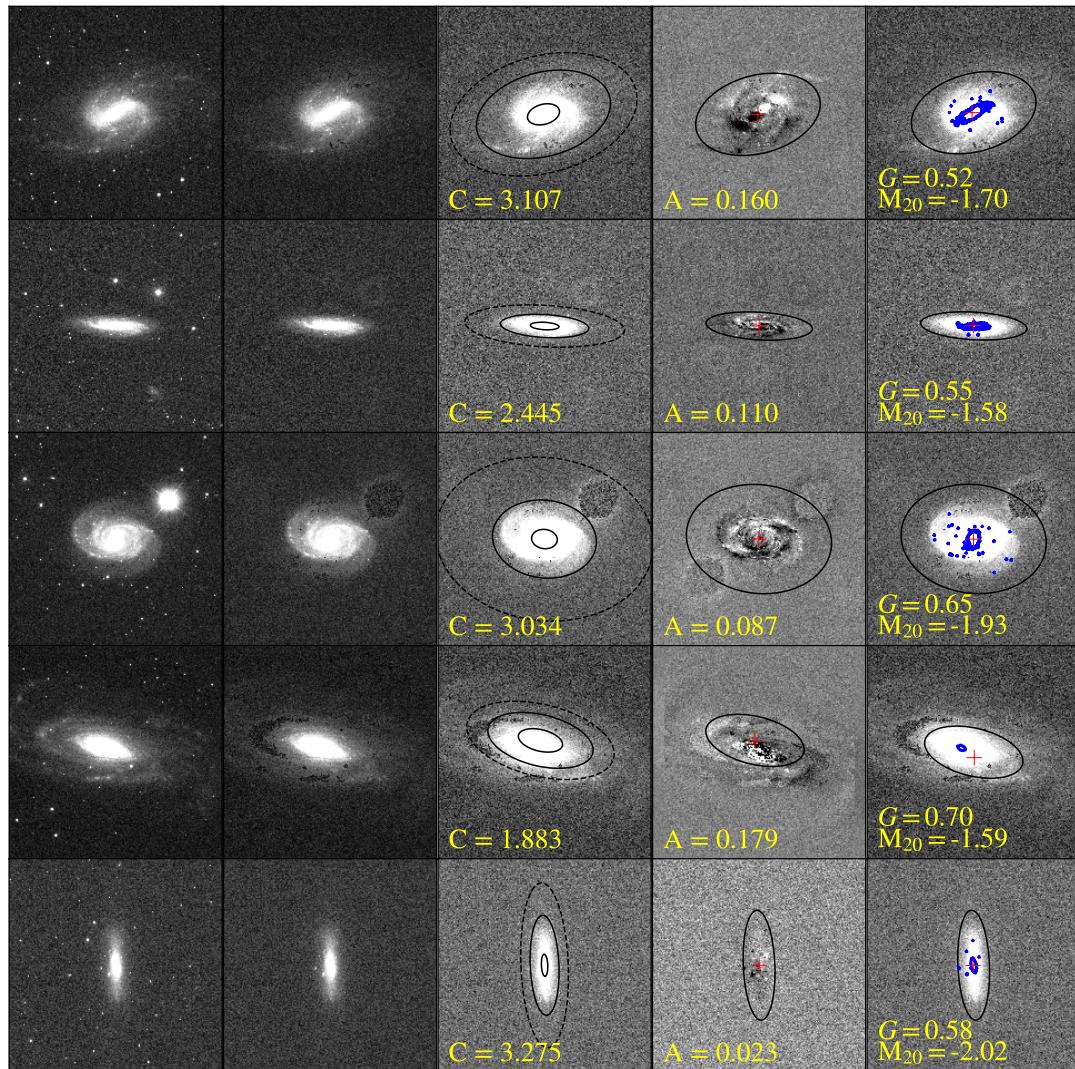


Figure 2.6 Examples of the postage stamp cleaning and morphology diagnostic measuring processes for a random sample of the largest and most difficult galaxies. See caption from Figure 2.4.

other sources, resulting in an η that never crosses the 0.2 threshold. To correct for this, we model the excess light in the outer component with a linear relation and subtract a constant value from the surface brightness profile. Even with these measures, some failures persist, however, we are only unable to obtain R_p for 16 galaxies, a vanishingly small fraction of our sample. We use these values of the Petrosian radius to compute the following morphology diagnostics.

Concentration is defined in several slightly different ways with the aim being to measure the ratio of the light within an inner aperture to that within an outer aperture. Small values of this ratio typically indicate disky galaxies, while larger values correlate with early-type ellipticals. We use the definition of Bershady et al. (2000):

$$C = 5 \log \left(\frac{r_{80}}{r_{20}} \right) \quad (2.4)$$

where r_{80} and r_{20} are the radii containing 80% and 20% of the total galaxy light respectively where we define the total flux as that within R_p and the galaxy center is that determined by the asymmetry minimization (described below, Lotz et al., 2004). A random sample of concentration measurements can be seen in the middle column of Figures 2.4 to 2.6.

The asymmetry parameter, A , quantifies the degree of rotational symmetry in the galaxy light distribution, not necessarily the physical shape of the galaxy as this is not highly sensitive to low surface brightness features. A is measured by subtracting the galaxy image rotated by 180°. A correction for background noise is applied (as in e.g., Conselice et al., 2000; Lotz et al., 2004), i.e.,

$$A = \frac{\sum_{x,y} |I(i,j) - I_{180}(i,j)|}{2 \sum |I(i,j)|} - B_{180} \quad (2.5)$$

where I is the galaxy flux in each pixel (x,y) , I_{180} is the image rotated by 180 degrees about the galaxy's central pixel, and B_{180} is the average asymmetry of the background. A is summed over all pixels within R_p of the galaxy's center and then normalized by a corresponding measure in the original image. The center is determined by minimizing A as described in Conselice et al. (2000). Briefly, an initial central pixel is chosen and A computed. Then asymmetry is calculated again in a 3×3 grid about that central pixel. If one of these produces a lower value of A , it becomes the new center and the process repeats until a minimum is found. This is a crucial step as Conselice et al. (2000) find

that a small difference can change the asymmetry by up to 50%. Additionally, the effects due to noise must also be corrected. This is accomplished by selecting a blank portion of the cutout with the same pixel area as defined for the A measurement. A is computed as before, including the minimization, and this value then constitutes B_{180} . A random sample displaying the apertures and asymmetry centers are shown in the fourth columns of Figures 2.4 to 2.6.

The Gini coefficient, G , has long been used in econometrics as a measure of inequality by estimating the concentration of wealth in a nation's population. It is based on the Lorenz curve (Lorenz, 1905) which is constructed by mapping the cumulative proportion of the population ranked by wealth onto the corresponding cumulative proportion of the size of their wealth. More formally, if X is a positive random variable with cumulative distribution function $F(x)$ then the Lorenz curve goes as

$$L(p) = \frac{1}{\bar{X}} \int_0^p F^{-1}(u) du, \quad (2.6)$$

where p is the percentile of the poorest denizens, and \bar{X} is the mean over all values of X_i , a random deviate drawn from X . G is then a summary statistic of this curve describing the mean of the absolute difference between all combinations of X_i :

$$G = \frac{1}{2\bar{X}n(n-1)} \sum_{i=1}^n \sum_{j=1}^n |X_i - X_j|. \quad (2.7)$$

This statistic was first applied to the distribution of a galaxy's light by Abraham et al. (2003) and further developed by Lotz et al. (2004). In these terms, G is 0 when the galaxy's flux is distributed homogeneously among all assigned pixels, and 1 if the light is concentrated within a single pixel. G can be calculated more efficiently if the X_i are first sorted into increasing order and then computing (Glasser, 1962)

$$G = \frac{1}{|\bar{X}|n(n-1)} \sum_i^n (2i - n - 1)|X_i|, \quad (2.8)$$

where n is the number of pixels assigned to the galaxy. Here we follow Lotz et al. (2004) by taking the absolute value of the flux, X_i , because in pixels with low signal-to-noise the flux is scattered to values below the mean sky level resulting in negative flux values for the faintest pixels.

Assigning pixels to each galaxy must be given due consideration. As Lotz et al. (2004) point out, including too many sky pixels will systematically increase G , while exclusion of low surface brightness galaxy pixels will decrease G . Abraham et al. (2003) measure G for pixels above a given surface brightness threshold but this will fall prey to cosmological effects. In this work, we compute G from all pixels within R_p . This will exclude some low surface brightness features, especially for galaxies with faint, extended elements. However, this definition puts every galaxy on equal footing, and as we discuss in Chapter 4, systematics are easily handled by the machine learning algorithm we exploit. The last column in Figures 2.4 to 2.6 shows examples of G .

M_{20} (Lotz et al., 2004) is the second order moment of the brightest 20% of the galaxy flux. It traces the spacial distribution of any bright galactic features such as nuclei, bars, spiral arms, or star-forming clusters. It is computed by first calculating the total moment as

$$M_{\text{tot}} = \sum_i^n M_i = \sum_i^n f_i[(x_i - x_c)^2 + (y_i - y_c)^2] \quad (2.9)$$

where f_i is the flux in pixel x_i , y_i , and x_c , y_c is the galaxy's center which is determined by minimizing M_{tot} in a similar fashion as is done for the asymmetry. The galaxy pixels are then ranked by flux in descending order and M_i is summed over the brightest pixels until that sum equals 20% of the total galaxy flux, normalized by M_{tot} :

$$M_{20} = \log_{10} \left(\frac{\sum_i M_i}{M_{\text{tot}}} \right), \quad \text{while } \sum_i f_i < 0.2 f_{\text{tot}} \quad (2.10)$$

where f_{tot} is the total flux defined within R_p . For centrally concentrated objects, M_{20} correlates with C but is also sensitive to bright off-centre knots of light.

It's worth noting that both M_{20} and G are correlated with C , but with key differences. Because G is a measure of concentration it correlates strongly with C , especially for local galaxies Abraham et al. (2003). However, G is independent of spatial distribution: whereas C measures the concentration in the central region of a galaxy, G is sensitive to any concentration of light: centralized or not. M_{20} takes this a step further with its key difference being a strong r^2 dependence. It is strongly weighted by the spatial distribution of bright regions but these need not be centralized either. Indeed, when G and M_{20} are taken together they are highly adept at identifying merging galaxies (Lotz et al., 2004; ?).

Table 2.1

Morphology measurement summary

	Number	% Success	Notes
Full Galaxy Zoo 2 sample	295 305		
Postage stamps	282 350	95.6	% of full sample
Petrosian radius	282 334	99.99	% of postage stamps
Concentration	281 927	99.85	% of postage stamps
Asymmetry	282 334	99.99	% of postage stamps
Gini coefficient	282 323	99.99	% of postage stamps
M_{20}	282 194	99.94	% of postage stamps
Ellipticity ($1 - b/a$)	282 350	100.0	% of postage stamps

For our final morphology diagnostic, we use the ellipticity, $\epsilon = 1 - b/a$, of the light distribution as measured by SExtractor which computes the semi-major axis a and semi-minor axis b from the second-order moments of the galaxy light. The ellipticity of a galaxy correlates strongly with edge-on galaxies.

In total, we measure morphological indicators for 282,350 SDSS galaxies. Some galaxies are lost at each stage of the measurement process due to various failures which we discuss in detail below. For example, if our measurement of the Petrosian radius is not successful, no morphology diagnostics are computed for that galaxy. The number of galaxies with successful measurements at each stage is listed in Table 2.1. The relations between these diagnostics for the full sample is shown in Figure 3.11. The code developed to clean and compute these morphology indicators is open source and can be found at https://github.com/melaniebeck/measure_morphology.

2.2.4 Quality and consistency

We perform several checks to determine the quality and consistency of our morphology diagnostics. Due to the wealth of information provided by the SDSS pipeline, we can compare some of our diagnostics against SDSS values. Obviously there are some instances where our measurements will outperform the SDSS pipeline and vice versa. We first check our values of the Petrosian radius as shown in Figure 2.7. Our R_p are on average $\sim 35\%$ larger than those computed by SDSS. The biggest reason for this discrepancy is due to aperture shape: SDSS compute R_p using circular annuli while we

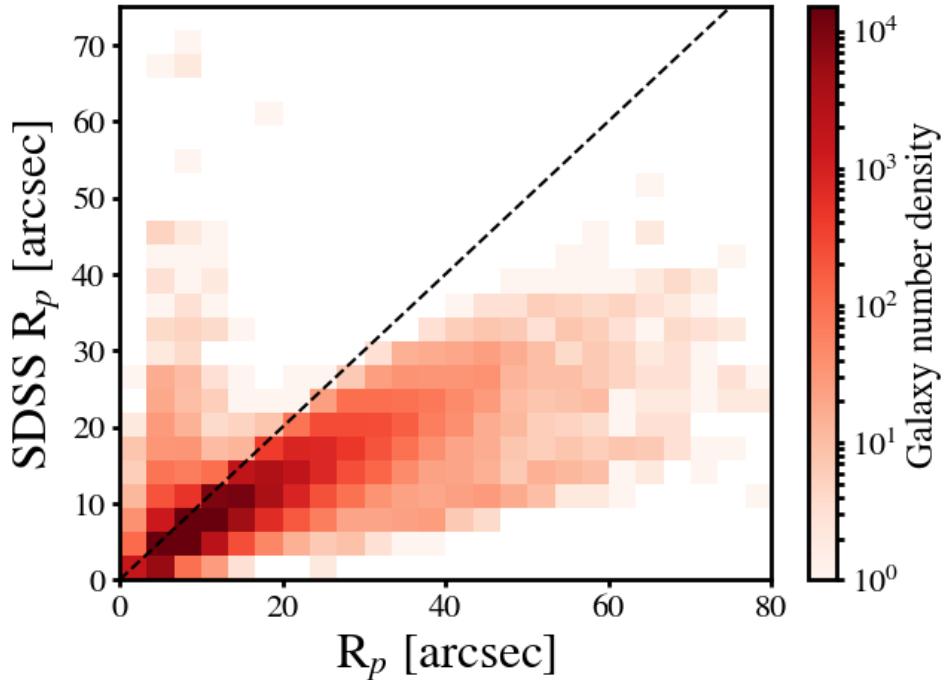


Figure 2.7 Comparison of the Petrosian radius R_p computed in the SDSS pipeline to that calculated in this work. Most of the discrepancy between these values is due to the apertures used: the SDSS pipeline defines R_p in circular annuli while we use elliptical.

use elliptical. [Mention the weird ones where our R_p is tiny but SDSS's is huge? Show some examples where my code does better than SDSS? And where mine fails?]

Of greater importance, however, are the values of the morphology diagnostics as these are used as features to train our machine learning algorithm in Chapter 4. Though SDSS does not measure all of the structural indicators we tackle here, they do provide a means to compute the concentration index via PETROR50 and PETROR90, the Petrosian radii containing 50% and 90% of the galaxy total light, respectively. The left panel of Figure 2.8 shows our C against that computed from SDSS, C_{SDSS} . Our values are systematically ~ 1 point larger than SDSS values but otherwise retain a strikingly tight correlation. This is surprising considering the drastically different ways these values are computed. Besides the obvious difference in radii (SDSS uses 50% and 90% vs our 20% and 80%), SDSS values are again computed using circular apertures whereas we use elliptical for all measurements. The latter are preferable as it has been shown that

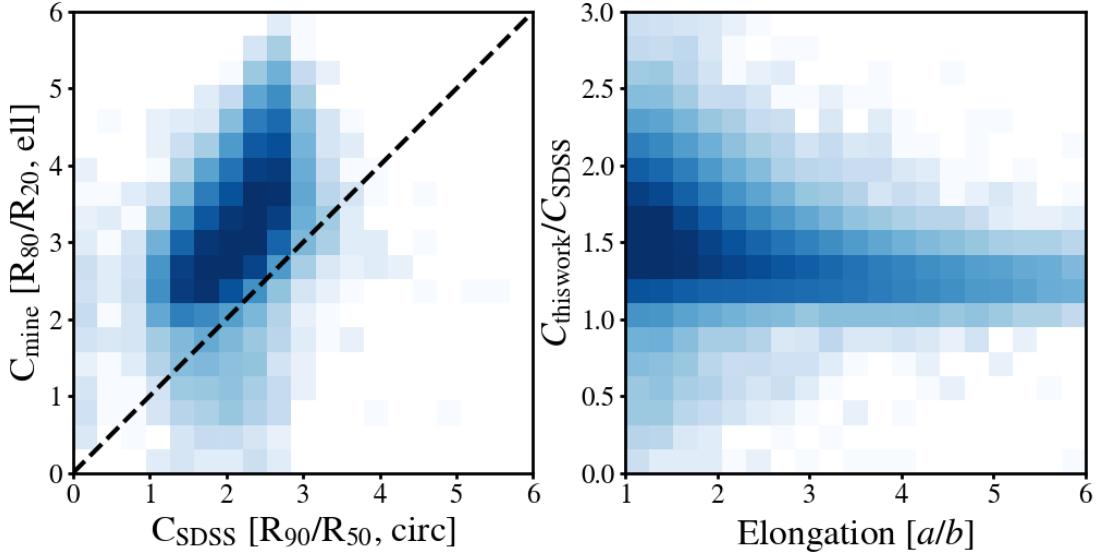


Figure 2.8 The left panel shows our measured concentration index against that computed from values taken from SDSS DR7. We find that our C is systematically ~ 1 point larger than SDSS. This is quite remarkable considering C_{SDSS} is computed using Petrosian radii containing 50% and 90% of the total galaxy light, while we use radii containing 20% and 80% of the galaxy total light. Additionally, SDSS use circular annuli when computing their radii whereas we use elliptical. The right panel shows the ratio C/C_{SDSS} as a function of galaxy elongation.

aperture shape affects C (Bershady et al., 2000; Andrae et al., 2011). This is hinted at in the right panel of Figure 2.8 which shows the ratio C/C_{SDSS} as a function of galaxy elongation. Andrae et al. (2011) demonstrate a bias such that the concentration index is artificially inflated when measured from circular apertures for highly elongated galaxies. Instead the opposite should be seen, that is, early-type galaxies tend to have much higher concentrations and are generally less elongated (compared to edge-on disks!). This is the trend we see in the right panel due to the fact that we use elliptical apertures to compute C and thus our C values are less prone to this potential bias.

SDSS does not provide any other direct morphology diagnostics comparable to what we measure here so we next assess the quality of our structural indicators by estimating the fraction that are poorly measured. We consider the galaxy central coordinates assigned at various stages during the cleaning and measuring process and compare those to the original SDSS galaxy coordinates. SExtractor assigns galaxy central coordinates

for each postage stamp based on the first order moment of light. Additionally, the asymmetry and M_{20} measurements each determine their own galaxy center via the minimization of their respective quantity and thus do not necessarily overlap. We compute the coordinate difference in arcseconds for each of these three centers as compared to the SDSS coordinates and determine that only 3.0%, 1.9%, and 1.5% of galaxies have SExtractor, asymmetry and M_{20} centers, respectively, that differ by more than $1''$ from the original SDSS coordinates. Thus we find very good agreement between SDSS and SExtractor galaxy centers, indicating that we correctly identify the galaxy of interest in each postage stamp. We also find good agreement between SDSS and both the A and M_{20} centers indicating that our minimizing algorithm fails only in a handful of instances.

Finally we consider the overall distribution of our measured structural indicators. It is well known that different galaxy types live in different parts of morphology parameter space. For example, in the $G-M_{20}$ plane, early-type galaxies are typically found in the upper right with late-type galaxies residing in the lower left, and merging systems in the upper left. Similar relations exist in the $C-G$, and $G-M_{20}$ planes. In Figure 2.9 we select a random sample of 5000 galaxies and plot them in each possible planar combination of our morphology diagnostics where the points are colored according to their Galaxy Zoo 2 debiased f_{smooth} fraction. Because f_{smooth} correlates well with early-type galaxies we see the expected trends discussed above.

2.3 Catalog of morphological indicators for 282350 SDSS galaxies

Put in a big table here with a subsample of the catalog? Link to somewhere online where the rest of the catalog can be found?

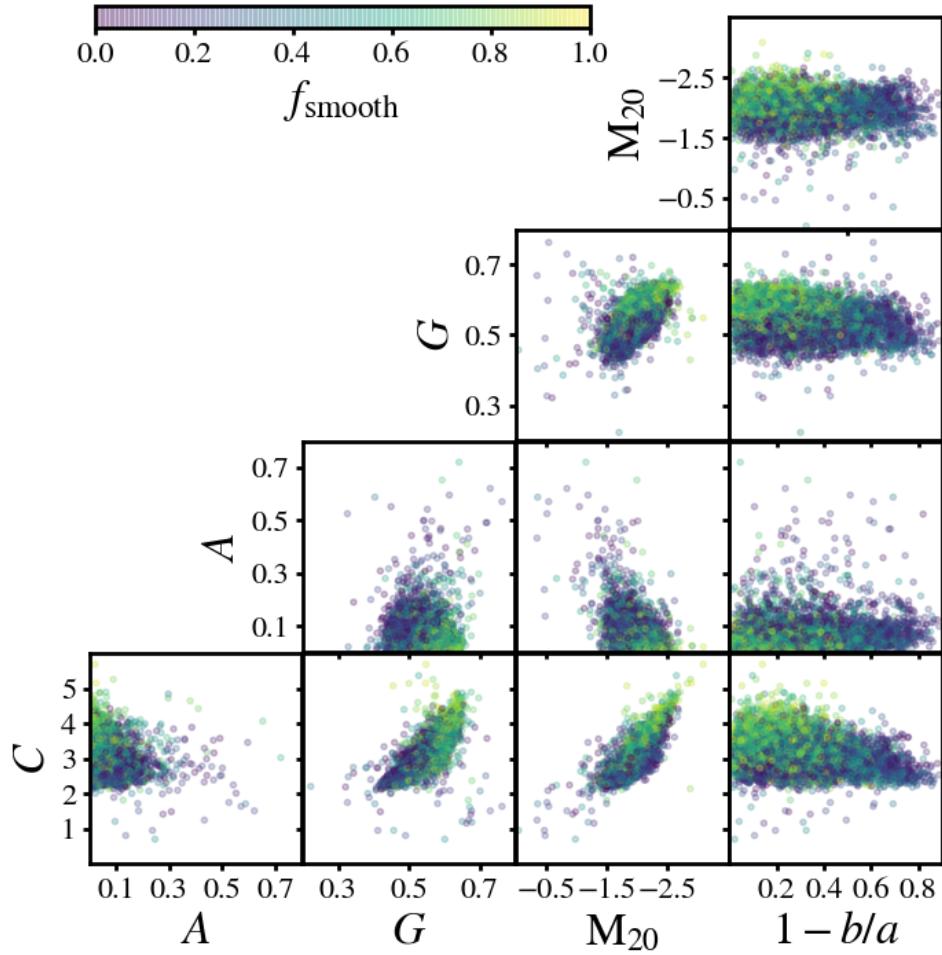


Figure 2.9 Here a random sample of 5000 galaxies are plotted in every possible planar combination of our morphology structural indicators where the color denotes the galaxy's debiased f_{smooth} fraction from GZ2. As expected, galaxies with higher f_{smooth} are found in locations where early-types are typically seen: notably they have larger C and G , and smaller M_{20} .

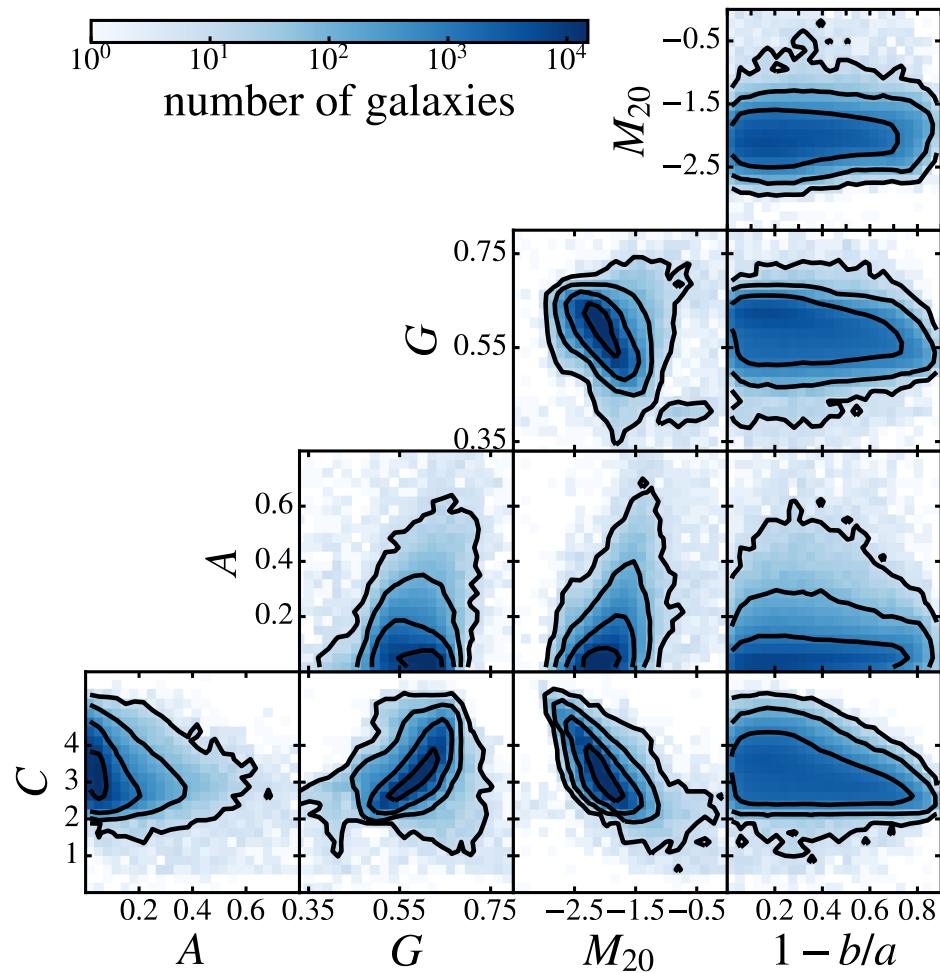


Figure 2.10 *Right*. Relation between measured morphology diagnostics for more than 280K SDSS galaxies. Correlations between several diagnostics are immediately obvious and not all relations are likely to be linear. These points will be revisited in Chapter 4 when we discuss machine learning algorithms.

Table 2.2. Morphology Diagnostics for $\sim 283K$ SDSS Galaxies in the GZ2 Sample

DR7 objID	RA	DEC	R_p	C	A	M_{20}	G	e	flags
587722981736054938	170.749	-1.25746	44.09	2.556	0.0243	-1.888	0.539	4.289	

Chapter 3

Intelligent management of visual classifications

3.1 Galaxy Zoo 2 Classification Data

Our simulations utilize original classifications made by volunteers during the GZ2 project. These data¹ are described in detail in Willett et al. (2013) and the preceding Chapter, though we provide a brief overview here. The GZ2 subject sample consists of 285,962 galaxies identified as the brightest 25% (r -band magnitude < 17) residing in the SDSS North Galactic Cap region from Data Release 7 and included subjects with both spectroscopic and photometric redshifts out to $z < 0.25$. Subjects were shown as colour composite images via a web-based interface² wherein volunteers answered a series of questions pertaining to the morphology of the subject. With the exception of the first question, subsequent queries were dependent on volunteer responses from the previous task creating a complex decision tree³. Using GZ2 nomenclature, a *classification* is the total amount of information about a subject obtained by completing all tasks in the decision tree. A subject is *retired* after it has achieved a sufficient number of classifications.

For our current analysis, we choose the first task in the tree: “Is the galaxy simply

¹ data.galaxyzoo.org

² www.galaxyzoo.org

³ A visualization of this decision tree can be found at https://data.galaxyzoo.org/gz_trees/gz_trees.html

smooth and rounded, with no sign of a disk?” to which possible responses include “smooth”, “features or disk”, or “star or artifact”. This choice serves two purposes: 1) this is one of only two questions in the GZ2 decision tree that is asked of every subject thus maximizing the amount of data we have to work with, and 2) our analysis assumes a binary task and this question is simple enough to cast as such. Specifically, we combine “star or artifact” responses with “features or disk” responses.

