Integrating Human and Machine Intelligence in Galaxy Morphology Classification Tasks

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

December, 2017

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.

First and foremost, I *literally* would not be where I am today without my advisor, Claudia Scarlata. She unwittingly saved me from a life of drudging Statistics by asking me to join her research group moments before I was to accept the department transfer, and for this I am eternally grateful. I have the utmost respect for you as a researcher, a leader, a scientist, and a collaborator. Thank you for helping me gain the confidence in my abilities and myself without resorting to bullshit. Thank you for pushing me to finish what I start, as much as I may loathe it. And thank you for sticking with me even during our rockier moments. I would be honored to consider you not just my advisor and collaborator, but also my friend.

For the past three years I've had the delightful fortune to be part of the Galaxy Zoo science team and the insight and guidance provided by collaborators Chris Lintott, Brooke Simmons, Karen Masters, Steven Bamford, and Becky Smethurst has been invaluable. Particular thanks go to Lucy Fortson as my unofficial co-adivsor. Thank you for your steady support, even hand, cool head, and constant (but gentle!) pressure. You and Claudia make an amazing team and I am so grateful to have been part of it.

Of course I am also indebted to the support provided by those within MIfA. Many thanks to Hugh Dickinson, John Phillips, Kyle Willett (ex-MIfA), Evan Skillman, and Terry J. Jones. Discussions with each of you helped shaped my research and my direction through graduate school in myriad, subtle ways. Finally, thanks to my additional committee members, Tom Jones and Yong Qian, for your expertise and time commitment.

Obviously, no grad student is an island and I am so thankful that the grads in our

program have always cared more for inclusion and camaraderie over advancement and competition. Thanks to all of you for the experiences that make grad school bearable: game nights, D&D campaigns, roof beers, and camping trips. Thanks to my office mates Miceala Bagley and Vihang Mehta. You two have taught me more things and listened to me bitch more times than I can count. Thanks to Michael Gordon for Computer Club, for teaching me all the Pythons, and for daily hilarious stories. Thanks to my dancing buddy, Brian O'Neill; to my Dranks Club: Tony Young, Matt Gomer, and Trevor Knuth; to the best traveling partner one could hope for and without whom I could not be Melanie^2, Dr. Melanie Galloway; to Evan Tyler for relieving me of Outreach Coordinator duties (sorry, bro), and to Karlen Shahinyan. If you wait for Karlen, you die; but look who graduated first.

Last but certainly not least, I couldn't have done this without all the support and love from my friends and family. Mom, thanks for letting me vent every Saturday morning. You never failed to cheer me on. And Max, thanks for helping me see (and believe!) what my skills are truly worth. PhD or GTFO.

This work could not have been completed without the financial support of several institutions. In particular I am extremely grateful to the University of Minnesota Graduate School for both the Doctoral Dissertation Fellowship and the Thesis Research Travel Grant. Additionally, I am eternally indebted to the University of Oxford and New College, Oxford for the opportunity to conduct visiting research through the Balzan Fellowship.

Chapters 3 and 4 of this thesis are modified versions of a paper submitted to MNRAS. For flow and clarity, the acknowledgments for that paper are placed below.

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.

Dedication

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

Abstract

The large flood of data flowing from observatories presents significant challenges to astronomy and cosmology – challenges that will only be magnified by projects currently under development. Growth in both volume and velocity of astrophysics data is accelerating: whereas the Sloan Digital Sky Survey (SDSS) has produced 60 terabytes of data in the last decade, the upcoming Large Synoptic Survey Telescope (LSST) plans to register 30 terabytes per night starting in the year 2020. Additionally, the *Euclid* Mission will acquire imaging for $\sim 5 \times 10^7$ resolvable galaxies. The field of galaxy evolution faces a particularly challenging future as complete understanding often cannot be reached without analysis of detailed morphological galaxy features. Historically, morphological analysis has relied on visual classification by astronomers, accessing the human brains capacity for advanced pattern recognition. However, this accurate but inefficient method falters when confronted with many thousands (or millions) of images. In the SDSS era, efforts to automate morphological classifications of galaxies (e.g., Conselice et al., 2000; Lotz et al., 2004) are reasonably successful and can distinguish between elliptical and disk-dominated galaxies with accuracies of $\sim$80%. While this is statistically very useful, a key problem with these methods is that they often cannot say which 80% of their samples are accurate. Furthermore, when confronted with the more complex task of identifying key substructure within galaxies, automated classification algorithms begin to fail.

The Galaxy Zoo project uses a highly innovative approach to solving the scalability problem of visual classification. Displaying images of SDSS galaxies to volunteers via a simple and engaging web interface, www.galaxyzoo.org asks people to classify images by eye. Within the first year hundreds of thousands of members of the general public had classified each of the $\sim$1 million SDSS galaxies an average of 40 times. Galaxy Zoo thus solved both the visual classification problem of time efficiency and improved accuracy by producing a distribution of independent classifications for each galaxy. While crowd-sourced galaxy classifications have proven their worth, challenges remain before establishing this method as a critical and standard component of the data

processing pipelines for the next generation of surveys. In particular, though innovative, crowd-sourcing techniques do not have the capacity to handle the data volume and rates expected in the next generation of surveys. These algorithms will be delegated to handle the majority of the classification tasks, freeing citizen scientists to contribute their efforts on subtler and more complex assignments.

This thesis presents 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 210,803 galaxies in just 32 days of GZ2 project time with 93.1% 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

List of Tables

List of Figures

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 a window to its past. The insights gleaned through the study of galaxy morphology have radically changed our view of the 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; de Vaucouleurs, 1959; Sandage, 1961; van den Bergh, 1976; Nair & Abraham, 2010; Baillard et al., 2011). The division between early-type and late-type (see Section 1.2) 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 objects, 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, though highly accurate, visual endeavor and only in the past twenty years have automated morphological assignment been possible. While current state-of-the-art image analysis routines can deliver rapid high-level morphologies, this efficiency comes at the expense of requiring prohibitively large training sets to achieve human-equivalent classification accuracy. 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 accurate 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 hundreds of 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 a suite of non-parametric morphology indicators widely used for automated morphologies. This thesis begins with a detailed account of the data utilized in this work as well as the methodology used to obtain these morphology indicators (Chapter **??**). Chapter 3 demonstrates how crowd-sourcing techniques can be optimized by applying a Bayesian approach to the aggregation of Galaxy Zoo classifications, while Chapter **??** 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 analysis of Galaxy Zoo: Hubble classifications (Chapter 5). This Introduction provides a brief overview of galaxy formation, the science achieved through the study of morphology, and galaxy classification techniques. It concludes with an overview of the work to be presented in this thesis, Galaxy Zoo: Express, a framework that integrates human and machine galaxy classifiers in order to scale the collection of galaxy morphologies for the next generation of astrophysical surveys.

1.1 Galaxy formation

Modern cosmology begins with the premise of the cosmological principle – that the universe is homogeneous and isotropic on large scales. Early observations revealed that the universe is expanding (Hubble, 1929; Hubble & Humason, 1931) and that it was much hotter and denser in the past (Penzias & Wilson, 1965; Dicke et al., 1965). The primordial universe is thought to have undergone a very early epoch of exponential expansion (Guth, 1981) during which quantum fluctuations gave rise to small inhomogeneities that are detected in observations of the Cosmic Microwave Background (Hinshaw et al., 2013; Planck Collaboration et al., 2016). These observations combined with those of Type Ia supernovae (Riess et al., 1998; Perlmutter et al., 1998) give rise to the standard Λ Cold Dark Matter (ΛCDM) model. In this model, normal matter accounts for only 4% of the energy density of the universe while cold, collisionless dark matter represents nearly 25%. The remainder is thought to be composed of mysterious dark energy, described by Einstein's famous "cosmological constant," and is currently causing the expansion of the universe to accelerate.

The initial seeds of inhomogeneity produced during the epoch of inflation are thought to cause the dark matter component of the early universe to develop small density perturbations characterized as over- and underdensities. These perturbations exanded with the expansion of the universe while the mean, or background, density decreased. When an overdensity critically exceeds that of the local background it ceases to expand and becomes gravitationally self-bound (Gunn & Gott, 1972), creating what is referred to as a dark matter halo in which every galaxy is born. These halos can then continue to grow through hierarchical clustering, i.e., mergers with other halos. In the current paradigm, these halos also contain baryonic matter which subsequently cools and condenses in the halo's potential well eventually creating the luminous content of galaxies (White & Rees, 1978).

In addition to gravity, there are several physical processes that are crucial to the success of galaxy formation: cosmological accretion, various mechanisms of galaxy feedback and quenching, and structural transformation through mergers and environmental processes. We touch on a few of these mechanisms here and in Section 1.3.

1.1.1 Gas accretion

When an overdense region of gas and dark matter collapses strong shocks form increasing the temperature of the gas (Binney, 1977; Rees & Ostriker, 1977). How the gas cools will determine the subsequent evolution of the system which is, in turn, dependent on the temperature of the gas. Gas hotter than $T > 10^7$ cools predominantly through bremsstrahlung (free-free emission), while gas with $10^4 < T < 10^7$ cools by ionized atoms decaying to their ground state or via electron recombination. Gas below these temperatures can cool via collisional excitation and de-excitation of heavy metals or molecules. This cooled gas can eventually condense and collapse thus triggering star formation.

An active area of current research is determining how galaxies accrete this gas from the interstellar medium. Two flavors are currently popular: "cold" and "hot" accretion modes. In hot-mode accretion, radiative cooling is inefficient and a hot, pressure-supported gaseous halo forms. This halo will gradually lose its thermal energy eventually collapsing and settling into a centrifugally supported disk. This mode of accretion is thought to be more important at lower redshift (Faucher-Giguère et al., 2011; van de Voort et al., 2011). On the other hand, cold-mode accretion is characterized by a gas cooling time much shorter than the dynamical time allowing cold gas to accrete directly onto the proto-galaxy in narrow streams and filaments (White & Frenk, 1991; Birnboim & Dekel, 2003; Kereš et al., 2005; Dekel et al., 2009a). This process is thought to dominate at higher redshifts (Katz et al., 2003) though direct detections of this mechanism are scarce (Steidel et al., 2010; Beck et al., 2016; Fumagalli et al., 2016; Martin et al., 2015, 2016).

1.1.2 Feedback and quenching

Gas cooling is predicted to be very efficient yet observations show that only 10% of baryonic matter is contained in stars or cold gas (Fukugita & Peebles, 2004). This "overcooling problem" implies that other physical processes must be invoked in order to *quench* star formation, i.e., inhibit gas from cooling or reheat it after it has cooled. These measures typically fall into either "preventative" or "ejective" feedback (Gabor et al., 2010; Kereš et al., 2009). Two major candidates include star formation feedback

in the form of supernovae, and that from accretion and jets of active galactic nuclei (AGN).

Stellar feedback via supernova driven winds, which heat and expel gas, are often invoked as the most likely mechanism to suppress star formation in low mass galaxies (Dekel & Silk, 1986; White & Rees, 1978). This has the added benefit of enriching the intergalactic medium with metals. In addition, several other processes associated with massive stars contribute to the inefficiency of star formation including photo-heating, photo-ionization and winds from massive stars (e.g., review by Hopkins et al., 2012). Stellar feedback mechanisms are expected to be more efficient at high redshift when galaxies had star formation rates much higher than that of today.

AGN are typically invoked for massive galaxies as there is strong evidence that all spheroidal galaxies likely contain a supermassive black hole (Kormendy & Ho, 2013), and studies have shown that the energy released via supernovae are not enough to sufficiently stifle star formation in these systems (Springel et al., 2005). These supermassive black holes accrete gas and can provide strong feedback by way of high-velocity winds which eject the cold interstellar medium or giant radio jets which prevent or slow the cooling of the surrounding hot gaseous halo (Fabian, 2012; Heckman & Best, 2014).

In addition to these largely *in situ* feedback mechanisms, several external processes can also lead to the suppression of star formation via galaxy-galaxy interactions. These environmental effects are discussed in more depth in Section 1.3.3.

1.1.3 Mergers

The basic picture of galactic structure is that smooth gas accretion produces disks (the formation of which is discussed in Section 1.1.4) while mergers destroy them (Toomre, 1977). Mergers are ubiquitous in the hierarchical paradigm of CDM with equal-mass mergers thought to completely destroy the disk, building dispersion-dominated spheroidal galaxies by efficiently removing angular momentum (Barnes, 1992; Mihos & Hernquist, 1996). Even unequal mergers can thicken disks and build spheroidal components (Moster et al., 2010). These two major galactic archetypes (disk- and spheroid-dominated) are explored in depth later in this chapter.

1.1.4 The importance of gas kinematics

The conventional approach to disk formation posits that accreting gas from the halo conserves angular momentum, eventually settling into a disk (Fall & Efstathiou, 1980; Mo et al., 1998). Unfortunately, simulated disks rotate too quickly at a given luminosity preventing them from lying on the observed Tully-Fisher relation (e.g, review by Brooks, 2010). This "angular momentum catastrophe" is the result of observed galaxies having a deficit of low angular momentum gas (Bullock et al., 2001; van den Bosch et al., 2001). Incorporating feedback via stellar winds allows for the redistribution of low angular momentum gas and keeps galaxies more gas-rich thus making the disk more resilient to mergers (Governato et al., 2009; Robertson et al., 2006; Brook et al., 2012).

The discovery of clumpy disks at $z \sim 2$ (Elmegreen & Elmegreen, 2005) presented additional challenges to disk formation models as these systems are found to be puffier and with a rotational velocity and gas dispersion inconsistent with local disk galaxies (Förster Schreiber et al., 2009). These clumps contain a substantial fraction of the disk star formation (Guo et al., 2012), a significant portion of the total stellar mass (Wuyts et al., 2012), and are generating outflows (Genzel et al., 2011). Understanding the origins of these properties and the evolution of such systems up to present day is ongoing and is addressed in greater detail in Chapter 5.

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

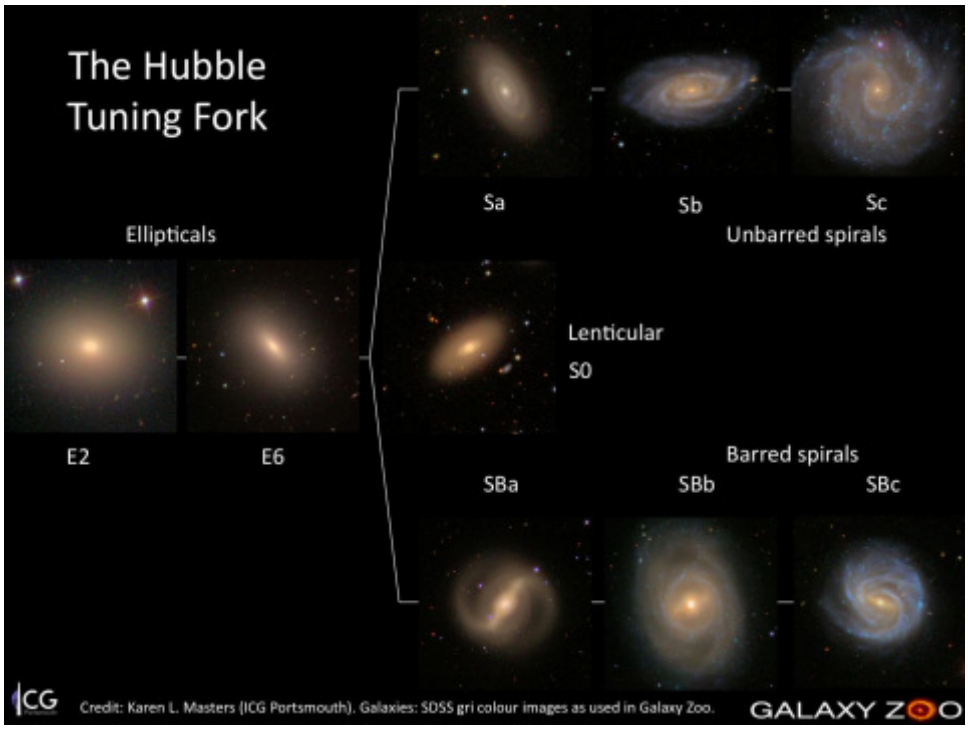

Figure 1.1 The Hubble "tuning fork" with example images for each galaxy type created from *gri*-composite SDSS imaging. Credit: Karen L. Masters and the Sloan Digital Sky Survey (SDSS) Collaboration.

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 a 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 color bi-modality

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. In the optical, it is well known that the color-luminosity distribution of galaxies is strongly bi-modal (Baldry et al., 2004a). Most galaxies fall into either the "red sequence" or the "blue cloud." There is a distinct gap between these two galaxy populations but recent studies have shown that, though sparse, this "green valley" could be a region of active galactic evolution related to a possible morphological transition (Schawinski et al., 2007).

This bimodality has resulted in the now ubiquitous color-magnitude relation (CMR) (Baldry et al., 2004b; Bell et al., 2004). These distinct populations are now believed to be due to differences in stellar populations as determined by spectroscopic age indicators as well as UV and IR photometry, with the red sequence being largely composed of *passive* or *quiescent* galaxies characterized by old, red stellar populations and little to no star formation, while the blue cloud is full of actively star forming galaxies with young, blue stellar populations (Brinchmann et al., 2004; Kauffmann et al., 2003; Salim et al., 2007; Schiminovich et al., 2007). A galaxy's morphology is tightly correlated with its color, with elliptical galaxies typically residing in the red sequence and disk galaxies inhabiting the blue cloud (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 ubiquitous 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). 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 (adapted from Kormendy & Bender, 2012).

The CMR is tighter for early-type galaxies in the red sequence suggesting that this population could be coeval, having had their last epoch of star-formation at some point in the past and evolving passively thereafter (Bower et al., 1992). Studying the CMR can provide insights into the last epoch of star formation for this galaxy population (Sandage & Visvanathan, 1978; Tully et al., 1982). There exists a degeneracy between the age and metallicity of stellar populations: older stars tend to be redder but this effect

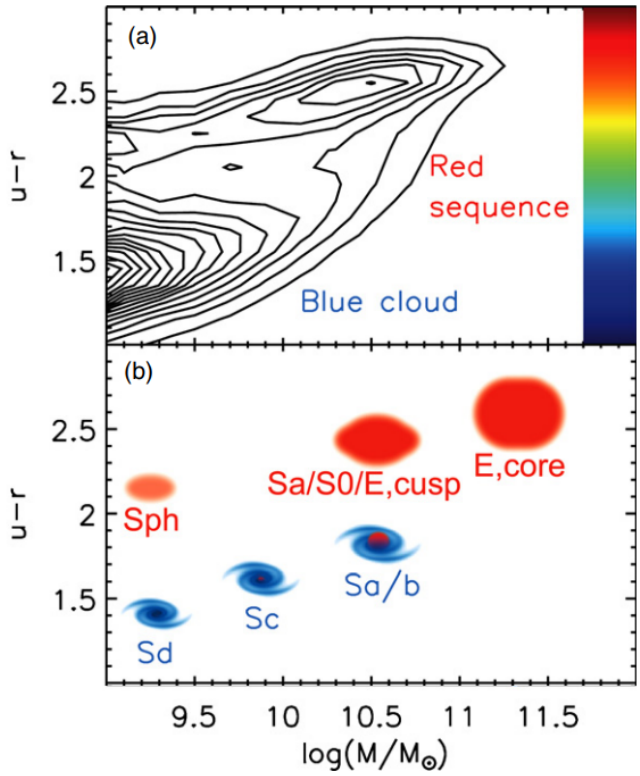

Figure 1.2 Color-magnitude relation adapted from Kormendy & Bender (2012). Though the x-axis is in units of log stellar mass, mass correlates strongly with a galaxy's luminosity. The top panel shows the contours of galaxy number density (Baldry et al., 2004a) with the rainbow bar denoting the associated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

can also be achieved through an increase in stellar metallicity. The color dependence of the red sequence is consistent with evolution models whereby the slope is due primarily to metallicity effects (Bower et al., 1992; Kodama & Arimoto, 1997). This allows for an age estimate to be placed on the last episode of significant star formation in early-type galaxies of typically between 3-7 Gyr ago (e.g., López-Cruz et al., 2004).

1.3.2 Mass assembly and the emergence of the Hubble sequence

Perhaps two of the most fundamental characteristics of a galaxy are its mass and its star formation rate (SFR). Though more challenging than luminosity to measure accurately, it is now thought that several processes such as feedback (e.g., Gabor et al., 2010) and gas accretion (e.g., Conselice et al., 2013) are directly tied to galaxy stellar mass. Thus, stellar mass functions (SMFs) are a key observable that can statistically trace the formation of stars in the universe. A cottage industry for decades, constructing mass and lumainosity functions have provided insights into the build-up of baryonic mass over cosmic time (Steidel et al., 1999; Ouchi et al., 2004; Giavalisco et al., 2004; Drory et al., 2005; Fontana et al., 2006; Marchesini et al., 2009; Caputi et al., 2011; González et al., 2011; Lee et al., 2012; Ilbert et al., 2013; Bernardi et al., 2013; Duncan et al., 2014).

It is well known that the cosmic star formation rate 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). Furthermore, Madau & Dickinson (2014) estimate that half of the stellar mass observed today was formed more than 8 Gyr ago. Understanding how this star formation was distributed among galaxy morphological types not only provides insight on the emergence of the Hubble sequence we see today, but also probes galaxy evolution and formation mechanisms as we discuss below.

Recent studies have shown that star-forming and quiescent galaxy types can be identified clearly up to $z \sim 2$ (Brammer et al., 2011; Muzzin et al., 2013). Morphologically speaking, most of the stellar mass at $z > 2$ seems to reside in irregular galaxies with the existence of disk galaxies being questionable (Dickinson, 2000; Papovich et al., 2005; Cameron et al., 2011; Conselice et al., 2005, 2011; Buitrago et al., 2013). Between $1 < z < 2$ the stellar mass density resides more in traditional disk systems but has

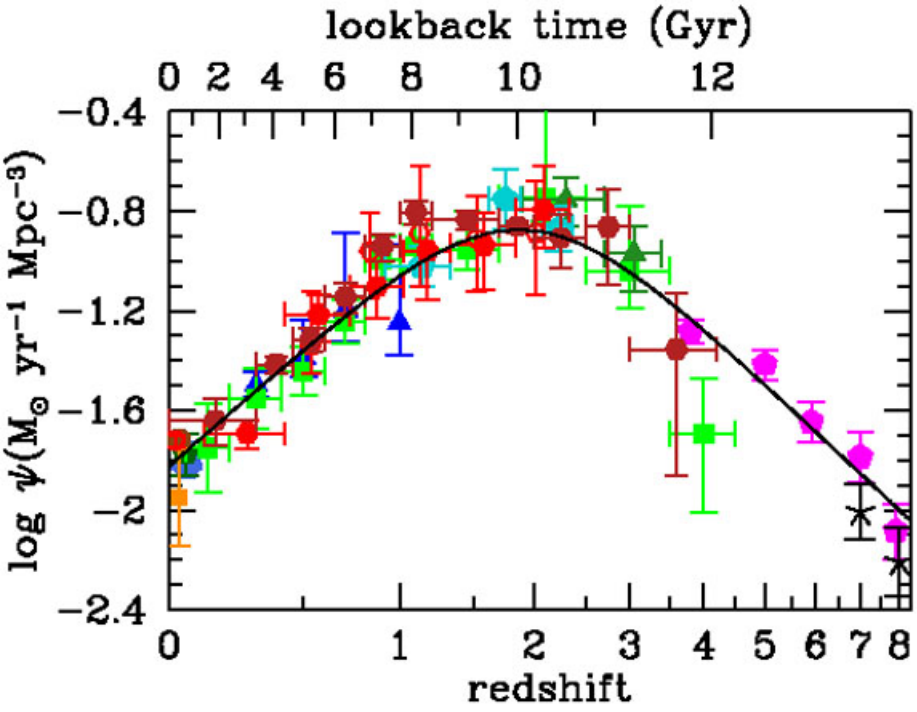

Figure 1.3 The history of cosmic star formation as determined by FUV+IR rest-frame measurements from several galaxy samples (credit: Madau & Dickinson, 2014, and references therein). This shows that cosmic star formation peaked around $z \sim 2$ when the universe was approximately 3.5 Gyr old followed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

stayed roughly constant since then (Bell et al., 2004; Brammer et al., 2011; Faber et al., 2007; Muzzin et al., 2013). Instead, SMFs indicate that the comoving number and mass density of quiescent systems has steadily increased over cosmic time with early-type systems residing as some of the most massive galaxies in the local universe (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). This is peculiar as it is the star-forming population that is expected to gain in stellar mass through the birth of new stars. This finding implies that more star-forming galaxies must have their star formation quenched.

Hierarchical galaxy formation 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 gas accretion from the intergalactic medium 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 hierarchical theory, is nevertheless favored observationally (Cowie et al., 1996; Brinchmann & Ellis, 2000; Bundy et al., 2005; Cimatti et al., 2006).

1.3.3 Morphology as a function of environment

Several studies have demonstrated that the probability for a galaxy to be quiescent depends on its stellar mass and large-scale environment (Balogh et al., 2004; Hogg et al., 2004; Peng et al., 2010; Woo et al., 2013). That a galaxy's environment has a direct relationship with its morphology was first quantified by Dressler (1980), and is known as the morphology-density relation. This empirical relation finds that elliptical galaxies tend to reside in the densest environments, i.e., rich groups and clusters, at all stages of cosmic time, while disks reside in isolated environments. This strongly indicates that environment plays a crucial role in galaxy evolution (e.g., Fasano et al., 2000; Smith et al., 2005; Peng et al., 2010).

Of particular interest, Peng et al. (2010) demonstrate that the fraction of quiescent galaxies in dense regions relies on different physical mechanisms depending on whether galaxies were central to the group or merely a satellite. Specifically, they find that the fraction of quiescent centrals is dependent solely on stellar mass, while that for satellites is dependent on both stellar mass and environment. There are several processes that could preferentially suppress star formation in satellite galaxies including harassment (Moore et al., 1996), tidal stripping or "strangulation" (Kawata & Mulchaey, 2008), and ram pressure stripping (Gunn & Gott, 1972). Another curious observational result is that of "galactic conformity," in which halos with red central galaxies preferentially have red satellites (Weinmann et al., 2006). There is currently little consensus on the physics that drive this phenomena (Kauffmann et al., 2013; Hearin et al., 2015, 2016; Pahwa & Paranjape, 2017) but it is clear that the relationship between morphology and environment continues to advance our understanding of galaxy evolution.

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 discussed below.

Identifying finer galactic structures has become easier with the advent of cameras such as the Wide Field Camera 3 (Dressel, 2012) aboard the Hubble Space Telescope, which has imaged the universe in unprecedented detail. However, it has only been since the development of large scale surveys like, for example, the Sloan Digital Sky Survey (SDSS, York et al., 2000; Abazajian et al., 2003), the Cosmic Assembly and Near-infrared Deep Extragalactic Legacy Survey (CANDELS, Grogin et al., 2011; Koekemoer et al., 2011), and the Dark Energy Survey (DES, The Dark Energy Survey Collaboration, 2005) 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 (Lintott et al., 2008; Cardamone et al., 2009). These oxygen-rich objects are observed to have some of the largest star formation rates with some of the smallest masses (Amorín et al., 2010). It is surmised that these galaxies were commonplace in the early universe and likely played a large role in the reionization of the universe (Izotov et al., 2016). 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 (Elmegreen & Elmegreen, 2005; Elmegreen et al., 2013). As previously discussed, galaxies of the past were largely irregular with star formation rates much higher than today (Madau & Dickinson, 2014). Much of the peculiar shapes of these galaxies are due to massive knots of star formation (Guo et al., 2015). 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 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 Canada-France-Hawaii Telescope Legacy Survey (Gwyn, 2012), 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 hundreds 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 CANDELS (Grogin et al., 2011; Koekemoer et al., 2011) team for 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 **??**.

One of the first and most popular parametric approaches for modeling a galaxy's

light distribution is the Sérsic profile:

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

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 (de Vaucouleurs, 1948). 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 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 Sersic 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 another 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; Domínguez Sànchez et al., 2017).

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 thesis 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

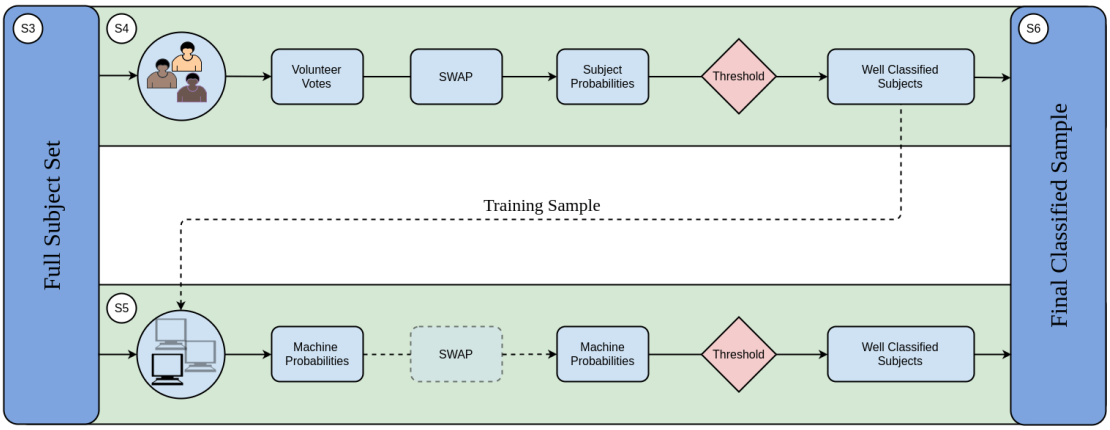

Figure 1.4 Schematic of our hybrid system. Humans provide classifications of galaxy images via a web interface. We simulate this with the Galaxy Zoo 2 classification data described in Chapter **??**. Human classifications are processed with an algorithm described in Chapter 3. Subjects that pass a set of thresholds are considered human-retired (fully classified) and provide the training sample for the machine classifier as described in Chapter **??**. 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.

algorithm that trains on a suite of non-parametric morphology indicators widely used for automated morphologies. We demonstrate that our method provides a factor of 11.4 increase in the rate of galaxy morphology classification while maintaining at least 93.1% classification accuracy as compared to Galaxy Zoo 2 published data. Here we present an overview of our framework, which also serves as a blueprint for this thesis.

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 **??**, we incorporate a machine classifier. We have developed a Random Forest algorithm that trains on measured morphology indicators (discussed in Chapter **??**) such as concentration, asymmetry, Gini coefficient, and M_{20}, well-suited for the top-level question of the GZ2 decision tree. 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 the 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 morpholgies

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 Schawinksi, Galaxy Zoo is a crowd-sourced initiative to visually classify large numbers of galaxies by enlisting members of the general public. The original project (GZ1, Lintott et al., 2008) 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 was an immediate success both in terms of the public interest and the resulting science: following its completion, over a dozen peer-reviewed articles were published which utilized GZ1 classifications.[1] In addition to explorations of galaxy

[1] `https://www.zooniverse.org/about/publications`

morphology and its dependence on color and environment (Skibba et al., 2009; Bamford et al., 2009), significant results also included the discovery of substantial populations of red disks (Masters et al., 2010) and blue ellipticals (Schawinski et al., 2009), as well as discoveries of rare objects such as the "green peas" (Cardamone et al., 2009) and Hanny's Voorwerp (Lintott et al., 2009), the first observation of an AGN ionization echo.

The early success 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 as well as simulated galaxy images using classifications provided from over a hundred thousand volunteers. The research presented in this thesis utilizes data from the Galaxy Zoo 2 project (GZ2, 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 morphologies such as the existence 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 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 `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 ∼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* and volunteers are allowed only one vote per task. 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×`petror90_r`. Throughout the life of GZ2 these images were randomly served to a web interface. 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 classification distribution. This resulted in a median of 44 classifications per galaxy with a wide spread, as shown in Figure 2.2. The full project spanned just over 14 months with the final dataset consisting of over 16 million classifications from over 80 thousand volunteers.

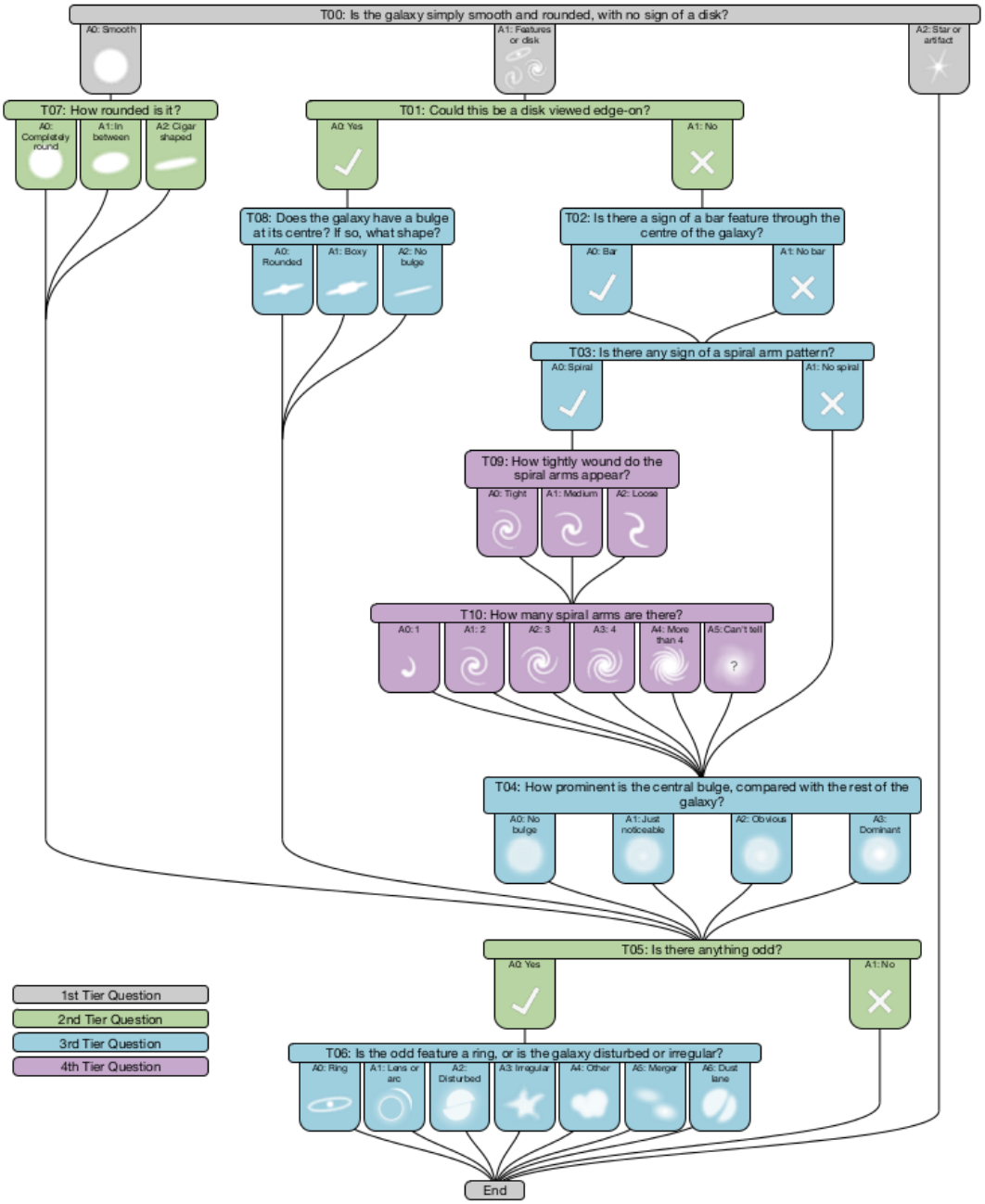

Figure 2.1 Galaxy Zoo 2 decision tree structure.

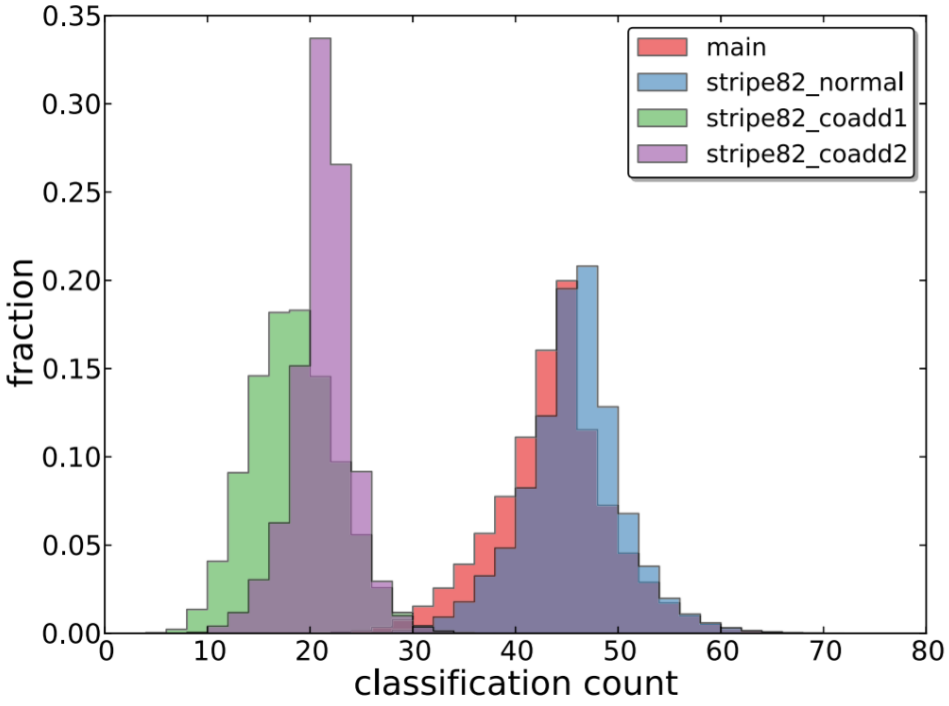

Figure 2.2 Distribution of the classification counts for the various subsets of the GZ2 galaxy sample (credit: Willett et al., 2013). In this thesis we use the main sample with a median of 44 votes per galaxy.

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

$$\kappa = \frac{1}{N_r} \sum_{i=1}^{N_r} \kappa_i \tag{2.1}$$

where N_r is the total number of possible responses to a task, and $\kappa_i = f_r$ if the volunteer's vote corresponds to response i, otherwise $\kappa_i = (1 - f_r)$. 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

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

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, vote fractions are adjusted for classification bias: a change in the observed morphology as a function of redshift that is independent of any true galaxy evolution. The galaxies in the GZ2 sample have $0.005 < z < 0.25$, a range that is shallow enough to justify an assumption of no significant evolution. Thus, the presumed culprit is that more distant galaxies are, on average, smaller and dimmer making fine features more difficult to identify. This effect is not unique to visual classifications and can also affect automated morphologies. 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. 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 for each task in the tree is computed in the lowest redshift bin for those galaxies with confirmed spectroscopic redshifts and with a sufficient number of votes to yield statistically reliable classifications, i.e., at least ten votes for a given task. 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 classes. In the case of galaxy morphology possible features could include pixel values, spectral indices, magnitude, color, etc. Choosing which features are most appropriate for a given task is a difficult undertaking as the field of machine learning provides little in the way of 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 galaxies 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 identifiers such as `CAMCOL`, `RUN`, and `FIELD` are used to select over 151,987 SDSS fields. Because the original GZ2 project

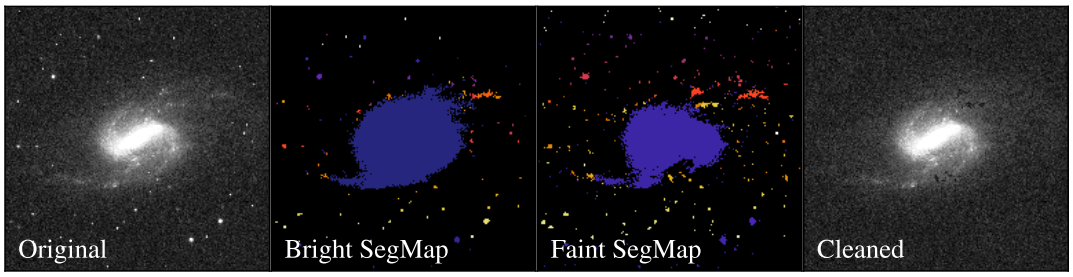

Figure 2.3 Example of SExtractor segmentation maps generated during the postage stamp cleaning process. The left panel shows the original i-band postage stamp, the middle two panels show the bright and faint segmentation maps where individual objects detected by SExtractor are shown on an arbitrary rainbow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

obtained imaging from DR7, we surmise that the route to some galaxies in DR12 have 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×Petrosian radius as measured by the SDSS pipeline. Galaxies located within 4 Petrosian radii of the edge of a field were excluded as image mosaicking 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 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 (SExtractor, 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 the object detection process. Two sets of parameters are used as it is unfeasible to find a single set of parameters that properly identifies all 282K galaxies. 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," from easy to hard, 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. Additionally, a prescription is required that measures a constant fraction of the total light while mitigating biases due to galaxy position and distance. 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.,

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

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

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 profiles are shown in Figure 2.8. In some cases the cleaning algorithm did

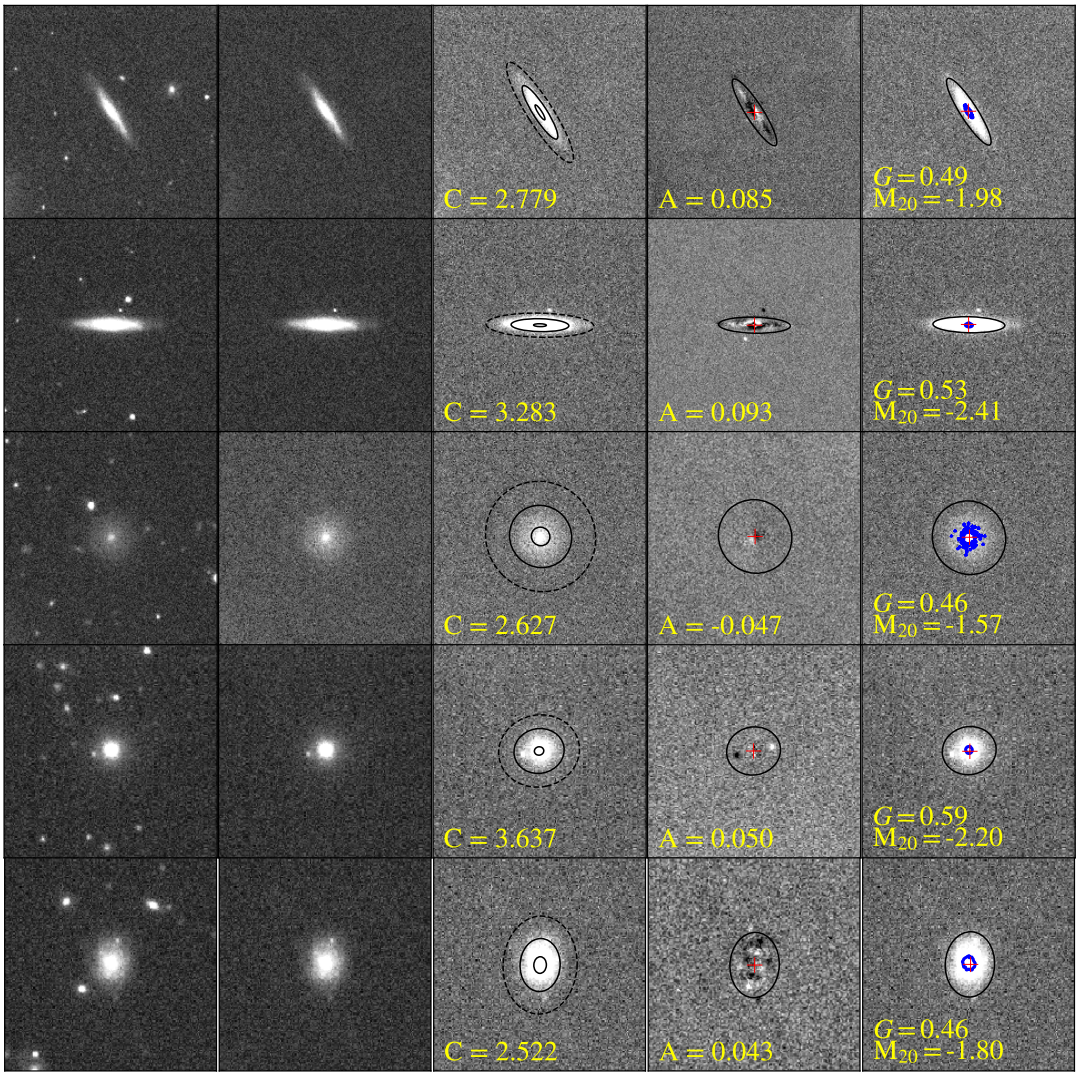

Figure 2.4 Examples of the postage stamp cleaning and morphology diagnostic measuring processes. The first and second columns show the original and cleaned i-band postage stamps. The third column shows the apertures used to calculate the concentration index where the two solid ellipses represent the apertures enclosing 20% and 80% of the galaxy total light which is defined as $1.5\times \mathrm{r}_p$ 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 $1\mathrm{r}_p$ within which all morphology diagnostics are computed.