We assign each subject a descriptive label in order to validate our classification output with GZ2. GZ2 classifications are composed of volunteer vote fractions for each response to every task in the decision tree, denoted as f_{response} . They are derived from the fraction of volunteers who voted for a particular response and are thus approximately continuous. A common technique is to place a threshold on these vote fractions to select samples with an emphasis on purity or completeness, depending on the science case. For our current analysis we choose a threshold of 0.5, that is, if $f_{\text{featured}} + f_{\text{artifact}} > f_{\text{smooth}}$, the galaxy is labelled ‘Featured’, otherwise it is labelled ‘Not’. We note that only 512 subjects in the GZ2 catalogue have a majority f_{artifact} , contributing less than half a percent contamination when combining the “star or artifact” with “features or disk” responses.

The GZ2 catalogue publishes three types of vote fractions for each subject: raw, weighted, and debiased. Debiased vote fractions are calculated to correct for redshift bias, a task that GZX does not perform. The weighted vote fractions account for inconsistent volunteers. The SWAP algorithm (described below) also has a mechanism to weight volunteer votes, however, the two methods are in stark contrast. For consistency, we thus derive labels from the raw vote fractions (GZ2_{raw}); those that have received no post-processing whatsoever. In total, the data consist of over 14 million classifications from 83,943 individual volunteers.

The labels we compute from GZ2 vote fractions are used solely to validate our classification method and are thus considered “ground truth,” though this is, of course, subjective. Furthermore, we envision our framework being applied to never-before-classified image sets for which “ground truth” labels would not yet exist. Nevertheless, in Appendix XXX we show how different choices of our descriptive GZ2 labels change the perceived quality of our classification system and demonstrate that our method yields robust galaxy classifications.

3.2 Efficiency through intelligent human-vote aggregation

Galaxy Zoo 2 had a brute-force subject retirement rule whereby each galaxy was to receive approximately forty independent classifications. Once the project reached completion, inconsistent volunteers were down-weighted (Willett et al., 2013), a process that does not make efficient use of those who are exceptionally skilled. To intelligently manage subject retirement and increase classification efficiency, we adapt an algorithm from the Zooniverse project Space Warps (Marshall et al., 2016), which searched for and discovered several gravitational lens candidates in the CFHT Legacy Survey (More et al., 2016). Dubbed SWAP (Space Warps Analysis Pipeline), this algorithm computed the probability that an image contained a gravitational lens given volunteers' classifications and experience after being shown a training sample consisting of simulated lensing events. We provide a brief overview here.

The algorithm assigns each volunteer an *agent* which interprets that volunteer's classifications. Each agent assigns a 2×2 confusion matrix to their volunteer which encodes that volunteer's probability to correctly identify feature A given that the subject exhibits feature A ; and the probability to correctly identify the absence of feature A (denoted N) given that the subject does not exhibit that feature. The agent updates these probabilities by estimating them as

$$P(\text{"}X\text{"}|X, \mathbf{d}) \approx \frac{\mathcal{N}_{\text{"}X\text{"}}}{\mathcal{N}_X} \quad (3.1)$$

where X is the true classification of the subject and "X" is the classification made by the volunteer upon viewing the subject. Thus $\mathcal{N}_{\text{"}X\text{"}}$ is the number of classifications the volunteer labelled as type X , \mathcal{N}_X is the number of subjects the volunteer has seen that were actually of type X , and \mathbf{d} represents the history of the volunteer, i.e., all subjects they have seen. Therefore the confusion matrix for a single volunteer goes as

$$\mathcal{M} = \begin{bmatrix} P(\text{"}A\text{"}|N, \mathbf{d}) & P(\text{"}A\text{"}|A, \mathbf{d}) \\ P(\text{"}N\text{"}|N, \mathbf{d}) & P(\text{"}N\text{"}|A, \mathbf{d}) \end{bmatrix} \quad (3.2)$$

where probabilities are normalised such that $P(\text{"}A\text{"}|A) = 1 - P(\text{"}N\text{"}|A)$.

Each subject is assigned a prior probability that it exhibits feature A : $P(A) = p_0$. When a volunteer makes a classification, Bayes' theorem is used to compute how that

subject's prior probability should be updated into a posterior using elements of the agent's confusion matrix. As the project progresses, each subject's posterior probability is updated after every volunteer classification, nudged higher or lower depending on volunteer input. Upper and lower probability thresholds can be set such that when a subject's posterior crosses the upper threshold it is highly likely to exhibit feature *A*; while if it crosses the lower threshold it is highly likely that feature *A* is absent. Subjects whose posteriors cross either of these thresholds are considered retired.

3.2.1 Gold-standard sample

A key feature of the original Space Warps project was the training of individual volunteers through the use of simulated images. These were interspersed with real imaging and were predominantly shown at the beginning of a volunteer's association with the project, allowing that volunteer's agent time to update before classifying real data. Volunteers were provided feedback in the form of a pop-up comment after classifying a training image. GZ2 did not train volunteers in such a way, presenting a challenge when applying SWAP to GZ2 classifications. Though we cannot retroactively train GZ2 volunteers, we develop a gold standard sample and arrange the order of gold standard classifications in order to mimic the Space Warps system.

We create a gold standard sample by selecting 3496 SDSS galaxies representative of the relative abundance of T-Types, a numerical index of a galaxy's stage along the Hubble sequence, at $z \sim 0$ by considering galaxies that overlap with the Nair & Abraham (2010) catalogue, a collection of $\sim 14K$ galaxies classified by eye into T-Types. We generate expert labels for these galaxies that are consistent with the labels we defined for GZ2 classifications. These are obtained through the Zooniverse platform⁴ from 15 professional astronomers, including members of the Galaxy Zoo science team. The question posed was identical to the original top-level GZ2 question and at least five experts classified each galaxy. Votes are aggregated and a simple majority provides an expert label for each subject. This ensures that our expert labels are defined in exactly the same manner as the labels we assign the rest of the GZ2 sample. Our final dataset consists of the GZ2 classifications made by those volunteers who classify at least one of these gold standard subjects. We thus retain for our simulation 12,686,170 classifications

⁴ The Project Builder template facility can be found at <http://www.zooniverse.org/lab>.

from 30,894 unique volunteers. When running SWAP, classifications of gold standard subjects are always processed first.

3.2.2 Volunteer Bias

By comparing expert and GZ2 volunteer votes we uncover indications of volunteer bias towards labeling galaxies as ‘Smooth’ (‘Not Featured’). For each galaxy in our gold standard sample we identify the five most skilled volunteers who classified that galaxy, where skill is determined by the volunteer confusion matrix in Figure 3.4. We compute f_{smooth} from those five volunteers. In Figure 3.1 we plot the volunteer f_{smooth} against the expert f_{smooth} sample, where the size of the circle is proportional to the number of subjects wherein both volunteers and experts agreed. Circles representing more than ten gold standard subjects are also labelled with the corresponding number of galaxies at that location in f_{smooth} space. For example, there are 1316 subjects that both experts and volunteers labeled ‘Featured’ with $f_{\text{smooth}} = 0.0$. Though the vast majority of galaxies received five expert classifications, some received more than that resulting in blue circles at one sixth intervals instead of the majority at fifths. The large circles at 0.0 and 1.0 indicate that both experts and volunteers agree for more than half of the gold standard sample. However, that the circles extend almost solely into the upper left hand corner indicates that volunteers consistently label galaxies as ‘Smooth’ more so than experts. Of this small sample, 40% of galaxies are given a larger f_{smooth} by volunteers than by experts.

It is well known that redshift and surface brightness affect the apparent strength of morphological features in CCD imaging in that finer details are harder to identify. GZ2 mitigate these issues by “debiasing” volunteer vote fractions (Willett et al., 2013). We compare both the original raw and debiased f_{smooth} vote fractions for the gold standard sample and find that it cannot account for the pronounced bias seen here. The reason for this is unsurprising as this sample was selected from the Nair & Abraham (2010) catalog, which consists of bright, low-redshift galaxies. Effects due to cosmic distance are thus minor.

This then suggests that the bias could be inherent to the question posed to volunteers: “Is the galaxy simply smooth and rounded, with no sign of a disk?” Terminology such as “disk” could be misinterpreted by those without specific astrophysics training.

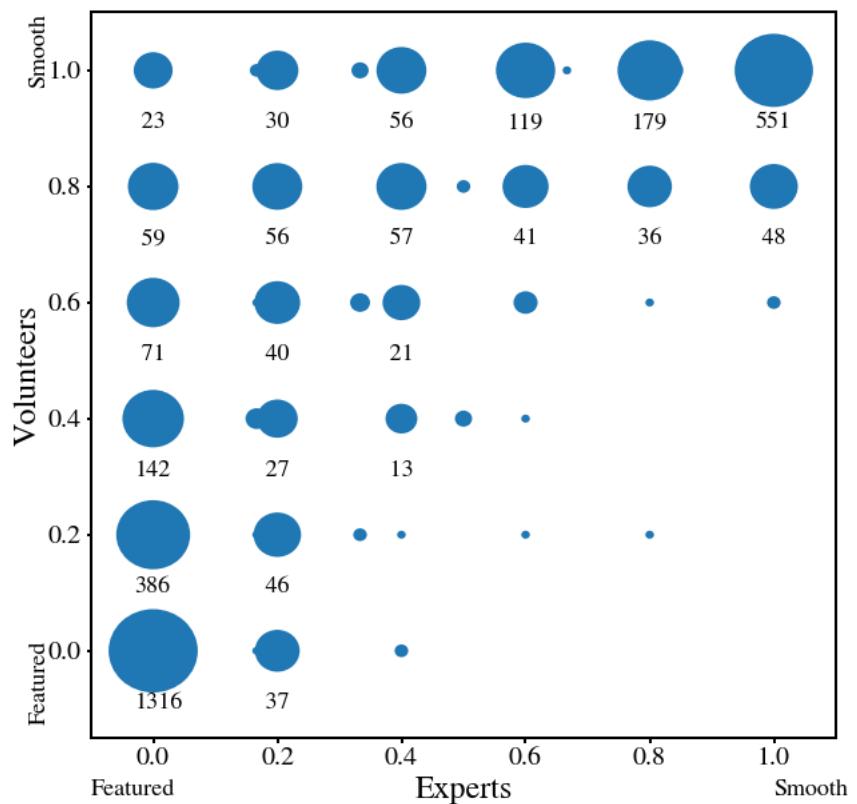


Figure 3.1 Volunteer bias of labeling galaxies ‘Smooth’ (‘Not’) as compared to expert classifications for the gold standard sample.

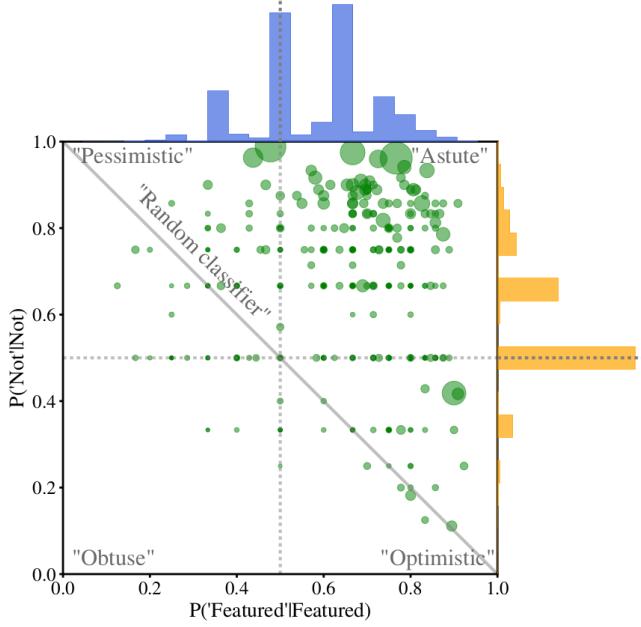


Figure 3.2 Confusion matrices for 1000 randomly selected GZ2 volunteers after fiducial SWAP assessment. Circle size is proportional to the number of gold standard subjects each volunteer classified. The histograms on top and right represent the distribution of each component of the confusion matrix for all volunteers. A quarter of GZ2 volunteers are “Astute”; they correctly identify both ‘Featured’ and ‘Not’ subjects more than 50% of the time. The peaks at 0.5 in both distributions are due primarily to volunteers who see only one training image: only half of their confusion matrix is updated.

Experienced astronomers are more adept at identifying galactic disks, even those which possess few other obvious features. On the other hand, if volunteers do not see obvious features such as spiral arms or a bar, they could well mark galaxies as “smooth.” A similar issue is also observed in the GZ: CANDELS project in which Simmons et al. (2017) compares volunteer classifications to expert classifications collected by the CANDELS team (Kartaltepe et al., 2015). This comparison is, however, not ideal as the question posed to the CANDELS team was not identical to that given to GZ volunteers. With the small sample presented here, it is difficult to determine how much this bias could affect the full GZ2 galaxy sample though we touch on this issue again in Chapter 4.

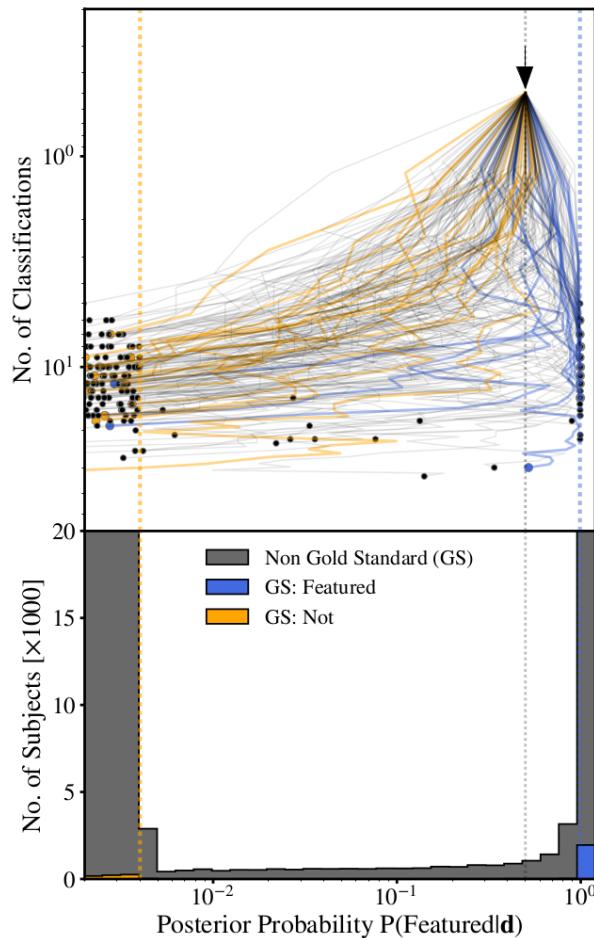


Figure 3.3 Posterior probabilities for GZ2 subjects. The top panel depicts the probability trajectories of 200 randomly selected GZ2 subjects. All subjects begin with a prior of 0.5 denoted by the arrow. Each subject's probability is nudged back and forth with each volunteer classification. From left to right the dotted vertical lines show the 'Not' threshold, prior probability, and 'Featured' threshold. Different colours denote different types of subjects. The bottom panel shows the distribution in probability for all GZ2 subjects by the end of our simulation, where the y axis is truncated to show detail.

3.2.3 Fiducial SWAP simulation

Before we run a simulation, a number of SWAP parameters must be chosen: the initial confusion matrix for each volunteer’s agent, ($P(\text{“F”}|F)$, $P(\text{“N”}|N)$); the subject prior probability, p_0 ; and the retirement thresholds, t_F and t_N . For our fiducial simulation, we initialize all confusion matrices at (0.5, 0.5), and set the subject prior probability, $p_0 = 0.5$. We set the ‘Featured’ threshold, t_F , i.e., the minimum probability for a subject to be retired as ‘Featured’, to 0.99. Similarly, we set the ‘Not’ threshold, $t_N = 0.004$. In Section XXX we show that varying these parameters has only a small affect on the SWAP output. To simulate a live project, we run SWAP on a time step of $\Delta t = 1$ day, during which SWAP processes all volunteer classifications with timestamps within that range. This is performed for three months worth of GZ2 classification data.

Figure 3.2 (adapted from Figure 4 of Marshall et al. 2016) demonstrates the volunteer assessment we achieve, and shows confusion matrices for 1000 randomly selected volunteers. The circle size is proportional to the number of gold standard subjects each volunteer classified. The histograms represent the distribution of each component of the confusion matrix for all volunteers. Nearly 25% of volunteers are considered “Astute” indicating they correctly identify both ‘Featured’ and ‘Not’ subjects more than 50% of the time. Furthermore, as long as a volunteer’s confusion matrix is different from random, they provide useful information to the project. The spikes at 0.5 in the histograms are due to volunteers who see only one gold standard subject (i.e., ‘Featured’), leaving their probability in the other (‘Not’) unchanged. Additionally, 4% of volunteers have a confusion matrix of (0.5, 0.5) indicating these volunteers classified two gold standard subjects of the same type, one correctly and one incorrectly.

Figure 3.3 (adapted from Figure 5 of Marshall et al. 2016) demonstrates how subject posterior probabilities are updated with each classification. The arrow in the top panel denotes the prior probability, $p_0 = 0.5$. With each classification, that prior is updated into a posterior probability creating a trajectory through probability space for each subject. The blue and orange lines show the trajectories of a random sample of ‘Featured’ and ‘Not’ subjects from our gold standard sample, while the black lines show the trajectories of a random sample of GZ2 subjects that were not part of the gold standard sample. The similarly coloured vertical dashed lines correspond to the retirement thresholds, t_F and t_N . The lower panel shows the full distribution of GZ2 subject

		GZX Prediction	
		Predicted “Featured”	Predicted “Not”
GZ2 classification	Labelled “Featured”	True Positives (TP) (Both methods agree)	False Negatives (FN) (Methods disagree)
	Labelled “Not”	False Positives (FP) (Methods disagree)	True Negatives (TN) (Both methods agree)

Figure 3.4 Confusion matrix for comparing GZ2 classifications to our method. True positives (TP) and true negatives (TN) indicate that the predictions from our method agree with GZ2 for subjects labelled ‘Featured’ and ‘Not’, respectively. When the two classification methods disagree, the result is a sample of false negatives (FN) and false positives (FP). This allows us to easily compute quality metrics like accuracy, completeness, and purity with respect to GZ2 as shown in Equations 3.

postiors at the end of our simulation, where the y-axis has been truncated to show detail. An overwhelming majority of subjects cross one of these retirement thresholds.

Our goal is to increase the efficiency of galaxy classification. We therefore use as a metric the cumulative number of retired subjects as a function of the original GZ2 project time. We define a subject as GZ2-retired once it achieves at least 30 volunteer votes, encompassing 98.6% of GZ2 subjects (explored in depth in Section XXX). In contrast, a subject is considered SWAP-retired once its posterior probability crosses either of the retirement thresholds defined above.

However, it is important not to prioritize efficiency at the expense of quality. Because we have a binary classification we can construct a confusion matrix from which we can compute the quality metrics of accuracy, completeness and purity as a function of GZ2 project time by comparing our predicted labels to the GZ2_{raw} labels. Figure 3.4 graphically depicts the elements of this confusion matrix. From this we compute:

$$\begin{aligned}
\text{accuracy} &= \frac{TP + TN}{TP + FP + TN + FN} \\
\text{completeness} &= \frac{TP}{TP + FN} \\
\text{purity} &= \frac{TP}{TP + FP}
\end{aligned} \tag{3}$$

Thus, a complete sample recovers *all* subjects labelled ‘Featured’ by GZ2, whereas a pure sample recovers *only* subjects labelled ‘Featured’ by GZ2. For example, by Day 20, SWAP retires 120K subjects with 96% accuracy, 99.7% completeness, and 92% purity.

Figure 3.5 and Table 4.1 detail the results of our fiducial SWAP simulation compared to the original GZ2 project. The bottom panel shows the cumulative number of retired subjects as a function of GZ2 project time. By the end of our simulation, GZ2 (dashed dark blue) retires \sim 50K subjects while SWAP (solid light blue) retires 226,124 subjects. We thus classify 80% of the entire GZ2 sample in three months. Processing volunteer classifications through SWAP presents nearly a factor of 5 increase in classification efficiency. The top panel of Figure 3.5 demonstrates the quality of those classifications as a function of time and establishes that our full SWAP-retired sample is 95.7% accurate, 99% complete, and 86.7% pure. We discuss these small discrepancies in Section XXX.

3.2.4 Intelligent subject retirement

That SWAP achieves a classification rate nearly 5 times faster than GZ2 comes with a caveat: we consider only the top-level question of the GZ2 decision tree, which, it can be argued, required that GZ2 secure enough votes to populate the subqueries. In order to put SWAP and GZ2 on equal footing we determine the minimum number of votes, N , that the GZ2 project would need to achieve the same top-level classification for 95% of its sample.

We compute the raw vote fractions (f_{featured} , f_{smooth} , and f_{artifact}) for every subject in the GZ2 sample using only the first N classifications for $N \in [10, 15, 20, 25, 30, 35]$. From this, we compute descriptive labels as described in Section A.3. Because our SWAP simulation did not retire every subject in the GZ2 sample, we select 100 random subsamples each consisting of 226,124 subjects, and compute the average accuracy and the average number of total GZ2 classifications necessary retire that subsample. These

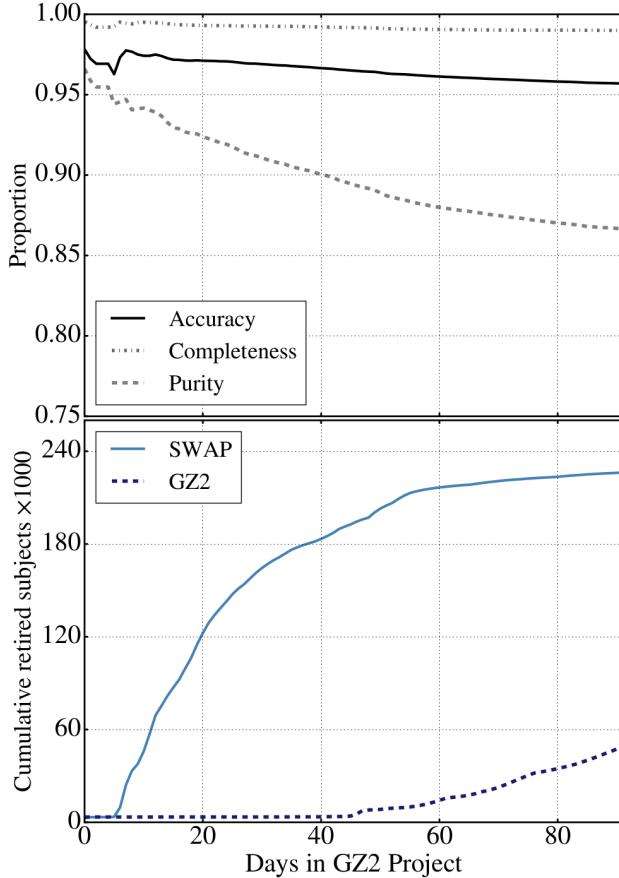


Figure 3.5 Fiducial SWAP simulation demonstrates a factor of 4.7 increase in the rate of subject retirement as a function of GZ2 project time (bottom panel, light blue) compared with the original GZ2 project (dashed dark blue). After 92 days, SWAP retires over 226K subjects, while GZ2 retires \sim 48K. The top panel displays the quality metrics (greys). These are calculated by comparing labels predicted by SWAP to GZ2_{raw} labels (Section A.3) for the subject sample retired by that day of the simulation. Thus, on the final day, SWAP retires 226,124 subjects with 95.7% accuracy, and with completeness and purity of ‘Featured’ subjects at 99% and 86.7% respectively. The decrease in purity as a function of time is due, in part, to the fact that more difficult to classify subjects are retired later in the simulation (see Section 3.2.4).

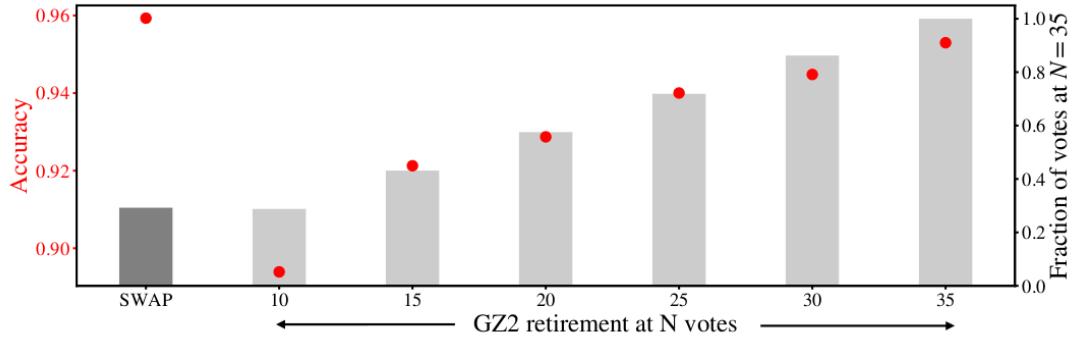


Figure 3.6 SWAP’s intelligent retirement mechanism allows it to use 30% fewer classifications than GZ2 for the top-level question. To see this, we determine the minimum number of votes GZ2 needs in order to achieve consistent top-level classifications 95% of the time, allowing us to compare the total number of votes in our SWAP simulation to the total number required for GZ2. We compute the vote fractions f_{smooth} , f_{featured} , and f_{artifact} using only the first N votes in GZ2 and then compute the resulting class label. We compare these labels to those we originally computed in Section A.3 using the full GZ2 vote fractions and compute the accuracy and the total number of votes necessary for subject retirement. We do this for 100 randomly selected subsamples from GZ2 with the same number of subjects as retired during our SWAP simulation. The red dots show the average accuracy while the grey bars denote the average fraction of votes compared to that required for $N = 35$ (statistical error bars on these quantities are too small to be seen.) GZ2 needs at least 35 votes per subject in order to retain consistent subject classifications for 95% of the GZ2 sample. SWAP can surpasses this accuracy using only 30% as many votes.

results are shown in Figure 3.6 for each value of N along with the accuracy and total classifications for our SWAP simulation. We see that GZ2 needs at least 35 votes per subject in order to achieve consistent class labels 95% of the time. SWAP achieves the same accuracy with 30% fewer classifications. Furthermore, this justifies our choice of defining a subject as GZ2-retired once it reached at least 30 classifications.

SWAP’s performance can be explained through its retirement mechanism. GZ2 retired subjects randomly and required an arbitrary number of classifications per subject. In contrast, SWAP retires “easier” subjects first while harder subjects remain in the system for longer (requiring many more classifications to nudge that subject’s posterior across a retirement threshold). Evidence for this can be seen in the top two panels of Figure 3.7. The top left panel shows the distribution of f_{smooth} for the entire GZ2 sample (orange), the SWAP-retired sample (blue), and the sample of subjects which SWAP has not yet retired, of which there are $\sim 19K$ at the end of our simulation. The SWAP-retired sample generally follows the same distribution as GZ2-full except for the noticeable dip around $f_{\text{smooth}} = 0.6$. In contrast, the SWAP-not-yet-retired sample peaks at $f_{\text{smooth}} = 0.6$. These subjects can be interpreted as being the most difficult to classify which can be understood intuitively: galaxies with $f_{\text{smooth}} \leq 0.5$ are easily identified as having features, while galaxies with $f_{\text{smooth}} \geq 0.8$ are obviously elliptical.

This is further corroborated in the top right panel which shows the distribution of the number of classifications a subject had at the time of retirement. The solid lines show this distribution from the original GZ2 project for the same subsamples as the top left panel. For comparison, the dashed line shows the number of classifications at retirement realized during our SWAP simulation. Again, we see that the SWAP-retired sample is representative of GZ2 as a whole. However, the distribution for the SWAP-not-yet-retired sample is skewed toward fewer total classifications. To understand this, consider the following: GZ2 served subject images at random with the exception that, towards the end of the project, subjects with low numbers of classifications were shown at a higher rate to ensure that each galaxy had enough responses to accurately characterize the likelihood of the classification (Willett et al., 2013). The median number of classifications was 44 with the full distribution shown in orange in the top right panel of Figure 3.7. Our SWAP simulation processes these classifications in the same order as the original project (with the exception that gold-standard subject classifications

are processed first as described in Section 3.2.1). Because our simulations cycle through only 92 days of GZ2 data, there are three general scenarios for why a subject has not yet been retired through SWAP: 1) SWAP has seen only a few of the many classifications for a given subject and it is not yet enough to retire it, 2) SWAP has seen many of the classifications for a subject but that subject is difficult; if we ran the simulation longer to process the remaining GZ2 classifications, SWAP would eventually retire it, and 3) SWAP has seen most or all of the classifications for a subject but it is difficult and there are few or no remaining GZ2 classifications; without additional volunteer input, these subjects will never be retired by SWAP. It is this third category that skews the red distribution towards fewer GZ2 votes. These are difficult-to-classify subjects that have only 30 - 40 GZ2 classifications, all of which are processed by SWAP, but these subjects remain unretired.