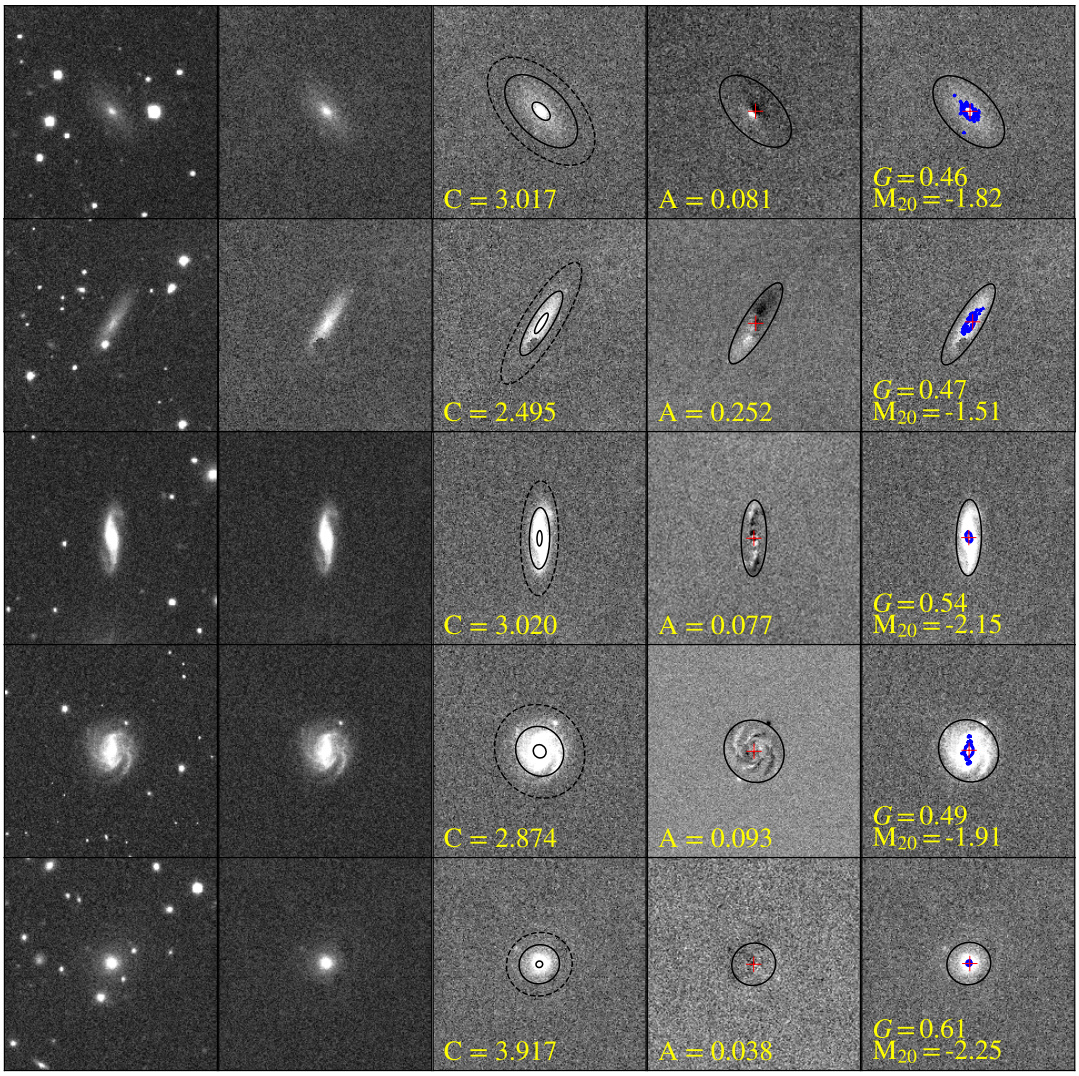

Figure 2.5 Additional examples of the postage stamp cleaning and morphology diagnostic measuring processes. See caption from Figure 2.4.

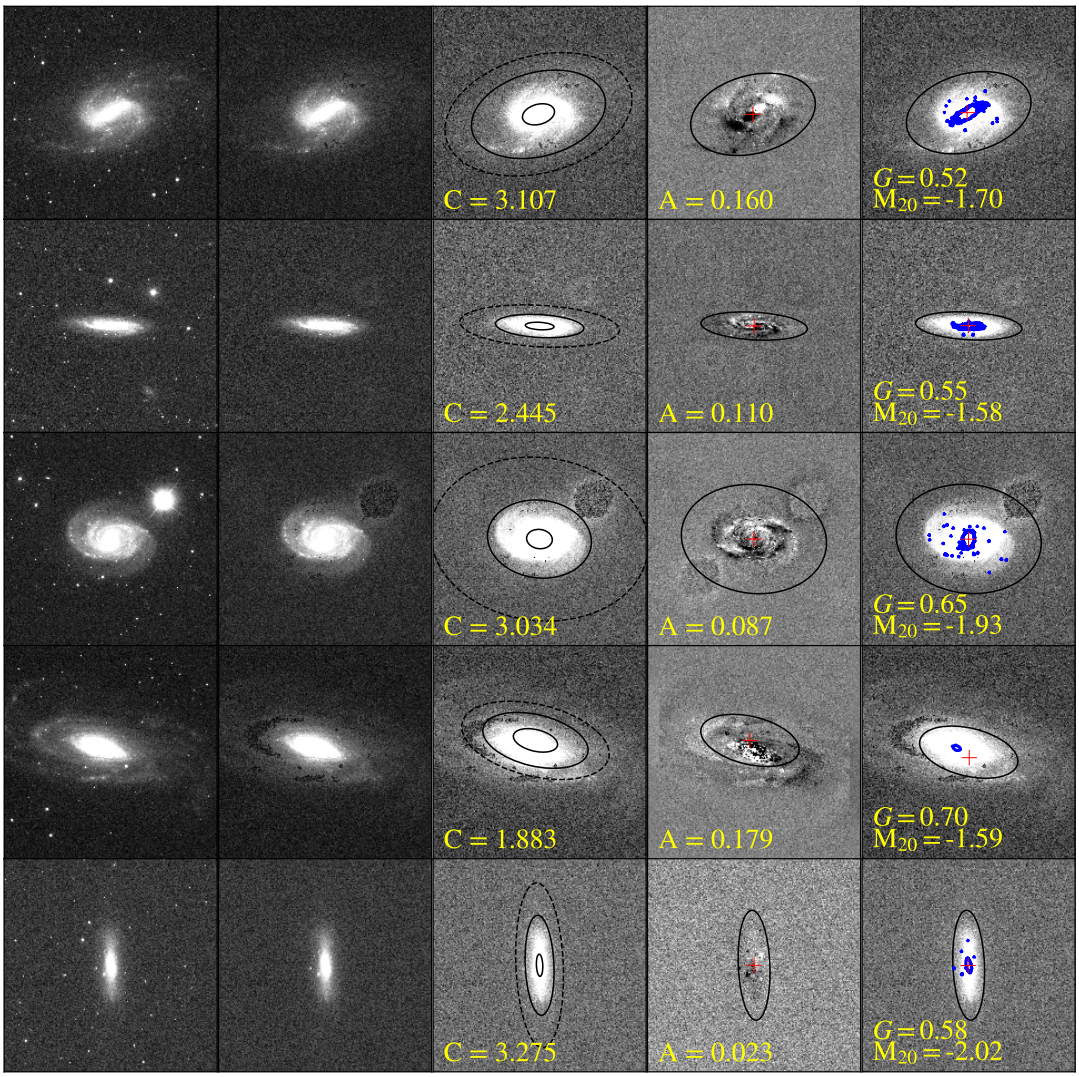

Figure 2.6 Examples of the postage stamp cleaning and morphology diagnostic measuring processes for a random sample of the largest and most difficult galaxies. See caption from Figure 2.4.

not fully remove nearby contaminating light from 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. Using our measured values of r_p, we compute the following morphology diagnostics in an elliptical aperture with semi-major axis of one Petrosian radius.

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):

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

where r_{80} and r_{20} are the radii containing 80% and 20% of the total galaxy light respectively. We define the total flux as that within 1.5 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, though not necessarily the physical shape of the galaxy, as A is not highly sensitive to low surface brightness features. A is measured by subtracting the galaxy image rotated by 180° from the original image itself. A correction for background noise is applied (as in e.g., Conselice et al., 2000; Lotz et al., 2004), i.e.,

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

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. Asymmetry is also calculated 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 generating a "background cutout": an image with the same pixel area as defined for the measurement of A, wherein each pixel is assigned a random value based on the statistics of the background in the original image as defined by the SExtractor segmentation maps. A is computed as before, including the minimization, and this value then constitutes B_{180}. A random sample of the residual A images, including 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

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

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:

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

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)

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

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 **??**, 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} is the second order moment of the brightest 20% of the galaxy flux (Lotz et al., 2004). It traces the spatial 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

$$M_{\mathrm{tot}} = \sum_{i}^{n} M_i = \sum_{i}^{n} f_i[(x_i - x_c)^2 + (y_i - y_c)^2] \tag{2.9}$$

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}:

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

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-center 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

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
All morphologies successful	281 801	95.4	% of full sample

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, 2008).

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 correlates strongly with edge-on galaxies.

In total, we successfully measure all morphological indicators for 281,801 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 2.10. 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

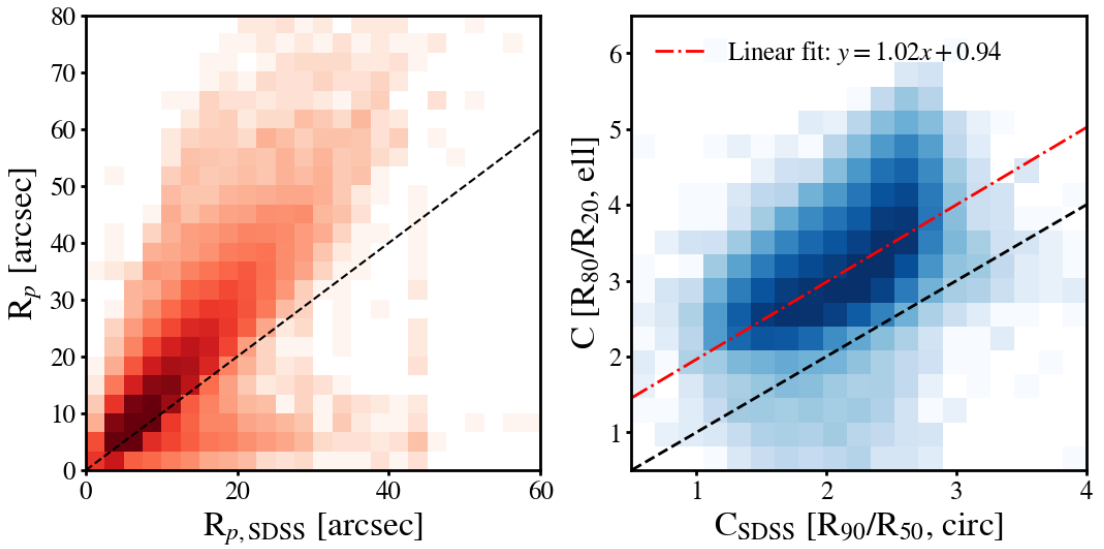

Figure 2.7 Comparison of our measured r_p and C to SDSS values. In both panels the black dashed line represents a 1-to-1. The left panel shows the Petrosian radius we measure against that computed in the SDSS pipeline. Most of the discrepancy is due to the apertures used: the SDSS pipeline defines r_p in circular annuli while we use elliptical. The right panel shows the relationship between concentration indices, where the red dash-dotted line is a linear fit. We find that our C is systematically $\sim$1 point larger than SDSS. This is 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%, in addition to the disparate aperture shapes already mentioned.

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 use elliptical. In Figure 2.8 we show examples wherein our measurement of r_p is much larger than that given by the SDSS pipeline, where the first column is the original postage stamp, the second column is the cleaned stamp, and the third column is the surface brightness and η profiles our algorithm computes. In the middle column, the red represents SDSS's r_p as circular apertures, while the blue represent our r_p as the semi-major axis of an elliptical aperture. The third panel is composed of two figures wherein the top plot shows $\mu(r_i)$ and $\bar{\mu}(r < r_i)$ in blue and orange dots, respectively.

The bottom plot shows η in green dots with the horizontal black dashed line representing 0.2. The vertical dash-dotted line marks the location of r_p according to our algorithm. The top row is an example of the rare case in which our algorithm seems to slightly overestimate r_p with no indications as to the cause, whereas in the middle two rows, the obvious culprit is poor cutout cleaning. However, in the final row, our algorithm outperforms SDSS and is better suited to edge-on galaxies, as is expected.

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 **??**. 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. Figure 2.7 shows our C against that computed from SDSS, $C_{\rm SDSS}$. Our values are systematically $\sim$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 measured using circular apertures, whereas we use elliptical. The latter is preferable as it has been shown that aperture shape affects C such that the concentration index is artificially inflated when measured within circular apertures, especially for highly elongated galaxies, i.e., edge-on disks (Bershady et al., 2000; Andrae et al., 2011).

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 A 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 1.8%, 1.4%, and 0.9% of galaxies have SExtractor, A, and M_{20} centers, respectively, that differ by more than $1.4''$ (the SDSS PSF) from the original SDSS coordinates. We thus find very good agreement between SDSS and

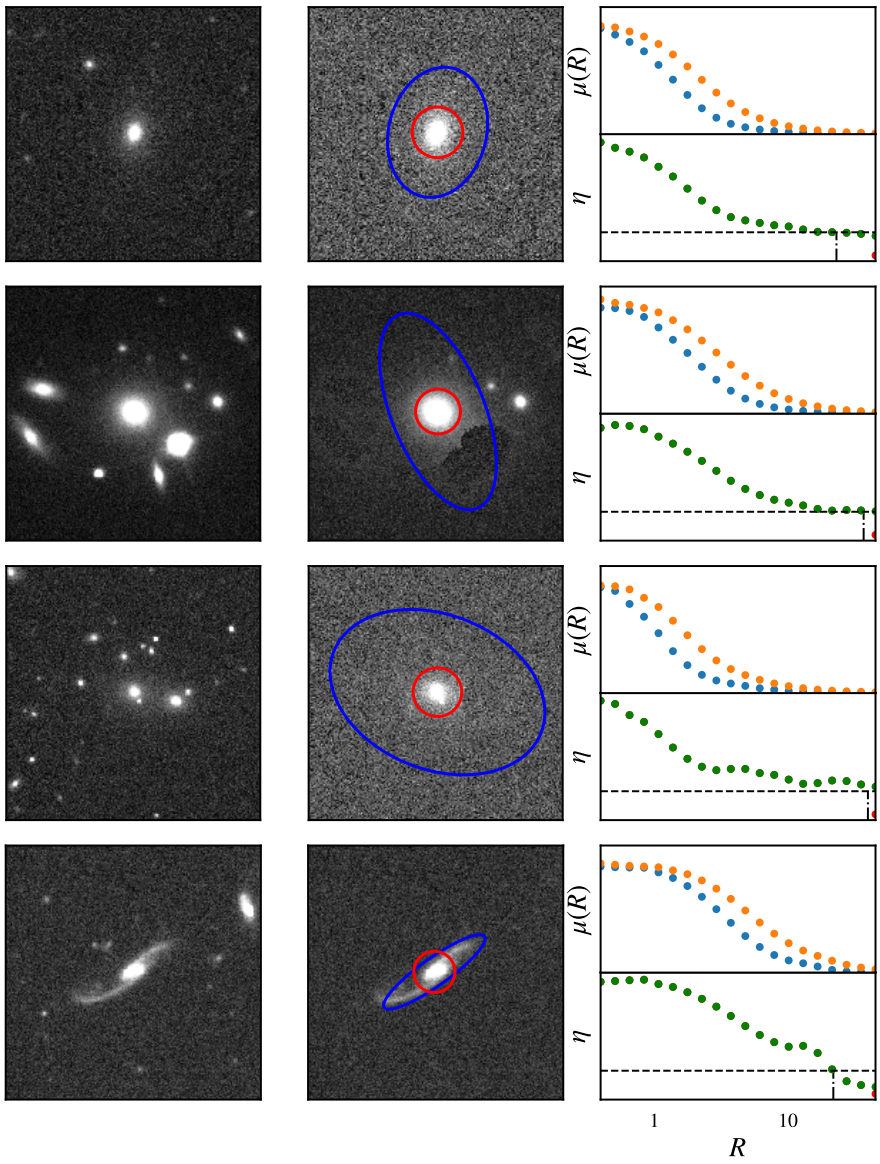

Figure 2.8 Here we show the original postage stamp, the cleaned stamp, and the surface brightness and η profiles our algorithm computes. In the middle column, the red represents SDSS's r_p as circular apertures, while the blue represents our r_p as the semi-major axis of an elliptical aperture. In the third column, the top panels display $\mu(r_i)$ and $\bar{\mu}(r < r_i)$ in blue and orange dots, respectively. The bottom panel shows η as green dots and the horizontal black dashed line is 0.2. The vertical dash-dotted line marks our measurement of r_p. The first three rows demonstrate various failures by our algorithm, however we outperform SDSS when it comes to edge on galaxies.

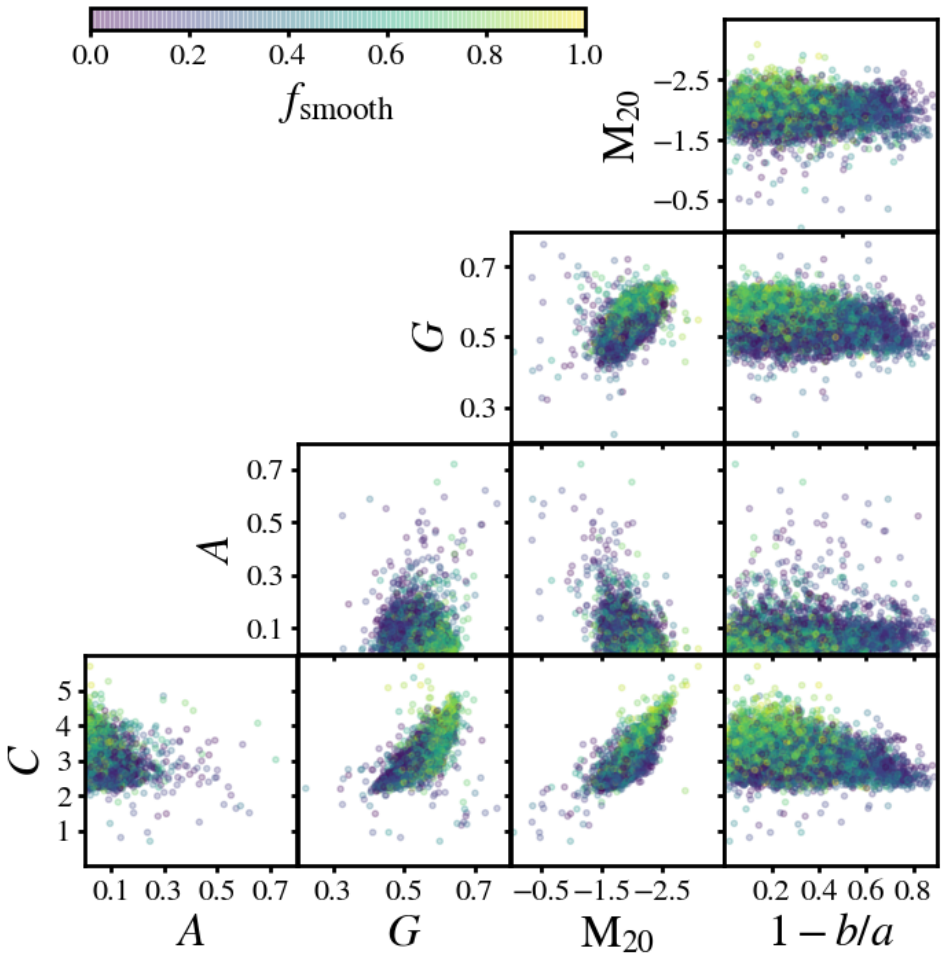

Figure 2.9 Here a random sample of 5000 galaxies are plotted in every possible planar combination of our morphology structural indicators where the color denotes the galaxy's debiased f_{smooth} fraction from GZ2. As expected, galaxies with higher f_{smooth} are found in locations where early-types are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

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 in only 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 (Abraham et al., 1996, 2003; Conselice et al., 2000; Lotz et al., 2004, 2008). 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, for clarity, 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 $f_{\rm smooth}$ fraction. Because $f_{\rm smooth}$ correlates well with early-type galaxies we see the expected trends discussed above. Automated morphologies for the full GZ2 sample is shown in Figure 2.10, where the color and contours correspond to the number density of galaxies.

2.3 Catalog of morphology diagnostics

Below we present an excerpt of the catalog of morphology diagnostics we measured and described in this chapter. The full catalog will be available online as soon as an appropriate hosting site has been acquired.

Table 2.2. Morphology Diagnostics for ∼283K SDSS Galaxies in the GZ2 Sample

DR7 objID	RA (°)	DEC (°)	R_p ($''$)	C	A	G	M_{20}	e	flag
587725550667628769	145.05811	60.57904	7.688	3.843	0.024	0.604	-2.169	1.050	0
587722984438300950	211.61574	0.84794	5.436	2.817	0.173	0.590	-1.857	1.375	0
587722982285574398	199.63151	-0.84205	6.236	3.377	0.048	0.595	-2.040	1.332	0
587722982300188997	233.01370	-0.83720	7.520	3.085	0.006	0.568	-1.903	1.372	0
587725550664220860	133.98611	55.43282	4.611	3.089	0.059	0.581	-1.784	1.228	0
587725083580367045	153.43677	-2.12333	17.687	2.865	0.054	0.500	-2.014	2.599	0
587725471212044550	146.47262	60.21963	6.118	3.281	0.057	0.577	-2.097	2.040	0
587725474417868988	151.83835	60.88533	11.891	3.394	-0.024	0.531	-2.219	1.519	0
587725471211389091	144.04851	59.35270	8.090	2.529	0.051	0.512	-1.906	2.744	0
587722982813204559	178.52952	-0.42060	8.073	3.581	-0.007	0.595	-2.083	1.099	0
587725550136459379	171.90446	65.68197	19.407	2.420	0.003	0.496	-1.890	5.162	0
587722983362134374	206.06578	-0.00380	7.243	3.024	0.048	0.583	-2.266	2.596	0
587725082507870286	156.26072	-3.02887	10.261	2.681	0.024	0.504	-1.822	1.224	0
587725471206670541	129.87227	52.06361	12.570	3.477	0.043	0.539	-2.197	1.437	0
587725040100704461	200.04925	-2.82914	7.173	3.432	0.039	0.590	-1.943	1.246	0
587725550143668309	210.94442	63.89091	4.833	2.963	0.058	0.556	-1.869	1.096	0
587725075528941624	139.16755	0.42750	17.521	2.841	0.092	0.517	-2.037	3.459	0

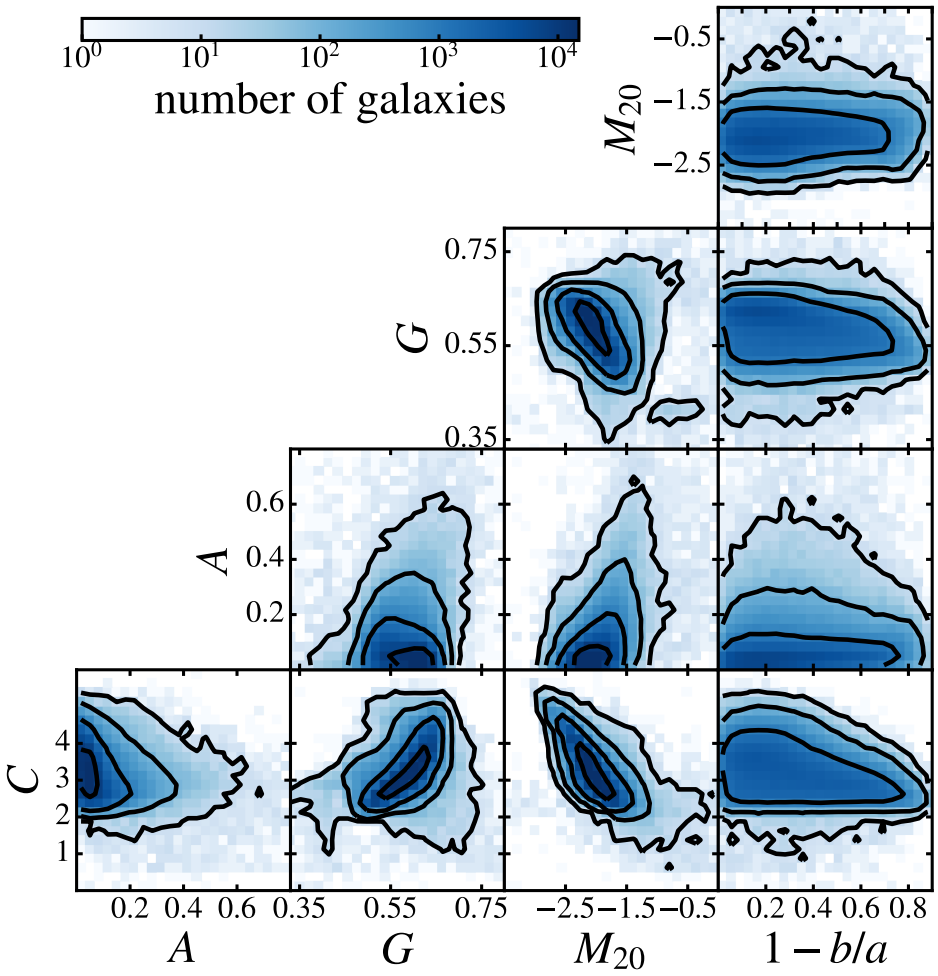

Figure 2.10 Relation between measured morphology diagnostics for more than 280K SDSS galaxies. Correlations between several diagnostics are immediately obvious and not all relations are likely to be linear. These points will be revisited in Chapter **??** when we discuss machine learning algorithms.

Chapter 3

Intelligent management of visual classifications

This chapter, albeit with some modifications, was submitted as part of a publication to the Monthly Notices of the Royal Astronomical Society and is currently under review with authorship Melanie R. Beck, Claudia Scarlata, Lucy F. Fortson, Chris J. Lintott, Melanie A. Galloway, Kyle W. Willett, B. D. Simmons, Hugh Dickinson, Karen L. Masters, Philip J. Marshall, and Darryl Wright; "Integrating human and machine intelligence in galaxy morphology classification tasks."

In this chapter we demonstrate that visual galaxy morphologies obtained through crowd-sourcing can be optimized by an intelligent vote aggregation system. We demonstrate this through a re-analysis of Galaxy Zoo 2 volunteer classifications where each re-analysis is dubbed a "simulation."

3.1 Galaxy Zoo 2 Classification Data

Our simulations utilize original classifications made by volunteers during the GZ2 project. These data[1] are described in detail in Willett et al. (2013) and the preceding Chapter,

[1] `data.galaxyzoo.org`

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 color composite images via a web-based interface[2] 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 tree[3] . 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 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 $f_{\rm response}$. 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 $f_{\rm featured}+f_{\rm artifact} > f_{\rm smooth}$, the galaxy is labeled 'Featured', otherwise it is labeled 'Not'. We note that only 512 subjects in the GZ2 catalog have a majority $f_{\rm artifact}$, contributing less than half a percent contamination when combining the "star or artifact" with "features or disk" responses.

The GZ2 catalog publishes three types of vote fractions for each subject: raw, weighted, and debiased. Debiased vote fractions are calculated to correct for redshift

[2] `www.galaxyzoo.org`

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

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 Section 3.3 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

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

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 labeled 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

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

where probabilities are normalized 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) catalog, a collection of ∼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 platform[4] 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 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 3.2 (and described in the next section). We compute $f_{\rm smooth}$ from those five volunteers. In Figure 3.1 we plot the volunteer $f_{\rm smooth}$ against the expert $f_{\rm smooth}$ 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 labeled with the corresponding number of galaxies at that location in $f_{\rm smooth}$ space. For example, there are 1316 subjects that both experts and volunteers labeled 'Featured' with $f_{\rm smooth} = 0.0$. 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

[4] The Project Builder template facility can be found at `http://www.zooniverse.org/lab`.

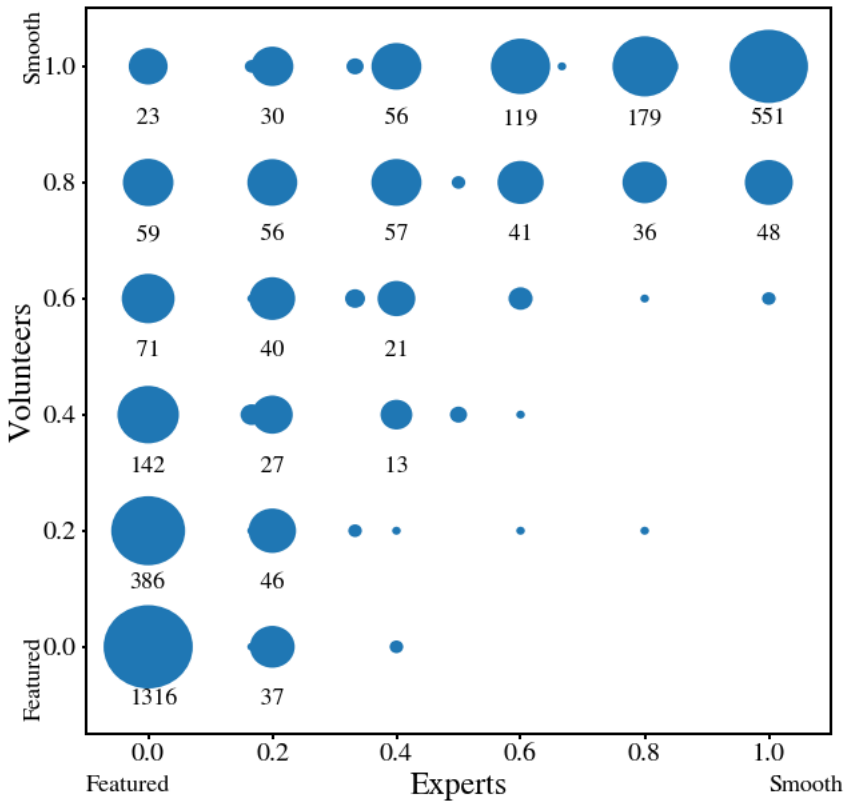

Figure 3.1 Volunter bias of labeling galaxies 'Smooth' ('Not') as compared to expert classifications for the gold standard sample.

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' moreso than experts. Of this small sample, 40% of galaxies are given a larger $f_{\rm smooth}$ 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 mitigates these issues by "debiasing" volunteer vote fractions (Willett et al., 2013). We compare both the original raw and debiased $f_{\rm smooth}$ 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. Experienced astronomers are more adept at identifying galactic disks, even those which posses 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 **??**.

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, ($P(\text{“}F\text{”}|F)$, $P(\text{“}N\text{”}|N)$); 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

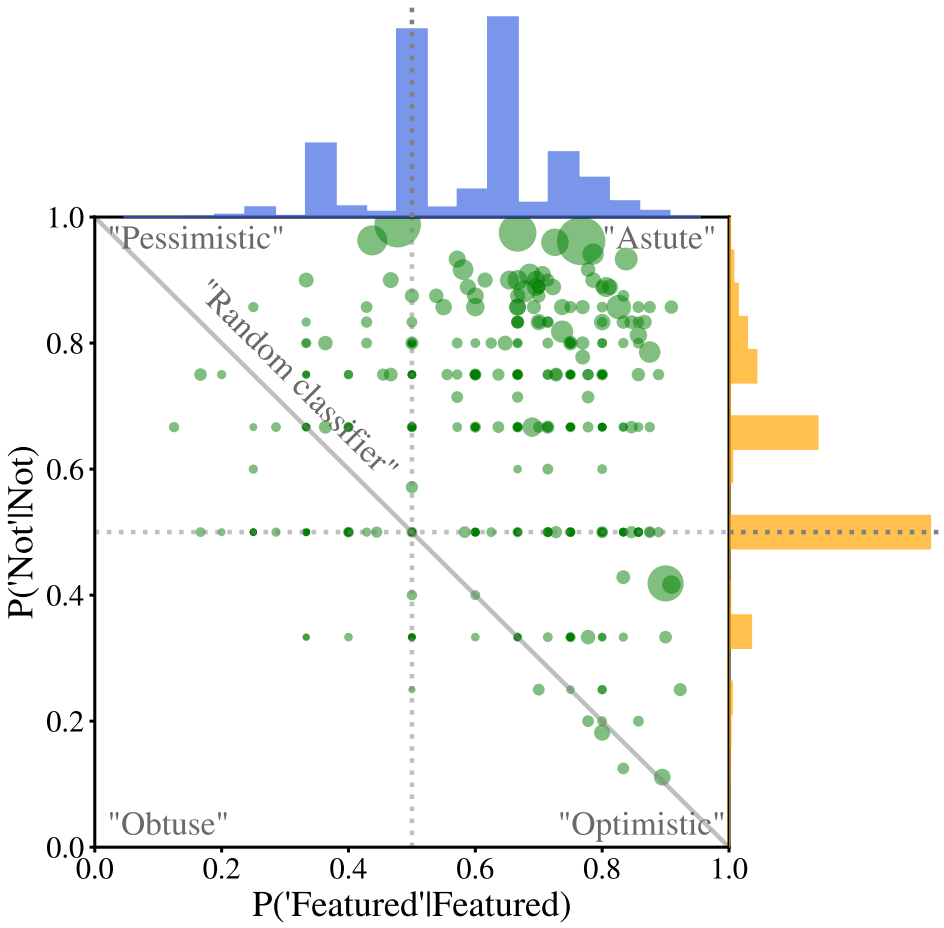

Figure 3.2 Confusion matrices for 1000 randomly selected GZ2 volunteers after fiducial SWAP assessment. Circle size is proportional to the number of gold standard subjects each volunteer classified. The histograms on top and right represent the distribution of each component of the confusion matrix for 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.

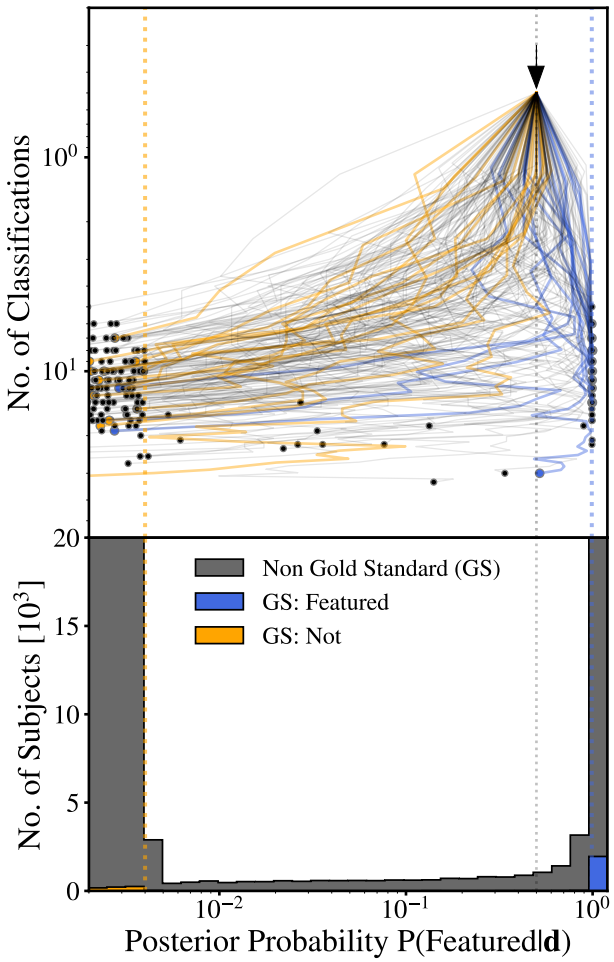

Figure 3.3 Posterior probabilities for GZ2 subjects. The top panel depicts the probability trajectories of 200 randomly selected GZ2 subjects. All subjects begin with a prior of 0.5 denoted by the arrow. Each subject's probability is nudged back and forth with each volunteer classification. From left to right the dotted vertical lines show the 'Not' threshold, prior probability, and 'Featured' threshold. Different colors 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.

to be retired as 'Featured', to 0.99. Similarly, we set the 'Not' threshold, $t_N = 0.004$. In Section 3.4 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 colored vertical dashed lines correspond to the retirement thresholds, t_F and t_N. The lower panel shows the full distribution of GZ2 subject posteriors 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 3.2.4). In

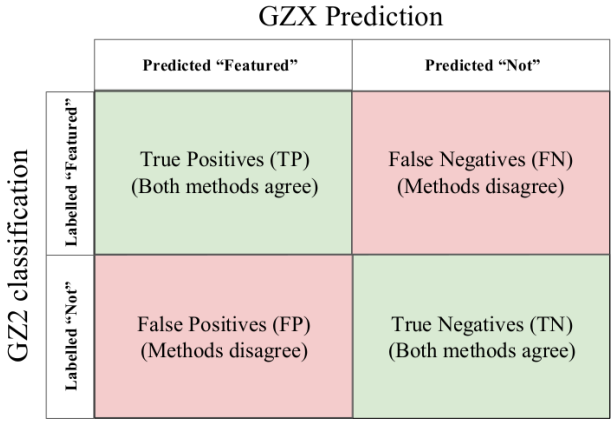

Figure 3.4 Confusion matrix for comparing GZ2 classifications to our method. True positives (TP) and true negatives (TN) indicate that the predictions from our method agree with GZ2 for subjects labeled 'Featured' and 'Not', respectively. When the two classification methods disagree, the result is a 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.

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 $\mathrm{GZ2_{raw}}$ 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 labeled 'Featured' by GZ2, whereas a pure sample recovers *only* subjects labeled '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

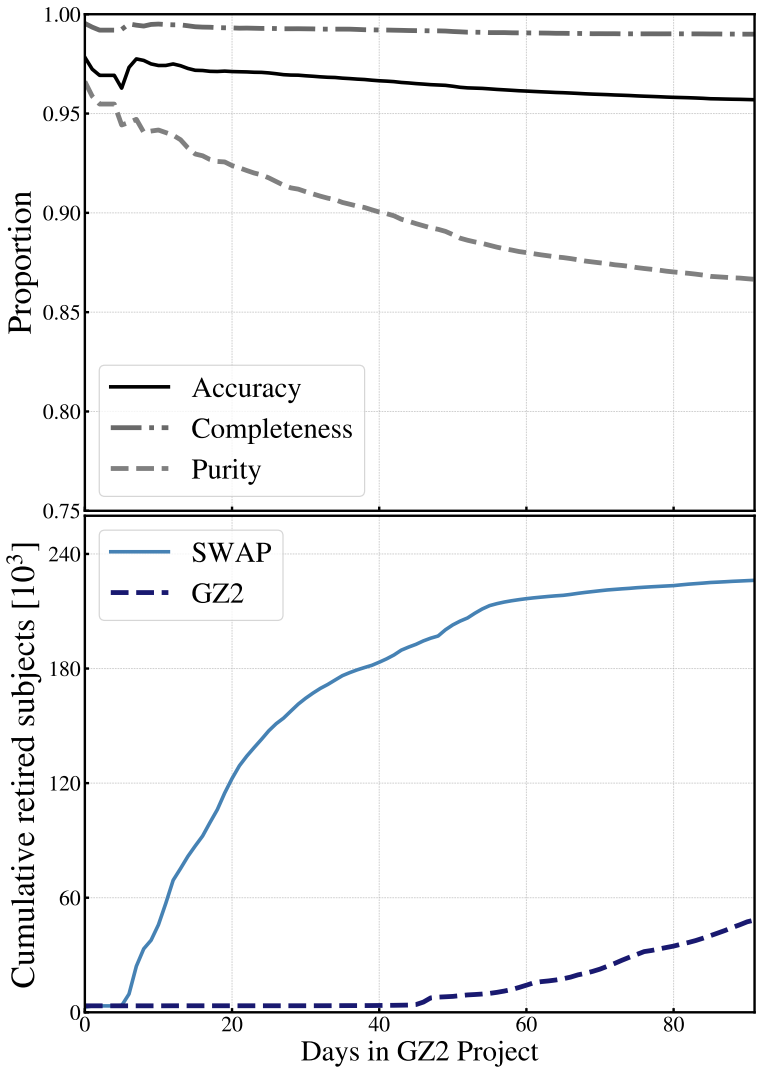

Figure 3.5 Fiducial SWAP simulation demonstrates a factor of 4.7 increase in the rate of subject retirement as a function of GZ2 project time (bottom panel, light blue) compared with the original GZ2 project (dashed dark blue). After 92 days, SWAP retires over 226K subjects, while GZ2 retires ∼48K. The top panel displays the quality metrics (greys). These are calculated by comparing labels predicted by SWAP to $GZ2_{raw}$ labels (Section 3.1) 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).

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 ∼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 3.2.6.

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. It can be argued that GZ2 did not need ∼40 votes per subject to achieve exquisite sampling for the top-level question but rather adequate sampling for the subqueries. It might therefore be the case that the top-level question could be accurately resolved with far fewer classifications. In order to put SWAP and GZ2 on equal footing we determine the minimum number of votes, N, that the GZ2 project would need in order to replicate the original GZ2 outcome for the top-level classification task for a canonical 95% of its sample.

We compute the raw vote fractions ($f_{\rm featured}$, $f_{\rm smooth}$, and $f_{\rm artifact}$) 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 3.1. Our SWAP simulation did not retire every subject in the GZ2 sample. We therefore select 100 random subsamples each consisting of 226,124 subjects, and compute the accuracy and the total number of GZ2 classifications necessary to retire each subsample. These 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, a full 3.5 times more classifications than SWAP needs to achieve the same accuracy. 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 did not have a predictive retirement rule, rather the project was declared complete when

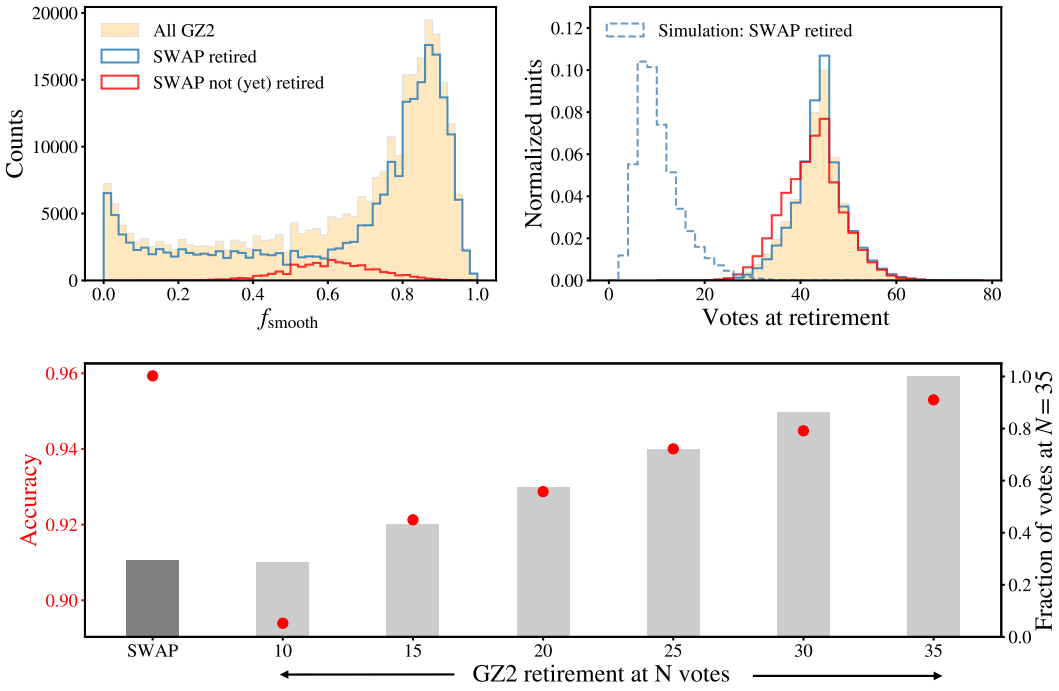

Figure 3.6 *Top panels:* The left panel shows $f_{\rm smooth}$ for the entire GZ2 sample (shaded orange), the subjects retired by SWAP (blue), and subjects that SWAP has not yet retired by the end of our simulation (red). The latter peaks at $f_{\rm smooth} \sim 0.6$, which we interpret as the most difficult to classify subjects: those with $f_{\rm smooth} \leq 0.5$ are easily identified as 'Featured', while those with $f_{\rm smooth} \geq 0.8$ are more obviously 'Not'. The right panel shows 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. *Bottom panel:* We compare SWAP to results of simulations of GZ2 run with a lower retirement limit. Solid bars show the number of classifications required to retire the same number of galaxies as SWAP (dark grey) for different fixed retirement limits in GZ2 (light grey). The height of the bars are normalized to show the counts relative to the highest simulated GZ2 retirement limit we test ($N = 35$, right vertical axis). The accuracy of the classifications for these simulated GZ2 runs against the full GZ2 project are shown as red points (left vertical axis). If GZ2 retirement were set at a level ($N = 10$) that reproduces the total number of classifications logged by SWAP, the accuracy would be below 90% (versus SWAP's 96%). Instead, GZ2 requires at least 3.5 times more votes to approach the same accuracy as SWAP.

the median classification count for the ensemble reached a value that was deemed to be sufficient for accurate characterization of the classification. In contrast, SWAP retires "easier" subjects first while harder subjects remain in the system for longer (requiring many more votes to nudge that subject's posterior across a retirement threshold). Evidence for this can be seen in the top two panels of Figure 3.6. The top left panel shows the distribution of $f_{\rm smooth}$ 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 $f_{\rm smooth}= 0.6$. In contrast, the SWAP-not-yet-retired sample peaks at $f_{\rm smooth}= 0.6$. These subjects can be interpreted as being the most difficult to classify which can be understood intuitively: galaxies with $f_{\rm smooth} \leq 0.5$ are easily identified as having features, while galaxies with $f_{\rm smooth} \geq 0.8$ are obviously elliptical.

This is further corroborated in the top right panel of Figure 3.6 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 (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 3.6. 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. This is an indication that such subjects should have continued to accrue classifications in order to reach strong consensus.