We have demonstrated that SWAP retires subjects intelligently: quickly retiring easy-to-classify subjects while allowing those that are more difficult to collect additional classifications. SWAP thus needs 30% as many votes to retire nearly 5 times as many subjects during the three months of GZ2 project time that we include in our simulation.

3.2.5 Reducing human effort

SWAP’s intelligent retirement mechanism is due, in large part, to how SWAP estimates volunteer classification ability which, in turn, allows for a dramatic reduction in the amount of human effort (votes) required. To see this, we consider a toy model wherein we simulate volunteers with fixed confusion matrices. We simulate 1000 ‘Featured’ subjects and 1000 ‘Not’ subjects each with prior, $p_0 = 0.5$. We simulate 100 volunteer agents all with the same fixed confusion matrix of (0.63, 0.65), where these values are computed as the average $P(\text{“}F\text{”}|F)$ and $P(\text{“}N\text{”}|N)$ from our assessment of real volunteers, excluding the spikes at 0.5. We generate volunteer classifications based on this confusion matrix (i.e., volunteers will correctly identify ‘Featured’ subjects 63% of the time) and update the subject’s posterior probability with each classification. We track how many classifications are required for each subject’s posterior to cross either the ‘Featured’ or ‘Not’ retirement thresholds.

The results are presented in Figure 3.8. The filled blue and orange histograms show the number of classifications per subject achieved from our SWAP simulation, where

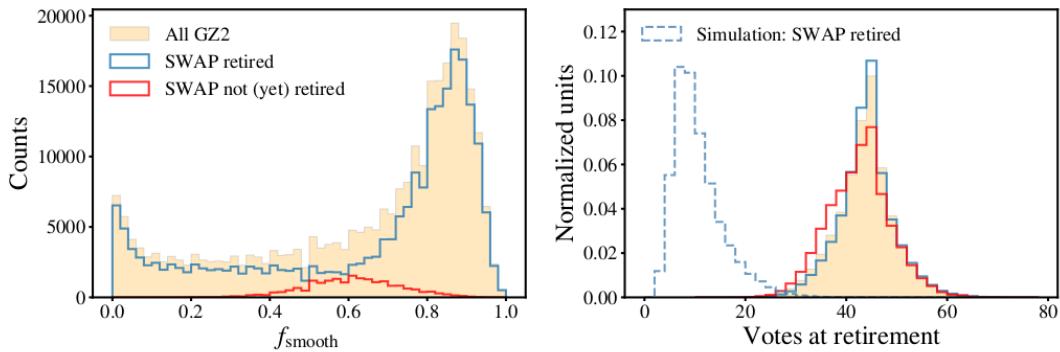


Figure 3.7 SWAP uses 30% fewer classifications to retire subjects due to its ability to retire easier subjects quickly, while more difficult subjects remain in the system to acquire additional classifications. The top left panel shows f_{smooth} for the entire GZ2 sample (orange), the subjects retired by SWAP (blue), and subjects that SWAP has not yet retired by the end of our simulation (red). The latter distribution peaks at $f_{\text{smooth}} \sim 0.6$, which can intuitively be understood as the most difficult to classify subjects: those with $f_{\text{smooth}} \leq 0.5$ are easily identified as ‘Featured’, while those with $f_{\text{smooth}} \geq 0.8$ are more obviously ‘Not’. The top right panel provides additional evidence where here we show the number of votes at retirement for both the original GZ2 project (solid lines) and our SWAP simulation (dashed blue). The left-skew inherent in the red SWAP-not-yet-retired sample is due to difficult-to-classify subjects that received only 30-40 classifications during the GZ2 project. Even after processing all available classifications, SWAP cannot retire these subjects without additional volunteer input.

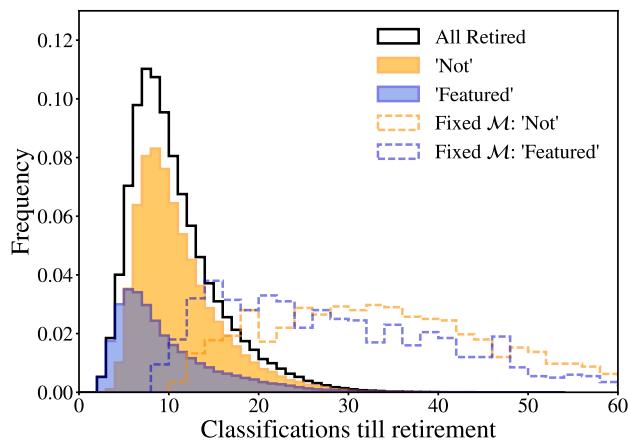


Figure 3.8 SWAP’s volunteer-weighting mechanism provides a factor of three reduction in the human effort required to retire GZ2 subjects. The filled histograms show the number of volunteer classifications per subject achieved during our SWAP simulation broken down by class label, where the solid black line is the total. The dashed histograms are results from our toy model in which we simulate volunteers with fixed confusion matrices, effectively disengaging SWAP’s volunteer-weighting mechanism. These broad distributions require ~ 3 times more classifications per subject to reach the same retirement thresholds.

volunteer agent confusion matrices are those from Figure 3.2. The dashed blue and orange distributions are the results from our toy model. When SWAP accounts for volunteer ability, most subjects are retired with between 6 and 15 votes, with a median of 9 votes. In contrast, when every volunteer is given equal weighting, subjects require 16 to 45 votes with a median of 30 votes before crossing one of the retirement thresholds. Thus the volunteer weighting scheme embedded in SWAP can reduce the amount of human effort required to retire subjects by a factor of three.

This reduction will be, in part, a function of the number of gold standard subjects each volunteer sees. Our gold standard sample was chosen to be representative of morphology rather than evenly distributed among GZ2 volunteers. We thus find that half of our volunteers classify only one or two gold standard subjects. That we achieve a factor of three reduction when only half of our volunteer pool has seen ≥ 2 gold standard subjects suggests that an additional reduction of human effort is possible with more extensive volunteer training.

3.2.6 Disagreements between SWAP and GZ2

Galaxy Zoo’s strength comes from the consensus of dozens of volunteers voting on each subject. Processing votes with SWAP reduces the number of classifications to reach consensus. Though we typically recover the GZ2_{raw} label, SWAP disagrees about 5% of the time. We thus examine the false positives (subjects SWAP labels as ‘Featured’ but GZ2_{raw} labels as ‘Not’) and false negatives (subjects SWAP labels as ‘Not’ but GZ2_{raw} labels as ‘Featured’). We explore these subjects in redshift, magnitude, physical size, and concentration but find no correlation with any of these variables, suggesting that, at least for this galaxy sample, the reliability of morphology depends on factors that are not captured by these coarse measurements. This is perhaps unsurprising since GZ2 subjects were selected from the larger GZ1 sample to be the brightest, largest and nearest galaxies: precisely those subjects most accessible for visual classification.

Instead we consider the stochastic nature of GZ2 vote fractions, which can be estimated as binomial. Let success be a response of “smooth” and failure be any other response. The 68% confidence interval on a subject with $f_{\text{smooth}} = 0.5$ is then $(0.42, 0.57)$

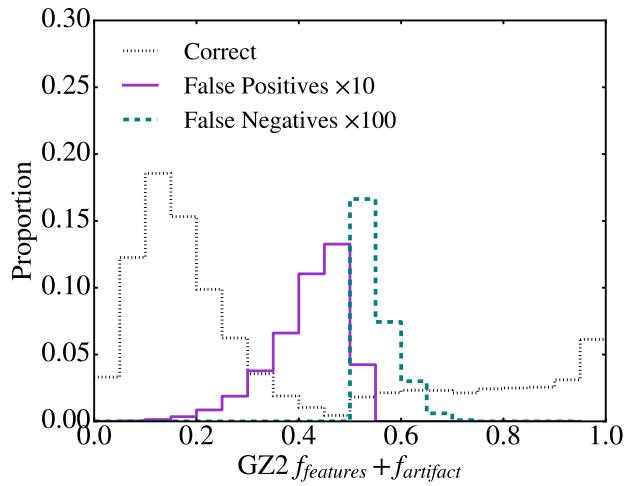


Figure 3.9 Distribution of GZ2 $f_{\text{features}} + f_{\text{artifact}}$ vote fractions for subjects correctly identified by SWAP (dotted grey), along with those identified as false positives (solid purple), and false negatives (dashed teal). The false positives and false negatives are scaled by factors of 10 and 100 respectively for easier comparison. From Section A.3, subjects with values > 0.5 are defined as ‘Featured’, however, the teal distribution indicates that SWAP labels them as ‘Not’. This is not a flaw of SWAP: 68.9% of incorrectly identified subjects have $0.4 \leq f_{\text{features}} + f_{\text{artifact}} \leq 0.6$ suggesting that GZ2_{raw} labels are simply too uncertain. The overlap between the false positives and negatives is due to subjects that are exactly 50-50; by default these are labelled ‘Not’.

assuming 40 classifications, each with a probability of 0.5. Figure 3.9 shows the distribution of $f_{\text{featured}} + f_{\text{artifact}}$ for the false positives (solid purple), and the false negatives (dashed teal) compared to the subjects where SWAP and GZ2 agree (dotted grey). Recall that if this value is greater than 0.5, the subject is labeled ‘Featured’. The majority of disagreements between SWAP and GZ2 are for subjects that have $0.4 < f_{\text{featured}} + f_{\text{artifact}} < 0.6$. It is thus unsurprising that SWAP and GZ2 disagree most within the approximate confidence interval of our selected GZ2 threshold. We note that the distribution overlap between false positives and false negatives is due to subjects that do not have a majority; these are labeled ‘Not’ by default.

Two other effects contribute to the disagreement between SWAP and GZ2. First, as the number of classifications used to retire a galaxy decreases, the likelihood of misclassification by random chance increases. Second, disagreement arises due to expert-level volunteers whose confusion matrices are close to 1.0. These volunteers are essentially more strongly weighted, allowing that subject’s posterior to cross a retirement threshold in as few as two classifications. In rare cases, despite training, some expert-level volunteers get it wrong compared to the gold-standard labels. These issues can be mitigated by requiring each subject reach a minimum number of classifications in addition to its posterior probability crossing a retirement threshold, thus combining the best qualities of GZ2 and SWAP.

3.2.7 Summary

We demonstrate nearly a factor of five increase in the classification rate, a reduction of at least a factor of three in the human effort necessary to maintain that increased rate, all while maintaining 95% accuracy, nearly perfect completeness of ‘Featured’ subjects, and with a purity that can be controlled by careful selection of input parameters to be better than 90% (see Appendix 3.3). Exploring those subjects wherein SWAP and GZ2 disagree, we conclude that the majority of this disagreement stems from the stochastic nature of GZ2_{raw} labels. We now turn our focus towards incorporating a machine classifier utilizing these SWAP-retired subjects as a training sample.

3.3 Exploring SWAP’s Parameter Space

The entirety of our analysis thus far has assumed the most basic SWAP parameters. In this section we explore how SWAP’s classification output changes as a function of varying the initial agent confusion matrices, prior probability, and retirement thresholds.

3.3.1 Initial agent confusion matrix.

In our fiducial simulation each volunteer was assigned an agent whose confusion matrix was initialized at $(0.5, 0.5)$, which presumes that volunteers are no better than random classifiers. We perform two simulations wherein we initialize agent confusion matrices as $(0.4, 0.4)$, slightly obtuse volunteers; and $(0.6, 0.6)$, slightly astute volunteers, with everything else remaining constant. Results of these simulations compared to the fiducial run are shown in the left panel of Figure 3.10. We find that SWAP is largely insensitive to the initial confusion matrix both in terms of the subject retirement rate and classification quality.

We retire $\sim 225K \pm 3.5\%$ subjects as shown by the light blue shaded region in the bottom left panel of Figure 3.10, where the dashed blue line denotes the fiducial run. Predictably, when the confusion matrix probabilities are low, we retire fewer subjects than when these probabilities are high for a given period of time. This is easy to understand since it takes longer for volunteers to become astute classifiers when they are initially given values denoting them as obtuse. Regardless, most volunteers become astute classifiers by the end of the simulation. The top left panel demonstrates our usual quality metrics as computed in Section 3.2.3. The dashed lines again denote the fiducial run. We maintain $\sim 95\%$ accuracy, 99% completeness, and $\sim 84\%$ purity; and no metric changes by $> 2\%$ regardless of initial confusion matrix values.

This spread is due to three effects: 1) subjects can receive an alternate SWAP label in different simulations, 2) subjects can be retired in a different order, and 3) the set of retired subjects is not guaranteed to be common to all runs. We find SWAP to be highly consistent: more than 99% of retired subjects are the same among all simulations, and, of these, 99% receive the same label. Instead we find that the order in which subjects are retired changes between runs. When the confusion matrix is low, subjects take longer to classify compared to the fiducial run (i.e., they retire on a later date in GZ2

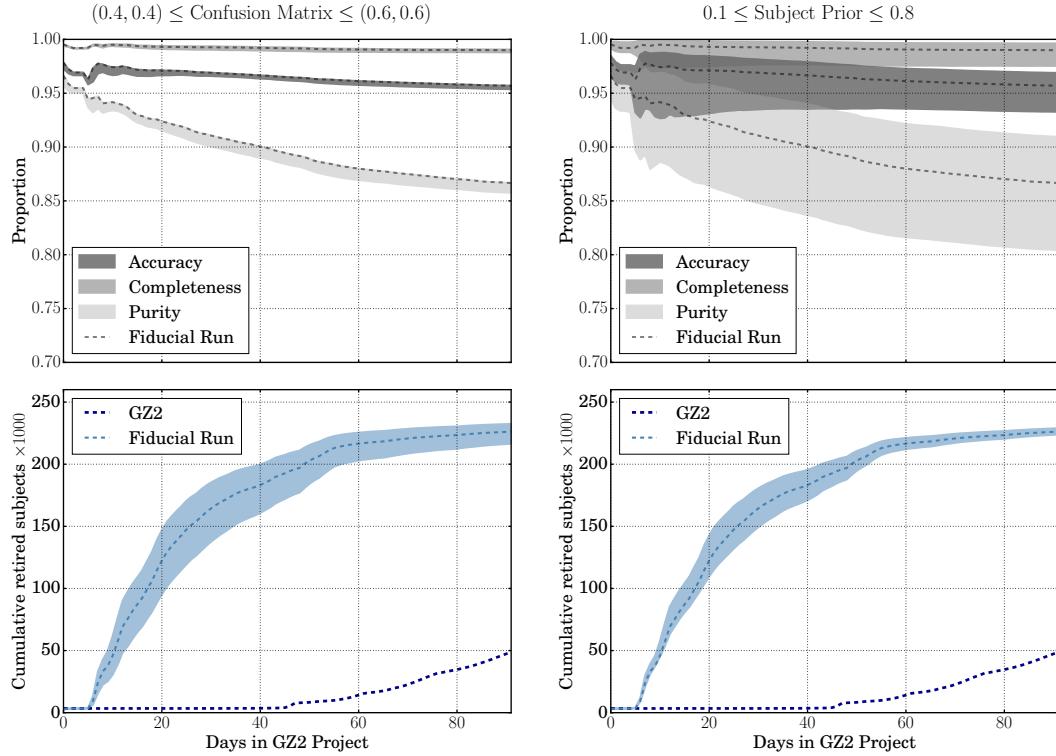


Figure 3.10 SWAP performance does not dramatically change even with a range of input parameters (shaded regions) as compared to the fiducial run of Section 3.2.3 (dashed lines). *Left.* The quality (top) and retirement rate (bottom) when the confusion matrix is initialized as $(0.4, 0.4)$ and $(0.6, 0.6)$, with all other input parameters remaining constant. *Right.* Same as the left panel but allowing the subject prior probability, $p_0 = 0.2, 0.35$ and 0.8 . Changing the confusion matrix has little impact on the quality of the labels but varies the total number of subjects retired. In contrast, changing the subject prior is more likely to affect the classification quality rather than the total number of subjects retired.

project time). Likewise, subjects retire sooner when the confusion matrix is high. This can cause quality metrics to vary since they are calculated on a day to day basis. These effects each contribute less than one per cent variation and thus we see a high level of consistency between simulations.

Of interest, perhaps, is that the quality metrics for these simulations are not symmetric about the fiducial run. However, in the Bayesian framework of SWAP, an agent with confusion matrix (0.4, 0.4) contributes as much information as an agent with confusion matrix (0.6, 0.6). The quality metrics computed are thus within a per cent of each other. In either case, we find that initializing agents at (0.5, 0.5) provides optimal performance for the ‘training’ we simulate with our current approach. Further assessment would require a live project with real-time training and feedback.

3.3.2 Subject prior probability, p_0 .

The prior probability assigned to each subject is an educated guess of the frequency of that characteristic in the scope of the data at hand. For galaxy morphologies, this number should be an estimate of the probability of observing a desired feature (bar, disk, ring, etc.). In our case, we desire simply to find galaxies that are ‘Featured’; however, this is dependent on mass, redshift, physical size, etc. The original GZ2 sample was selected primarily on magnitude and redshift. As there was no cut on galaxy size (with the exception that each galaxy be larger than the SDSS PSF), the sample includes a large range of masses and sizes. Designating a single prior is not clear-cut; we thus explore how various p_0 values effect the SWAP outcome.

We run simulations allowing p_0 to take values 0.2, 0.35, and 0.8 and compare these to the fiducial run, with everything else remaining constant. The results are shown in the right panels of Figure 3.10. We again find that SWAP is consistent in terms of subject retirement which varies by only 1%. However, as can be seen in the top panel, the variation in our quality metrics is more pronounced. Firstly, though we retire nearly the same number of subjects over the course of each simulation, they are less consistent than our previous runs. That is, only 95% of retired subjects are common to all simulations. Secondly, of those that are common, only 94% receive the same label from SWAP indicating that hanging the prior is more likely to produce a different label for a given subject than changing the initial agent confusion matrix. Finally, there is also

a larger spread for the day on which a subject is retired as compared to the fiducial run. These trends all contribute to a broader spread in accuracy, completeness, and purity as a function of project time. We stress, however, that although more substantial than the previous comparison, these variations are all within $\pm 5\%$.

We can understand these variations more intuitively by considering the following. Recall that our retirement thresholds, t_F and t_N , have not changed in these simulations. When p_0 is small, the subject’s probability is already closer to t_N in probability space, and thus more subjects are classified as ‘Not’ compared to the fiducial run. Similarly, when p_0 is large, some of these same subjects can instead be classified as ‘Featured’ because p_0 is already closer to t_F . Obviously, both outcomes cannot be correct. We find that the simulation with $p_0 = 0.8$ performs the worst of any run; this is a direct reflection of the fact that this prior is not suitable for this question or this dataset. Indeed, the best performance is achieved when $p_0 = 0.35$. This reflects the distribution of ‘Featured’ subjects as determined by GZ2_{raw} labels and is more characteristic of the expected proportion of ‘Featured’ galaxies in the local universe. As a value far from the correct value can have a significant impact on the classification quality, it is important to choose a prior wisely.

3.3.3 Retirement thresholds, t_F and t_N .

Retirement thresholds are directly related to the time that a subject will spend in SWAP before retirement. If we lower t_F (and/or raise t_N), more subjects will be retired compared to the fiducial run as each subject will have a smaller swath of probability space in which to fluctuate before crossing one of these thresholds. On the other hand, if we raise t_F (and/or lower t_N), it will take longer for subjects to cross one of these thresholds. This also increases the likelihood of some subjects never crossing either threshold, instead oscillating indefinitely through probability space.

What thresholds should one choose? To answer this question, we consider the left panel of Figure 3.11, which depicts the receiver operating characteristic (ROC) curve for our fiducial simulation, an illustration of performance as a function of a threshold for a binary classifier. ROC curves display the true positive rate against the false positive rate for a discriminatory threshold or score with a perfect classifier achieving 100% true positives and no false positives. The value of the threshold optimal for predicting class

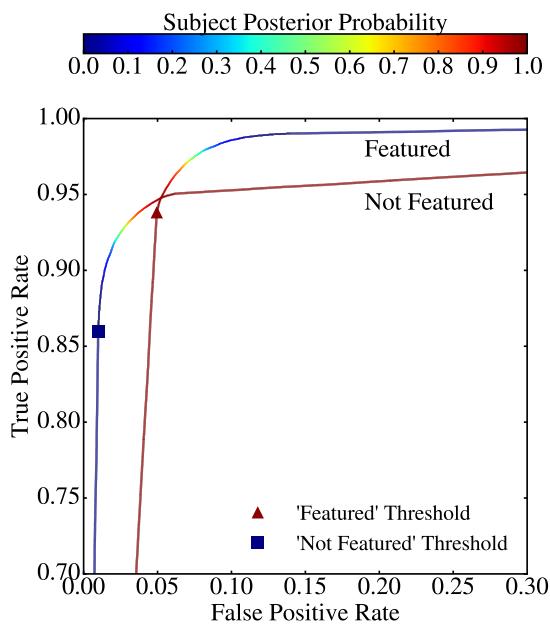


Figure 3.11 Identifying ‘Featured’ subjects is independent of identifying ‘Not’ subjects. Both ROC curves use all subjects processed by SWAP where the score used to create the ROC curve is simply each subject’s achieved posterior probability. The Featured curve demonstrates how well we identify ‘Featured’ subjects with a threshold of 0.99, while the Not Featured curve demonstrates how well we identify ‘Not’ subjects with a threshold of 0.004. Typically, best performance is achieved by the score associated with the upper-left-most part of the curve. Our ‘Featured’ threshold is nearly optimal, while our ‘Not’ threshold could be improved since the blue square is not as close to the upper left hand corner as other possible values of the subject posterior.

labels would be that which allows the ROC curve to reach the upper-left-most point in the diagram. We have two thresholds to consider and thus we plot the curve twice: once under the assumption that “true positives” denote correctly identified ‘Featured’ subjects; and again under the assumption that “true positives” instead denote correctly identified ‘Not’ subjects. In both cases, the color of the line corresponds to the subject posterior probability. We mark the location of $t_F = 0.99$ and $t_N = 0.004$ from our fiducial run with a red triangle and blue square respectively. We see that t_F is nearly optimal but t_N could be improved upon.

Chapter 4

Incorporating machine intelligence

4.1 Efficiency through incorporation of machine classifiers

We construct the full Galaxy Zoo Express by incorporating supervised learning, the machine learning task of inference from labelled training data. The training data consist of a set of training examples, and must include an input feature vector and a desired output label. Generally speaking, a supervised learning algorithm analyses the training data and produces a function that can be mapped to new examples. A properly optimized algorithm will correctly determine class labels for unseen data. By processing human classifications through SWAP, we obtain a set of binary labels by which we can train a machine classifier. We briefly outline the technical details of our machine below, turning towards the decision engine we develop in Section 4.1.4.

4.1.1 Random Forests

We use a Random Forest (RF) algorithm (Breiman, 2001), an ensemble classifier that operates by bootstrapping the training data and constructing a multitude of individual decision tree algorithms, one for each subsample. An individual decision tree works by deciding which of the input features best separates the classes. It does this by performing splits on the values of the input feature that minimize the classification error.

These feature splits proceed recursively. Decision trees alone are prone to over-fitting, precluding them from generalizing well to new data. Random Forests mitigate this effect by combining the output labels from a multitude of decision trees. Specifically, we use the `RandomForestClassifier` from the Python module `scikit-learn` (Pedregosa et al., 2011).

4.1.2 Grid Search and Cross-validation

Of fundamental importance is the task of choosing an algorithm’s hyperparameters, values which determine how the machine learns. For a RF, key quantities include the maximum depth of individual trees (`max_depth`), the number of trees in the forest (`n_estimators`), and the number of features to consider when looking for the best split (`max_features`). The goal is to determine which values will optimize the machine’s performance and thus these values cannot be chosen *a priori*. We perform a grid search with k -fold cross-validation whereby the training sample is split into k subsamples. One subsample is withheld to estimate the machine’s performance while the remaining data are used to train the machine. This is performed k times and the average performance value is recorded. The entire process is repeated for every combination of the hyperparameters in the grid space and values that optimize the output are chosen. In this work we let $k = 10$, however, we leave this as an adjustable input parameter. In the interest of computational speed, we set `n_estimators` = 30 and perform the grid search for `max_depth` over the range [5, 16], and `max_features` over the range [\sqrt{D} , D], where D is the number of features in the feature vector, described below.

4.1.3 Feature Representation and Pre-Processing

The feature vector on which the machine learns is composed of D individual numeric quantities associated with the subject that the machine uses to discern that subject from others in the training sample. To segregate ‘Featured’ from ‘Not’, we draw on ZEST (Scarlata et al., 2007) and compute concentration, asymmetry, Gini coefficient, and M_{20} , the second-order moment of light for the brightest 20% of galaxy pixels as measured from SDSS DR12 i -band imaging (see Appendix ??). Coupled with SExtractor’s measurement of ellipticity (Bertin & Arnouts, 1996), we provide the machine with

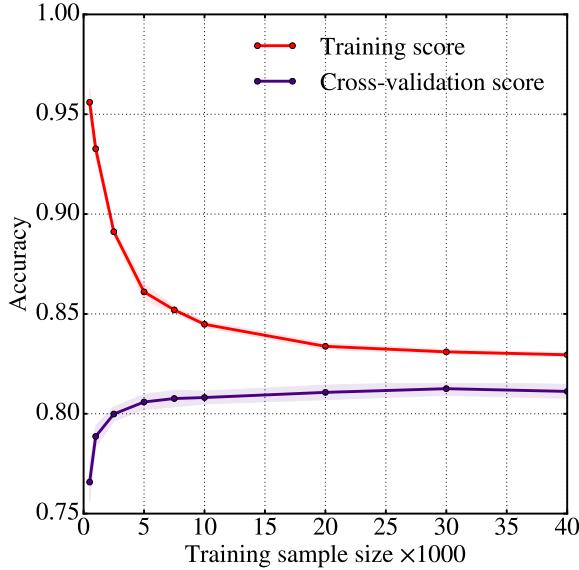


Figure 4.1 Learning curve for a Random Forest with fixed hyperparameters. These curves show the mean accuracy computed during cross-validation and on the training sample, where the shaded regions denote the standard deviation. When the training sample size is small, the machine accurately identifies its own training sample but is unable to generalize to unseen data as evidenced by a low cross-validation score. As the training sample size increases, the cross-validation score increases. This behaviour plateaus indicating that larger training samples provide little in additional performance.

a $D = 5$ dimensional morphology parameter space. These non-parametric diagnostics have long been used to distinguish between early- and late-type galaxies in an automated fashion (e.g., Abraham et al., 1996; Bershady et al., 2000; Conselice et al., 2000; Abraham et al., 2003; Conselice, 2003; Lotz et al., 2004; Snyder et al., 2015). Because the RF algorithm handles a variety of input formats, the only pre-processing step we perform is the removal of poorly-measured morphological indicators, i.e. catastrophic failures.

4.1.4 Decision Engine

A number of decisions must be addressed before attempting to train the machine. In particular, which subjects should be designated as the training sample? When should

the machine attempt its first training session? When has the machine’s performance been optimized such that it will successfully generalize to unseen subjects? The field of machine learning provides few hard rules for answering these questions, only guidelines and best practices. Here we briefly discuss our approach for the development of our decision engine.

As discussed in detail in Section 3.2, SWAP yields a probability that a subject exhibits the feature of interest. While some machine algorithms can accept continuous input labels, the RF requires distinct classes. We thus use only those subjects which have crossed either of the retirement thresholds. Though we find that SWAP consistently retires 35-40% ‘Featured’ subjects on any given day of the simulation, a balanced ratio of ‘Featured’ to ‘Not’ isn’t guaranteed. Highly unbalanced training samples should be resampled to correct the imbalance; however, as we exhibit only a mild lopsidedness, we allow the machine to train on all SWAP-retired subjects.