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 requires only 30% of the votes that GZ2 needs and retires 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 3.7. The filled blue and orange histograms show the number of classifications per subject achieved from our SWAP simulation, where 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

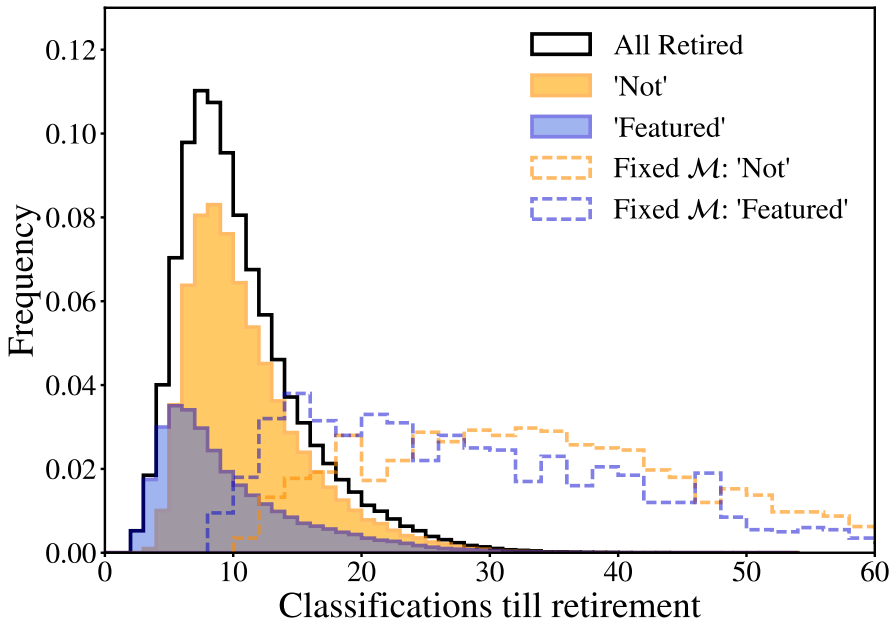

Figure 3.7 SWAP's volunteer-weighting mechanism provides a factor of three reduction in the human effort required to retire GZ2 subjects. The filled histograms show the number of volunteer classifications per subject achieved during our SWAP simulation broken down by class label, where the solid black 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 $\sim$3 times more classifications per subject to reach the same retirement thresholds.

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 $\mathrm{GZ2_{raw}}$ label, SWAP disagrees about 5% of the time. We thus examine the false positives (subjects SWAP labels as 'Featured' but $\mathrm{GZ2_{raw}}$ labels as 'Not') and false negatives (subjects SWAP labels as 'Not' but $\mathrm{GZ2_{raw}}$ 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 $f_{\mathrm{smooth}} = 0.5$ is then $(0.42, 0.57)$ assuming 40 classifications, each with a probability of 0.5. Figure 3.8 shows the distribution of $f_{\mathrm{featured}} + f_{\mathrm{artifact}}$ 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 $0.4 < f_{\mathrm{featured}} + f_{\mathrm{artifact}} < 0.6$. It is thus unsurprising that SWAP and GZ2 disagree

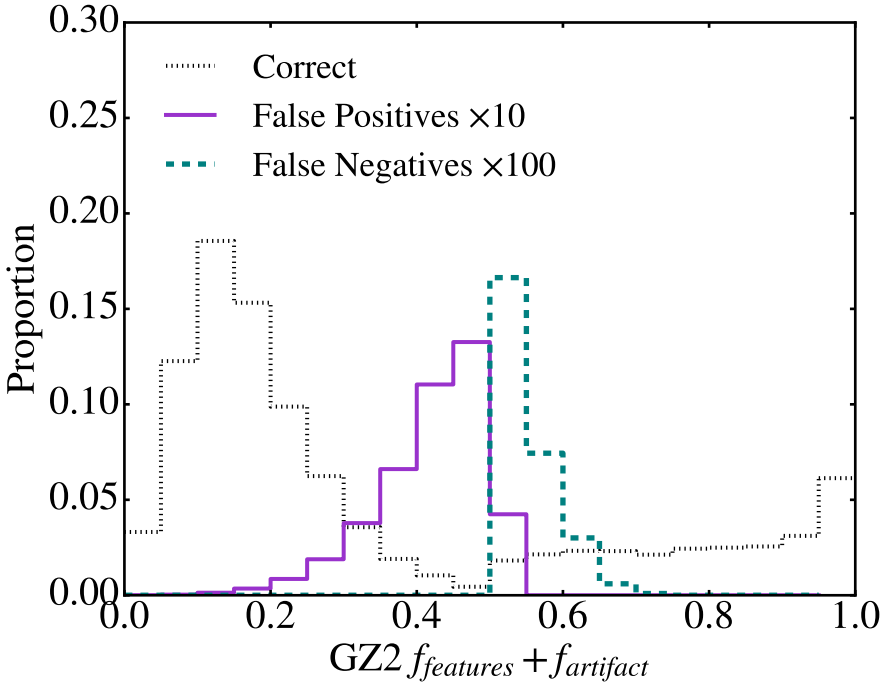

Figure 3.8 Distribution of GZ2 $f_{\text{featured}}+f_{\text{artifact}}$ vote fractions for subjects correctly identified by SWAP (dotted grey), along with those identified as false positives (solid purple), and false negatives (dashed teal). The false positives and false negatives are scaled by factors of 10 and 100 respectively for easier comparison. From Section 3.1, 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 $0.4 \leq f_{\text{featured}}+f_{\text{artifact}} \leq 0.6$ 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 labeled 'Not'.

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 Section 3.4). Exploring those subjects wherein SWAP and GZ2 disagree, we conclude that the majority of this disagreement stems from the stochastic nature of $\mathrm{GZ2_{raw}}$ labels. We now turn our focus towards incorporating a machine classifier utilizing these SWAP-retired subjects as a training sample.

3.3 Exploring the quality of Galaxy Zoo: Express

In this section we consider the robustness of GZX by computing several sets of "ground truth" labels from the GZ2 catalog. Recall in Section 3.1 we defined a subject as 'Featured' if $f_{\mathrm{featured}} + f_{\mathrm{artifact}} \geq f_{\mathrm{smooth}}$, a threshold, t, of 0.5. Here we compute new descriptive labels by allowing that threshold to vary where $t \in [0.2, 0.3, 0.4, 0.5]$. Any subject with $f_{\mathrm{featured}} + f_{\mathrm{artifact}} \geq t$ is labeled 'Featured', otherwise it is labeled 'Not'. We recalculate the quality metrics of accuracy, purity, and completeness on the sample

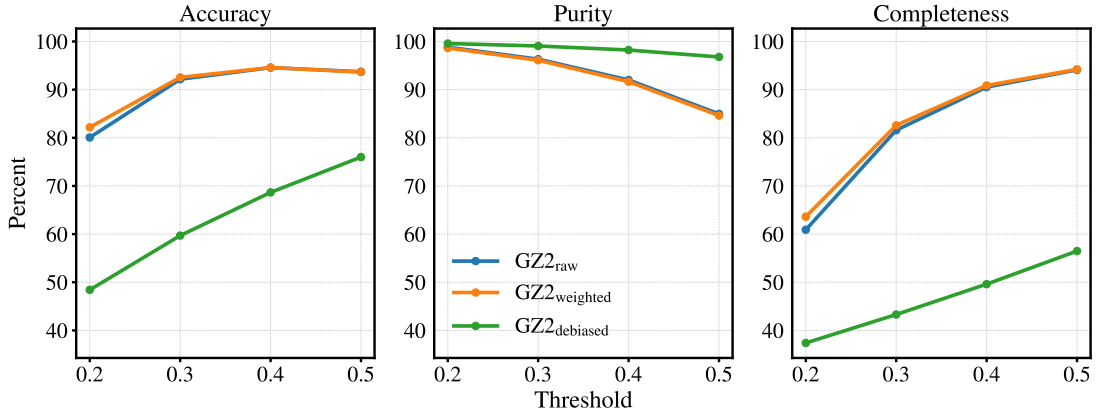

Figure 3.9 Quality metrics computed on the subjects retired during the GZX simulation for a range of thresholds and GZ2 vote fraction types.

of galaxies retired during the full GZX simulation (discussed in Chapter 4) for each threshold and each type of GZ2 vote fraction: raw, weighted, and debiased. The results are shown in Figure 3.9. GZX classifications are quite robust, with accuracy fluctuating by only a few percent for $\mathrm{GZ2_{raw}}$ and $\mathrm{GZ2_{weighted}}$ labels computed using a threshold between 0.3 and 0.5. Instead we see a trade off between purity and completeness. Decreasing the threshold results in more subjects labeled as 'Featured', which in turn increases sample purity while simultaneously decreasing completeness.

That the $\mathrm{GZ2_{debiased}}$ labels perform poorly is not surprising. These vote fractions are computed after considerable post-processing of the raw volunteer votes in order to remove the effects of redshift and surface brightness. As with any set of visual classifications, these biases must be accounted for and this is traditionally done *a posteriori*. It is also unsurprising that the $\mathrm{GZ2_{raw}}$ and $\mathrm{GZ2_{weighted}}$ classifications are in such tight agreement. $\mathrm{GZ2_{weighted}}$ vote fractions are computed by down-weighting inconsistent volunteers of which there are relatively few. These two sets of vote fractions are thus very similar.

When applying GZX to future imaging programs, there will be no "ground truth" labels for comparison. In some sense, these thresholds can be interpreted as a prior for the SWAP p_0 value, the initial probability for a subject to be 'Featured'. As we show in Section 3.4, changing the prior has little affect on the retirement rate but does result in considerable variability in the completeness and purity of the resulting classifications.

The choice of whether to optimize SWAP to recover pure or complete samples is a decision for a given science team.

3.4 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.

3.4.1 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,

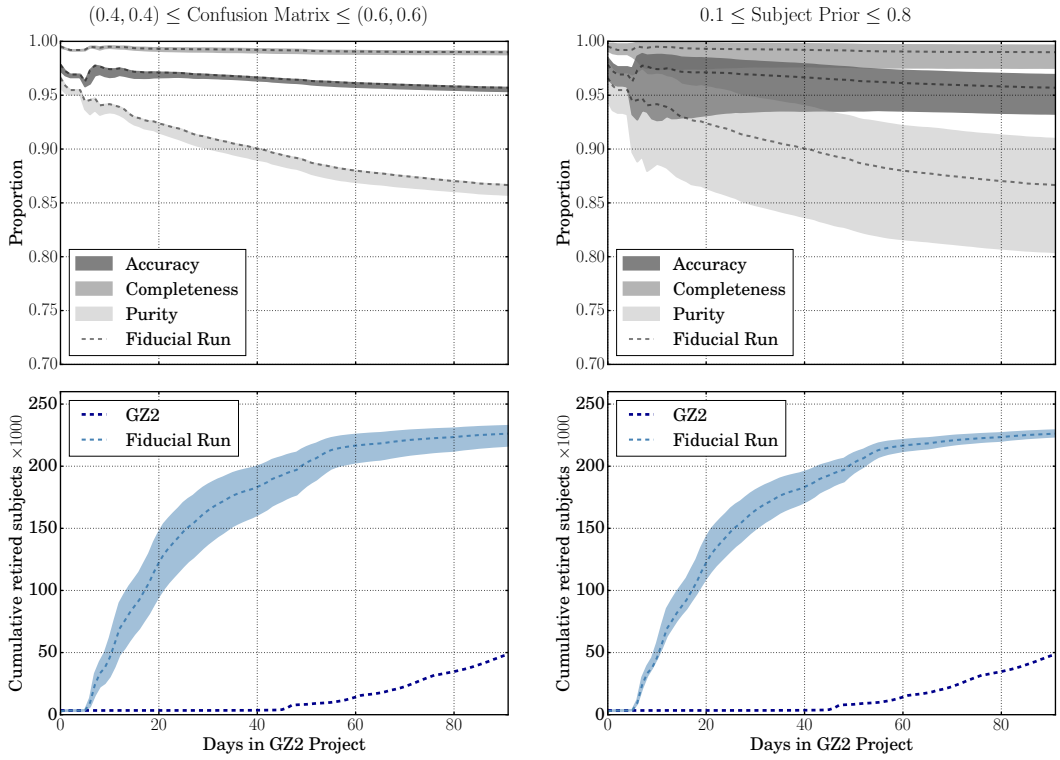

Figure 3.10 SWAP performance does not dramatically change even with a range of input parameters (shaded regions) as compared to the fiducial run of Section 3.2.3 (dashed lines). *Left.* The quality (top) and retirement rate (bottom) when the confusion matrix is initialized as (0.4, 0.4) and (0.6, 0.6), 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.

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 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.

3.4.2 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 $\mathrm{GZ2_{raw}}$ 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.

3.4.3 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

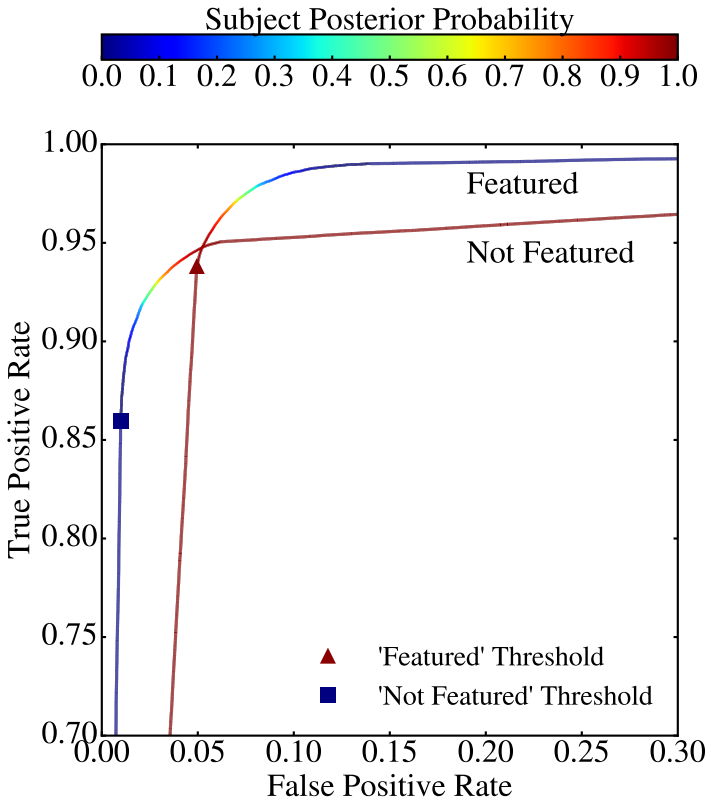

Figure 3.11 Identifying 'Featured' subjects is independent of identifying 'Not' subjects. Both ROC curves use all subjects processed by SWAP where the score used to create the ROC curve is simply each subject's achieved posterior probability. The Featured curve demonstrates how well we identify 'Featured' 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.

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 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

This chapter, albeit with some modifications, was submitted as part of a publication to the Monthly Notices of the Royal Astronomical Society and is currently under review with authorship Melanie R. Beck, Claudia Scarlata, Lucy F. Fortson, Chris J. Lintott, Melanie A. Galloway, Kyle W. Willett, B. D. Simmons, Hugh Dickinson, Karen L. Masters, Philip J. Marshall, and Darryl Wright; "Integrating human and machine intelligence in galaxy morphology classification tasks."

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 hyper-parameters 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 Chapter 2). Coupled with SExtractor's measurement of ellipticity (Bertin & Arnouts, 1996), we provide the machine with 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 GZX attempts 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 which allows GZX to dynamically train a machine classifier.

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.

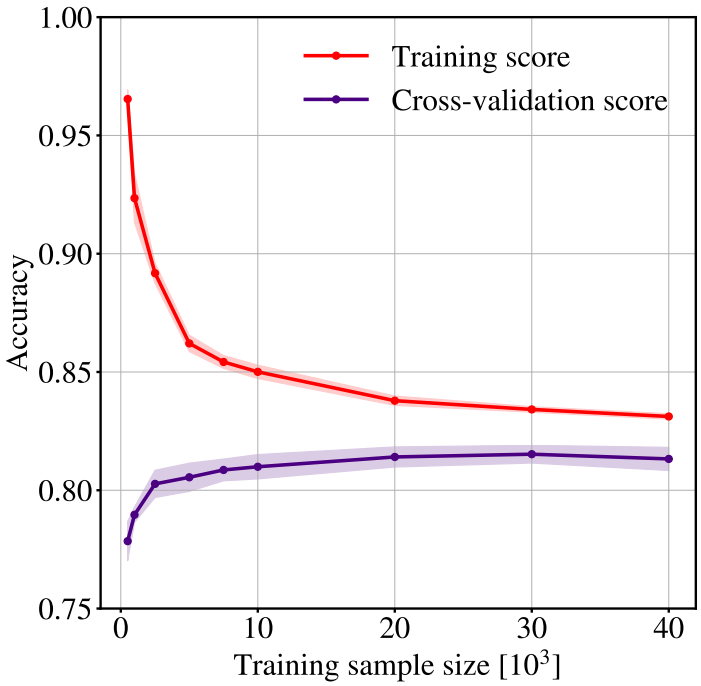

Figure 4.1 Learning curve for a Random Forest with fixed hyperparameters. These curves show the mean accuracy computed during cross-validation and on the training sample, where the shaded regions denote the standard deviation. When the training sample size is small, the machine accurately identifies 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 behavior plateaus indicating that larger training samples provide little in additional performance.

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

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

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 ∼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 machine retires over 70K, dramatically increasing the subset of retired subjects. We end the simulation after 32 days, having retired ∼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 behavior 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

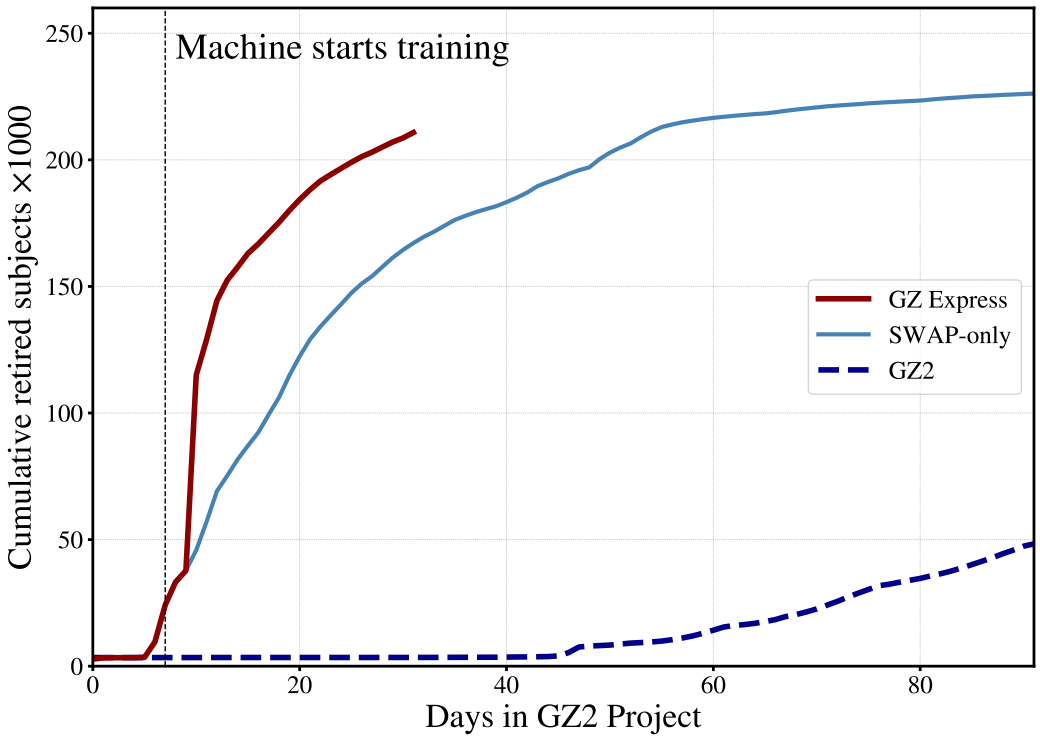

Figure 4.2 By incorporating a machine classifier, GZX (red) increases the classification rate by an order of magnitude compared to GZ2 (dashed dark blue) and out-performs the SWAP-only run (light blue), retiring more than 200K subjects in just 27 days of GZ2 project time. The dashed black line marks 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.

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 burden with the machine and SWAP responsible for ∼98K and ∼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 subjects identified as 'Featured' ('Not') according to their $\mathrm{GZ2_{raw}}$ 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 artifact of our choice

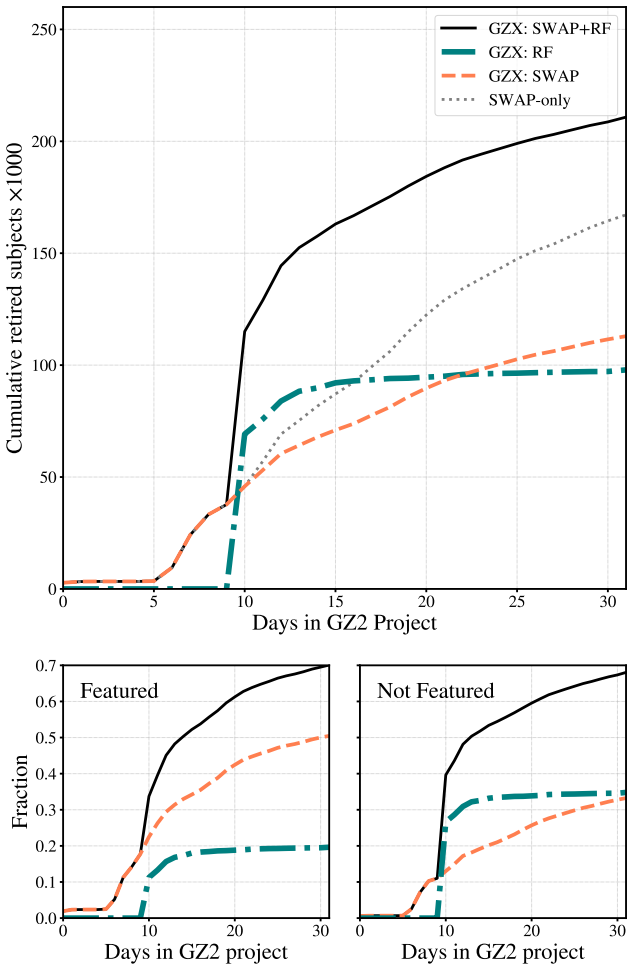

Figure 4.3 Contributions to subject retirement by both classifying agents of GZX: human (SWAP) and machine. The top panel shows cumulative subject retirement for GZX as a whole (solid black), along with that attributed to the RF (dash-dotted teal), and SWAP (dashed orange). The dotted grey line shows 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 behaviors: 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.