SWAP retires a few hundred subjects during the first days of the simulation. In principle, a machine can be trained with such a small sample, but will be unable to generalize to unseen data. We estimate a minimum number of training samples and the machine’s ability to generalize by considering a learning curve, an illustration of a machine’s performance with increasing sample size for fixed hyperparameters. Figure 4.1 demonstrates such a curve wherein we plot the accuracy from both the 10-fold cross-validation, and the trained machine applied to its own training sample for a random sample of GZ2 subjects required to be balanced between ‘Featured’ and ‘Not’. We fix the RF’s hyperparameters as follows: `max_depth = 8`, `n_estimators = 30`, and `max_features = 2`. When the sample size is small, the cross-validation score is low and the training score is high, a clear sign of over-fitting. However, as the training sample size increases, the cross-validation score increases and eventually plateaus, indicating that larger training sets will yield little additional gain.

We estimate this plateau begins when the training sample reaches 10,000 subjects and require SWAP retire at least this many before the machine attempts its first training. We estimate the machine has trained sufficiently if the cross-validation score fluctuates by less than 1% for three consecutive nights of training to ensure we have reached the plateau. This requires that we record the machine’s training performance each night, including how well it scores on the training sample, the cross-validation score, and the

best hyperparameters.

4.1.5 The Machine Shop

We can now describe a full GZX simulation, which begins with human classifications processed through SWAP for several days. Once at least 10K subjects have been retired, their feature vectors are passed to the machine for its inaugural training. A suite of performance metrics are recorded by a machine agent, similar in construction to SWAP’s agents. This agent determines when the machine has trained sufficiently by assessing the variation in performance metrics for all previous nights of training. Once the machine has been optimized, the agent introduces it to the test sample consisting of any subject that has not yet reached retirement through SWAP and is not part of the gold standard sample.

Analogous to SWAP, we generate a retirement rule for machine-classified subjects. In addition to the class prediction, the RF algorithm computes the probability for each subject to belong to each class. This probability is simply the average of the probabilities of each individual decision tree, where the probability of a single tree is determined as the fraction of subjects of class X on a leaf node. Only subjects that receive a class prediction of ‘Featured’ with $p_{\text{machine}} \geq 0.9$ ($p_{\text{machine}} \leq 0.1$ for ‘Not’) are considered retired. The remaining subjects have the possibility of being classified by humans or the machine on a future night of the simulation. This constitutes the core of our passive feedback mechanism. Subjects that are not retired by the machine can instead be retired by humans, thus providing the machine a more fully sampled morphology parameter space on future training sessions.

4.2 Results

We perform a full GZX simulation incorporating our RF with the fiducial SWAP run discussed in Section 3.2.3. The machine attempts its first training on Day 8 with an initial training sample of $\sim 20K$ subjects. It undergoes several additional nights of training, each time with a larger training sample. By Day 12, SWAP has provided over 40K subjects for training and the machine’s agent has deemed the machine optimized. The machine predicts class labels for the remaining 230K GZ2 subjects. Of those, the

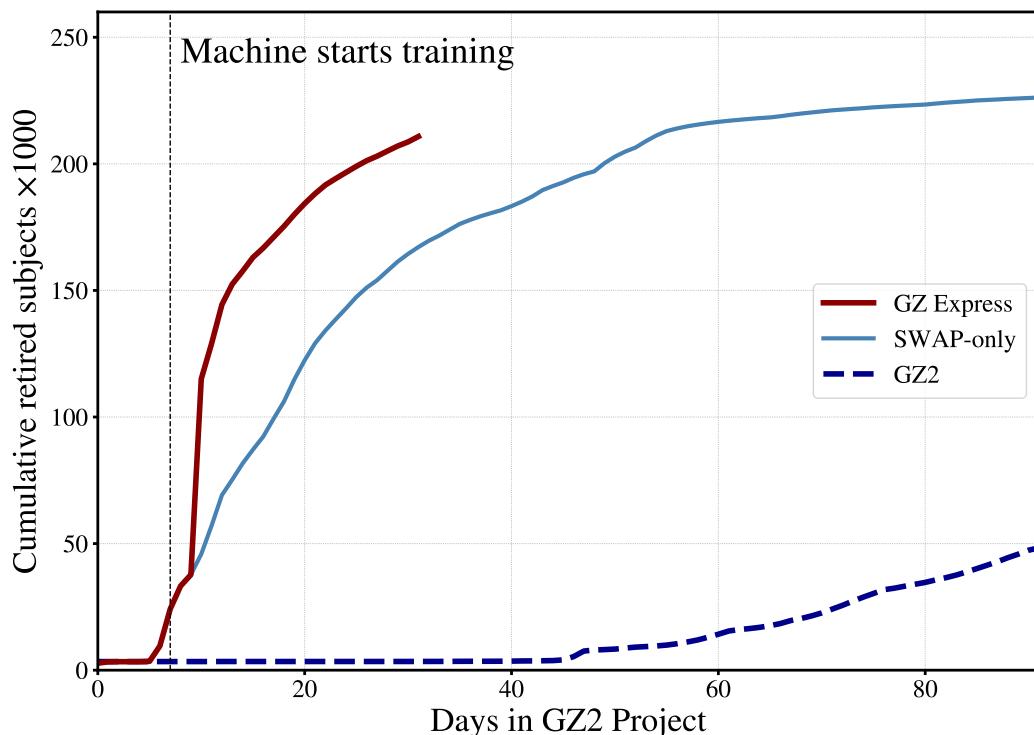


Figure 4.2 By incorporating a machine classifier, GZX (red) increases the classification rate by an order of magnitude compared to GZ2 (dashed dark blue) and out-performs the SWAP-only run (light blue), retiring more than 200K subjects in just 27 days of GZ2 project time. The dashed black line marks the first night the machine trains. After several additional nights of training, it is deemed optimized and allowed to retire subjects. Both humans and machine then contribute to retirement. We end the simulation after 32 days having retired over 210K galaxies. See Table 4.1 for details.

Table 4.1 Summary of key quantities for GZ2 and our various simulations. All quality metrics are calculated using GZ2_{raw} labels.

Simulation Summary						
	Days	Subjects Retired	Human Effort (classifications)	Accuracy (%)	Purity (%)	Completeness (%)
Galaxy Zoo 2	430	285962	16,340,298	—	—	—
SWAP only	92	226124	2,298,772	95.7	86.7	99.0
SWAP+RF	32	210803	936,887	93.1	83.2	94.0

machine retires over 70K, dramatically increasing the subset of retired subjects. We end the simulation after 32 days, having retired \sim 210K subjects as detailed in Table 4.1.

We present these results in Figure 4.2 where subject retirement with GZX (red) is compared to our fiducial SWAP-only run (light blue) and GZ2 (dashed dark blue). Using the GZ2_{raw} labels as before, we compute our usual quality metrics on the full sample of GZX-retired subjects; reported in Table 4.1. Accuracy and purity remain within a few percent of the SWAP-only run at 93.1% and 83.2% respectively. Instead we see a 5% decline in the completeness. While the SWAP-only run identified 99% of ‘Featured’ subjects, incorporation of the machine seems to miss a significant portion thus dropping GZX completeness to 94.0%. We discuss this behaviour below.

By dynamically generating a training sample through a more sophisticated analysis of human classifications coupled with a machine classifier, we retire more than 200K GZ2 subjects in just 27 days. Visual classification through SWAP alone retires as many in 50 days, while GZ2 requires a full year. Though our analysis considers only the top-level task of GZ2’s decision tree, GZX suggests a tantalizing potential to increase the classification rate by an order of magnitude over the traditional crowd-sourced approach. We next explore the composition of those classifications.

4.2.1 Who retires what, when?

In the top panel of Figure 4.3 we explore the individual contributions to GZX subject retirement from the RF (dash-dotted teal) and SWAP (dashed orange). The solid black line shows the total GZX retirement (SWAP+RF), while the dotted grey line depicts the fiducial SWAP-only run from Section 3.2.3 for reference. Two things are immediately obvious. First, each component shoulders approximately half of the retirement

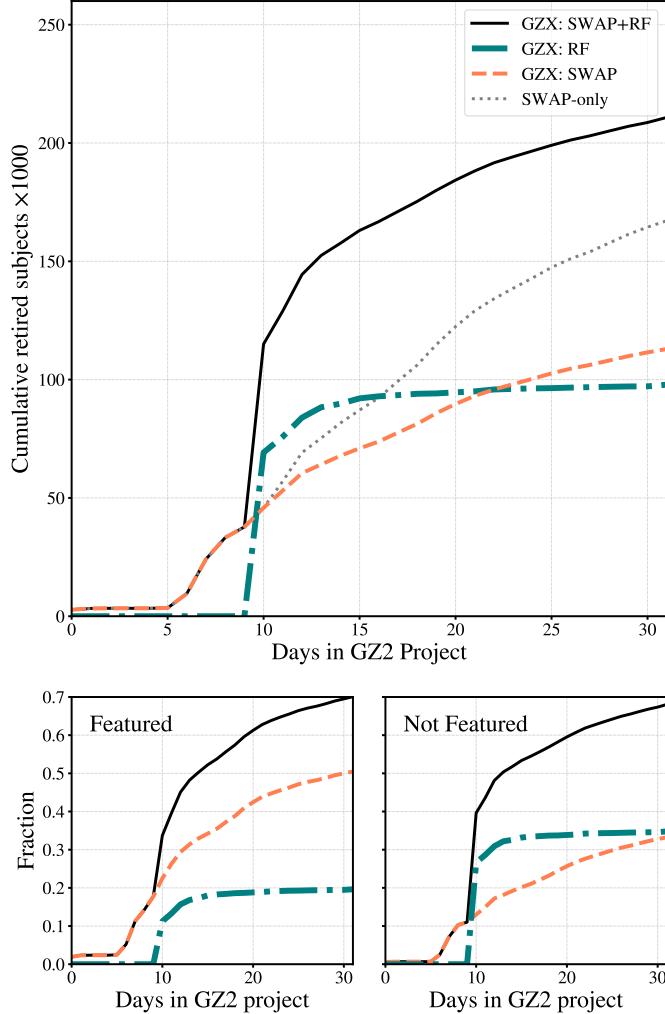


Figure 4.3 Contributions to subject retirement by both classifying agents of GZX: human (SWAP) and machine. The top panel shows cumulative subject retirement for GZX as a whole (solid black), along with that attributed to the RF (dash-dotted teal), and SWAP (dashed orange). The dotted grey line shows the fiducial SWAP-only run for comparison. Retirement totals for humans and machine are nearly equal over the course of the simulation but display different behaviours: SWAP’s retirement rate is almost constant while the RF contributes substantially after its initial application and then plateaus. The bottom panels show what fraction of GZ2 subjects are retired, separated by class label. Overall, GZX retires 73.7% of the entire GZ2 sample in 32 days, retiring the same proportion of ‘Featured’ and ‘Not’ subjects as indicated by the black lines. However, humans retire 30% more ‘Featured’ subjects than the machine, while both components retire a similar proportion of ‘Not’ subjects.

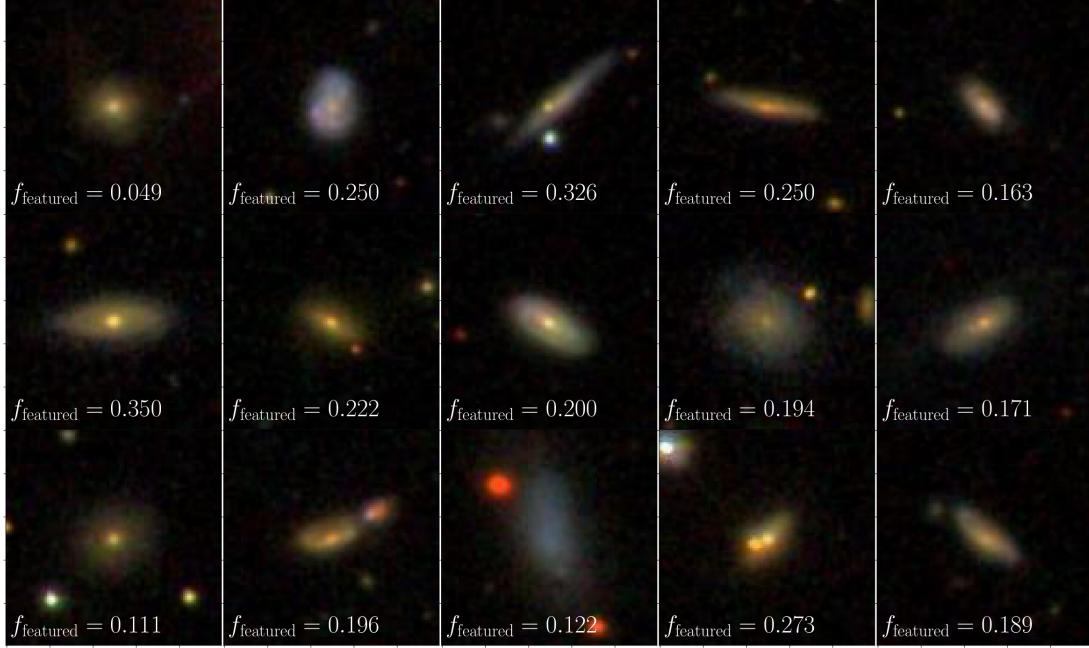


Figure 4.4 A random subsample of subjects identified as false positives: labelled by machine as ‘Featured’, but as ‘Not’ according to GZ2_{raw}. We display f_{smooth} in the lower left corner, that is, the fraction of volunteers who classified the subject as ‘smooth’ (‘Not’). Values are typically between 0.5 and 0.65 indicating that GZ2 volunteers did not reach a strong consensus. Fortunately, the machine is able to identify these subjects as ‘Featured’ due to their measured morphology diagnostics.

burden with the machine and SWAP responsible for $\sim 98K$ and $\sim 112K$ subjects respectively. Secondly, the rate of retirement exhibited by the two components is in stark contrast. SWAP retires at a relatively constant rate while the machine retires dramatically at the beginning of its application, quickly surpassing the human contribution, and plateaus thereafter. We thus clearly see three epochs of subject retirement. In the first phase, humans are the only contributors to subject retirement. Once the machine is optimized, it immediately contributes more to retirement than humans. However, the machine’s performance plateaus quickly; the third phase is again dominated by human classifications.

In the bottom panels of Figure 4.3, we consider the class composition of subjects retired by SWAP and the RF. The left (right) panel shows the retired fraction of GZ2

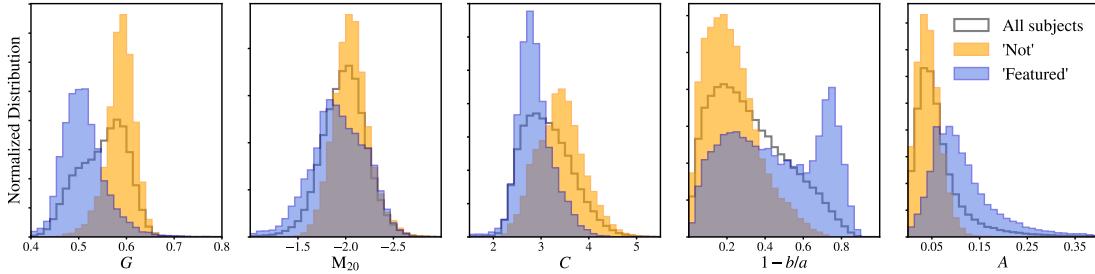


Figure 4.5 The RF is trained on a 5-dimensional morphology parameter space. We show the distribution of each morphology indicator for machine-retired ‘Featured’ (blue) and ‘Not’ (orange) subjects compared to the full GZ2 subject sample (black). The difference between ‘Featured’ and ‘Not’ subjects is in stark contrast for all distributions except, perhaps, M_{20} .

subjects identified as ‘Featured’ (‘Not’) according to their GZ2_{raw} labels as a function of GZ2 project time. Overall, GZX retires 73.7% of the GZ2 subject sample and this is almost evenly distributed between ‘Featured’ and ‘Not’ subjects as indicated by the solid black lines in both panels. However, SWAP retires 50% of all ‘Featured’ subjects while the machine retires only 20%. This divergence does not exist for ‘Not’ subjects where each component contributes 33-34%.

What is the source of this discrepancy? Each night the machine trains on a sample composed consistently of 30-40% ‘Featured’ subjects but does not retire a similar proportion, indicating that the 30% of non-retired ‘Featured’ subjects do not receive high p_{machine} . In the following section we explore whether this is an artefact of our choice in machine or in the human-machine combination implemented here.

4.2.2 Machine performance

Throughout our analysis we have defined ‘Featured’ and ‘Not’ subjects by their GZ2_{raw} labels as this was the most compatible choice for comparison with SWAP output. However, the machine does not learn in the same way, nor is it presented with the same information. Machine and human classifications each provide valuable and complementary information for identifying ‘Featured’ galaxies.

We isolate the 7,060 subjects that were deemed false positives, i.e., galaxies retired by the machine as ‘Featured’ that have ‘Not’ GZ2_{raw} labels, a sample that comprises

7.2% of all subjects the machine retires. We visually examine several hundred and assess that, to the expert eye, a majority are in fact ‘Featured’. A random sample is shown in Figure 4.4.

That the machine strongly identifies these galaxies as ‘Featured’ ($p_{\text{machine}} \geq 0.9$) where humans instead classify them as ‘Not’ ($f_{\text{featured}} < 0.5$) has several contributing factors: 1) as discussed in Section 3.2.6, the threshold we chose carries with it a confidence interval such that subjects with $0.4 < f_{\text{featured}} + f_{\text{artifact}} < 0.6$ are most likely to receive disagreeing labels from other classifying agents, 2) the first task of the GZ2 decision tree asks a question that does not necessarily correlate with a split between early- and late-type galaxies, and 3) the machine learns on morphology diagnostics that are very different from visual inspection.

We find that 40% of these false positives have $0.4 \leq f_{\text{featured}} + f_{\text{artifact}} < 0.5$ indicating that the disagreement between humans and machine is likely due to the labels we assign at our given threshold. However, we also find that 45% of false positives have $f_{\text{featured}} + f_{\text{artifact}} \leq 0.35$, and this discrepancy is not as easily explained. In Figure 4.4 we examine a random sample of false positives in this regime where, for clarity, we display only the f_{featured} value in the lower left corner. The majority of these subjects are disks lacking features such as spiral arms or strong bars. Whether this is the reason the majority of volunteers classify these objects as “smooth” is beyond the scope of this paper, however, this behaviour might be modified by providing actual training images and live feedback as performed in Marshall et al. (2016). We suggest that, at least for this particular question, if either human or machine identifies a subject as ‘Featured’, it is likely the subject is disk and worth further investigation.

Accordingly, this suggests that, in some cases, the morphology indicators we measure are sufficient for the machine to recognize ‘Featured’ galaxies regardless of the labels humans provide. Figure 4.5 shows the distribution of each morphology indicator for all subjects the machine retires as ‘Featured’ (blue) and ‘Not’ (orange) compared to the full GZ2 subject set. The difference between ‘Featured’ and ‘Not’ is stark in all but the M_{20} distribution. This can be seen explicitly in Figure 4.6 in which we show the RF’s ranked feature importances, where large values indicate higher importance. Feature importance is computed as how much each feature decreases the impurity of a split in a tree. The impurity decrease from each feature is then averaged over all trees

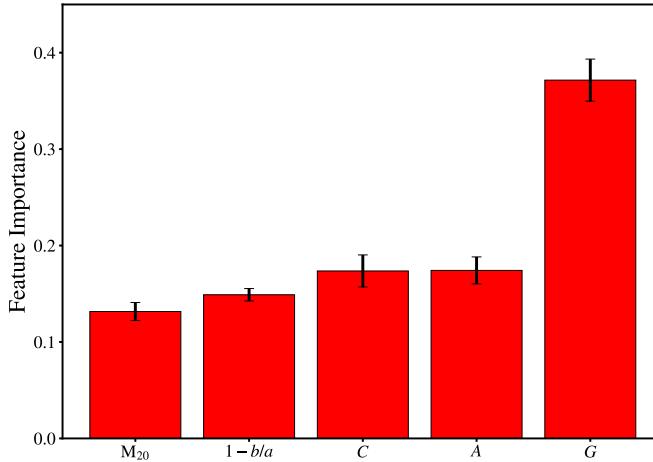


Figure 4.6 The RF’s ranked feature importance averaged over all nights of training with black bars indicating the standard deviation. A larger value corresponds to higher importance. The machine computes feature importance according to how much each feature increases the purity of the resulting split averaged over all trees in the forest. The RF places great importance in the Gini coefficient though we note that it can under-represent the importance of highly correlated features such as concentration.

and ranked. We show the feature importance averaged over all nights of training with black bars indicating the standard deviation. The machine finds the Gini coefficient most important for class prediction, placing little emphasis on M_{20} . It is well known that the Gini coefficient is more sensitive to noise than other diagnostics, however, we point out that when a machine is faced with two or more correlated features any of them can be used as the predictor. Once chosen, the importance of the others is reduced. This explains why Concentration is ranked much lower than Gini even though they are strongly correlated as seen in Figure 3.11. That the machine relies heavily on these two morphology diagnostics is unsurprising as concentration has long been an automated predictor between early- and late-type galaxies (Abraham et al., 1994, 1996; Shen et al., 2003).

The complementary nature of human and machine classification can best be utilized by a feedback mechanism in which a portion of machine-retired subjects are reviewed by humans. Subjects that display excessive disagreement should be verified by an expert (or expert-user). In the same way that humans increase the machine’s training sample

over time, subjects that the machine properly identifies can become part of the humans' training sample.

4.3 Looking Forward

We have demonstrated the first practical framework for combining human and machine intelligence in galaxy morphology classification tasks. While we focus below on a brief discussion of our next steps and potential applications to large upcoming surveys, we note that our results have implications for the future of citizen science and Galaxy Zoo in particular.

GZX is perhaps one of the simplest ways to combine human and machine intelligence and its impressive performance motivates a higher level of sophistication. A first step will be an implementation of SWAP that can handle a complex decision tree. In addition, we envision multiple forms of active feedback in addition to our passive feedback mechanism. SWAP allows us to leverage the most skilled volunteers to review galaxies difficult for either human or machine to classify. Additionally, machine-retired subjects should contribute to the training sample for humans in an analogous fashion to what we have already implemented.

Secondly, our RF can be improved by providing it information equal to what humans receive: multi-band morphology diagnostics will be included in our future feature vector. However, the Random Forest algorithm is not easily adapted to handle measurement errors or class labels with continuous distributions. To fully utilize the information provided by SWAP, sophisticated algorithms should be considered such as deep convolutional neural networks (CNN) or Latent Dirichlet allocation (LDA), an algorithm that is frequently used in document processing. Furthermore, there is no reason to limit to a single machine. As hinted at in Figure 1.4, several machines could train simultaneously, their predictions aggregated through SWAP, creating an on-the-fly machine ensemble.

With the above upgrades implemented, we expect performance of both the classification rate and quality to further increase. However, even our current implementation can cope with upcoming data volumes from large surveys. By some estimates, *Euclid* is expected to obtain measurable morphology with its visual instrument (VIS) for approximately $10^6 - 10^7$ galaxies (Laureijs et al., 2011). Visual classification at the rate

achieved with Galaxy Zoo today would require 12–120 years to classify.¹ If the *Euclid* sample is on the high end, GZX as currently implemented could classify the brightest 20% during the six years of its observing mission. As currently implemented, we obtain accuracy around 95% potentially leaving hundreds of thousands of galaxies with unreliable classifications. In a companion paper that seeks to identify supernovae, Wright et al. (submitted) demonstrate a dramatic increase in accuracy through an entirely different human-machine combination whereby the scores from human and machine are averaged together with the combined score yielding the most reliable classification. Again, a combination of both approaches will allow us to take full advantage of legacy output from large scale surveys.

4.3.1 Conclusions

In this paper we design and test Galaxy Zoo Express, an innovative system² for the efficient classification of galaxy morphology tasks that integrates the native ability of the human mind to identify the abstract and novel with machine learning algorithms that provide speed and brute force. We demonstrate for the first time that the SWAP algorithm, originally developed to identify rare gravitational lenses in the Space Warps project, is robust for use in galaxy morphology classification. We show that by implementing SWAP on GZ2 classification data we can increase the rate of classification by a factor of 4-5, requiring only 90 days of GZ2 project time to classify nearly 80% of the entire galaxy sample.

Furthermore, we have implemented and tested a Random Forest algorithm and developed a decision engine that delegates tasks between human and machine. We show that even this simple machine is capable of providing significant gains in the classification rate when combined with human classifiers: GZX retires over 70% of GZ2 galaxies in just 32 days of GZ2 project time. This represents a factor of 11.4 increase in the classification rate as well as an order of magnitude reduction in human effort compared to the original GZ2 project. This is achieved without sacrificing the quality of classifications as we maintain accuracy well above 90% throughout our simulations.

¹ We note that the classification rate of GZ2 was 4 times higher than GZ’s current steady rate.

² Our code can be found at <https://github.com/melaniebeck/GZExpress>

Additionally, we have shown that training on a 5-dimensional parameter space of traditional non-parametric morphology indicators allows the machine to identify subjects that humans miss, providing a complementary approach to visual classification. The gain in classification speed allows us to tackle the massive amount of data promised from large surveys like *LSST* and *Euclid*.

4.4 Acknowledgements

MB thanks Steven Bamford and Boris Häußler for insightful discussions on citizen science and Galaxy Zoo; and John Wallin and Marc Huertas-Company for several enlightening conversations on machine learning and classification. We are grateful to Elisabeth Baeten, Micaela Bagley, Karlen Shahinyan, Vihang Mehta, Steven Bamford, Kevin Schawinski, and Rebecca Smethurst for providing expert classifications in addition to those provided by the authors. PJM acknowledges Aprajita Verma and Anupreeta More for their ongoing collaboration on the Space Warps project.

MB, CS, LF, KW, and MG gratefully acknowledge support from the US National Science Foundation Grant AST-1413610. MB acknowledges additional support through New College and Oxford University’s Balzan Fellowship as well as the University of Minnesota Doctoral Dissertation Fellowship. Travel funding was supplied to MB, in part, by the University of Minnesota Thesis Research Travel Grant. CJL recognizes support from a grant from the Science & Technology Facilities Council (ST/N003179/1). BDS acknowledges support from Balliol College, Oxford, and the National Aeronautics and Space Administration (NASA) through Einstein Postdoctoral Fellowship Award Number PF5-160143 issued by the Chandra X-ray Observatory Center, which is operated by the Smithsonian Astrophysical Observatory for and on behalf of NASA under contract NAS8-03060. The work of PJM is supported by the U.S. Department of Energy under contract number DE-AC02-76SF00515.

Chapter 5

A population of “clumpy” galaxies in local Universe

Simultaneous to the development of this research was the publishing of the Galaxy Zoo: Hubble galaxy morphology classification catalog. A portion of the original GZ2 SDSS galaxy sample was included in Galaxy Zoo: Hubble because the decision tree for the latter was more complex than that for GZ2. This enabled a comparison between different decision trees and led to the identification of a rare sample of low redshift clumpy galaxies. This chapter details the identification of such a sample, preliminary analysis of the star-forming clumps, and methods to identify more such galaxies in the local universe as these may be analogs of high redshift galaxies with similar properties.

5.1 Introduction

Structural properties of typical galaxies today were forged over the 5 billion years between the peak of the cosmic star-formation history at $z = 1.5$ and now. From the theoretical perspective, galaxies growth is predominantly governed by the baryonic physics of gas accretion and energy feedback, rather than by mergers that induce starbursts (e.g., Somerville and Davé, 2015). Gas inflow modulates galaxies star formation history and plays a crucial role in the dynamical state of the galaxy. A varying gas fraction changes the amount of fragmentation within the gaseous disk, the star-formation rates (SFR), strengths and lifetimes of disk structural components, and possibly the fueling rate

of the AGN. As a consequence of the gas accretion process, the small-scale structure within disk galaxies is also predicted to change over time, with the appearance of stable grand design spiral arms only at later epochs (e.g. Oppenheimer et al., 2010; Bouché et al., 2010; Davé et al., 2011a,b; Lilly et al., 2013; Hirschmann et al., 2013).

In the theoretical paradigm discussed above, the origin of these massive clumps is still an open debate. Two main physical scenarios have been proposed. On one side (hereafter referred to as in-situ models), giant clumps are expected to form inside pre-existing galaxies. In these scenarios, haloes located at the center of multiple filaments continuously accrete gas from the cosmic web. This accretion results in the formation of gas-dominated disks that eventually fragment into giant clumps by gravitational instability (Bournaud et al., 2007; Dekel et al., 2009; Behrendt et al., 2015). On the other side (hereafter referred to as ex-situ models) clumps may have an external origin, as in the case of mergers with small star-forming companions (Ceverino et al., 2010; Bournaud, 2016).

Observationally, the situation is still unclear. The large gas-to-baryonic fraction of 20 to 80 recently measured with interferometric studies in clumpy galaxies lends support to the in-situ models (e.g., Erb et al., 2006; Tacconi et al., 2008, 2010). Similarly, the kinematics in the majority of these galaxies is characterized by large, regularly-rotating disks, with no signs of on-going or recent mergers (Genzel et al., 2006; Förster Schreiber et al., 2009; Epinat et al., 2012; Newman et al., 2012). These studies, however, are 1) limited to galaxies with $SFR \gtrsim 10 M_{\odot} \text{ yr}^{-1}$ and $M \gtrsim 10^{10} M_{\odot}$ and 2) include only a few galaxies with high spatial resolution kinematic and interferometric data (Genzel et al., 2014).

The lack of existing observational constraints below $10^{10} M_{\odot}$ is of particular concern. It is precisely in this mass range that most of the constraining power resides. In fact, state-of-the-art simulations that directly resolve the interstellar medium of individual galaxies while capturing their cosmological environment show that low mass galaxies are mostly affected by preventive feedback (e.g., FIRE Muratov et al. 2015; GASOLINE Christensen et al. 2015; and MU-FASA Davé et al. 2016). These simulations show that in galaxies with $M \lesssim 10^{10} M_{\odot}$ supernova-driven fast outflows can lower the gas inflow rate, and consequently the rate at which clumps are formed in-situ, compared to more massive galaxies. If clumps are the result of minor merging, on the other hand, the

dependency with stellar mass would depend on the specifics of the dynamics of the mergers.

Throughout this chapter we assume a flat Planck cosmology with $H_0 = 67.8$ and $\Omega_m = 0.308$ where appropriate.

5.2 Sample Selection & Data

We identify a sample of clumpy galaxies in the local universe by considering classifications from both the Galaxy Zoo 2 (GZ2 Willett et al., 2013) and Galaxy Zoo: Hubble (GZH Willett et al., 2017) projects. We provide a brief overview of each of these projects here.

The GZ2 subject sample consists of 285,962 galaxies identified as the brightest 25% (r -band magnitude < 17) residing in the SDSS North Galactic Cap region from Data Release 7 and included subjects with both spectroscopic and photometric redshifts out to $z < 0.25$. In addition to the DR7 Legacy catalog, galaxies were included from Stripe 82, a multiply-imaged strip along the celestial equator in the Southern Galactic Cap. Galaxies in this region were selected to have $m_r \leq 17.7$ and $\text{petror90_r} > 3$, where petror90_r is the radius containing 90% of the r-band Petrosian aperture flux.

The GZH project contains images drawn from several dedicated surveys and sample selection criteria. Most samples are drawn from various Hubble fields such as AEGIS, GOODS and COSMOS; however also included are the single-epoch and co-added images from Stripe 82 of the SDSS Data Release 7 that were part of the GZ2 project. The single-epoch imaging allows for local comparison to higher redshift galaxies, while the co-added imaging allows for image depth analyses.

The most notable difference between these two projects is the decision tree presented to volunteers. The goal of GZH was to collect detailed morphologies for galaxies in Hubble imaging which extends to much higher redshift than those of the SDSS sample. Galaxies at higher redshift do not necessarily follow the traditional Hubble sequence and thus a new branch of questioning was added to the decision tree for this project, shown in Figure 5.1. This included several questions concerning the clumpy nature of galaxies, a morphological feature well known to exist at higher redshift. Because GZH included galaxies from the GZ2 project, we can draw on both sets of decision trees to

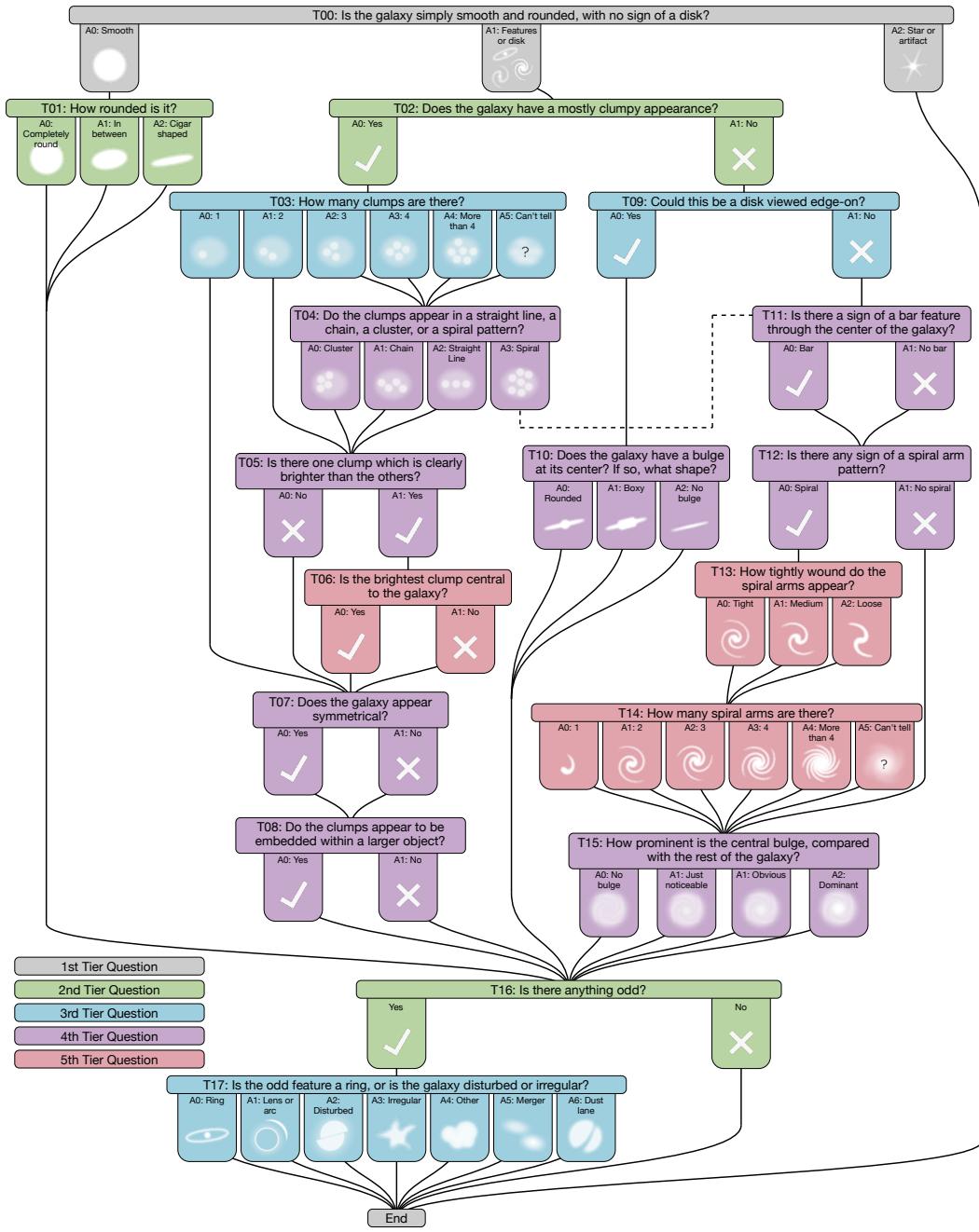


Figure 5.1 Galaxy Zoo: Hubble decision tree. The most notable difference between this decision tree and that used during the Galaxy Zoo 2 project is the “clumpy” branch of tasks.

identify clumpy features for galaxies in Stripe 82.

To select a sample of “clumpy” galaxies from the GZH Stripe 82 sample, we consider only those subjects with large featured (f_{featured}) and clumpy (f_{clumpy}) vote fractions: the fraction of volunteers who voted a subject as ‘featured or disk’ in response to the first question “Is the galaxy simply smooth and rounded, with no sign of a disk?” and who answered ‘yes’ to the question “Does the galaxy have a mostly clumpy appearance?” Specifically, we select galaxies which satisfy $f_{\text{featured}} \geq 0.5$ and $f_{\text{clumpy}} \geq 0.5$. Additionally, we require $N_{\text{votes}} \geq 20$, where N_{votes} is the number of volunteers who answered the clumpy question. This insures that f_{clumpy} is statistically significant and not a product of too few votes. This yields a sample of 629 galaxies: 273 single-epoch imaging and 356 from the co-added imaging. After visual inspection we find that this is hardly a pure sample of clumpy galaxies in the traditional sense, instead including a sizeable sample of tight groups of elliptical galaxies, as well as galaxies in various merging states, possessing multiple nuclei. [show example image?] After excluding these and duplicate imaging, we retain 92 coadd-depth and 105 single-depth clumpy galaxies, of which 156 are unique systems (some objects are common to both the single-epoch and coadd-depth imaging). Finally, we exclude galaxies with $z > 0.06$ in order to retain a sample wherein the physical scale as observed by SDSS is similar to Hubble’s at $z \sim 3$ to allow for comparison with high redshift samples. Our final sample contains 105 unique galaxies.

We obtain SDSS Data Release 12 (DR12) *ugriz* coadd imaging, as well as all optical spectra associated with each galaxy. The wavelength range of the SDSS spectra is $3800 - 9200\text{\AA}$ with a resolution of $R \sim 1500$ at $\sim 3800\text{\AA}$. The $3''$ SDSS fibers cover 2.9 kpc at $z = 0.05$. We visually inspect all spectra to verify that they are associated with an object in our sample as opposed to a nearby object or overlapping star. Through this inspection we determine that one clumpy galaxy is actually a juxtaposition of three galaxies at disparate redshifts flagged as clumpy in GZH due to the low resolution of SDSS imaging. We exclude this subject from our sample. Our final list includes 171 spectra. Approximately half of the galaxies in our sample have more than one spectrum with a handful having three or four spectra. Figure 5.2 shows some examples of our clumpy galaxies where the first column is the *gri* color composite SDSS jpeg image with red squares denoting the location of SDSS spectra, the middle column is

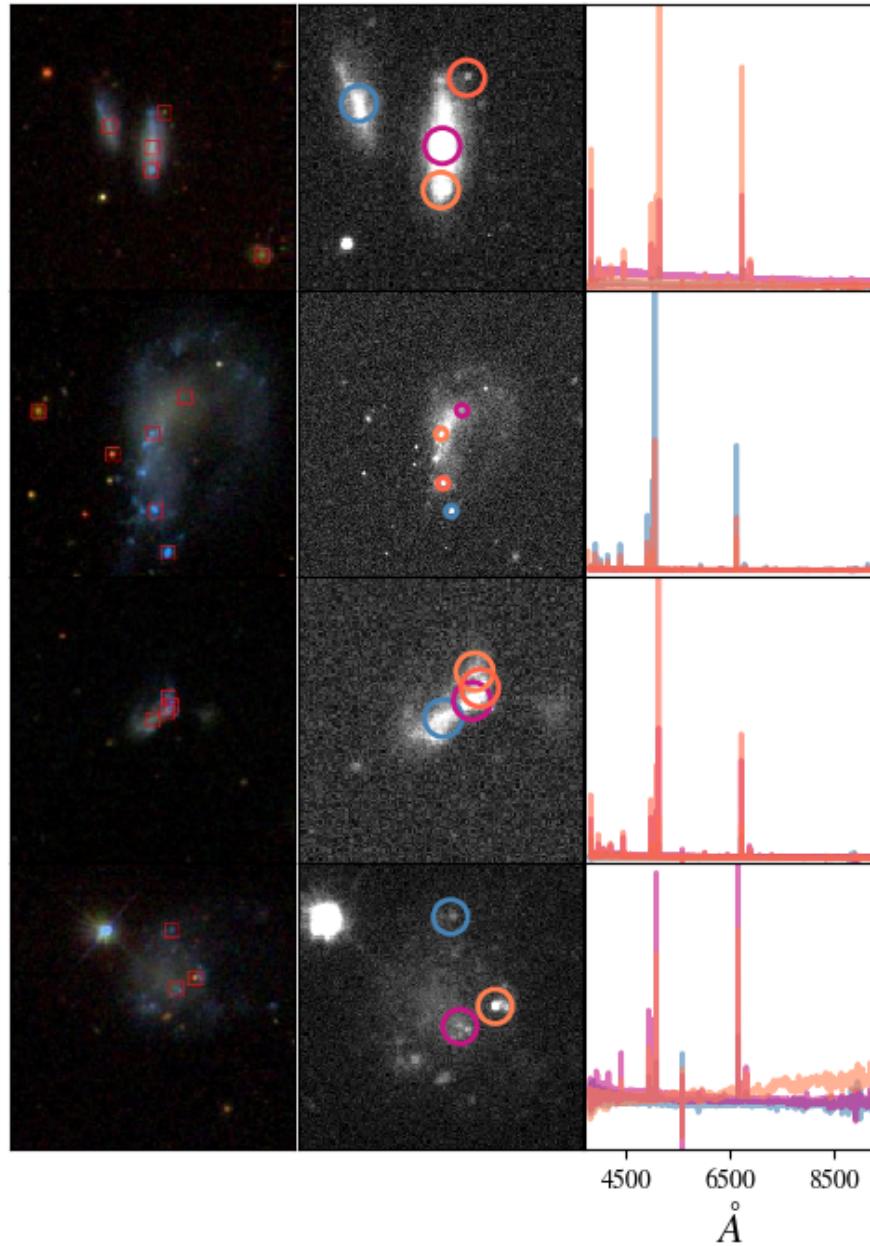


Figure 5.2 Sample of “clumpy” galaxies. The first column shows the SDSS *gri* color composite image where the red squares denote locations of SDSS spectroscopy. The middle column shows our postage stamp for the object where the colored circles show the same spectral locations, however these circles are color coded to the third panel which shows the spectra associated with each object.

the r -band postage stamp (described below) where the colored circles again denote the location of SDSS spectra and reflect the size of the fiber. The third column shows the spectra associated with each object, color-coded to the apertures in the middle column.

We create postage stamps of each galaxy from the r -band SDSS fields. The size of the postage stamp is taken to be $3 \times \text{petrorad}_r$, the Petrosian radius as measured in the r -band by the SDSS pipeline. We then run Source Extractor (Bertin & Arnouts, 1996) on each postage stamp. During visual inspection of these cutouts and the associated SExtractor segmentation maps we discover that the SDSS coordinates are occasionally incorrectly assigned, that is, star-forming clumps are mistaken for individual galaxies likely due to the low surface brightness of some of these systems. An example is shown in the final row of Figure 5.2. It is clear in the first panel that the central coordinates are not well aligned with the galactic center. The middle panel depicts our postage stamp in which we have corrected the galaxy's coordinates using that determined by SExtractor.

We also draw on the *GALEX-SDSS-WISE* Legacy Catalog (GSWLC, Salim et al., 2016) which provides stellar mass, star formation rates (SFR) and dust attenuations for 700,000 low-redshift galaxies. Specifically we obtain the GSWLC-X version which contains measurements using the deepest imaging available for each galaxy in the sample. We provide a brief overview of their methodology here. Galaxy physical properties were obtained from UV and optical spectral energy distribution (SED) fitting utilizing imaging data from the *Galaxy Evolution Explorer*, SDSS, and *Wide-field Infrared Survey Explorer* surveys for galaxies with $z < 0.3$, covering up to 90% of the SDSS footprint. Salim et al. (2016) use a Bayesian fitting methodology that includes corrections for photometry blending and emission-line. We match our sample against this catalog to objects within $15''$, of which we find 95. Of these, 6 are flagged as having failed the SED fitting and are thus excluded. Thus we retain 89 galaxies with estimates of global galactic parameters. Figure 5.3 shows the distribution of our sample wherein we plot the galaxy log stellar mass as a function of redshift with color denoting the log SFR. The median galaxy log stellar mass and SED SFR are $\sim 9 M_\odot$ and XXX, respectively.

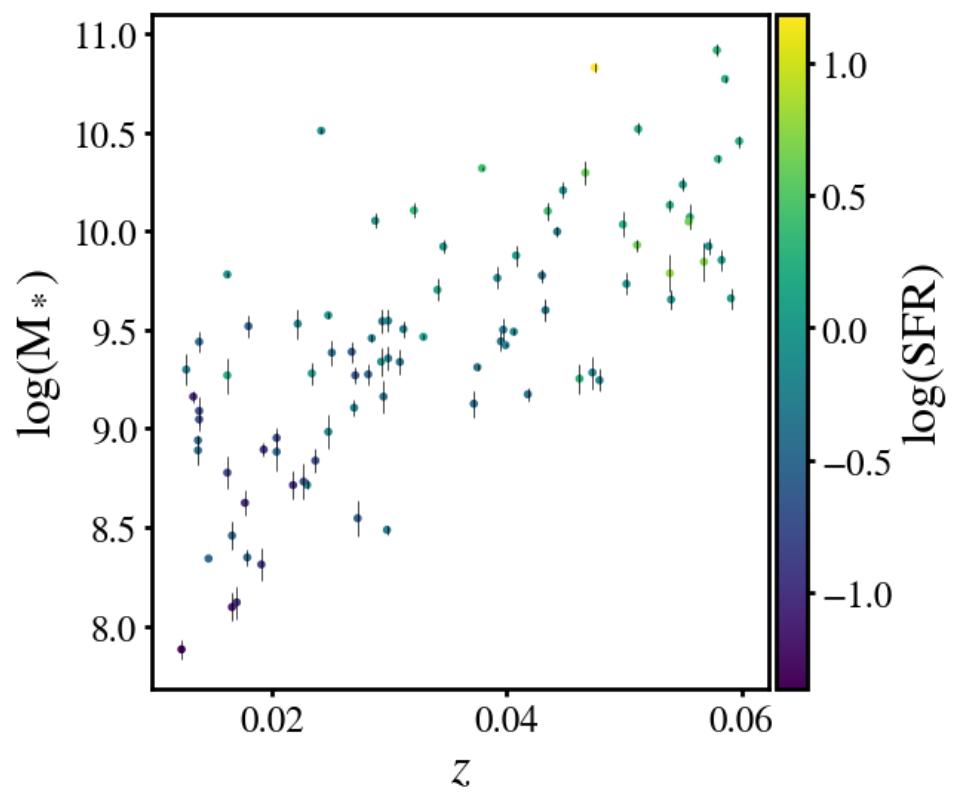


Figure 5.3 Stellar mass, redshift, and star-formation rate (SFR) for a subset of 89 “clumpy” galaxies as measured by the GSWLC (Salim et al., 2016).

5.3 Are these star-forming regions analogs of high-redshift clumps?

Many studies have been conducted (cite cite cite) comparing local HII regions to high-redshift star-forming clumps and the debate between whether these regions arise due to the same physical processes is still under debate. In this section we use the GSWLC and spectral features as measured by SDSS to explore various properties of the clumps in our sample.

5.3.1 Clump galactic radial distance

Most SDSS fibers are either centered on a bright star-forming region or on/near the galactic center. We thus compute the galactocentric distance of each fiber from the galaxy center and normalized by the galaxy’s half light radius (`FLUX_RADIUS`) as determined by SExtractor. The resulting distribution is shown in the third panel of Figure 5.4. That the distribution is obviously skewed towards the small galactocentric values is likely due to the several spectra that are not covering a clump but instead probe the central region.

The H_{α} equivalent width is a rough stand-in for specific SFR since it is the ratio of a strong star formation indicator (H_{α} line flux) and a reasonable proxy for stellar mass, i.e., the stellar continuum at H_{α} (Mármol-Queraltó et al., 2016). However, we find no relation between $\text{EW}(H_{\alpha})$ and clump radial distance. More telling would be a trend between clump radial distance and stellar age as this diagnostic will potentially distinguish between *in situ* and external clump origin theories. The high quality of the SDSS spectra will allow for detailed stellar age estimates in a future analysis.

5.3.2 Clump luminosity

We use the H_{α} flux as measured by a Gaussian fit to the H_{α} emission line to compute the star-formation rate using the well known calibration (e.g., Calzetti, 2013, and references therein)

$$\text{SFR}(\text{M}_{\odot}\text{yr}^{-1}) = 5.5 \times 10^{-42} L(H_{\alpha})(\text{ergs s}^{-1}) \quad (5.1)$$

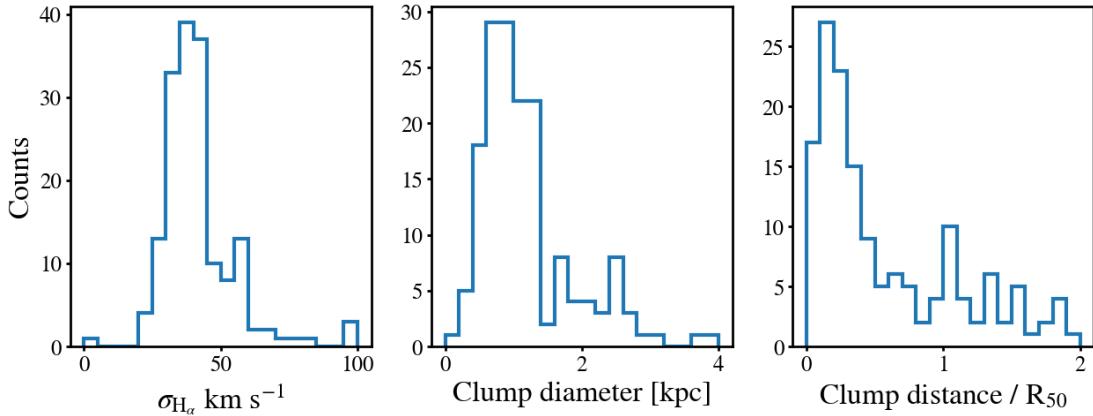


Figure 5.4 Clump properties.

We can then compare SFRs between those derived from the $\text{H}\alpha$ emission and the global SED SFR. For each galaxy in the sample, we sum the $\text{SFR}_{\text{H}\alpha}$ from each spectrum and compare that to SFR_{SED} , for those 89 galaxies that were matched to the GSWLC. Of those, we find the median ratio of the two star-formation measures is $\sim 10\%$, with a majority of galaxies having only one spectrum which we interpret as one clump. This rough estimate is not dissimilar to giant star-forming regions found at high redshift where it is typical for individual clumps to contribute 10% or more to the total star formation of the galaxy (Genzel et al., 2011; Guo et al., 2012; Wisnioski et al., 2012)

5.3.3 Velocity dispersion and clump diameters

In the first panel of Figure 5.4 we show the distribution of the $\text{H}\alpha$ velocity dispersion as measured by the SDSS spectroscopic pipeline, where the instrument resolution is subtracted in quadrature. The median for our sample is $\sigma_{\text{H}\alpha} \sim 40 \text{ km/s}$ and while this is slightly smaller than giant star-forming regions observed at $z \sim 2$ (Sanchez et al., 2013; Elmegreen and Elmegreen, 2010), it is also significantly larger than typical HII regions observed in local galaxies (cite cite cite).

Wisnioski et al. (2012) develop empirical relations between clump size, luminosity, and velocity dispersion for a sample of local HII regions along with a sample of high redshift clumps. They demonstrate that these relations hold over three orders of magnitude in clump diameter and five orders of magnitude in $\text{H}\alpha$ luminosity for thousands

of local HII regions and high-redshift clumps taken from 11 different studies. Specifically, we invert the following relation to estimate the clump diameter from the SDSS measurement of the H_{α} velocity dispersion

$$\log(\sigma) = (0.42 \pm 0.03) \times \log(d) + (0.33 \pm 0.09) \quad (5.2)$$

The resulting clump diameter distribution is shown in the middle panel of Figure 5.4. The median diameter is ~ 1 kpc which is larger than the average local HII region but slightly smaller than high redshift clumps. Keep in mind that some of these “clumps” could actually be the central bulge region of some of these systems as we do not specifically separate the two so as to bias against the possibility of older (and hence redder, less star formation) clumps at smaller galactic radius. The central regions will typically have much lower rates of star formation as measured by H_{α} and subsequently, smaller velocity dispersion, and thus they will have significantly smaller sizes according to this relation. In a more detailed analysis, clump sizes should be confirmed independently through a “core” analysis whereby clump sizes are measured by fitting a Gaussian to the 1D radial surface brightness profile of each star-forming region.

In future analyses, the quality of the SDSS spectra will allow for the derivation of accurate stellar ages from fitting of the continuum, as well as gas metallicity (e.g., Henry et al., 2015). With these measurements we can begin to study statistical properties that correlate with clump galactocentric distance.

5.4 Summary

Though preliminary, this analysis demonstrates that star-forming regions in this galaxy sample are likely more similar to high-redshift clumps than to typical local HII regions. This motivates not only a more in depth analysis of this sample but also justifies the search for similar galaxies in the local universe. These clumpy galaxies were found as a consequence of SDSS Stripe 82 galaxies being included in the GZH project. We next detail a new project that will potentially find an order of magnitude more galaxies similar to the sample presented here.

5.5 Clump Scout

The above analysis was performed on a subsample of galaxies determined as “clumpy” from the SDSS Stripe 82 sample included in the GZH project. The area covered by Stripe 82 was only a fraction of the full SDSS sky coverage. In this section we describe the *Clump Scout* project, a citizen-science initiative to discover more “clumpy” galaxies in the remaining SDSS footprint. We thus consider the remaining non-Stripe 82 SDSS galaxies that were originally part of the GZ2 project. We first exclude any galaxies with $z > 0.06$ in order to satisfy the resolution requirements discussed above. We also make a cut such that $f_{\text{smooth}} \leq 0.8$ in order to exclude those galaxies which are obviously elliptical and thus would likely not have travelled down the clumpy track of the GZH decision tree. These criteria yield a sample of $\sim 63K$ galaxies as shown in Figure 5.5 where we depict galaxy number density contours in the $z-f_{\text{smooth}}$ plane. The red dashed lines denotes the *Clump Scout* sample region. Based on the statistics of the clumpy galaxies found in the sky area of Stripe 82, we expect to find approximately 1000 such systems in the non-Stripe 82 SDSS sky coverage.

The goal of the project is to collect additional volunteer morphology information on par with that collected for the Stripe 82 sample in GZH. We will display jpeg images to volunteers via the Zooniverse Project Builder web platform which also hosts the Galaxy Zoo project. However, because we have such a large galaxy sample to go through, we will ask volunteers a slightly modified version of the top level question of the clumpy branch of the GZH decision tree in order to increase the speed of classification. Specifically, we ask one question, “Does the galaxy have a mostly clumpy appearance? If so, use the Clump Clicker tool to click the clumps you see.” This takes advantage of the Zooniverse Point Tool which marks the location on the image with each click a volunteer makes. This will provide an affirmation of a clumpy galaxy, a preliminary clump count, and coarse clump localization information. We have already created a pilot version of this project which consisted of 1000 galaxy images in two different “zoom” levels and received favorable reviews from volunteers who participated.