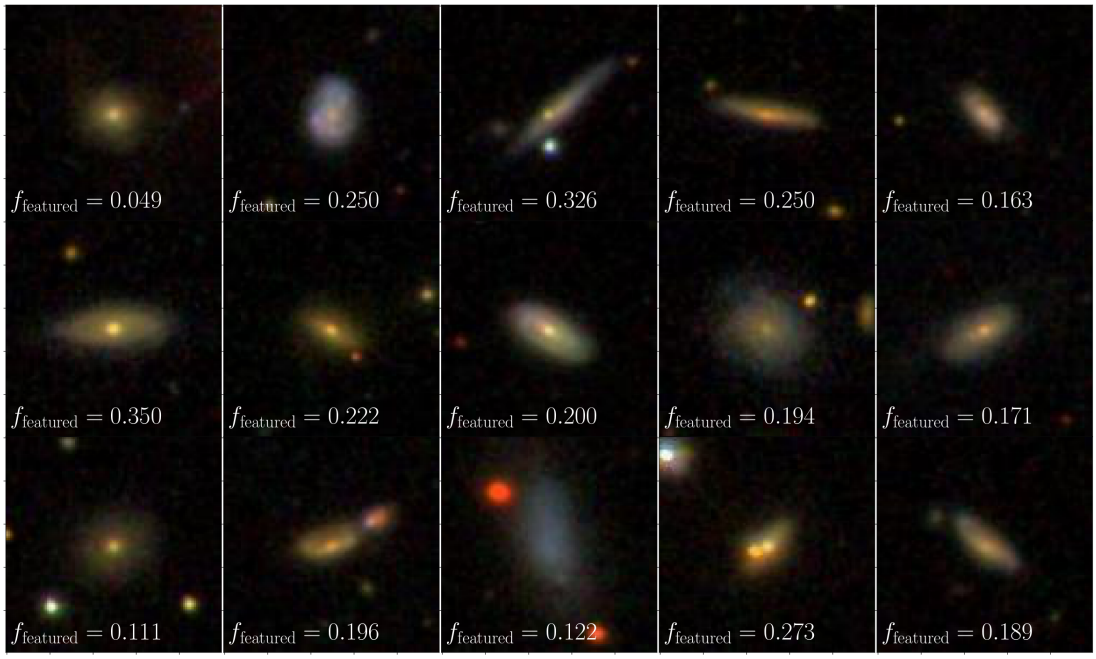

Figure 4.4 A random subsample of subjects identified as false positives: labeled by machine as 'Featured' but as 'Not' according to $\mathrm{GZ2_{raw}}$. We display $f_{\rm featured}$ in the lower left corner and choose a random sample of subject with $f_{\rm featured} < 0.35$, indicating that GZ2 volunteers strongly identified these as 'smooth' ('Not'). Fortunately, the machine is able to identify these subjects as 'Featured' due to their measured morphology diagnostics.

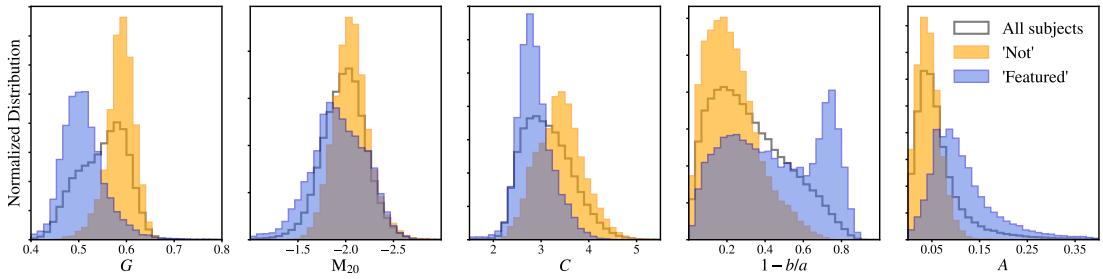

Figure 4.5 The RF is trained on a 5-dimensional morphology parameter space. We show the distribution of each morphology indicator for machine-retired 'Featured' (blue) and 'Not' (orange) subjects compared to the full GZ2 subject sample (black). The difference between 'Featured' and 'Not' subjects is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

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' ($p_{\rm machine} \geq 0.9$) where humans instead classify them as 'Not' ($f_{\rm featured} < 0.5$) 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 $0.4 < f_{\rm featured} + f_{\rm artifact} < 0.6$ 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 $0.4 \leq f_{\rm featured} + f_{\rm artifact} < 0.5$ 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 $f_{\rm featured} + f_{\rm artifact} \leq 0.35$, 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 $f_{\rm featured}$ 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 behavior 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',

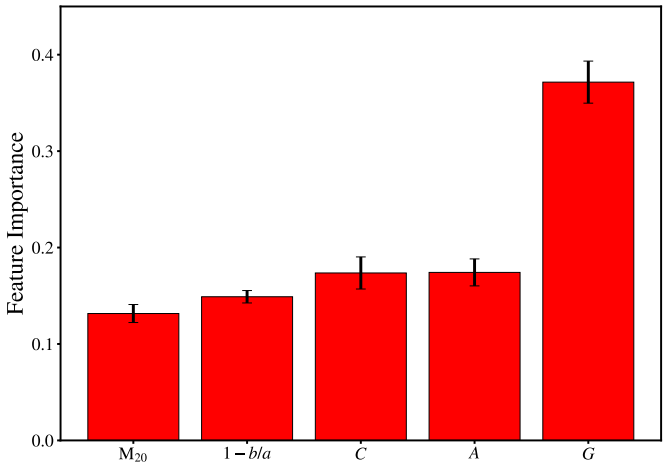

Figure 4.6 The RF's ranked feature importance averaged over all nights of training with black bars indicating the standard deviation. A larger value corresponds to higher importance. The machine computes feature importance according to how much each feature increases the purity of the resulting split 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.

it is likely the subject is disky 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 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. (2017) 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 system[2] 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

[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`

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. 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*.

Chapter 5

A population of "clumpy" galaxies in the 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 cosmic star-formation at $z \sim 1.5 - 2$ and now. Theoretically, galaxy growth is predominantly governed by the baryonic physics of gas accretion and energy feedback, rather than by mergers that induce starbursts (Somerville & Davé, 2015). Gas inflow modulates galaxies' star formation histories 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 (Oppenheimer et al., 2010; Bouché et al., 2010; Davé et al., 2011a,b; Lilly et al., 2013; Hirschmann et al., 2013).

Galaxies along the so-called star-forming main sequence at $z > 1.5$ (Noeske et al., 2007) tend to be thick, turbulent disks, with specific SFRs of the order of a Gyr, substantially higher than what is observed in the thin, quiescent, Milky-Way-like disks typical of the local universe (Genzel et al., 2008). The SFRs in these "clumpy" galaxies is concentrated in a few massive star-forming regions, usually dubbed as star-forming clumps, with sizes of approximately $100 - 1000$pc, stellar masses of $10^7 - 10^9 M_\odot$, and stellar ages typically younger than the age of the host galaxy (e.g., Elmegreen & Elmegreen, 2005; Genzel et al., 2011; Guo et al., 2012, 2015)

The origin of these massive clumps is still an open debate. Two main physical scenarios have been proposed. On the one hand, giant clumps are expected to form inside pre-existing galaxies. In these scenarios, halos located at the center of multiple filaments continuously accrete gas from the cosmic web, resulting in the formation of gas-dominated disks that eventually fragment into giant clumps through gravitational instability (Bournaud et al., 2007; Dekel et al., 2009b; Behrendt et al., 2015). Alternatively, 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. Large gas-to-baryonic fractions of 20% to 80% recently measured with interferometric studies in clumpy galaxies lends support to the in-situ models (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 $> 10 M_\odot$ yr^{-1} and $M > 10^{10} M_\odot$ 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 (Muratov et al., 2015; Christensen, 2011; Davé et al., 2016). These simulations show that in galaxies with $M < 10^{10}$ 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 hinge on the specifics of the dynamics of the mergers.

In this work we utilize the high quality data accumulated by SDSS for a sample of local clumpy galaxies discovered during the preparation of the Galaxy Zoo: Hubble morphology catalog which collected morphological data for galaxies in Stripe 82. Taking advantage of the increased depth of the Stripe 82 coadd imaging for over 100 galaxies, as well as high quality SDSS spectra for more than 150 "clumps", we explore the properties of this initial sample to assess how similar these objects are to their high redshift counterparts. Additionally, we seek to differentiate between possible clump formation mechanisms, especially for low mass galaxies of which our sample peaks around $9M_{\odot}$, precisely the range for which observations are sorely needed. In Section 5.2 we detail the galaxy selection process and data utilized in this work, followed by a preliminary analysis of various clump properties in Section 5.3. We next discuss the *Clump Scout* project in Section 5.5: an initiative to discover up to 1000 more low redshift clumpy galaxies hiding in SDSS imaging data. Finally, in Sections 5.6 and 5.7 we present our conclusions and future work. Throughout this chapter we assume a flat Planck cosmology (Planck Collaboration et al., 2016) 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 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 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 sizable sample of tight groups of elliptical galaxies, as well as galaxies in various merging states, possessing multiple nuclei. After excluding these and duplicate imaging, we

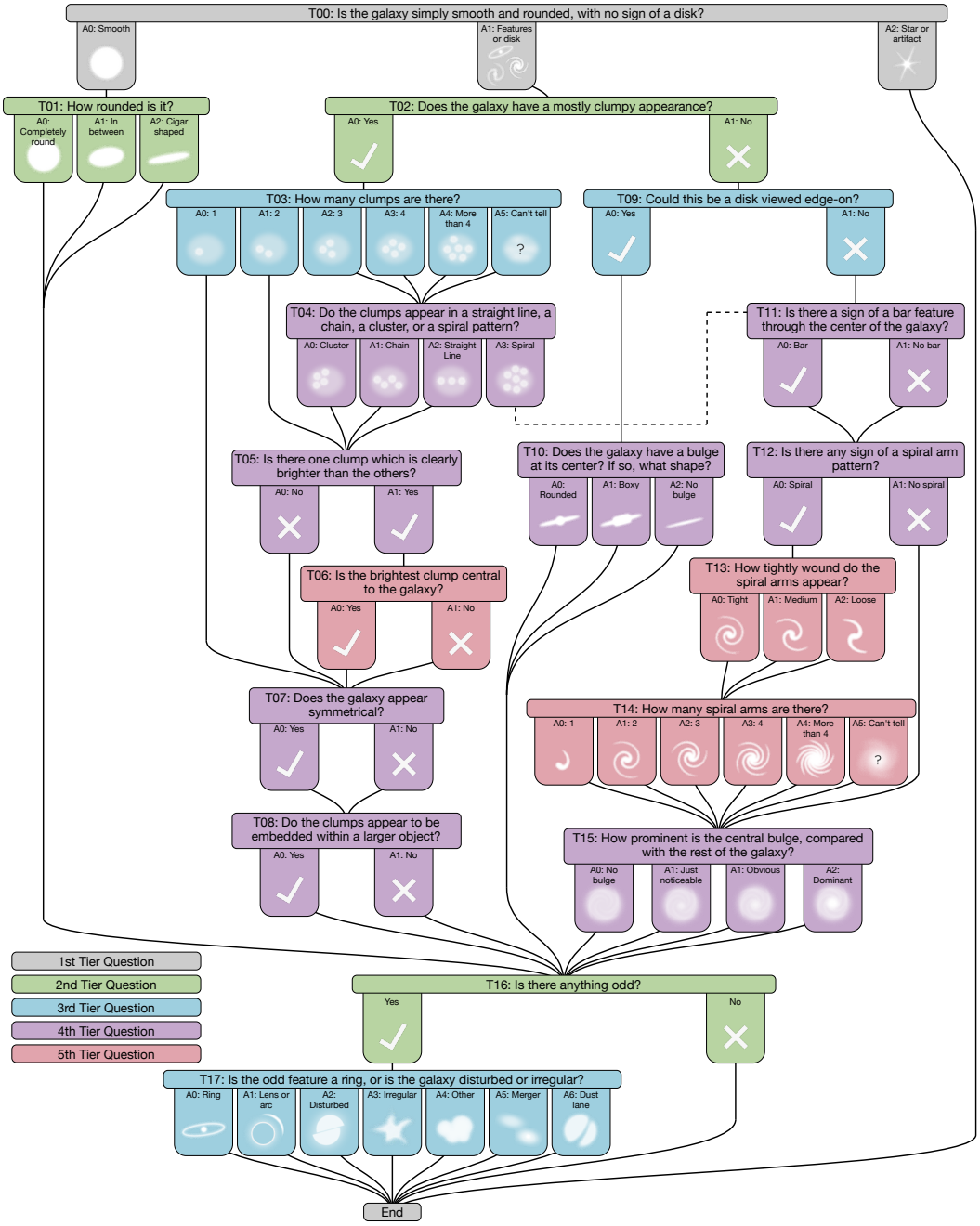

Figure 5.1 Galaxy Zoo: Hubble decision tree. The most notable difference between this decision tree and that used during the Galaxy Zoo 2 project is the "clumpy" branch of tasks.

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 104 unique galaxies. SDSS *gri* color-composite jpegs for all galaxies can be found at the end of this work in Section 5.8 where SDSS's spectroscopy flag is enabled, generating a red square around each spectroscopic target.

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 $3800 - 9200$Å with a resolution of $R \sim 1500$ at ~ 3800Å. 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 shows 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 $3\times$`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

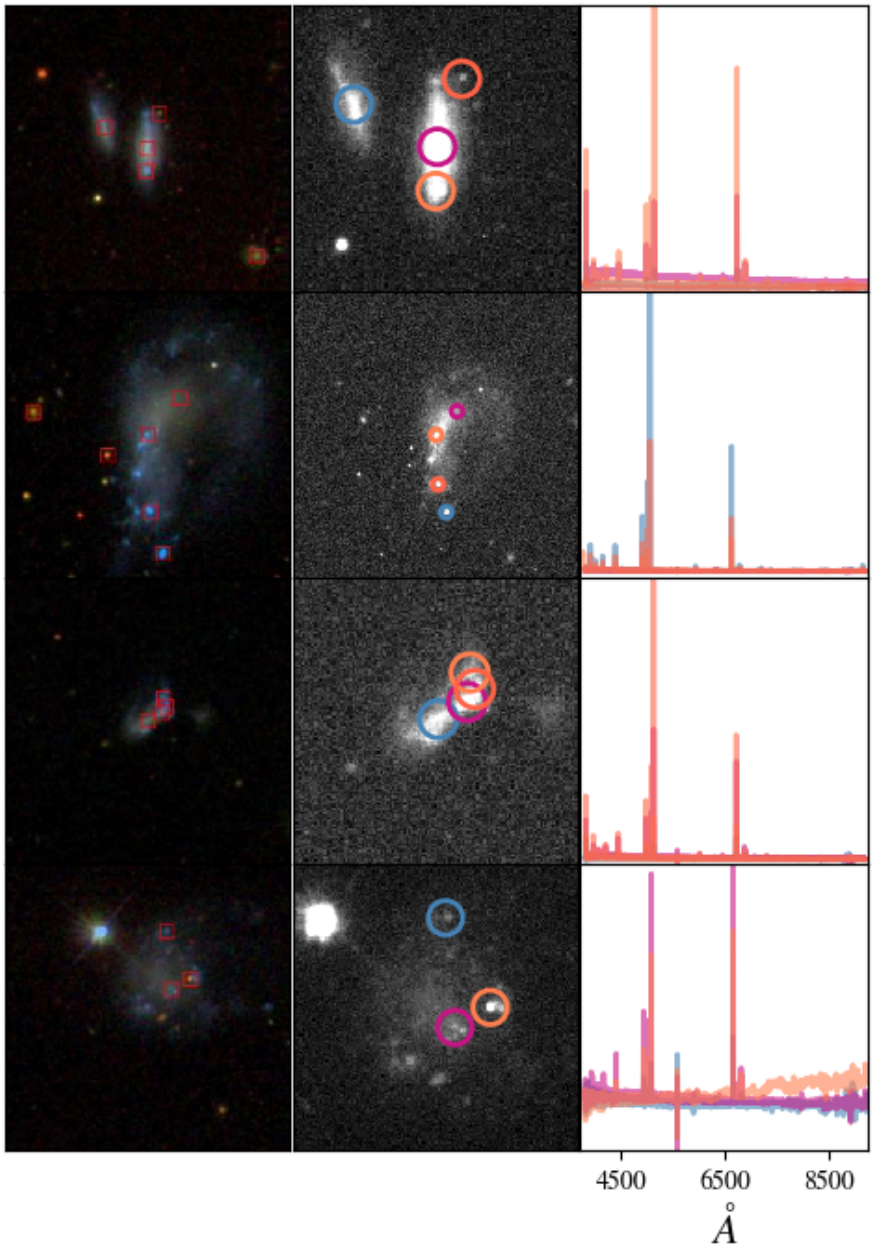

Figure 5.2 Sample of "clumpy" galaxies. The first column shows the SDSS *gri* color composite image where the red squares denote locations of SDSS spectroscopy. The middle column shows our postage stamp for the object where the colored circles show the same spectral locations, however these circles are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

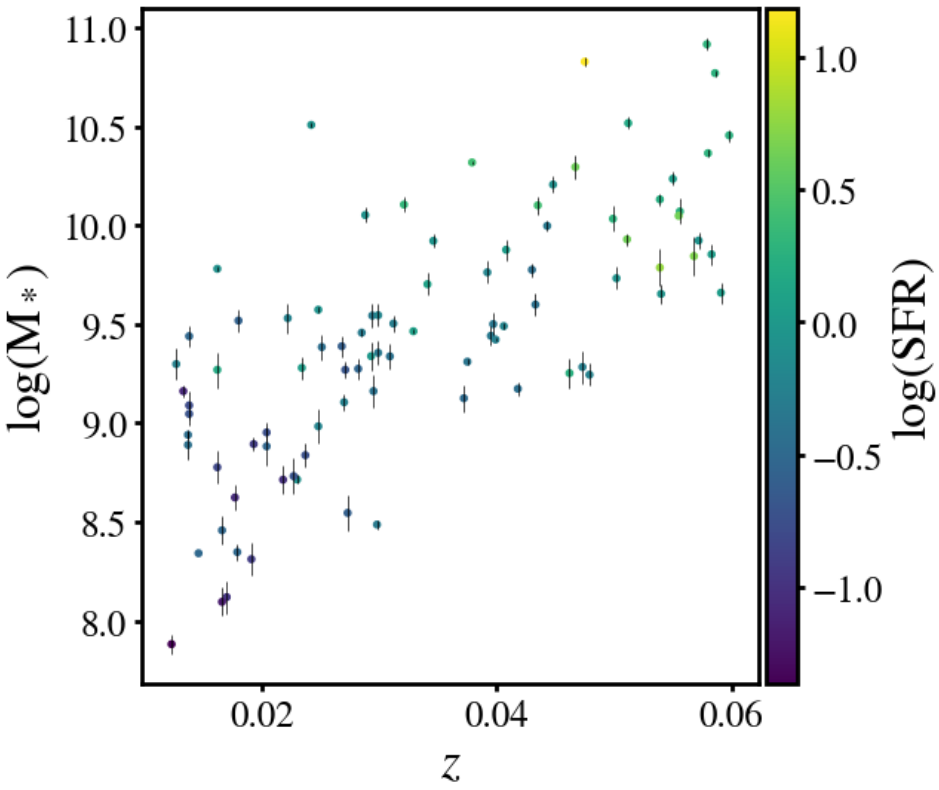

Figure 5.3 Stellar mass, redshift, and star-formation rate (SFR) for a subset of 89 "clumpy" galaxies as estimated by Salim et al. (2016) in the GSWLC. With a median log $M_{\odot} = 9.4$, our sample includes dozens of low mass galaxies that will be crucial for helping to constrain models of star-formation in low mass galaxies.

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-lines. 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 for our sample is $\sim 9.4 M_{\odot}$.

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

Several studies have been conducted which examine the properties of both local HII regions and high-redshift star-forming clumps (e.g., Monreal-Ibero et al., 2007; Swinbank et al., 2009; Wisnioski et al., 2012) but the debate between whether these regions arise due to the same physical processes remains unresolved. That these regions arise through different physical mechanisms would not be surprising given the stark changes in the global environments of galaxies between those at Cosmic Noon and of today. In particular, gas accretion rates which drive up turbulence and sustain disk instabilities are expected to dissipate on cosmic timescales, at least for massive galaxies (Dekel & Birnboim, 2006). However, Wisnioski et al. (2012) show that despite these changes, similarities between local HII regions and clumps at $z \sim 1.3$ persist in the form of several elegant scaling relations. In this section we revisit some of these relations with the goal of ascertaining whether the star forming regions in our sample could pose as analogs of high redshift clumps.

Specifically, we exploit the high quality emission line measurements produced through the SDSS spectroscopic pipeline. Most of these fibers are either centered on a bright star-forming region which we interpret as a "clump," or on/near the galactic center (see Section 5.8 for reference). If clump lifetimes are longer than the dynamical time of the galaxy it is possible that they migrate to the galactic center, contributing to bulge formation. We thus do not remove those spectra that are close to the central regions of their host galaxy so as not to bias ourselves against this possibility.

5.3.1 H_α luminosity and velocity dispersion

Wisnioski et al. (2012) develop empirical scaling 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 H_α luminosity for thousands of local HII regions and high-redshift clumps taken from 11 different studies. We briefly describe two of these relations here.

If a star-forming region can be idealized as a Strömgren sphere of radius, r, then it is straightforward that a scaling relation between H_α luminosity and the size of the region should exist. A Strömgren sphere describes the region of ionized HII surrounding hot, young O-B stars and the H_α luminosity of such a region scales as r^3, where this radius is assumed to be the clump radius. Under this assumption, (Wisnioski et al., 2012) fit the following relation:

$$\log(L_{\mathrm{H}_\alpha}\ [\mathrm{ergs}^{-1}]) = 2.72 \times \log(\mathrm{d}\ [\mathrm{pc}]) + 31.99 \tag{5.1}$$

where d is the clump diameter in units of pc.

They derive a similar relation between velocity dispersion and size under the assumption that these regions form out of a Jeans collapse in an isothermal disk. Under this scenario, the clump radius is assumed to correspond to $\lambda_\mathrm{J}/2$, where λ_J is the Jeans length which scales with velocity dispersion as σ^2. With this assumption, Wisnioski et al. (2012) determine a velocity dispersion-size relation according to

$$\log(\sigma\ [\mathrm{km\ s}^{-1}]) = 0.42 \times \log(\mathrm{d}\ [\mathrm{pc}]) + 0.33 \tag{5.2}$$

where again d is the clump diameter in units of pc.