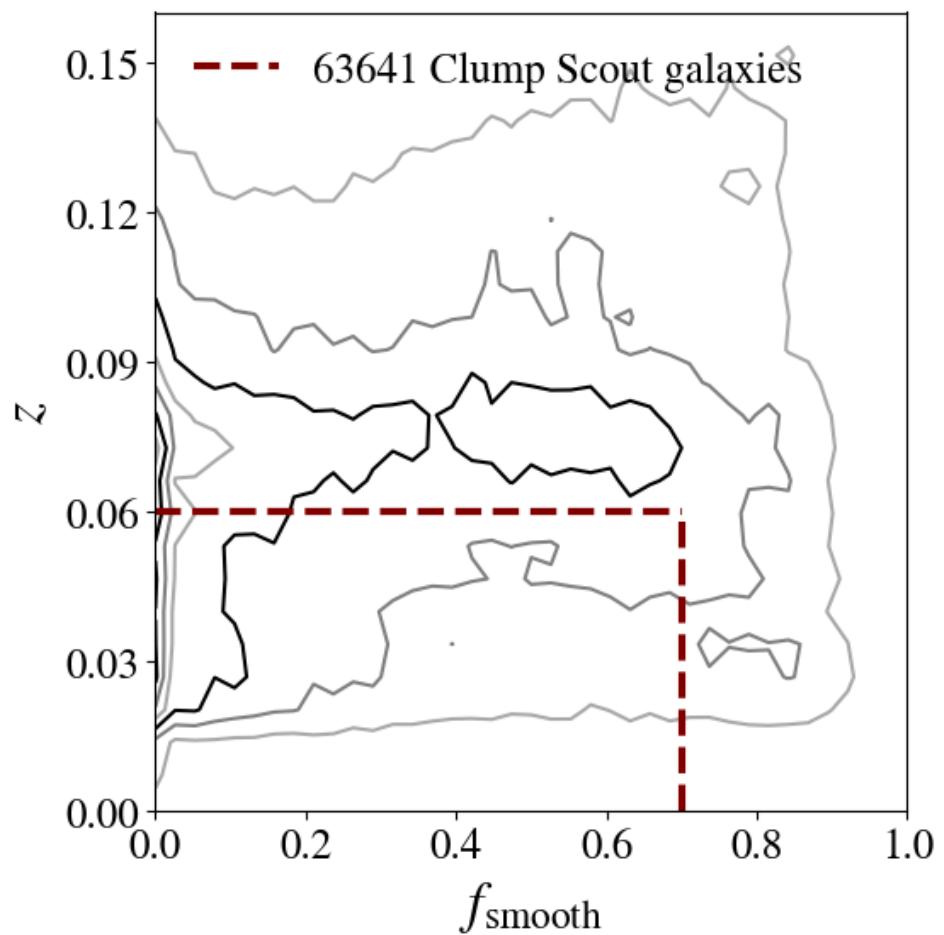


Figure 5.5 *Clump Scout* sample selection criteria from non-Stripe 82 GZ2 galaxy sample.

5.5.1 Redshift evolution of f_{clumpy}

With so many more clumpy galaxies at low redshift, we will be able to constrain clump origin formation theories. Specifically, we will investigate the evolution of the fraction of clumpy galaxies as a function of cosmic time, f_{clumpy} (not to be confused with the GZH clumpy vote fraction).

Guo et al. (2015) recently showed that the fraction of star-forming galaxies that have at least one off-center clump (f_{clumpy}) can be used to place constraints on theoretical models. By focusing on the redshift range between $0.5 < z < 3$, they find that f_{clumpy} changes with the stellar mass of the galaxies. Low-mass ($M < 10^{10} M_{\odot}$) galaxies keep an almost constant f_{clumpy} of $\sim 60\%$ from $z \sim 3$ to $z \sim 0.5$, while massive galaxies drop their f_{clumpy} from 55% at $z \sim 3$ to 15%, at $z \sim 0.5$, as shown in Figure XXX (adapted from Guo et al., 2015). Guo et al. (2015) argue that these observations support a model in which the clumpy star-formation results from multiple processes. In massive galaxies, the evolutionary trends are consistent with violent disk instability, however the apparent lack of f_{clumpy} evolution in low mass galaxies is more consistent with a minor merger original. However, this conclusion, is based primarily on the lowest redshift bin probed by the Guo et al. (2015) data. In fact, when these data are combined with f_{clumpy} measured from a variety of different surveys at $z < 1$, the result is not so convincing anymore. The variation in the low-z measurements, however, is clearly large mostly as a consequence of non-uniform selection criteria and uncontrolled-for biases. *Clump Scout* promises to provide the largest sample to date selected in a uniform fashion and with biases properly accounted for, especially in the low-mass regime where measurements of f_{clumpy} can provide the strongest model constraints.

5.6 Summary and conclusions

In this work, we have isolated a sample of galaxies with morphologies which resemble those of star-forming clumpy galaxies of the high-redshift universe. We identify these galaxies in the local universe through the Galaxy Zoo: Hubble project which included imaging from Stripe 82 of the SDSS. We isolate 105 galaxies that have a traditional clumpy morphology and acquire SDSS imaging and spectroscopic data for these galaxies, including spectra for over 150 clumps. We obtain stellar mass and star-formation rates

for these galaxies through the GSWLC catalog. Through a preliminary analysis of the SDSS spectra we determine the clumps in these galaxies are in many ways consistent with high redshift clumps. However a more detailed analysis is necessary.

Additionally, we describe the *Clump Scout* project designed to search for additional clumpy galaxies in the local universe by presenting color-composite SDSS non-Stripe 82 imaging to the general public. We have made preliminary tests of this project with highly favorable reviews from volunteers and we expect to find an additional 1000 clumpy galaxies. This will provide a large statistical sample of local clumpy galaxies which will constrain the fraction of clumpy galaxies at low redshift allowing us to distinguish between various clump original theories. Additionally, because these are all SDSS objects, there will undoubtedly be spectra for thousands of clumps allowing for large statistical studies of star formation in the nearby universe.

Chapter 6

Summary & Future Work

The goal of this thesis was to study the mass-loss histories of hypergiants stars using new capabilities in near-IR imaging and polarimetry, and airborne mid-IR imaging. With LBT/LMIRCam and MMT-Pol on the MMT we have imaged the nebulae of VY CMa and IRC +10420 with sub-arcsecond resolution, revealing recent mass-loss in the close environments around these stars. To probe further into these hypergiants' past history, we have used mid-infrared imaging with SOFIA/FORCAST to search for cold dust and performed 1-D radiative transfer modeling of their SEDs and resolved profiles.

Our $2 - 5 \mu\text{m}$ adaptive optics imaging of the cool hypergiant VY CMa penetrates deeper into its dusty nebula than in the optical, probing its recent mass-loss history. In Chapter 2 we analyzed the resolved images of its peculiar “Southwest” Clump, which has no obvious counterpart on the opposite side of the star. The distinct shape of the SW Clump is suggestive of a short-lived, localized event and may be analogous to a coronal mass ejection (CME) from a single location on the Sun’s surface. A short-lived ejection event is consistent with the SW Clump appearing as a confined, coherent shape several hundred years after ejection. Using adaptive optics imaging polarimetry, in Chapter 3 we demonstrated the Clump is optically thick through at least $3.1 \mu\text{m}$ and reaffirmed the lower limit mass of $5 \times 10^{-3} M_\odot$ based on modeling its surface brightness as optically thick scattered light. Our $1.3 \mu\text{m}$ polarimetry detects several other prominent features of VY CMa’s nebula include the NW Arc, Arc 2, S Knot, and S Arc. Using the polarized intensity as a lower limit on total scattered light intensity, we found each of these features to be optically thick as well. Their relatively high intrinsic polarizations

are consistent with their high scattering optical depths since the depolarizing effect of multiple scatters is reduced for typical silicate grain albedos. In Chapter 4, we used 20 – 37 μm infrared imaging with SOFIA/FORCAST to search for evidence of earlier mass loss. VY CMa’s morphology at the longest wavelengths coincides with the general shape of the highly asymmetric nebulae seen in the visual, suggesting thermal emission from dust associated with the expanding arcs to the northwest and southwest. Modeling its SED we computed an average mass-loss rate of $6 \times 10^{-4} M_{\odot} \text{ yr}^{-1}$ over the past ~ 1200 years, with no clear evidence of mass loss much farther in its past.

Our study of the warm hypergiant IRC +10420 similarly traced its mass-loss over a range of angular scales. At the sub-arcsecond scale using adaptive optics, in Chapter 3 we used 2.2 μm polarimetry to reveal a relatively uniform nebula largely in the plane of the sky extending out to 2''.5 from the star. This low-latitude ejecta is optically thick. Combining the polarimetry with 3 – 5 μm imaging that shows extended emission, we modeled the flux of this nebula and found its emission is an order of magnitude brighter than can be explained by simple extrapolation of the scattered light seen at 2.2 μm . We hypothesized grains warmed to a temperature higher than the expected grain equilibrium temperature, but consistent with the local gas temperature in this region. In Chapter 4 we presented 8 – 12 μm adaptive optics images that reveal spatially extended emission spanning nearly the same range as the 2.2 μm polarimetry. Applying radiative transfer modeling to the intensity profile of this extended emission and the SED, we found that IRC +10420’s mass-loss history is divided into two distinct periods. Our best-fit model showed that it lost mass at a high average rate of $2 \times 10^{-3} M_{\odot} \text{ yr}^{-1}$ from 6000 – 2000 yr ago during its presumed RSG stage, followed by an order of magnitude decrease to an average rate of $1 \times 10^{-4} M_{\odot} \text{ yr}^{-1}$ in the past 2000 yr.

In addition to VY CMa and IRC +10420, our SOFIA/FORCAST program included the RSG μ Ceph and the warm hypergiant ρ Cas. Our study of μ Cep probes a mass-loss history extending back $\sim 13,000$ yr. To match the observed intensity profile of its resolved nebula in FORCAST 25 – 37 μm images, our radiative transfer modeling requires a declining mass-loss rate. We find that over the 13,000 yr dynamical age of the shell, its mass loss-rate has declined from 5×10^{-6} to $1 \times 10^{-6} M_{\odot} \text{ yr}^{-1}$. In contrast to the high mass-loss rate of VY CMa, μ Cep’s rate is significantly lower than RSGs of comparable luminosity. In the case of ρ Cas, we demonstrated that our new 19.7 –

$37.1\text{ }\mu\text{m}$ SOFIA/FORCAST photometry are consistent with the continued expansion and thinning of the dust shell formed as a result of the 1946 eruption. We did not find evidence of new dust formation from more recent eruptions.

In the future, infrared imaging over a range of angular scales will be used to expand the sample of RSGs and post-RSGs with resolved mass-loss. The SOFIA/FORCAST Cycle 3 program 03_0082 (PI: R. Humphreys) included three additional RSG targets (NML Cyg, VX Sgr and S Per) and contingent approval has been granted for observations of the yellow hypergiant HR 5171A during a Cycle 4 Southern deployment if flight scheduling permits. In Fall 2016, VY CMa is scheduled for observation with NOMIC, the $8 - 13\text{ }\mu\text{m}$ imager on the Large Binocular Telescope. With NOMIC's $0.^{\prime\prime}27$ angular resolution (for single dish observing mode), VY CMa's SW Clump should be resolved. These observations will better sample the Clump's SED and could help resolve the puzzle of its non-detection by ALMA reported in ?.

The work presented in this thesis has demonstrated the capabilities of MMT-Pol and LMIRCam to resolve circumstellar ejecta around RSGs and post-RSGs. These capabilities will be extended to several recently identified RSG clusters, each of which contains substantial coeval RSG populations. These include RSGC1 (??), RSGC2 = Stephenson 2 (?), NGC 7419 (?) and the Per OB1 association, which contains many RSGs in its halo including S Per (?). Imaging these clusters of RSGs in polarized intensity in the near-infrared with MMT-Pol will cleanly separate light scattered by dusty nebulae from the stars' light, and when combined with adaptive optics imaging from $3 - 13\text{ }\mu\text{m}$ with LMIRCam/NOMIC will enable us to trace the RSGs mass-loss. Since the RSG cluster populations are coeval, this will allow the role of high mass loss events on their evolution to be studied.

References

- Abazajian, K., Adelman-McCarthy, J. K., Agüeros, M. A., et al. 2003, AJ, 126, 2081
- Abazajian, K. N., Adelman-McCarthy, J. K., Agüeros, M. A., et al. 2009, ApJS, 182, 543
- Abraham, R. G., Tanvir, N. R., Santiago, B. X., et al. 1996, MNRAS, 279, L47
- Abraham, R. G., Valdes, F., Yee, H. K. C., & van den Bergh, S. 1994, ApJ, 432, 75
- Abraham, R. G., van den Bergh, S., & Nair, P. 2003, ApJ, 588, 218
- Adelman-McCarthy, J. K., Agüeros, M. A., Allam, S. S., et al. 2008, ApJS, 175, 297
- Andrae, R., Jahnke, K., & Melchior, P. 2011, MNRAS, 411, 385
- Appenzeller, I., Fricke, K., Fürtig, W., et al. 1998, The Messenger, 94, 1
- Baillard, A., Bertin, E., de Lapparent, V., et al. 2011, A&A, 532, A74
- Baldry, I. K., Balogh, M. L., Bower, R., Glazebrook, K., & Nichol, R. C. 2004a, in American Institute of Physics Conference Series, Vol. 743, The New Cosmology: Conference on Strings and Cosmology, ed. R. E. Allen, D. V. Nanopoulos, & C. N. Pope, 106–119
- Baldry, I. K., Glazebrook, K., Brinkmann, J., et al. 2004b, ApJ, 600, 681
- Ball, N. M., Loveday, J., Fukugita, M., et al. 2004, MNRAS, 348, 1038
- Bamford, S. P., Nichol, R. C., Baldry, I. K., et al. 2009, MNRAS, 393, 1324
- Banerji, M., Lahav, O., Lintott, C. J., et al. 2010, MNRAS, 406, 342

- Basu-Zych, A., & Scharf, C. 2004, ApJ, 615, L85
- Bell, E. F., McIntosh, D. H., Katz, N., & Weinberg, M. D. 2003, ApJS, 149, 289
- Bell, E. F., Wolf, C., Meisenheimer, K., et al. 2004, ApJ, 608, 752
- Bershady, M. A., Jangren, A., & Conselice, C. J. 2000, AJ, 119, 2645
- Bertin, E., & Arnouts, S. 1996, A&AS, 117, 393
- Blanton, M. R., Brinkmann, J., Csabai, I., et al. 2003a, AJ, 125, 2348
- Blanton, M. R., Hogg, D. W., Bahcall, N. A., et al. 2003b, ApJ, 594, 186
- Bluck, A. F. L., Conselice, C. J., Bouwens, R. J., et al. 2009, MNRAS, 394, L51
- Bower, R. G., Lucey, J. R., & Ellis, R. S. 1992, MNRAS, 254, 601
- Brandt, J. C., & Chamberlain, J. W. 1959, ApJ, 130, 670
- Brasken, M., & Kyrola, E. 1998, A&A, 332, 732
- Breiman, L. 2001, Machine Learning, 45, 5
- Brinchmann, J., & Ellis, R. S. 2000, ApJ, 536, L77
- Buitrago, F., Trujillo, I., Conselice, C. J., & Häufner, B. 2013, MNRAS, 428, 1460
- Bundy, K., Ellis, R. S., & Conselice, C. J. 2005, ApJ, 625, 621
- Bunker, A., Smith, J., Spinrad, H., Stern, D., & Warren, S. 2003, Ap&SS, 284, 357
- Buta, R. J. 2011, arXiv:1102.0550
- Calzetti, D. 2013, Star Formation Rate Indicators, ed. J. Falcón-Barroso & J. H. Knapen, 419
- Cameron, E., Carollo, C. M., Oesch, P. A., et al. 2011, ApJ, 743, 146
- Cantalupo, S., Lilly, S. J., & Haehnelt, M. G. 2012, MNRAS, 425, 1992
- Cardamone, C., Schawinski, K., Sarzi, M., et al. 2009, MNRAS, 399, 1191

- Casteels, K. R. V., Conselice, C. J., Bamford, S. P., et al. 2014, MNRAS, 445, 1157
- Chapman, S. C., Scott, D., Windhorst, R. A., et al. 2004, ApJ, 606, 85
- Cirasuolo, M., McLure, R. J., Dunlop, J. S., et al. 2007, MNRAS, 380, 585
- Colbert, J. W., Scarlata, C., Teplitz, H., et al. 2011, ApJ, 728, 59
- Conselice, C. J. 2003, ApJS, 147, 1
- . 2006, MNRAS, 373, 1389
- . 2014, ARA&A, 52, 291
- Conselice, C. J., Bershady, M. A., & Jangren, A. 2000, ApJ, 529, 886
- Conselice, C. J., Blackburne, J. A., & Papovich, C. 2005, ApJ, 620, 564
- Conselice, C. J., Mortlock, A., Bluck, A. F. L., Grützbauch, R., & Duncan, K. 2013, MNRAS, 430, 1051
- Conselice, C. J., Bluck, A. F. L., Buitrago, F., et al. 2011, MNRAS, 413, 80
- Cowie, L. L., Songaila, A., Hu, E. M., & Cohen, J. G. 1996, AJ, 112, 839
- Darg, D. W., Kaviraj, S., Lintott, C. J., et al. 2010, MNRAS, 401, 1552
- de Vaucouleurs, G. 1959, Handbuch der Physik, 53, 275
- . 1963, ApJS, 8, 31
- de Vaucouleurs, G., de Vaucouleurs, A., Corwin, Jr., H. G., et al. 1991, Third Reference Catalogue of Bright Galaxies. Volume I: Explanations and references. Volume II: Data for galaxies between 0^h and 12^h . Volume III: Data for galaxies between 12^h and 24^h .
- Dickinson, M. 2000, in Philosophical Transactions of the Royal Society of London Series A, Vol. 358, Astronomy, physics and chemistry of H $^{+}$, 3, 2001
- Dieleman, S., Willett, K. W., & Dambre, J. 2015, MNRAS, 450, 1441
- Dijkstra, M., & Kramer, R. 2012, MNRAS, 424, 1672

- Dijkstra, M., & Loeb, A. 2008, MNRAS, 386, 492
- . 2009, MNRAS, 400, 1109
- Dressler, A. 1980, ApJ, 236, 351
- Driver, S. P., Robotham, A. S. G., Bland-Hawthorn, J., et al. 2013, MNRAS, 430, 2622
- Elmegreen, B. G., Bournaud, F., & Elmegreen, D. M. 2008, ApJ, 688, 67
- Elmegreen, B. G., Elmegreen, D. M., Sánchez Almeida, J., et al. 2013, ApJ, 774, 86
- Fardal, M. A., Katz, N., Gardner, J. P., et al. 2001, ApJ, 562, 605
- Fasano, G., Poggianti, B. M., Couch, W. J., et al. 2000, ApJ, 542, 673
- Faucher-Giguère, C.-A., Kereš, D., Dijkstra, M., Hernquist, L., & Zaldarriaga, M. 2010, ApJ, 725, 633
- Francis, P. J., Woodgate, B. E., Warren, S. J., et al. 1996, ApJ, 457, 490
- Freeman, P. E., Izbicki, R., Lee, A. B., et al. 2013, MNRAS, 434, 282
- Fukugita, M., Nakamura, O., Okamura, S., et al. 2007, AJ, 134, 579
- Furlanetto, S. R., Schaye, J., Springel, V., & Hernquist, L. 2005, ApJ, 622, 7
- Galloway, M. A., Willett, K. W., Fortson, L. F., et al. 2015, MNRAS, 448, 3442
- Geach, J. E., Smail, I., Chapman, S. C., et al. 2007, ApJ, 655, L9
- Geach, J. E., Matsuda, Y., Smail, I., et al. 2005, MNRAS, 363, 1398
- Geach, J. E., Alexander, D. M., Lehmer, B. D., et al. 2009, ApJ, 700, 1
- Geach, J. E., Bower, R. G., Alexander, D. M., et al. 2014, ApJ, 793, 22
- Genzel, R., Newman, S., Jones, T., et al. 2011, ApJ, 733, 101
- Glasser, G. J. 1962, Journal of the American Statistical Association, 57, 648
- GMTO Corporation. 2012, Giant Magellan Telescope Scientific Promise and Opportunities

- Griffith, R. L., Cooper, M. C., Newman, J. A., et al. 2012, ApJS, 200, 9
- Grogin, N. A., Kocevski, D. D., Faber, S. M., et al. 2011, ApJS, 197, 35
- Guo, Y., Giavalisco, M., Ferguson, H. C., Cassata, P., & Koekemoer, A. M. 2012, ApJ, 757, 120
- Guo, Y., Ferguson, H. C., Bell, E. F., et al. 2015, ApJ, 800, 39
- Haiman, Z., & Rees, M. J. 2001, ApJ, 556, 87
- Haiman, Z., Spaans, M., & Quataert, E. 2000, ApJ, 537, L5
- Hayes, M., & Scarlata, C. 2011, in SF2A-2011: Proceedings of the Annual meeting of the French Society of Astronomy and Astrophysics, ed. G. Alecian, K. Belkacem, R. Samadi, & D. Valls-Gabaud, 129–133
- Hayes, M., Scarlata, C., & Siana, B. 2011, Nature, 476, 304
- Holwerda, B. W., Muñoz-Mateos, J.-C., Comerón, S., et al. 2014, ApJ, 781, 12
- Hubble, E. P. 1926, ApJ, 64, doi:10.1086/143018
- . 1936, *The Realm of the Nebulae* (Yale University Press)
- Huertas-Company, M., Rouan, D., Tasca, L., Soucail, G., & Le Fèvre, O. 2008, A&A, 478, 971
- Huertas-Company, M., Gravet, R., Cabrera-Vives, G., et al. 2015, ApJS, 221, 8
- Huertas-Company, M., Bernardi, M., Pérez-González, P. G., et al. 2016, MNRAS, 462, 4495
- Humphrey, A., Vernet, J., Villar-Martín, M., et al. 2013, ApJ, 768, L3
- Jimenez, R., & Haiman, Z. 2006, Nature, 440, 501
- Kartaltepe, J. S., Mozena, M., Kocevski, D., et al. 2015, ApJS, 221, 11
- Kasper, M. E., Beuzit, J.-L., Verinaud, C., et al. 2008, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 7015, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, 1

- Kauffmann, G., Heckman, T. M., White, S. D. M., et al. 2003, MNRAS, 341, 54
- Keller, C. U., Schmid, H. M., Venema, L. B., et al. 2010, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 7735, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, 6
- Kelvin, L. S., Driver, S. P., Robotham, A. S. G., et al. 2014, MNRAS, 444, 1647
- Kodama, T., & Arimoto, N. 1997, A&A, 320, 41
- Koekemoer, A. M., Faber, S. M., Ferguson, H. C., et al. 2011, ApJS, 197, 36
- Kormendy, J. 1977, ApJ, 217, 406
- Kormendy, J., & Bender, R. 2012, ApJS, 198, 2
- Kormendy, J., & Kennicutt, Jr., R. C. 2004, ARA&A, 42, 603
- Kubo, M., Yamada, T., Ichikawa, T., et al. 2015, ArXiv e-prints, arXiv:1510.04816
- Land, K., Slosar, A., Lintott, C., et al. 2008, MNRAS, 388, 1686
- Laureijs, R., Amiaux, J., Arduini, S., et al. 2011, ArXiv e-prints, arXiv:1110.3193
- Lee, B., Giavalisco, M., Williams, C. C., et al. 2013, ApJ, 774, 47
- Lee, H.-W., & Ahn, S.-H. 1998, ApJ, 504, L61
- Lintott, C., Schawinski, K., Bamford, S., et al. 2011, MNRAS, 410, 166
- Lintott, C. J., Schawinski, K., Slosar, A., et al. 2008, MNRAS, 389, 1179
- Lintott, C. J., Schawinski, K., Keel, W., et al. 2009, MNRAS, 399, 129
- Loeb, A., & Rybicki, G. B. 1999, ApJ, 524, 527
- López-Cruz, O., Barkhouse, W. A., & Yee, H. K. C. 2004, ApJ, 614, 679
- Lorenz, M. O. 1905, Publications of the American Statistical Association, Volume 9, Number 70, p. 209-219, 9, 209
- Lotz, J. M., Primack, J., & Madau, P. 2004, AJ, 128, 163

- Madau, P., & Dickinson, M. 2014, ARA&A, 52, 415
- Man, A. W. S., Toft, S., Zirm, A. W., Wuyts, S., & van der Wel, A. 2012, ApJ, 744, 85
- Mármol-Queraltó, E., McLure, R. J., Cullen, F., et al. 2016, MNRAS, 460, 3587
- Marshall, P. J., Verma, A., More, A., et al. 2016, MNRAS, 455, 1171
- Masters, K. L., Mosleh, M., Romer, A. K., et al. 2010, MNRAS, 405, 783
- Masters, K. L., Nichol, R. C., Hoyle, B., et al. 2011, MNRAS, 411, 2026
- Matsuda, Y., Yamada, T., Hayashino, T., et al. 2004, AJ, 128, 569
- Matsuo, T., & Tamura, M. 2010, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 7735, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, 84
- McLinden, E. M., Malhotra, S., Rhoads, J. E., et al. 2013, ApJ, 767, 48
- Meert, A., Vikram, V., & Bernardi, M. 2016, MNRAS, 455, 2440
- Mei, S., Holden, B. P., Blakeslee, J. P., et al. 2009, ApJ, 690, 42
- More, A., Verma, A., Marshall, P. J., et al. 2016, MNRAS, 455, 1191
- Mori, M., & Umemura, M. 2006, Nature, 440, 644
- Mortlock, A., Conselice, C. J., Hartley, W. G., et al. 2013, MNRAS, 433, 1185
- Nair, P. B., & Abraham, R. G. 2010, ApJS, 186, 427
- Nakamura, O., Fukugita, M., Yasuda, N., et al. 2003, AJ, 125, 1682
- Neufeld, D. A. 1990, ApJ, 350, 216
- Nilsson, K. K., Fynbo, J. P. U., Møller, P., Sommer-Larsen, J., & Ledoux, C. 2006, A&A, 452, L23
- Odewahn, S. C., Cohen, S. H., Windhorst, R. A., & Philip, N. S. 2002, ApJ, 568, 539

- Palunas, P., Teplitz, H. I., Francis, P. J., Williger, G. M., & Woodgate, B. E. 2004, *ApJ*, 602, 545
- Papovich, C., Dickinson, M., Giavalisco, M., Conselice, C. J., & Ferguson, H. C. 2005, *ApJ*, 631, 101
- Patat, F., Moehler, S., O'Brien, K., et al. 2011, *A&A*, 527, A91
- Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, *Journal of Machine Learning Research*, 12, 2825
- Peng, C. Y., Ho, L. C., Impey, C. D., & Rix, H.-W. 2002, *AJ*, 124, 266
- Peng, Y.-j., Lilly, S. J., Kovač, K., et al. 2010, *ApJ*, 721, 193
- Pertenais, M., Neiner, C., Parès, L. P., et al. 2014, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 9144, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, 3
- Peth, M. A., Lotz, J. M., Freeman, P. E., et al. 2016, *MNRAS*, 458, 963
- Petrosian, V. 1976, *ApJ*, 209, L1
- Plaszczynski, S., Montier, L., Levrier, F., & Tristram, M. 2014, *MNRAS*, 439, 4048
- Prescott, M. K. M., Kashikawa, N., Dey, A., & Matsuda, Y. 2008, *ApJ*, 678, L77
- Prescott, M. K. M., Momcheva, I., Brammer, G. B., Fynbo, J. P. U., & Møller, P. 2015, *ApJ*, 802, 32
- Prescott, M. K. M., Smith, P. S., Schmidt, G. D., & Dey, A. 2011, *ApJ*, 730, L25
- Rosdahl, J., & Blaizot, J. 2012, *MNRAS*, 423, 344
- Rybicki, G. B., & Loeb, A. 1999, *ApJ*, 520, L79
- Saito, T., Shimasaku, K., Okamura, S., et al. 2006, *ApJ*, 648, 54
- Salim, S., Lee, J. C., Janowiecki, S., et al. 2016, *ApJS*, 227, 2
- Sandage, A. 1961, *The Hubble atlas of galaxies*

- Sandage, A., & Tammann, G. A. 1981, A revised Shapley-Ames Catalog of bright galaxies
- Sandage, A., & Visvanathan, N. 1978, ApJ, 223, 707
- Scarlata, C., Carollo, C. M., Lilly, S., et al. 2007, ApJS, 172, 406
- Scarlata, C., Colbert, J., Teplitz, H. I., et al. 2009, ApJ, 706, 1241
- Schawinski, K., Thomas, D., Sarzi, M., et al. 2007, MNRAS, 382, 1415
- Schawinski, K., Lintott, C., Thomas, D., et al. 2009, MNRAS, 396, 818
- Schawinski, K., Urry, C. M., Simmons, B. D., et al. 2014, MNRAS, 440, 889
- Seager, S., Cash, W. C., Kasdin, N. J., et al. 2014, in American Astronomical Society Meeting Abstracts, Vol. 224, American Astronomical Society Meeting Abstracts 224, 311.06
- Sersic, J. L. 1968, Atlas de galaxias australes
- Shen, S., Mo, H. J., White, S. D. M., et al. 2003, MNRAS, 343, 978
- Sheth, K., Elmegreen, D. M., Elmegreen, B. G., et al. 2008, ApJ, 675, 1141
- Simard, L., Mendel, J. T., Patton, D. R., Ellison, S. L., & McConnachie, A. W. 2011, ApJS, 196, 11
- Simmons, B. D., Melvin, T., Lintott, C., et al. 2014, MNRAS, 445, 3466
- Simmons, B. D., Lintott, C., Willett, K. W., et al. 2017, MNRAS, 464, 4420
- Simmons, J. F. L., & Stewart, B. G. 1985, A&A, 142, 100
- Skibba, R. A., Bamford, S. P., Nichol, R. C., et al. 2009, MNRAS, 399, 966
- Smethurst, R. J., Lintott, C. J., Simmons, B. D., et al. 2016, MNRAS, 463, 2986
- Smith, D. J. B., & Jarvis, M. J. 2007, MNRAS, 378, L49
- Smith, D. J. B., Jarvis, M. J., Simpson, C., & Martínez-Sansigre, A. 2009, MNRAS, 393, 309

- Smith, G. P., Treu, T., Ellis, R. S., Moran, S. M., & Dressler, A. 2005, ApJ, 620, 78
- Snyder, G. F., Torrey, P., Lotz, J. M., et al. 2015, MNRAS, 454, 1886
- Stapelfeldt, K. R., Brenner, M. P., Warfield, K. R., et al. 2014, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 9143, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, 2
- Steidel, C. C., Adelberger, K. L., Shapley, A. E., et al. 2000, ApJ, 532, 170
- Stenflo, J. O. 1980, A&A, 84, 68
- Strateva, I., Ivezić, Ž., Knapp, G. R., et al. 2001, AJ, 122, 1861
- Strauss, M. A., Weinberg, D. H., Lupton, R. H., et al. 2002, AJ, 124, 1810
- Taniguchi, Y., & Shioya, Y. 2000, ApJ, 532, L13
- Taylor, E. N., Hopkins, A. M., Baldry, I. K., et al. 2015, MNRAS, 446, 2144
- Thanjavur, K., Simard, L., Bluck, A. F. L., & Mendel, T. 2016, Monthly Notices of the Royal Astronomical Society, 459, 44
- Toomre, A. 1964, ApJ, 139, 1217
- Tully, R. B., Mould, J. R., & Aaronson, M. 1982, ApJ, 257, 527
- van den Bergh, S. 1976, ApJ, 206, 883
- van Dokkum, P. G. 2001, PASP, 113, 1420
- Vinokur, M. 1965, Annales d'Astrophysique, 28, 412
- Wardle, J. F. C., & Kronberg, P. P. 1974, ApJ, 194, 249
- Watanabe, M., Kodaira, K., & Okamura, S. 1985, ApJ, 292, 72
- Weidinger, M., Møller, P., & Fynbo, J. P. U. 2004, Nature, 430, 999
- Weijmans, A.-M., Bower, R. G., Geach, J. E., et al. 2010, MNRAS, 402, 2245
- Whitmore, B. C., Lucas, R. A., McElroy, D. B., et al. 1990, AJ, 100, 1489

- Willett, K. W., Lintott, C. J., Bamford, S. P., et al. 2013, MNRAS, 435, 2835
- Willett, K. W., Galloway, M. A., Bamford, S. P., et al. 2017, MNRAS, 464, 4176
- Wisnioski, E., Glazebrook, K., Blake, C., et al. 2012, MNRAS, 422, 3339
- Yang, Y., Zabludoff, A., Tremonti, C., Eisenstein, D., & Davé, R. 2009, ApJ, 693, 1579
- Yang, Y., Decarli, R., Dannerbauer, H., et al. 2012, ApJ, 744, 178

Appendix A

Spectro-polarimetry Confirms Central Powering in a Ly α Nebula at z = 3.09

A slightly modified version of this chapter has been published in The Astrophysical Journal with the following bibliographic reference: Beck, Melanie; Scarlata, Claudia; Hayes, Matthew; Dijkstra, Mark; and Jones, Terry J. 2016, ApJ, 818, 138.