We begin by estimating the H_α flux and velocity dispersion for each spectrum as measured by a Gaussian fit to the emission line in the SDSS spectroscopic pipeline. Additionally we correct the H_α velocity dispersion by subtracting in quadrature the SDSS instrument resolution. In the bottom left panel of Figure 5.4 we show the distribution of the H_α velocity dispersion for all spectra in our sample, where the median is $\sigma_{\mathrm{H}_\alpha} \sim 40$km/s. While this is slightly smaller than giant star-forming regions observed at $z \sim 2$ (Elmegreen & Elmegreen, 2010; Sánchez et al., 2013), it is also significantly larger

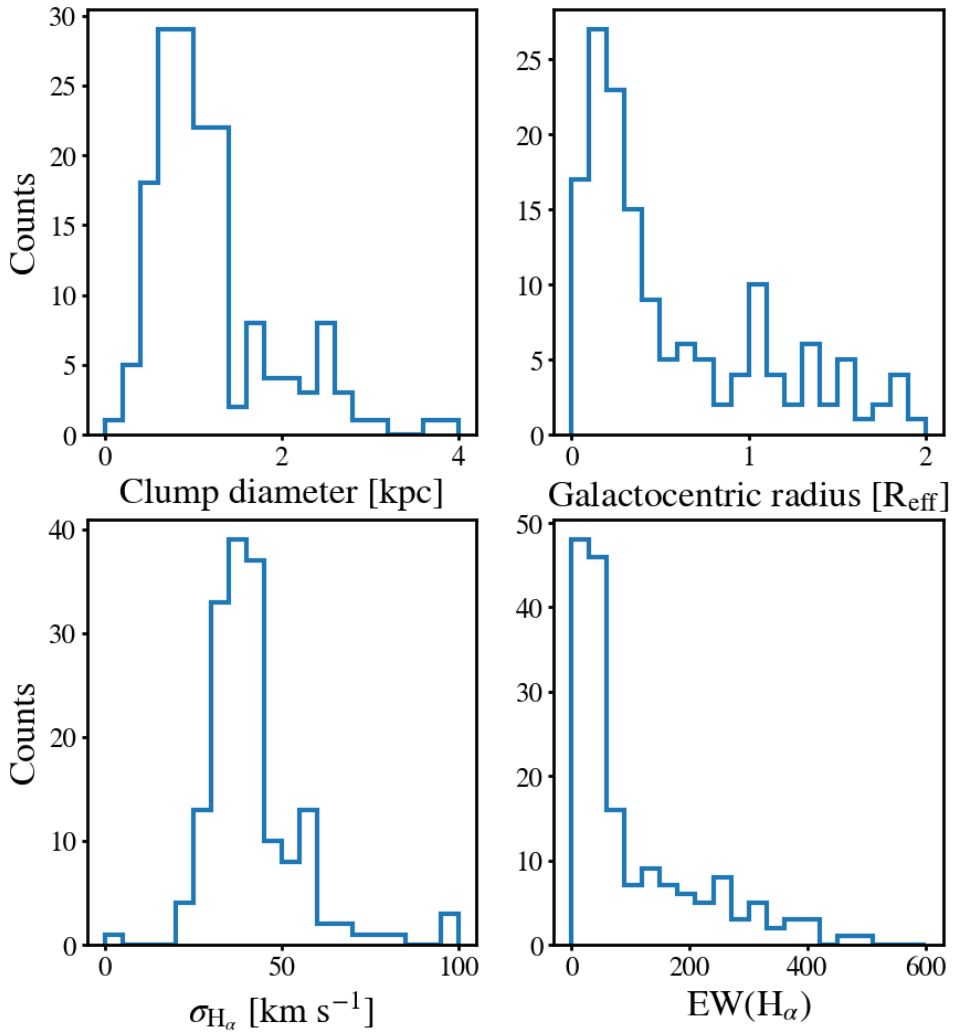

Figure 5.4 We present here distributions of several physical properties of the ∼170 star-forming regions in our sample. These properties (with the exception of galactocentric radius, see Section 5.4) are all derived from emission-line measurements produced by the SDSS spectroscopic pipeline.

than typical HII regions observed in local galaxies. From this, we estimate the clump diameter according to equation 5.2. The resulting clump diameter distribution is shown in the top left 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 systems since we do not specifically separate them as discussed earlier. The central regions will typically have much lower rates of star formation, smaller H_α velocity dispersion, and thus significantly smaller sizes according to this relation. If these regions were excluded, the median diameter would be more typical of high-redshift clumps. In a more detailed analysis, clump sizes will be confirmed independently through a "core" analysis whereby sizes are measured by fitting a Gaussian to the 1D radial surface brightness profile of each star-forming region.

The left panel of Figure 5.5 shows the relationship between the H_α luminosity and the clump diameters derived from the H_α velocity dispersion. The black dashed line represents the luminosity-size scaling relation of equation 5.1. The points are color-coded according to the stellar mass of the host galaxy where the dark purple points are those spectra whose corresponding galaxy's do not have a match in the GSWLC. The clumps in our sample lie reasonably well along the empirical relation of Wisnioski et al. (2012) though much of the sample has lower H_α luminosity. The H_α luminosity distribution overlaps with that of high redshift clumps, though the size distribution is intermediate between local and high-redshift star-forming regions, as previously discussed. There is also a strong galaxy stellar mass dependence along the relation. This isn't altogether surprising since lower mass galaxies typically have smaller SFRs in our sample (as evidenced by the global SED SFR in Figure 5.3) and, by extension, lower H_α luminosities.

Several observations have detected a relationship between luminosity and velocity dispersion of galaxies both locally and at high redshift (Dib et al., 2006; Green et al., 2010; Lehnert et al., 2009; Genzel et al., 2011). It has been suggested that the velocity dispersion is driven by star formation. One possibility is through released mechanical energy as discussed in Lehnert et al. (2009). Wisnioski et al. (2012) offer an alternative whereby the velocity dispersion is the result of the star formation density in a marginally stable galactic disk (i.e., Toomre parameter, $Q \sim 1$), assuming the Kennicutt-Schmidt

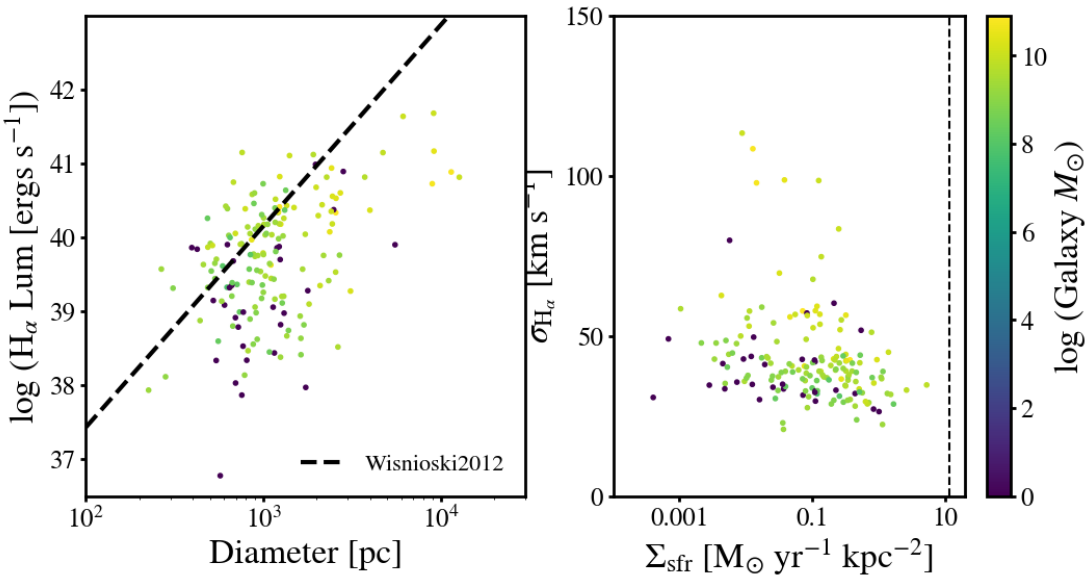

Figure 5.5 In the left panel we show the H_α luminosity versus clump size as estimated via the H_α velocity dispersion utilizing equation 5.2. The black dashed line represents the luminosity-size relation of Wisnioski et al. (2012), reproduced in equation 5.1. The left panel shows the H_α velocity dispersion versus the clump star formation rate density. In bath panels the points are colored according to host galaxy total stellar mass as measured by Salim et al. (2016) in the GSWLC. The dark purple points are those clump spectra whose host galaxy did not match to that catalog.

law, an empirical relation between SFR surface density and gas surface density (Schmidt, 1959; Kennicutt, 1998b). While Wisnioski et al. (2012) find a moderate correlation between $\sigma - \Sigma_{\rm SFR}$, Genzel et al. (2011) find a weak dependence concluding that feedback does not appear to directly drive the velocity dispersion of the ionized gas. Utilizing the clump sizes measured previously, we estimate the $\Sigma_{\rm SFR}$ after computing the H_α SFR using the well known calibration (Kennicutt, 1998a; Calzetti, 2013)

$$\mathrm{SFR}(\mathrm{M_\odot yr^{-1}}) = 5.5 \times 10^{-42} L(\mathrm{H}_\alpha)(\mathrm{ergs\ s^{-1}}) \tag{5.3}$$

The right panel of Figure 5.5 shows the resulting $\sigma_{\mathrm{H}_\alpha} - \Sigma_{\rm SFR}$ relation, where again the points are colored according to the host galaxy stellar mass as given by the GSWLC with purple points denoting those spectra for which the host galaxy is not found in that catalog. Instead, we observe a mild anti-correlation. If this finding holds, it could be an indication that feedback mechanisms in low mass galaxies are poorly understood since our sample probes much lower galaxy stellar mass and it is precisely this population that skews the relation.

Finally, we compare SFRs between those derived from the H_α emission and the global SED SFR from the GSWLC. For each galaxy, we sum the $\mathrm{SFR}_{\mathrm{H}_\alpha}$ from each spectrum and compare that to $\mathrm{SFR}_{\rm SED}$, for those 89 galaxies that were matched to the GSWLC. The median ratio of these two star-formation measures is $\sim$10%. Because a majority of galaxies have only one spectrum, we interpret this rough estimate as being 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).

5.4 Physical properties vs galactocentric radius

Two broad scenarios for giant clump formation make distinct predictions for the distribution of clump physical properties. If most clumps have an ex-situ origin then there should exist a random distribution of clump properties. On the other hand, most in-situ models suggest that clumps tend to be long-lived (few hundreds Myrs) and migrate to the galaxy center where they coalesce to form a bulge (Bournaud, 2016). An age gradient is therefore expected with older clumps preferentially located closer to the galactic

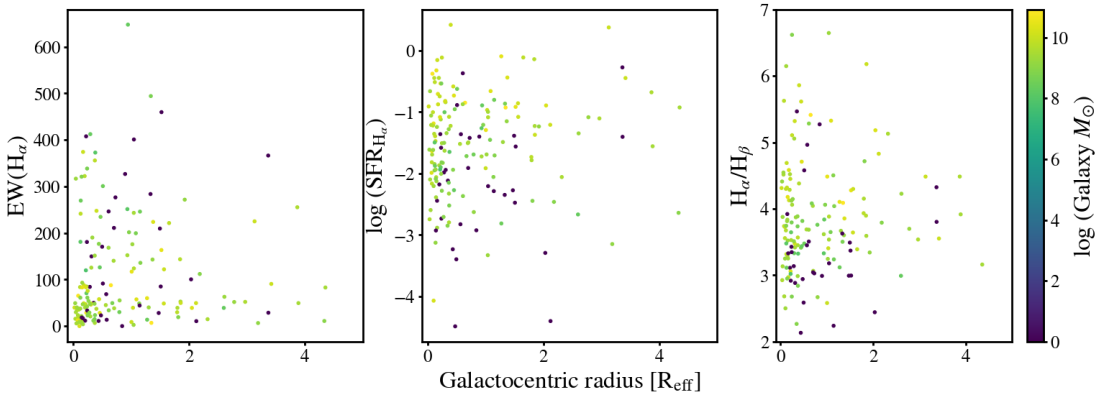

Figure 5.6 H_α equivalent width, star formation rate, and Balmer decrement as a function of clump galactocentric radius normalized by the host galaxy's R_{eff}. In all panels the points are colored according to host galaxy total stellar mass as measured by Salim et al. (2016) in the GSWLC. The dark purple points are those clump spectra whose host galaxy did not match to that catalog.

center. In this section we search for other possible trends with galactocentric radius though in Section 5.6 we discuss the next phase of this project which is to measure the metallicity and stellar ages of these clumps to further constrain clump formation theories.

We compute the galactocentric distance of each SDSS fiber from the galaxy center as measured by SExtractor, rather than the coordinates provided by SDSS since it was found that some of these coordinates were centered improperly, as discussed in Section 5.2. We compute the galactocentric distance both in kpc and normalized by the galaxy's half light radius (`FLUX_RADIUS`) as determined by SExtractor. The resulting distribution is shown in the top right panel of Figure 5.4. That the distribution is obviously skewed towards the small galactocentric values is likely due to the several spectra that probe the central region. In Figure 5.6 we show several physical properties of clumps as a function of galactocentric radius where the color scheme is the same as described in the previous section. Though no strong trends are seen, some curious insights can still be gleaned.

The first panel of Figure 5.6 shows the restframe H_α equivalent width versus galactocentric radius. This 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). We find that clumps with the strongest EW are more likely to be closer to the galactic center and that clumps far from the center are dominated by low EW. That we find few examples of high EW regions at large radius is interesting as SDSS should be sensitive to this region of the parameter space. In the middle panel we show the relation between the H_α SFR and galactocentric radius wherein we find that, at large radius, clumps are slightly more likely to have high SFRs. Similar results are found by Soto et al. (2017) where that trend was accompanied by a trend of young clump ages preferentially located at greater galactocentric radius. Such a scenario is consistent with clump migration theories. There also exists a mild trend with global galaxy stellar mass but this is expected as regional SFRs correlate with the global property as previously discussed. In the final panel, we show the Balmer decrement (H_α/H_β), which probes galactic attenuation. We find that the highest levels of attenuation are found in regions located closest to the galactic center but that significant levels are found in clumps located at large radii. Furthermore there is a mild galaxy stellar mass dependence wherein lower mass galaxies typically have regions with lower Balmer decrement.

Finally, we compare our sample with the full SDSS Data Release 8 spectroscopic catalog (black) on the BPT diagram in Figure 5.7. The BPT diagram utilizes a set of nebular emission lines to distinguish between various ionization mechanisms of nebular gas (Baldwin et al., 1981). Though several versions exist, we use the most common [OIII]/H_α versus [NII]/H_β. As both ratios are based on lines close in wavelength, effects due to reddening are negligible. Here our regions are colored according to galactocentric radius. As expected, the majority of our clumps lie comfortably in the star-forming region of the BPT diagram but there does not appear to be a correlation with clump galactocentric radius. This is not necessarily surprising since, to our knowledge, no trend on clump specific ionization state has been predicted.

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. However, the area covered by Stripe 82 was only a fraction of the full SDSS sky coverage. In this section we describe

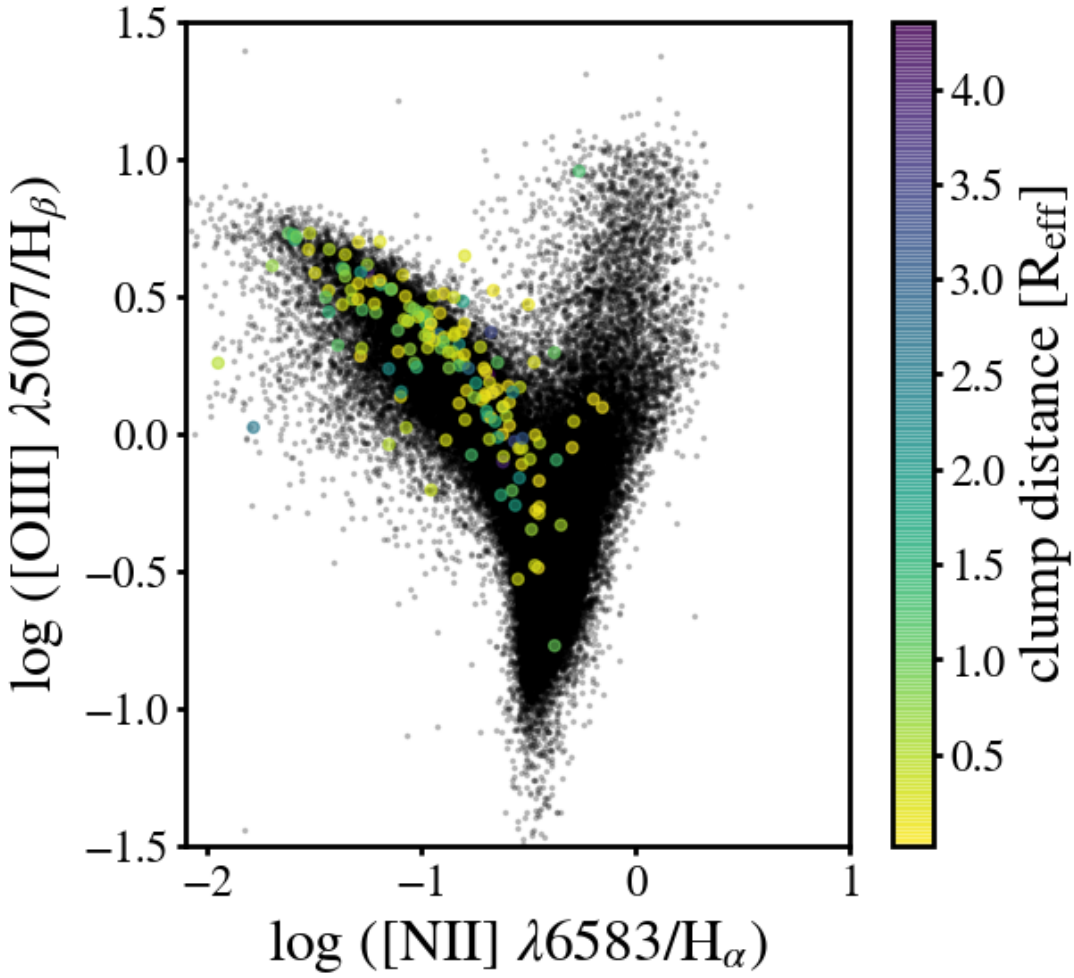

Figure 5.7 BPT diagram of the spectra in our sample color-coded according to clump galactocentric radius against all SDSS spectroscopic targets (black). As expected, the majority of the regions in our sample are well situated on the star-forming branch of the diagram though there does not appear to be any relation with galactocentric distance.

the *Clump Scout* project, a citizen-science initiative to discover more "clumpy" galaxies in the remaining SDSS footprint. We begin with the remaining non-Stripe 82 SDSS galaxies that were originally part of the GZ2 project. We exclude any galaxies with $z > 0.06$ in order to satisfy the resolution requirements discussed above. We also make a cut such that $f_{\rm smooth} \leq 0.8$ in order to exclude those galaxies which are obviously elliptical and would likely not have traveled down the clumpy track of the GZH decision tree. These criteria yield a sample of $\sim$ 63K galaxies as shown in Figure 5.8 where we depict galaxy number density contours in the z-$f_{\rm smooth}$ 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.

5.6 Future work

The analysis presented thus far is unable to constrain clump formation models. In this section we detail the next phase of this work for which we recently received an NSF grant. The in-situ and ex-situ models for giant clump formation make distinct predictions for the distribution of the clumps metallicities and stellar ages. If clumps really formed within their host galaxy, statistically some newly formed ones (age $<$ 20

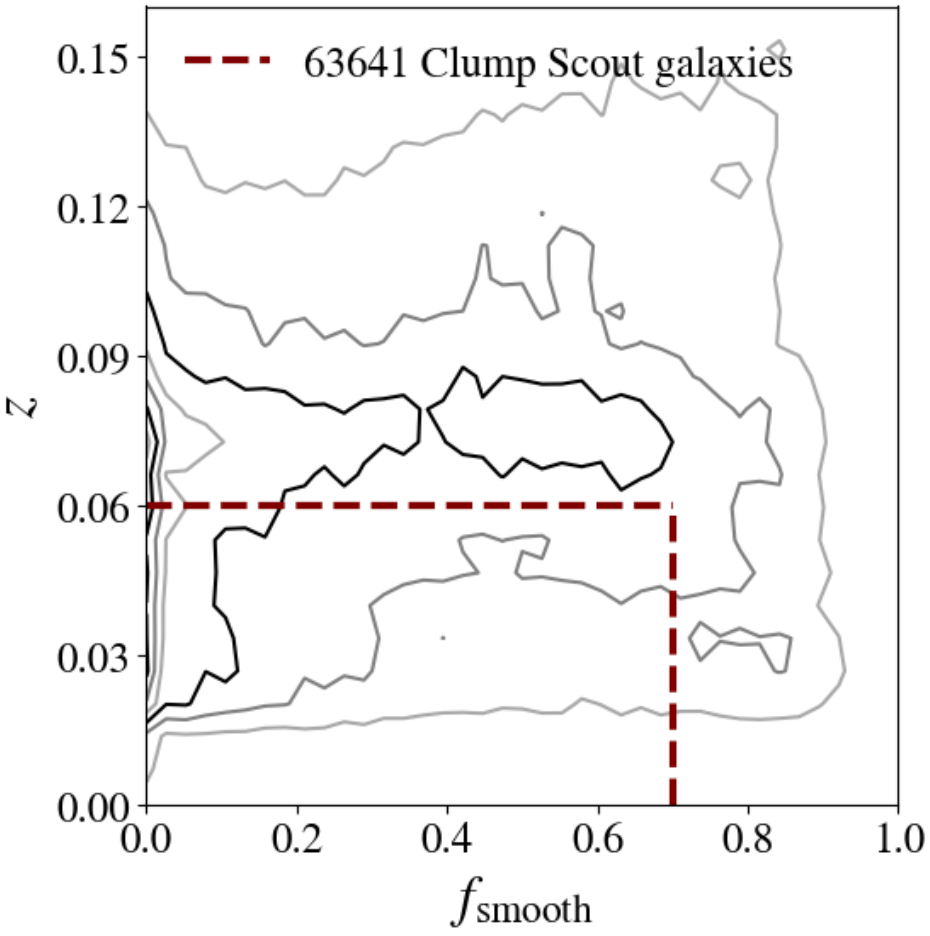

Figure 5.8 *Clump Scout* sample selection criteria from non-Stripe 82 GZ2 galaxy sample.

Myr) should be observed within the preexisting galactic disk. Also, the metallicity of the gas associated with the youngest clumps should be similar to the metallicity of the gas surrounding the clumps. Conversely, if the observed giant clumps are of external origin, they would be expected to show a broad range of stellar ages and metallicities, and often large deviations from the host disk velocity field (Bournaud et al., 2008; Wuyts et al., 2012). Additionally, these properties are expected to vary with the galactocentric radius because in in-situ models, clumps tend to be long-lived (few hundreds Myrs) and migrate to the galaxy center where they coalesce to form a bulge (Bournaud, 2016). An age gradient is therefore expected, with older clumps being preferentially located closer to the galaxy center. Recently, however, this paradigm has been questioned as simulations incorporating more realistic feedback find that clumps are not long-lived (Genel et al., 2012; Hopkins et al., 2012; Oklopčić et al., 2017). Although some observational constraints do exist, they are mostly based on studies of broad band images, and very limited spectroscopic data are available (Guo et al., 2012; Zanella et al., 2015).

The quality of the SDSS spectra will allow us to derive accurate stellar ages (from fitting of the continuum) as well as gas metallicity (Henry et al., 2015). By combining our current sample with those discovered through the *Clump Scout* project, we will be able to derive statistical trends in the clumps properties. We will group the clump properties according to galaxy stellar mass, distance from the galaxy center, and environmental density.