Abstract

We present a follow-up study to the imaging polarimetry performed by Hayes et al. (2011) on LAB1 in the SSA22 protocluster region. Arguably the most well-known Lyman- α “blob”, this radio-quiet emission-line nebula likely hosts a galaxy which is either undergoing significant star formation or hosts an AGN, or both. We obtain deep, spatially resolved spectro-polarimetry of the Ly α emission and detect integrated linear polarization of $9\text{--}13\% \pm 2\text{--}3\%$ at a distance of approximately 15 kpc north and south of the peak of the Ly α surface brightness with polarization vectors lying tangential to the galactic central source. In these same regions, we also detect a wavelength dependence in the polarization which is low at the center of the Ly α line

profile and rises substantially in the wings of the profile. These polarization signatures are easily explained by a weak out-flowing shell model. The spectral dependence of the polarization presented here provide a framework for future observations and interpretations of the southern portion of LAB1 in that any model for this system must be able to reproduce this particular spectral dependence. However, questions still remain for the northern-most spur of LAB1. In this region we detect total linear polarization of between 3 and 20% at the 5% significance level. Simulations predict that polarization should increase with radius for a symmetric geometry. That the northern spur does not suggest either that this region is not symmetric (which is likely) and exhibits variations in column density, or that it is kinematically distinct from the rest of LAB1 and powered by another mechanism altogether.

A.1 Introduction

First discovered over a decade ago during the course of deep optical narrowband imaging (Francis et al., 1996; Steidel et al., 2000), Lyman- α “blobs” (LABs) are large, rare, gaseous nebulae in the high-redshift Universe detectable by their extensive Ly α luminosity. Found predominantly in regions of galaxy overdensities (Palunas et al., 2004; Matsuda et al., 2004; Prescott et al., 2008; Yang et al., 2009), these objects are some of the most promising candidates for the study of ongoing galaxy formation (Mori & Umemura, 2006). Displaying a range of sizes from tens to hundreds of kiloparsecs and luminosities spanning $\sim 10^{43-44}$ erg s $^{-1}$, LABs are reminiscent of high-redshift radio galaxies, yet most are not associated with strong radio sources (Saito et al., 2006). Instead, it seems that LABs are singularly associated with galaxies of one variety or another as even the famed Nilsson’s Blob (Nilsson et al., 2006), widely cited as the most overt example of a host-less LAB, is now believed to be associated with an AGN (Prescott et al., 2015). Other LABs have been associated with an assortment of galaxy populations including Lyman break galaxies (LBGs) (Matsuda et al., 2004), luminous infrared and submillimeter galaxies (SMGs) (Geach et al., 2005, 2007; Yang et al., 2012), unobscured and obscured quasars (QSOs) (Bunker et al., 2003; Weidinger et al., 2004;

Basu-Zych & Scharf, 2004; Smith et al., 2009), as well as starbursting galaxies (Scarlata et al., 2009; Colbert et al., 2011).

Though most LABs seem to have in common a host galaxy or galaxies, the debate over the powering mechanism of the extended Ly α emission remains unresolved in part due to the fact that many of these galaxies seem unable to produce sufficient ionizing flux to light up the surrounding medium (Matsuda et al., 2004; Smith & Jarvis, 2007). In addition to photoionization from luminous AGN and/or young stars as a power source (Haiman & Rees, 2001; Jimenez & Haiman, 2006; Geach et al., 2009; Cantalupo et al., 2012), other possible mechanisms include mechanical energy injected by supernovae winds during powerful starbursts (Taniguchi & Shioya, 2000; Scarlata et al., 2009), and radiative cooling (Haiman et al., 2000; Fardal et al., 2001; Dijkstra & Loeb, 2009; Faucher-Giguère et al., 2010; Rosdahl & Blaizot, 2012). In reality, it is more than likely that LABs are powered by multiple mechanisms simultaneously (Furlanetto et al., 2005). Theoretical studies have shown that polarization of Ly α photons can be induced by scattering thus providing a potential diagnostic to probe these various powering mechanisms (Lee & Ahn, 1998; Rybicki & Loeb, 1999; Loeb & Rybicki, 1999; Dijkstra & Loeb, 2008).

In particular, we focus our attention on a giant LAB (dubbed LAB1) in the SSA22 protocluster region first discovered by Steidel et al. (2000). This nebula is one of the most well-studied with observations ranging from optical to X-ray. LAB1 is known to be loosely associated with an LBG (C11, Steidel et al., 2000; Matsuda et al., 2004), though the peak Ly α surface brightness (SB) is more likely associated with an 850 μ m source (Geach et al., 2014) with a weak radio counterpart (Chapman et al., 2004) as well as associated detections in the near infrared (Geach et al., 2007), all suggestive of a dust-obscured star-forming galaxy leaking Ly α photons which interact with the surrounding medium. Deep integral-field spectroscopy of the Ly α emission has been presented by Weijmans et al. (2010) supporting this conclusion for the brightest regions of LAB1 though they suggest this mechanism is less promising in regions of lower Ly α SB. Additionally, McLinden et al. (2013) perform longslit NIR spectroscopy of portions of LAB1 detecting [OIII] emission in the LBGs C15 and C11 thus determining their systemic velocity. Hayes et al. (2011, hereafter H11) perform narrow-band imaging polarimetry and report low polarization in the central region rising to $P=11.9\pm2\%$ within a radius

of $7''$ (45 kpc physical). Coupled with tangential polarization vectors around the central region, they conclude their observations are consistent with powering from an obscured galaxy resulting in scattered Ly α photons by HI.

Due to the observational expense involved, polarization measurements of spatially extended Ly α emission have so far been attempted only three times. In addition to the work of H11, Prescott et al. (2011) present narrow-band imaging polarimetry of a LAB associated with a radio-quiet galaxy at $z=2.66$ though polarization was not detected. Humphrey et al. (2013) present the spectro-polarimetry of the gas surrounding the $z=2.34$ radio galaxy TXS 0211-122 and report low polarization centrally, rising to $P=16.4\pm4.6\%$ in some parts of the nebula and conclude that at least a portion of the nebula is powered by the scattering of Ly α photons produced by the galaxy within. In this paper we present a follow-up to H11 with the first spectropolarimetric measurement of a radio-quiet LAB. In §A.2 we discuss Ly α radiative transfer, scattering, and polarization basics. In §A.3 we discuss the observations and data analysis. Our methods and initial results are presented in §A.4, and in §A.5 we present a discussion of our results in the context of recent observational work. Finally, in §A.6 we discuss the future of polarization as a diagnostic tool in relation to upcoming space-based polarimeters.

A.2 Ly α Polarization Basics

Ly α polarization requires photons be scattered imbuing them with a preferential direction or impact angle. Localized (*in situ*) production of Ly α photons, either from stars or gas, is not expected to have significant polarization as these photons will either not scatter sufficiently or have no preferential orientation. In this section we briefly review the necessary physics behind generating a significant Ly α polarization signal.

The detection of signifiant polarization fraction of Ly α emission depends on two crucial factors: wing vs resonant scattering and Doppler boosting by thermal atoms in the surrounding medium. Ly α is the transition between the first excited and ground states of hydrogen and is a resonant transition, a doublet consisting of two fine-structure lines: $1S_{1/2} - 2P_{1/2}$ and $1S_{1/2} - 2P_{3/2}$. The latter transistion can exhibit polarization while the former cannot as scattering through this transition does not retain information

on the scattering angle thus producing an isotropized photon. Scattering “near” this doublet is called resonant or core scattering and has been shown to be a superposition of Rayleigh and isotropic scattering producing a minimum level of polarization (Brandt & Chamberlain, 1959; Brasken & Kyrola, 1998). In most astrophysical circumstances Ly α undergoes this type of scattering and the Ly α photons are repeatedly absorbed and re-emitted until they are either destroyed by dust or escape the surrounding neutral medium. However, thermal motions within the gas cause ‘partially’ coherent scattering where the absorbed and emitted photons are equal only in the rest-frame of the scattering atom. To the outside observer, the Ly α photons are Doppler boosted with respect to the scattering atom and thus perform a random walk in both frequency and physical space (Neufeld, 1990; Loeb & Rybicki, 1999). This can cause the Ly α photons to scatter in the wing of the profile where it has been shown that the phase function and degree of polarization are qualitatively consistent with pure Rayleigh scattering (Stenflo, 1980). Furthermore, Stenflo (1980) has shown that wing scattering can produce three times more polarization than resonant scattering. Thus, photons scattering in the wing of the profile are those which are most highly polarized and which see the lowest optical depth in the surrounding medium, enabling them to escape preferentially.

Theoretical predictions have been made by Dijkstra & Loeb (2008) for the expected amount of polarization in the Ly α line for various astrophysical situations. They explore both an expanding shell and a collapsing cloud. The expanding shell is a simple model of backscattering off a galactic outflow and predicts polarization to increase with radial distance from the central source with total Ly α polarization as high as 40%, depending on the assumed column density and velocity of the outflow. Such large values of polarization can be understood due to the kinematics of the gas. Photons scattering off the “back” of the expanding shell are quickly shifted out of resonance with the gas and into the wing of the line profile thus allowing many to escape after a single wing-scattering. Similar levels of total polarization ($p \sim 35\%$) are expected in the case of cooling radiation from a collapsing, optically thick gas cloud with the polarization again increasing as a function of radius from the central source due to photons emitted over a spatially extended region within the cloud. .

In both cases, detecting a high level of polarization through narrow-band imaging polarimetry would be able to rule out *in situ* production of Ly α photons. However,

imaging polarimetry alone can not distinguish between outflows or inflows as both predict similar levels of polarization and increasing polarization as a function of radius from the central source. Instead, the frequency dependence of Ly α polarization is required. For the case of an outflowing thin shell, Dijkstra & Loeb (2008) predict that Ly α polarization will increase redwards of the line center. This is because the redder Ly α photons appear farther from resonance in the frame of the gas and scatter less thus achieving higher levels of polarization. In fact, Dijkstra & Loeb (2008) state that this frequency dependence can be interpreted as a “fingerprint” for outflows and predict that Ly α polarization could be as high as $\sim 60\%$ in the reddest part of the line profile. In stark contrast, polarization increases blueward of line center for a collapsing cloud. Thus the frequency dependence of Ly α polarization can also constrain the kinematic structure of the surrounding gas.

A.3 Observations, Reduction, and Calculations

We now turn our attention to the spectro-polarimetry of LAB1. Hayes et al. (2011) present imaging polarimetry of this nebula in which they find significant polarization increasing as a function of radius from the point of brightest Ly α SB as calculated in Voronoi bins. Furthermore they find polarization vectors which lie tangentially around this central point and conclude that LAB1 is indeed powered by a bright central galaxy obscured from our line of sight. In this portion of the paper we present follow-up spectro-polarimetry in order to confirm and further probe the kinematics of this enigmatic object. In this section we discuss the observations and data reduction methods, as well as the polarization and error calculations performed.

A.3.1 Observations

We choose as our target one of the largest known Ly α blobs located in the SSA22 protocluster region at $z=3.09$ (see Steidel et al. (2000)). Dubbed LAB1, this object was observed over the course of five consecutive half nights from 5-9 October 2010, using the FOcal Reducer and low dispersion Spectrograph (FORS2) (Appenzeller et al., 1998) instrument mounted on the Antu (UT1) node of the Very Large Telescope (VLT) European Southern Observatory (ESO). The first stage of the dedicated dual-beam

polarization optics is the introduction of a strip mask designed to avoid overlapping on the CCD of the two beams of polarized light. Six MOS slitlets, each 1'' wide and 20'' long are then positioned over the objects of interest. The light is passed through a super-achromatic half-wave plate (HWP) retarder mosaic (**RETA2+5**), which rotates the angle of the polarized light. We adopt the standard four angles for unambiguous recovery of the Q and U Stokes parameters: 0°, 22.5°, 45°, and 67.5°. The rotated beam is subsequently passed through a Wollaston prism (**WOLL_34+13**), which splits the randomly polarized and unpolarized light into two orthogonal, linearly polarized outgoing beams, arbitrarily denoted the ‘ordinary’ (**ord**) and ‘extraordinary’ (**ext**) beams. Finally, the beams are passed through a grism dispersion element (**GRISM_1400V+18**) with a central wavelength of 5200 Å (\sim 1271 Å rest frame), spectral range of 4560-5860 Å, and dispersion of .63 Å/pixel. The spectral resolution is approximately 2100 with our 1'' slit width. This produces simultaneous **ord** and **ext** spectra which are then projected onto the CCD. Thus each frame consisted of six spectra: three slits contained only sky, two contained stars which were used in the alignment process (see below), and one was positioned over LAB1. We observed LAB1 at a position angle of -3.09° in the standard N=0°– E=90° reference frame.

Throughout the course of each night the sequence of four HWP position angles was observed repeatedly with two full sequences being completed each night. Integration times were 1,800 seconds for each HWP position with the exception of the first night where each exposure had 1,600 and 2,000 seconds of integration time per HWP position for the first and second sequences respectively. Thus at each retarder position angle we obtain a total integration time of 18,000 seconds.

The entirety of the observing time was classified as clear or photometric with no clouds present on any given night. Since the **ord** and **ext** beams are obtained simultaneously, deviations from photometricity would anyway have no impact on the determination of the Stokes parameters. Observations were taken around the new moon in order to minimize the sky background at bluer wavelengths with moonrise not occurring until after observations were complete each night. Because we observe LAB1 with constant position angle, atmospheric dispersion will vary with airmass over the course of the exposure time. Airmass ranged from 1.1 to 1.53 and was thus well within the FORS instrument’s atmospheric dispersion corrector to compensate. The bulk of the

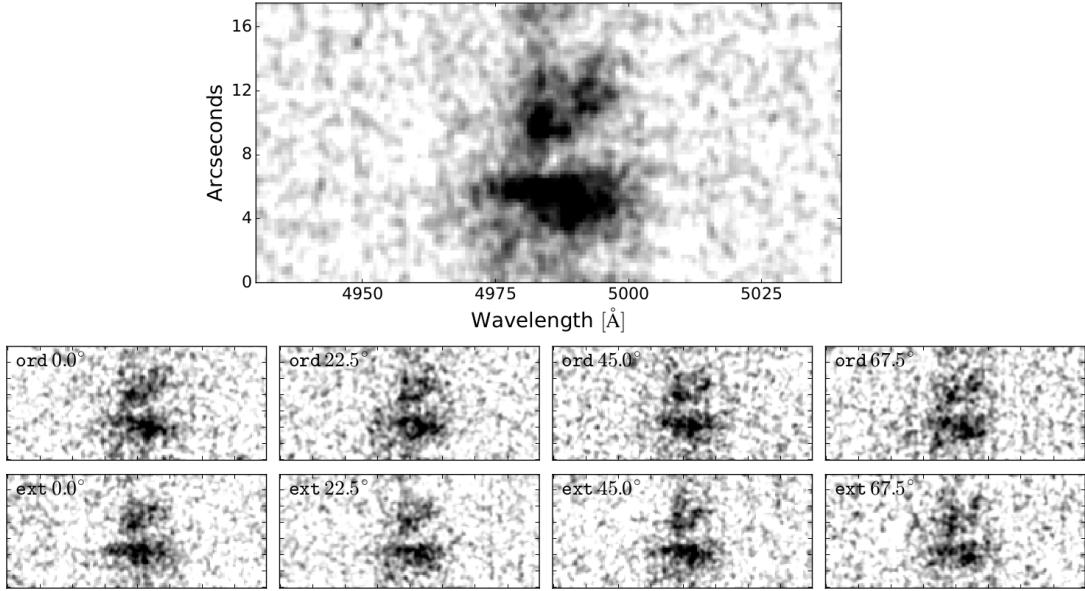


Figure A.1 *Top.* Spatial and wavelength distribution of the master total intensity Ly α spectrum. Co-added Ly α spectrum of 77 science spectra smoothed by a Gaussian with FWHM = 0''.5. Due to light loss at the edges of the slit, we present here only the central 18''. *Bottom.* Ordinary and extraordinary beams at each HWP angle which provide the basis for our polarization measurements. Each frame consists of several co-added frames taken over the 5 nights of observations.

20 hours of observation time experienced astronomical seeing which varied between 0.5 and 1.0 arcsecond with median seeing at $\sim 0.^{\prime\prime}75$. There were three observations which experienced seeing as high as 1.7''. These were excluded from the following analysis although their inclusion does not significantly alter our results – polarization fractions varied only by a few per cent. We thus obtain a total of 37 individual observations for a total of 12,600 seconds of integration.

A.3.2 Data Reduction

The initial steps of data reduction were carried out in the standard manner for spectroscopic observations. Individual frames were biased subtracted. Master flat frames were created from several dome flats using EsoRex¹, an ESO recipe execution tool,

¹ <http://www.eso.org/sci/software/cpl/esorex.html>

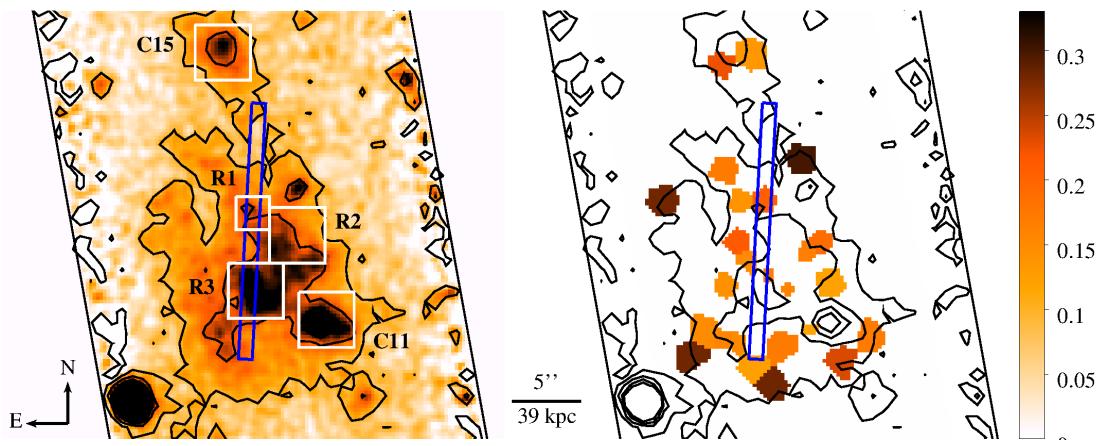


Figure A.2 Slit position over LAB1 as compared with results from H11. In the left panel we show the combined Ly α intensity frame from H11 adaptively smoothed to show detail including emission from LAB1 and nearby LBGs C11 and C15. Overlaid in white are boxes from Weijmans et al. (2010), regions in which they performed integral-field spectroscopy on the Ly α emission. As before, the slit shown here spans $\sim 18''$. Contours denote arbitrary flux levels. In the right panel we show fractional polarization results from H11 for those bins which were detected at or above 2σ (See H11 Fig. 2, panel e). Our slit passes over the region of brightest Ly α emission in the southern portion of the slit as well as a dimmer region in the north.

and applied to individual frames to correct for pixel-to-pixel variations. Each frame was then normalized by exposure time. Cosmic rays were thoroughly removed using L.A. Cosmic² (van Dokkum, 2001). To account for small variations in the spatial direction during observing, frames were aligned using the `shift_sub` function in IDL where the pixel shift was calculated as the average of fitted Gaussians of each stellar continua in the slits above and below the science spectra. At this stage all the individual frames were split into their `ord` and `ext` beams (37 observations \times 2 beams = 74 spectra). Only slits 3 (sky) and 4 (LAB1) were considered for the remainder of the analysis.

Each beam of the sky and LAB1 spectra was wavelength calibrated individually via a He-Ar arc lamp spectrum using NOAO/IRAF³ `onedspec` and `twodspec` packages. The `identify - reidentify - fitcoords - transform` sequence was used on the 2D spectra yielding a fit r.m.s. typically between 0.05 and 0.09 Å.

Sky subtraction was performed using the sky spectra in slit 3. For each `ord` and `ext` beam, the sky spectrum was median extracted, normalized to the spatial dimension of the 2D LAB1 spectra, and subtracted from the corresponding LAB1 beam. The residual sky background in the wavelength direction was modeled with a linear fit after masking the Ly α line. This was then subtracted from the spectra. This step was appropriate as there is no evidence of UV continuum according to our preliminary analysis of recently acquired MUSE data which will be published in Hayes et al., in prep. Atmospheric correction was applied using an extinction coefficient as a function of wavelength obtained from Patat et al. (2011) and the airmass at the midpoint of each observation. Spectra were then co-added using a mean combination with simple `minmax` rejection of the highest and lowest value at each pixel using IRAF's `imcombine`. As mentioned previously, we exclude 3 observations from further analysis. These include one 45° observation and two 67.5° observations whose delivered seeing was well above 1.0''.

We combine eight groups of 10 spectra to produce science frames for polarimetry measurements: `ord` and `ext` at each of the four angles, as shown in the bottom panel of Figure A.1. It is immediately obvious that there are variations between the `ord` and

² <http://www.astro.yale.edu/dokkum/lacosmic/>

³ IRAF is distributed by the National Optical Astronomy Observatories, which are operated by the Association of Universities for Research in Astronomy, Inc., under cooperative agreement with the National Science Foundation

`ext` beams at each HWP angle which fundamentally leads to a measurement of the polarization. Finally, a master total intensity frame is created by averaging these eight frames together, as shown in the top panel of Figure A.1.

A.3.3 Polarization and Error Calculations

Polarization of Ly α is expected to be linear, thus the decomposition of polarized light falls only into Q and U normalized Stokes parameters. The V parameter represents circular polarization and is not expected for Ly α radiation. The fourth parameter I is the total intensity which is equal to the sum of the `ord` and `ext` beams. For each HWP position, θ , the normalized flux difference, F_θ , is defined as:

$$F_\theta = \frac{f_\theta^{ord} - f_\theta^{ext}}{f_\theta^{ord} + f_\theta^{ext}} \quad (\text{A.1})$$

where f_θ^{ord} is the flux in the ordinary beam for a given θ and likewise of f_θ^{ext} for the extraordinary beam.

Once the four HWP angles have been obtained, Q , U , and I relate to the observables by

$$\begin{aligned} q &= \frac{Q}{I} = \frac{1}{2}F_{0.0} - \frac{1}{2}F_{45.0} \\ u &= \frac{U}{I} = \frac{1}{2}F_{22.5} - \frac{1}{2}F_{67.5} \end{aligned} \quad (\text{A.2})$$

From these, the polarization fraction, p , and the polarization angle, χ , can be calculated by

$$\begin{aligned} p &= \sqrt{q^2 + u^2} \\ \chi &= \frac{1}{2} \arctan \frac{u}{q} \end{aligned} \quad (\text{A.3})$$

However, we actually desire an estimate of the “true” polarization, p_0 . When it is assumed that the Stokes parameters q and u are drawn from Gaussian distributions centered around the true values (q_0 and u_0) each with variance σ , it can be shown (e.g. Plaszczynski et al., 2014) that the distribution of the polarization follows the Rice distribution:

$$f_p(p) = \frac{p}{\sigma^2} e^{-\frac{p^2+p_0^2}{2\sigma^2}} I_0\left(\frac{pp_0}{\sigma^2}\right) \quad (\text{A.4})$$

where p_0 is the true amplitude of the polarization and I_0 is the modified Bessel function of the zeroth order. Equation A.3 is considered the naive estimator for this distribution

and is known to be strongly biased at low polarization signal-to-noise ratio (SNR_p) in part because, in this regime, the Rice distribution can be approximated as a Rayleigh distribution which is highly skewed to larger values of polarization. Additionally, the naive estimator cannot take experimental noise into account. Several attempts have been made to produce an unbiased estimator for p_0 (see Simmons & Stewart, 1985, for a review) but most have other undesirable qualities such as being unphysical at very low signal-to-noise or containing discontinuities. Plaszczynski et al. (2014) develop a polarization estimator dubbed the Modified Asymptotic Estimator (MAS) which is less biased in both low (Rayleigh) and high (Gaussian) SNR_p regimes and is continuous between these regions. Furthermore, this estimator also takes into account measurement error in q and u . For these reasons we adopt this estimator for the polarization which goes as:

$$\hat{p}_{MAS} = p_i - b_i^2 \frac{1 - \exp^{-p_i^2/b_i^2}}{2p_i} \quad (\text{A.5})$$

where p_i is given by Equation A.3 and b^2 is the noise bias of the estimator given by

$$b_i^2 = \frac{q_i^2 \sigma_u^2 + u_i^2 \sigma_q^2}{q_i^2 + u_i^2} \quad (\text{A.6})$$

where each i is an individual binned measurement of the quantity of interest.

In practice, the quantities p_i , q_i and u_i are calculated as given in equations A.1, A.2, and A.3 in each bin (see below for a discussion on our binning strategy). The σ_q and σ_u are computed from Monte Carlo simulations whereby each of the eight science frames is allowed to deviate according to a Gaussian spread wherein the deviates are simply computed as the standard deviation of a large portion of the background of each individual science `ord` and `ext` frame. Ten thousand realizations are performed and for each realization q and u are computed. The resulting probability distributions of q and u are Gaussian as expected and from these we obtain σ_q and σ_u by measuring the spread of the distributions.

We consider the polarization SNR_p by computing

$$\text{SNR}_p = \frac{p_{MAS}}{\sqrt{\frac{1}{2}(\sigma_q^2 + \sigma_u^2)}}, \quad (\text{A.7})$$

where again we allow for measurement error by incorporating both σ_q and σ_u . Following the prescription of Plaszczynski et al. (2014), values of $\text{SNR}_p > 3.8$ indicate that the Rice

distribution is sufficiently Gaussian and thus unbiased. In this regime one may compute a point estimate along with the estimator variance. Values less than this, however, fall in the Rayleigh regime wherein one must instead rely on confidence intervals (CIs). As we show below, different binning techniques yield different SNR_p and thus in some cases we report point estimates of the fractional polarization while in others we provide the 95% CIs according to Eqn. 26 in Plaszczynski et al. (2014).

Finally, we consider the measurement and error estimates for the polarization angle. It has been shown (e.g. Wardle & Kronberg, 1974; Vinokur, 1965) that the distribution of χ is symmetric about the true value of the angle and thus the estimator presented in Equation A.3 is already unbiased. At large SNR_p this distribution also tends towards a Gaussian with a standard deviation of $\sigma_\chi \approx \sigma_p/2p$. However, at low SNR_p this approximation underestimates the error. In this work we follow Wardle & Kronberg (1974) and approximate the error by their Eqn. A6 (see also their Figure 3) which provides the most conservative error estimate for measurements with $\text{SNR}_p > 0.5$.

A.4 Polarization of Ly α in LAB1

A.4.1 Polarization integrated over the line profile

Because our data have low signal-to-noise per pixel ($\text{SNR}_p \lesssim 1$), binning of the science frames is a necessity. However, any Ly α polarization signal will result from the particular geometry inherent in the HI gas with a unique set of Stokes parameters and polarization angle. If these regions are not azimuthally resolved, one risks overlapping each region's polarization angles thus averaging the polarization signal and potentially washing it out entirely (Dijkstra & Loeb, 2008). Thus, some binning is necessary but overbinning will make it unmeasurable.

To aid in the determination of appropriate bins we examine the slit position over LAB1 as shown in Figure A.2. LAB1 fills the slit and contains regions of varying Ly α surface brightness (SB) as denoted by the set of arbitrary contours. Also shown in this figure are white boxes corresponding to Ly α emission integral-field spectroscopy as presented by Weijmans et al. (2010). Our slit overlaps their regions **R1** and **R3** and we adopt this nomenclature throughout. **R3** is situated over the brightest peak of the Ly α SB while **R1** is associated with a somewhat dimmer region. Between these two

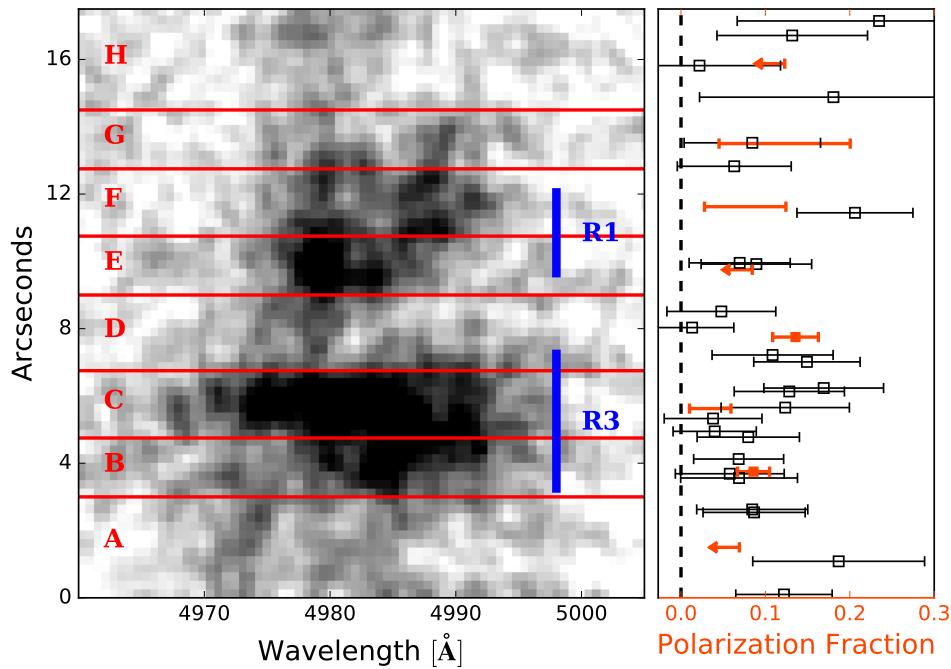


Figure A.3 Polarization of spectrally integrated Ly α . In the left panel we show the spatial binning of the Ly α emission spectrum with bins ranging from 2–3'' overlaid in red. Ly α emission is integrated over 4965–5000 Å. The blue lines indicate the spatial extent of Weijmans et al. (2010) IFU boxes. Depicted in the right panel are our polarization measurements in orange. Polarization point estimates are denoted by orange boxes with 1σ error bars; 95% CIs are shown as closed brackets; and upper limits are denoted with orange arrows where we define our upper limits as the upper 95% confidence bound for that spatial bin. Black squares show fractional polarization from H11 as measured in Voronoi bins which overlap with our slit. See text for discussion on differences and limitations between datasets.

Table A.1 Polarization signal-to-noise and fractional polarization measurements for spatial bins. For those bins with sufficient SNR_p , we report the polarization point estimate and corresponding $1\sigma_p$ error. Otherwise we report 95% confidence intervals for bins in which the lower 95% confidence bound is greater than zero.

Bin	SNR_p	p_{min}	p_{max}	p	σ_p
B	4.6			8.59	1.9
C	2.6	1.00	5.92		
D	4.9			13.6	2.7
F	2.8	2.78	12.4		
G	2.8	4.51	20.0		

features there exists a distinct gap that can also clearly be seen in the Ly α emission shown in Figure A.1. We determine to bin these regions separately as they can exhibit different polarization fractions as shown in the right panel of Figure A.2. Altogether, we bin the slit into 8 individual spatial regions, each spanning $2''$ – $3''$, as this is large enough to achieve adequate SNR_p in some bins yet small enough that we do not wash out any polarization signal. These spatial elements are labelled **A** through **H**, with **A** being the southern-most portion of the slit and **H** the northern-most. These spatial bins are shown explicitly in Figure A.3.

With these considerations in mind we first spectrally integrate over the Ly α emission to calculate the total p for comparison with H11. Integration is carried out over the wavelength range 4965–5000 Å. We note that though the range of the Ly α emission varies within each aperture, varying the integration range only changes the fractional polarization by a few percent difference for all but bin **H** which has an increase in p of 10%. However, since we are unable to place reasonable constraints on the polarization fraction in this bin we consider this to be moot. Only two of these spatial bins have $\text{SNR}_p > 3.8$ and for these we report the measured polarization and $1\sigma_p$ error. The remaining bins have $\text{SNR}_p < 3.8$ and for these we report 95% CIs for those bins in which the 95% lower confidence bound is greater than zero. Our results are summarized in Table 1 as well as in Figure A.3 along with the spatial binning pattern and the polarization measurements from H11 for those bins which our slit overlapped. The fractional polarization values roughly agree when one takes into account the distinct methods used between our two analyses. H11 utilize a Voronoi binning technique whereby the size of

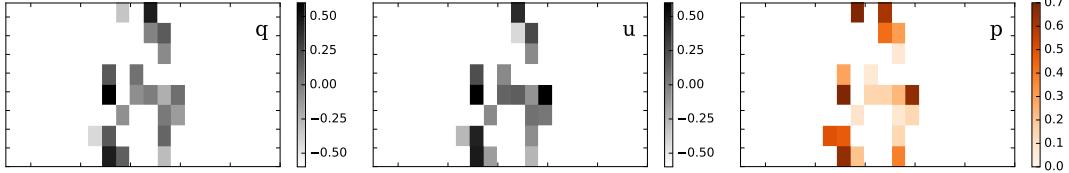


Figure A.4 2D Stokes parameters and polarization maps. The q and u maps are computed directly from the science frames shown in the bottom panel of Figure A.1 according to the prescription of Equation A.2 after binning as described in subsection A.4.2. The polarization map in the right panel is then computed via q and u according to Equation A.5. In all frames, only those bins are shown in which the polarization was deemed significant as discussed in subsection A.3.3. The variation between the q and u maps is readily seen by eye and is directly related to the amplitude of the polarization measured in that bin.

each bin is determined by the achieved SNR within that bin. This allows them to have bins of various sizes. Each of their bins only partially overlaps our slit and we include in Figure A.3 all such bins. The largest disparity between datasets occurs in Bin F. In this region, H11 measure the fractional polarization to be $p \sim 20\%$ whereas we find, at most, $\sim 12\%$. The reason for this discrepancy is not fully understood. In Table A.4.1 we report the 95% confidence intervals for those spatial bins whose lower 95% confidence bound is greater than zero. We see polarization in spatial bin **C** which is on the order of a few per-cent though not quite consistent with zero. The fractional polarization then increases to 13.6% north and 8.6% south of this region as seen in spatial bins **B** and **D**. The distance between bin **C** and these bins is roughly 15 kpc in either direction.

A.4.2 Polarization across the line profile

We next explore p as a function of wavelength. Using the same spatial elements, we further bin each into 5 Å increments in the wavelength direction and compute q and u as shown in Figure A.4. In this figure we show only those bins in which we detect a significant polarization signal. One can see by eye the differences in the q and u frames which is directly responsible for the strength of the polarization shown in the third panel.

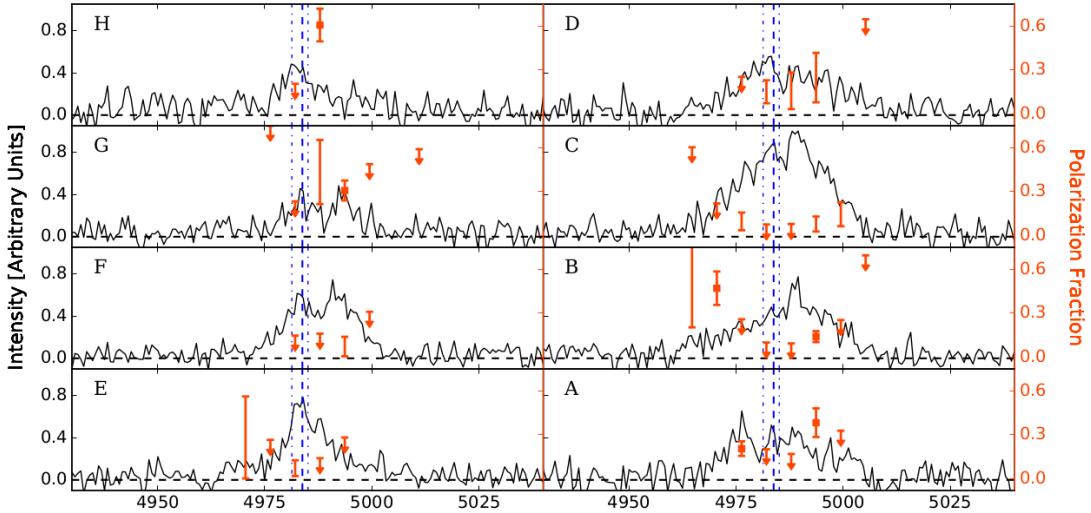


Figure A.5 Polarization of Ly α as a function of wavelength. Figures **A** through **H** contain the extracted 1D Ly α spectrum of the corresponding aperture from Figure A.3. All spectra have been scaled to show their relative intensity. Overplotted in orange we show the polarization fraction as a function of wavelength in bins of 5Å. Polarization point estimates are denoted by orange boxes with 1σ error bars; 95% CIs are shown as closed brackets; and upper limits are denoted with orange arrows where we define our upper limits as the upper 95% confidence bound for that bin. The dashed blue line represents the average systemic velocity as measured from [O III] of four galaxies which are associated with LAB1. The dash-dotted lines are the minimum and maximum of those objects. In bins **B-E**, we see a trend of low polarization associated with the core of the Ly α emission and higher polarization toward the wings of the line profile. This trend is less apparent in other bins, though the line profiles are not as well defined and many display a double peak. In general, p is typically highest at those wavelengths in which the Ly α intensity is relatively low.

In Figure A.5 we present the extracted 1D spectra for each spatial element along with the wavelength response of p in orange. For those bins with sufficient SNR_p (as shown in the middle panel of Figure A.6), we show the polarization point estimate along with $1\sigma_p$ errors. For those bins which have lower 95% confidence bounds greater than zero, we show the CI as orange closed brackets. Otherwise, we show upper limits defined as the upper 95% confidence bound for that bin. In general we see a trend of high (low) polarization corresponding to lower (higher) relative Ly α intensity. In particular, bin **B** displays p which is consistent with zero in near 4985Å but which rises substantially in the wings of the profile, reaching up to 45% bluewards and with upper limits as high as 65% redwards. In spatial bins **C** and **D** we see the suggestion of similar behavior with lower polarization in the core of the line and potentially higher polarization in the wings of the profile though the data do not allow us to further constrain the trend.

It is important to recall that these measurements inform us as to the fraction of the total intensity which is polarized. In the case of box **B**, for example, the wavelength bin at 4970Å is 45% polarized. The intensity in this portion of the line is quite low, however, relative to the peak at 4990Å. It is instructive to compare this to the integrated polarization in Figure A.3. In that figure, box **B** has fractional polarization of $\sim 9\%$. Thus we see that spectropolarimetry gives us more information than what can be gained solely through imaging or integrated polarimetry. Highly polarized individual wavelength bins are ‘washed out’ by the relative strength of the core of the profile which is typically not strongly polarized. The overall fraction of polarized photons decreases when integrating over the entire line and it is impossible to reconstruct from imaging alone which wavelength regimes are the most highly polarized.

In Figure A.6 we present 2D maps of the spectrally binned intensity, SNR_p , and polarization of LAB1. In the middle panel we show the SNR_p where we stress that the “noise” in this equation is not the equivalent of a polarization error. This figure instead serves to give the reader a feeling for the relative quality of our measurements. Comparing the middle and top panels we see that many areas of high intensity have very low polarization signal-to-noise indicating that these regions most likely have very low or no fractional polarization for us to detect with the current data. However, portions of the spectrum exhibiting relatively less intensity have much more significant SNR_p . The fraction of polarized light in these bins is relatively more substantial. We also point out

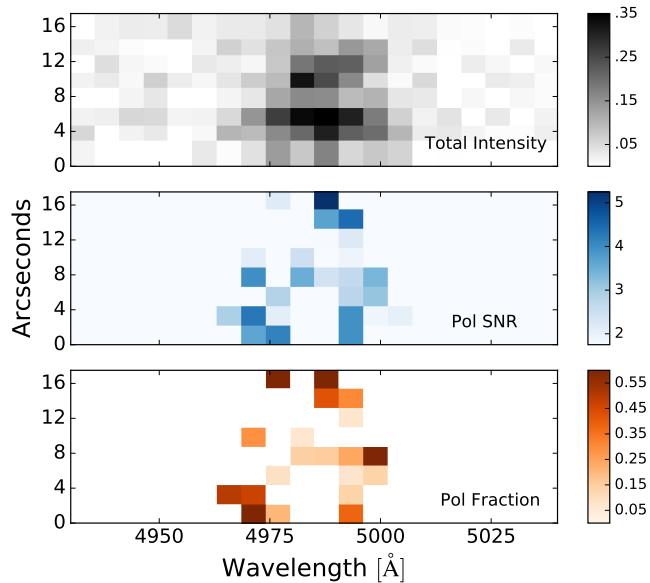


Figure A.6 *Top*. Total intensity of the 2D Ly α spectrum in bins of 5\AA by $2\text{-}3''$. *Middle*. SNR_p map which demonstrates the relative quality of our polarization measurements. $\text{SNR}_p > 2$ is generally enough for us to discriminate p statistically significant from zero with 95% confidence and we show CIs for such bins in Figure A.5. $\text{SNR}_p > 3.8$ indicates the Rice distribution is sufficiently Gaussian and we measure polarization with standard errors, also shown in Figure A.5. *Bottom*. Map of p determined to be statistically significant from zero with 95% confidence (lower 95% confidence bounds greater than zero). For those bins with sufficient SNR_p , the bin color reflects the point estimate of p , otherwise it represents the middle value of the CI associated with that bin.

that in this SNR_p map we see a range of values from ~ 2 to ~ 5 which indicates that for some bins we report point estimates of the fractional polarization while for others we instead provide 95% CIs. In the bottom panel, we show p in individual bins in which we either have sufficiently high SNR_p or in which we calculate a lower 95% confidence bound greater than zero. In general we see higher values of p in the reddest part of the line though we also note some substantial polarization in the blue wing as well. In most cases we see that the central emission region is characterized by low SNR_p and low fractional polarization.

Finally, in Figure A.7 we present the direction of the polarization vectors, χ , for those spatial elements (**B-D**, **F**, **G**) which have $\text{SNR}_p > 2$. Errors on the polarization angles range from $\sim 6\text{--}15^\circ$. In this figure we also plot polarization vectors from H11 for comparison, including only those which they measured at or above 2σ . We see that both sets are generally consistent. Like H11, our angles lie tangentially around the peak of Ly α SB.

A.5 Discussion and Conclusions

We have presented deep spectro-polarimetry of the LAB1 Ly α emission nebula. The data allow us to probe the kinematics and distribution of the neutral gas and reinforce the idea that LAB1 is likely composed of several smaller, more complex regions instead of one large, kinematic structure. In particular, our observations suggest at least a weak outflow in the southern portion of the LAB as we discuss below.

Simulations predict that polarization due to scattering should exhibit a radial dependence on the sky. The region of highest Ly α SB would not be strongly polarized but the polarization would rise with increasing radius (Dijkstra & Loeb, 2008). The observed polarization of Ly α in the southern portion of the slit is consistent with Ly α photons produced by a luminous galaxy (or galaxies) and scattered at large radii by the surrounding neutral hydrogen. As in H11, we see this signature here, most notably in spatial elements **B-D** where the peak of the Ly α SB has little observable polarization as shown in spatial element **C**. North of this location (**D**), we find $p = 13.6 \pm 2.7\%$ which is also consistent with the Voronoi bins from H11 lying on either side of our slit at approximately the same radial distance (see the right side of Figure A.2). Similarly

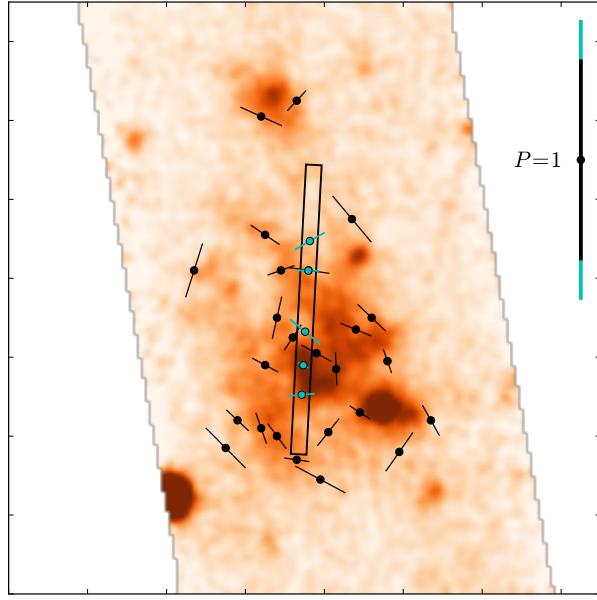


Figure A.7 Direction of Ly α polarization vectors. Smoothed total Ly α intensity from H11 overlaid with polarization vectors from H11 (black) and from this analysis (cyan) for those spatial bins from Figure A.3 which have $\text{SNR}_p > 2$. Because we measure small polarization amplitude in our bins we use a larger scaling for our vectors in order for the reader to more easily compare the angles we measure with those presented in H11. Both sets of angles generally lie tangentially about the region of highest Ly α SB.

to the south (**B**), we find total $p = 8.59 \pm 1.9\%$. Thus we have increasing polarization away from the peak Ly α emission with a radius of ~ 15 kpc. However, this general trend cannot determine between inflowing or outflowing gas as both are predicted to have this observational signature.

Dijkstra & Loeb (2008) also predict that, for an envelope of expanding gas, those photons in the bulk of the line profile should exhibit increasing polarization redward of the line core. The strength of this increase depends strongly on the distance from the center of Ly α SB as well as the column density and outflow velocity. In other words, for a given column density and outflow speed, polarization will increase weakly (up to $\sim 10\%$) in the reddest part of the wing at the peak of Ly α SB, but will increase substantially (up to $\sim 70\%$) at larger radii from this region. We caution that direct comparison of our data with these simulations comes with a caveat since we are not integrating over

the entire shell with our long-slit observations. Nevertheless, the spectrally resolved polarization in bins **C** and **B** at least suggest this behavior.

To see this trend, we first estimate the systemic velocity of the region as the average from four galaxies measured in [O III] and known to be components of LAB1 (Kubo et al., 2015; McLinden et al., 2013) (see Figure A.5). We note that these measurements are all within ~ 230 km/s. Given the width of the Ly α profile, this uncertainty has minimal impact. In this context we can see that the polarization in **C** is well constrained in the red wing to be at most $\sim 20\%$ polarized. Though we are unable to tightly constrain the red wing in bin **B**, large polarization values are suggested by the detection at 4995Å, and are not ruled out at longer wavelengths. This particular polarization pattern can be explained easily with a weak outflowing shell model whereby Ly α photons emitted from an embedded source interact with the expanding shell. Those photons which interact with the receding portion of the shell (from our point of view) are Doppler shifted into the red wing of the profile. Being in the wing of the line profile, these photons see a lower optical depth and thus preferentially escape the medium having only scattered a few times which preserves their polarization. Photons which remain in the core of the profile in the frame of the gas scatter many more times, effectively erasing any polarization signal.

The polarization data presented here provide a framework for future observations of the southern portion of LAB1. The imaging polarimetry of H11 coupled with a recent strong submillimeter source within 1.5'' of the peak Ly α intensity (Geach et al., 2014) provide the ‘smoking gun’ that there is indeed at least one powerful source embedded in this region, the photons from which are likely scattering at large radii. If there are indeed multiple sources within about 30 kpc of this region, observations must also provide a mechanism by which these sources can reproduce the spectral polarization signature presented here, namely, a configuration which is consistent with low polarization in the core of the Ly α profile and high polarization in the wings.

However, the picture remains obscure for **R1** corresponding to our spatial elements **E-F**. Were **R1** to be part of the same smooth, kinematic structure as **R3** we would expect its total Ly α polarization to increase relative to **R3** due to its increased radius from the galactic center. Instead, we see the total polarization drop and flatten across **R1**. It’s possible that the gas in this region is clumpier or denser than the southern

portion of the blob. An increase in the column density of the gas will decrease the observed polarization fraction as additional scatterings tend to isotropize the photons. Another possibility is that this region is powered by fluorescence from ionizing radiation emanating from the central source. This would naturally explain the lower polarization as this type of *in situ* production of photons is not expected to be highly polarized. Weijmans et al. (2010) present compelling evidence that suggests this region is kinematically distinct from the rest of LAB1 and thus a third possibility is that this region is instead powered by radiative cooling. Most likely is the possibility that this region is dominated by an embedded source of its own. Though interesting to speculate, the wavelength dependence of the polarization in these spatial bins is not sufficient for us to further probe the kinematics and polarization properties.

A.6 The Future of Ly α Polarization

With another successful detection of the spectral dependence of Ly α polarization the question arises: What does the future hold for Ly α polarization? Additionally, should emphasis be placed on imaging or spectral polarimetry? The integration times for either mode are similar in magnitude and require a substantial commitment so the choice between methods is not a trivial one.

We have explicitly demonstrated that much information can be gleaned from the spectral dependence of the polarization signal. In particular, features emerge which narrowband imaging polarimetry simply cannot detect. Not only are we able to detect the wavelength dependence for many of our spatial bins but we also find high polarization upwards of 60% in portions of the Ly α profiles – information which is completely lost in imaging polarimetry. While imaging polarimetry can confirm the presence of scattering and probe the overall geometry of the scattering medium, it cannot probe the kinematics of the system to determine potential outflows or inflows. Furthermore, the spectrally integrated polarization can still provide spatial clues as to any existing radial dependence with advantageous slit placement. Because the geometry of the blob is important to the overall detection of a polarization signal, spectropolarimetry should not be conducted blindly but instead be guided by spatial information obtained from the already existing narrowband surveys of LABs as well as IFUs.

Though it remains to be seen, suggestions have arisen that the next generation of ~ 30 m telescopes could extend Ly α polarization studies. This is supported in that all projected Extremely Large Telescopes (ELT) have proposed polarimetry as a necessary part of their instrument suite. On the E-ELT, the *Exo-Planet Imaging Camera and Spectrograph* (*EPICS*, Kasper et al., 2008) includes the *EPOL* polarimeter (Keller et al., 2010). The Thirty Meter Telescope has discussed plans to include the *Second-Earth Imager for TMT* (Matsuo & Tamura, 2010). And the Giant Magellan Telescope has discussed spectro-polarimetric capabilities as necessary to meeting their science goals (GMTO Corporation, 2012). Off the ground, several probe-scale NASA space missions have been proposed to study exoplanetary systems such as the AFTA and “EXO” missions (Stapelfeldt et al., 2014; Seager et al., 2014), with considerable emphasis given to polarimeters to enrich the science output. While many of these projects focus on imaging polarimetry, the UVMag consortium has proposed the Arago space mission which would be devoted to unprecedented spectropolarimetry from the FUV through the NIR (Pertenais et al., 2014). This field is currently driven almost exclusively by exo-planetary science for studying the scattering off planetary atmospheres and circumstellar disks but it is the development of such instrumentation which is of greatest importance. At this stage, we cannot say whether we will be able to point one of these instruments directly towards a Ly α blob without slight modification of the initial design or incorporation of additional settings but it is optically plausible (Hayes & Scarlata, 2011).

In the meantime, much can still be accomplished from the ground with 8 m class telescopes. To date, only three Ly α emitting sources have been studied in depth: LAB1 (Hayes et al., 2011, and this work), LABd05 (Prescott et al., 2011) and radio galaxy TXS 0211–122, known to be associated with a 100 kpc scale Ly α nebula (Humphrey et al., 2013). Radio-quiet LABs were first targeted due to their apparently controversial nature unlike high redshift radio galaxies (HzRGs) which did not pose an energy problem. With the discovery that a HzRG is at least partially polarized due to Ly α scattering, it behooves us to test further what relationship, if any, exists between radio-loud and -quiet nebulae. Compact Ly α sources also remain unexplored with the polarimeter. Though targeting resolved objects ensures that the Stokes parameters do not all cancel, symmetry is likely broken in most systems. Thus we may expect a measureable signal

from LAEs (Lee & Ahn, 1998).

Additionally, the interpretation of Ly α polarization is still a challenging prospect of its own. Most state-of-the-art simulations assume density and kinematic structures that are still unrealistic in that variations proceed smoothly. What is urgently needed is the implementation of Ly α polarization in all Ly α radiative transport codes to generate predictions for various applications including clumpy and filamentary media as well as non-spherically symmetric geometries. While some work has already been done in this area (Dijkstra & Kramer, 2012), there is still much to be explored in terms of predicting observable polarized Ly α line profiles. With the current limitation on instrumentation coupled with exacting observations, we need dedicated theoretical and observational developments that proceed in tandem.

We thank the anonymous referee for the useful comments that significantly improved the analysis and presentation of our results. Additionally, MB and CS are grateful to Jérémie Blaizot and Maxime Trebitsch for proofing the manuscript and providing feedback which helped to clarify the text.