Additionally, this large statistical sample will allow us to investigate the evolution of the fraction of clumpy galaxies as a function of cosmic time, $f_{\rm clumpy}$ (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 ($f_{\rm clumpy}$) can be used to place constraints on theoretical models. By focusing on the redshift range between $0.5 < z < 3$, they find that $f_{\rm clumpy}$ changes with the stellar mass of the galaxies. Low-mass (log M $<$ 10 M$_{\odot}$) galaxies keep an almost constant $f_{\rm clumpy}$ of $\sim$60% from $z \sim 3$ to $z \sim 0.5$, while massive galaxies drop their $f_{\rm clumpy}$ from 55% at $z \sim 3$ to 15%, at $z \sim 0.5$. 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 $f_{\rm clumpy}$ evolution in low mass galaxies is more consistent with a minor merger origin. 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 $f_{\rm clumpy}$ measured from a variety of different surveys at $z < 1$, the result is not so convincing anymore. The variation in the low-z measurements quite large and is primarily 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 $f_{\rm clumpy}$ can provide the strongest model constraints.

5.7 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 104 galaxies that have a traditional clumpy morphology and acquire SDSS imaging and spectroscopic data for these galaxies, including spectra for over 150 clumps. These galaxies have a median log stellar mass of 9.4, a sample that could provide crucial constraints on star formation models of low mass galaxies. We obtain stellar mass and star-formation rates for these galaxies through the GSWLC catalog. Our preliminary analysis of the spectra reveals that star-forming regions in these galaxies are in many ways similar to high-redshift clumps: our clumps have a median physical size of $\sim$ 1kpc and median H_α velocity dispersion of $\sim$ 40km/s, values that are only slightly lower than high redshift clumps. We also compare several spectroscopically derived properties with clump galactocentric radius but find no strong correlations.

These findings motivate not only a more in depth analysis of this sample but also justify the search for similar galaxies in the local universe. To that end we described 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 expect to find an additional 1000 clumpy galaxies, with upwards of $\sim$ 1500

associated spectra. This will provide a large statistical sample of local clumpy galaxies which we will use to constrain the fraction of clumpy galaxies at low redshift, as well as investigate the age and metallicity of individual clumps as a function of galaxy stellar mass, galactocentric distance, and environmental density, allowing us to distinguish between various clump formation theories.

5.8 Visual catalog of SDSS Stripe 82 clumpy galaxies

Here we present the full sample of 104 clumpy galaxies found as part of the selection process described in Section 5.2. These images are generated through SDSS's jpeg service and are *gri* color-composite images, where the galaxy of interest is generally at the center of the image. We include the spectroscopic flag which generates a red square over every spectroscopic target in SDSS. We note that not all of these spectra are kept. For each galaxy we visually inspect the spectra and match it to a region within a galaxy in our sample, excluding spectra of stars or of other nearby galaxies. This sample is organized by redshift, beginning with the closest objects.

Figure 5.9 SDSS *gri* color-composite images of all galaxies in our sample, organized by redshift with the closest objects first. The red squares are generated via SDSS's spectroscopic flag, identifying all spectroscopic targets in the SDSS sample. We exclude those spectra which are obviously centered on a galaxy outside of our sample or classified by SDSS as a star.

Figure 5.10 Same as Figure 5.9.

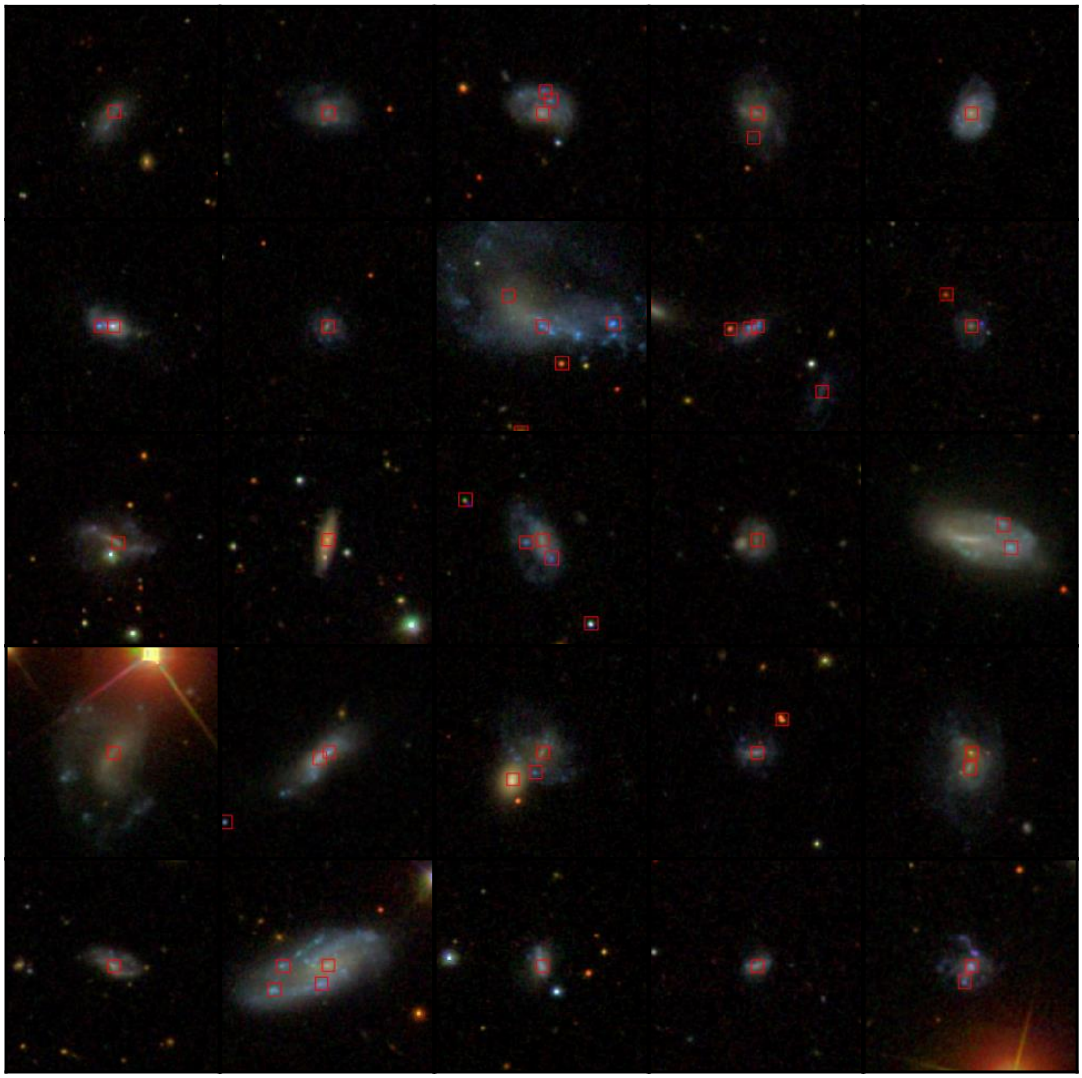

Figure 5.11 Same as Figure 5.9.

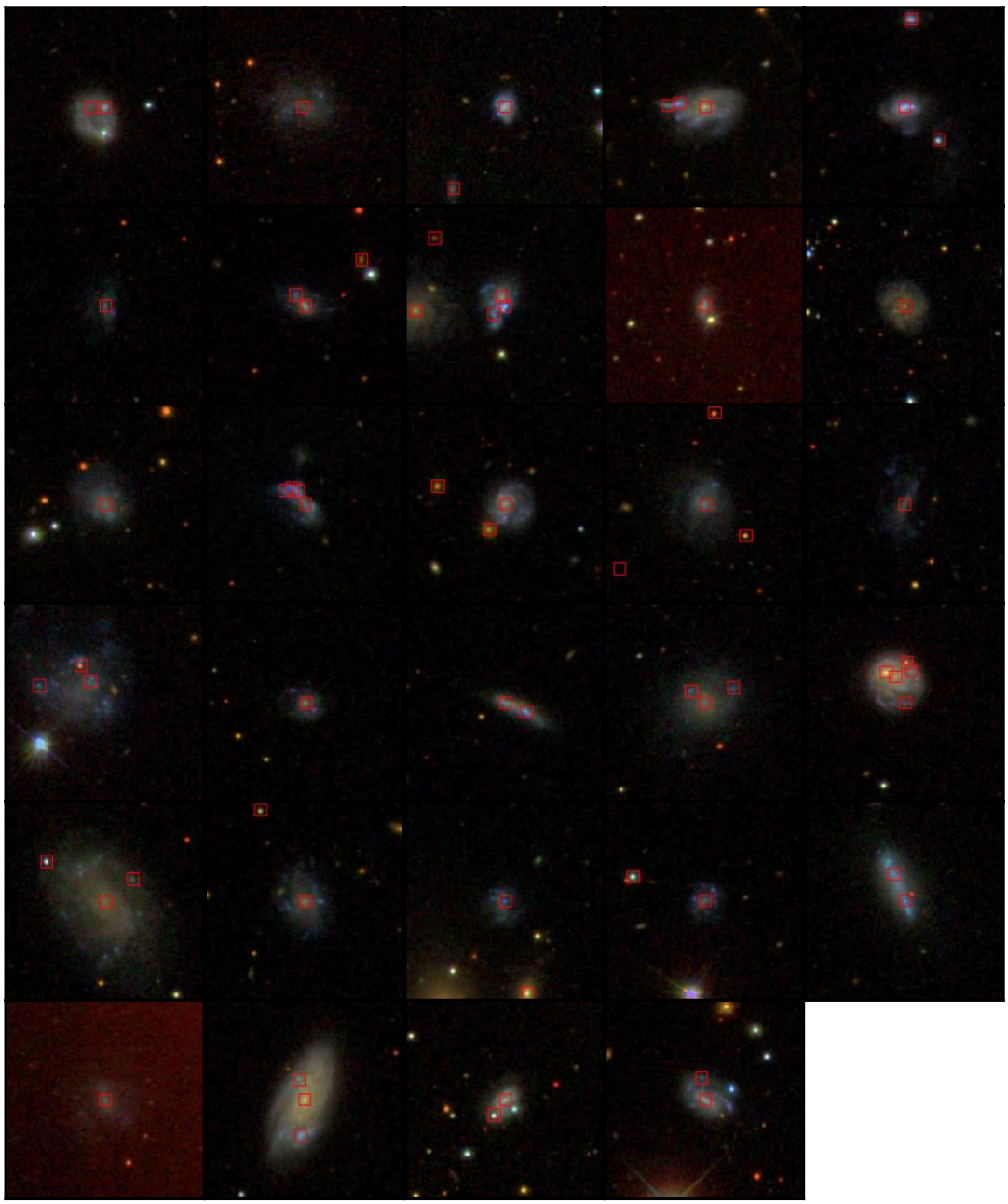

Figure 5.12 Same as Figure 5.9.

Chapter 6

Summary & Future Work

The goal of this thesis was to develop a framework that brings both human and machine intelligence to the task of galaxy morphology classifications to handle the scale and scope of next generation of astrophyical surveys. We demonstrate the effectiveness of our solution, Galaxy Zoo: Express (GZX), through a re-analysis of visual galaxy morphology classifications collected during the Galaxy Zoo 2 crowd-sourced 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.

We measure traditional automated morphologies for ∼280K SDSS galaxies to use as features in our Random Forest algorithm. Specifically we measure the concentration index, asymmetry, M_{20}, Gini coefficient, and ellipticity; parameters that are well known to correlate strongly with the distinction between early- and late-type galaxies. We present a catalog of all such measurements and demonstrate that they are generally robust against various possible failure mechanisms we explore.

We demonstrate that classification efficiency can be increased through intelligenit t management of visual classifications from volunteers. We examine this through several simulations whereby we re-process Galaxy Zoo 2 classifications with SWAP, a Bayesian code first developed for the Space Warps gravitational lens discovery project (Marshall et al., 2016). We demonstrate for the first time that the SWAP algorithm 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 achieve a reduction of at least a factor of three in the human effort necessary to maintain that increased rate, all while maintaining 95% classification 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%. Exploring those subjects wherein SWAP and GZ2 disagree, we conclude that the majority of this disagreement stems from the stochastic nature of GZ2 vote fractions from which we assign binary labels.

We implement and test a Random Forest algorithm and develop a decision engine that delegates tasks between human and machine. After a sufficient number of subjects have been classified by humans and processed through SWAP, the machine is trained and its performance assessed through cross-validation. 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. 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.

GZX is perhaps one of the simplest ways to combine human and machine intelligence and it(s 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. The most likely candidate are deep convolutional neural networks (CNNs) such as those employed in recent studies with high levels of accuracy (Domínguez Sànchez et al., 2017; Huertas-Company et al., 2008). The drawback to most deep CNNa is the computational cost required to train such a model. Our Random Forest requires a fraction of the computational power as a CNN and fewer training samples as well, thus demonstrating that it should be considered as a vaiable appropach to morphology classification. However, 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.

Finally, we identify a sample of "clumpy"vgalaxies in the local universe through the Galaxy Zoo: Hubble project which included imaging from Stripe 82 of the SDSS. We isolate 104 galaxies that have a traditional clumpy morphology and acquire SDSS imaging and spectroscopic data for these galaxies, including spectra for over 150 clumps. These galaxies have a median log stellar mass of 9.4, a sample that could provide crucial constraints on star formation models of low mass galaxies. We obtain stellar mass and star-formation rates for these galaxies through the GSWLC catalog (Salim et al., 2016). Our preliminary analysis of the spectra reveals that the star-forming regions in these galaxies are in many ways similar to high-redshift clumps: our clumps have a median physical size of $\sim$ 1kpc and median H_α velocity dispersion of $\sim$ 40km/s, values that are only slightly lower than high redshift clumps. We also compare several spectroscopically derived properties with clump galactocentric radius but find no strong correlations.

These findings motivate not only a more in depth analysis, but also justify the search for similar galaxies in the local universe. To that end we described 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 expect to find an additional 1000 clumpy galaxies, with upwards of $\sim$1500 associated spectra. This will provide a large statistical sample of local clumpy galaxies which we will use to conduct several additional studies for which we have recently recieved a NSF grant. In particular we will constrain the fraction of clumpy galaxies at low

redshift, as well as investigate the age and metallicity of individual clumps as a function of galaxy stellar mass, galactocentric distance, and environmental density, allowing us to distinguish between various clump formation theories.

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

Amorín, R. O., Pérez-Montero, E., & Vílchez, J. M. 2010, ApJ, 715, L128

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

Baldwin, J. A., Phillips, M. M., & Terlevich, R. 1981, PASP, 93, 5

Ball, N. M., Loveday, J., Fukugita, M., et al. 2004, MNRAS, 348, 1038

Balogh, M., Eke, V., Miller, C., et al. 2004, MNRAS, 348, 1355

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

Barnes, J. E. 1992, ApJ, 393, 484

Basu-Zych, A., & Scharf, C. 2004, ApJ, 615, L85

Beck, M., Scarlata, C., Hayes, M., Dijkstra, M., & Jones, T. J. 2016, ApJ, 818, 138

Behrendt, M., Burkert, A., & Schartmann, M. 2015, MNRAS, 448, 1007

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

Bernardi, M., Meert, A., Sheth, R. K., et al. 2013, MNRAS, 436, 697

Bershady, M. A., Jangren, A., & Conselice, C. J. 2000, AJ, 119, 2645

Bertin, E., & Arnouts, S. 1996, A&AS, 117, 393

Binney, J. 1977, ApJ, 215, 483

Birnboim, Y., & Dekel, A. 2003, MNRAS, 345, 349

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

Bouché, N., Dekel, A., Genzel, R., et al. 2010, ApJ, 718, 1001

Bournaud, F. 2016, in Astrophysics and Space Science Library, Vol. 418, Galactic Bulges, ed. E. Laurikainen, R. Peletier, & D. Gadotti, 355

Bournaud, F., Elmegreen, B. G., & Elmegreen, D. M. 2007, ApJ, 670, 237

Bournaud, F., Daddi, E., Elmegreen, B. G., et al. 2008, A&A, 486, 741

Bower, R. G., Lucey, J. R., & Ellis, R. S. 1992, MNRAS, 254, 601

Brammer, G. B., Whitaker, K. E., van Dokkum, P. G., et al. 2011, ApJ, 739, 24

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

Brinchmann, J., & Ellis, R. S. 2000, ApJ, 536, L77

Brook, C. B., Stinson, G., Gibson, B. K., et al. 2012, MNRAS, 419, 771

Brooks, A. 2010, in Astronomical Society of the Pacific Conference Series, Vol. 432, New Horizons in Astronomy: Frank N. Bash Symposium 2009, ed. L. M. Stanford, J. D. Green, L. Hao, & Y. Mao, 17

Buitrago, F., Trujillo, I., Conselice, C. J., & Häußler, B. 2013, MNRAS, 428, 1460

Bullock, J. S., Dekel, A., Kolatt, T. S., et al. 2001, ApJ, 555, 240

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

Caputi, K. I., Cirasuolo, M., Dunlop, J. S., et al. 2011, MNRAS, 413, 162

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

Ceverino, D., Dekel, A., & Bournaud, F. 2010, MNRAS, 404, 2151

Chapman, S. C., Scott, D., Windhorst, R. A., et al. 2004, ApJ, 606, 85

Christensen, C. R. 2011, PhD thesis, University of Washington

Cimatti, A., Daddi, E., & Renzini, A. 2006, A&A, 453, L29

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

Davé, R., Finlator, K., & Oppenheimer, B. D. 2011a, MNRAS, 416, 1354

Davé, R., Oppenheimer, B. D., & Finlator, K. 2011b, MNRAS, 415, 11

Davé, R., Thompson, R., & Hopkins, P. F. 2016, MNRAS, 462, 3265

de Vaucouleurs, G. 1948, Annales d'Astrophysique, 11, 247

—. 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.

Dekel, A., & Birnboim, Y. 2006, MNRAS, 368, 2

Dekel, A., Sari, R., & Ceverino, D. 2009a, ApJ, 703, 785

—. 2009b, ApJ, 703, 785

Dekel, A., & Silk, J. 1986, ApJ, 303, 39

Dib, S., Bell, E., & Burkert, A. 2006, ApJ, 638, 797

Dicke, R. H., Peebles, P. J. E., Roll, P. G., & Wilkinson, D. T. 1965, ApJ, 142, 414

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

Domínguez Sànchez, H., Huertas-Company, M., Bernardi, M., Ticcillo, D., & Fisher, J. L. 2017, ArXiv e-prints, arXiv:1711.05744

Dressel, L. 2012, Wide Field Camera 3 Instrument Handbook for Cycle 21 v. 5.0

Dressler, A. 1980, ApJ, 236, 351

Driver, S. P., Robotham, A. S. G., Bland-Hawthorn, J., et al. 2013, MNRAS, 430, 2622

Drory, N., Salvato, M., Gabasch, A., et al. 2005, ApJ, 619, L131

Duncan, K., Conselice, C. J., Mortlock, A., et al. 2014, MNRAS, 444, 2960

Elmegreen, B. G., Bournaud, F., & Elmegreen, D. M. 2008, ApJ, 688, 67

Elmegreen, B. G., & Elmegreen, D. M. 2005, ApJ, 627, 632

—. 2010, ApJ, 722, 1895

Elmegreen, B. G., Elmegreen, D. M., Sánchez Almeida, J., et al. 2013, ApJ, 774, 86

Epinat, B., Tasca, L., Amram, P., et al. 2012, A&A, 539, A92

Erb, D. K., Steidel, C. C., Shapley, A. E., et al. 2006, ApJ, 646, 107

Faber, S. M., Willmer, C. N. A., Wolf, C., et al. 2007, ApJ, 665, 265

Fabian, A. C. 2012, ARA&A, 50, 455

Fall, S. M., & Efstathiou, G. 1980, MNRAS, 193, 189

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

Faucher-Giguère, C.-A., Kereš, D., & Ma, C.-P. 2011, MNRAS, 417, 2982

Fontana, A., Salimbeni, S., Grazian, A., et al. 2006, A&A, 459, 745

Förster Schreiber, N. M., Genzel, R., Bouché, N., et al. 2009, ApJ, 706, 1364

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., & Peebles, P. J. E. 2004, ApJ, 616, 643

Fukugita, M., Nakamura, O., Okamura, S., et al. 2007, AJ, 134, 579

Fumagalli, M., O'Meara, J. M., & Prochaska, J. X. 2016, MNRAS, 455, 4100

Furlanetto, S. R., Schaye, J., Springel, V., & Hernquist, L. 2005, ApJ, 622, 7

Gabor, J. M., Davé, R., Finlator, K., & Oppenheimer, B. D. 2010, MNRAS, 407, 749

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

Genel, S., Naab, T., Genzel, R., et al. 2012, ApJ, 745, 11

Genzel, R., Tacconi, L. J., Eisenhauer, F., et al. 2006, Nature, 442, 786

Genzel, R., Burkert, A., Bouché, N., et al. 2008, ApJ, 687, 59

Genzel, R., Newman, S., Jones, T., et al. 2011, ApJ, 733, 101

Genzel, R., Förster Schreiber, N. M., Lang, P., et al. 2014, ApJ, 785, 75

Giavalisco, M., Dickinson, M., Ferguson, H. C., et al. 2004, ApJ, 600, L103

Glasser, G. J. 1962, Journal of the American Statistical Association, 57, 648

GMTO Corporation. 2012, Giant Magellan Telescope Scientific Promise and Opportunities

González, V., Labbé, I., Bouwens, R. J., et al. 2011, ApJ, 735, L34

Governato, F., Brook, C. B., Brooks, A. M., et al. 2009, MNRAS, 398, 312

Green, A. W., Glazebrook, K., McGregor, P. J., et al. 2010, Nature, 467, 684

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

Gunn, J. E., & Gott, III, J. R. 1972, ApJ, 176, 1

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

Guth, A. H. 1981, Phys. Rev. D, 23, 347

Gwyn, S. D. J. 2012, AJ, 143, 38

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

Hearin, A. P., Behroozi, P. S., & van den Bosch, F. C. 2016, MNRAS, 461, 2135

Hearin, A. P., Watson, D. F., & van den Bosch, F. C. 2015, MNRAS, 452, 1958

Heckman, T. M., & Best, P. N. 2014, ARA&A, 52, 589

Henry, A., Scarlata, C., Martin, C. L., & Erb, D. 2015, ApJ, 809, 19

Hinshaw, G., Larson, D., Komatsu, E., et al. 2013, ApJS, 208, 19

Hirschmann, M., Naab, T., Davé, R., et al. 2013, MNRAS, 436, 2929

Hogg, D. W., Blanton, M. R., Brinchmann, J., et al. 2004, ApJ, 601, L29

Holwerda, B. W., Muñoz-Mateos, J.-C., Comerón, S., et al. 2014, ApJ, 781, 12

Hopkins, P. F., Kereš, D., Murray, N., Quataert, E., & Hernquist, L. 2012, MNRAS, 427, 968

Hubble, E. 1929, Proceedings of the National Academy of Science, 15, 168

Hubble, E., & Humason, M. L. 1931, ApJ, 74, 43

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

Ilbert, O., McCracken, H. J., Le Fèvre, O., et al. 2013, A&A, 556, A55

Izotov, Y. I., Orlitová, I., Schaerer, D., et al. 2016, Nature, 529, 178

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

Katz, N., Keres, D., Dave, R., & Weinberg, D. H. 2003, in Astrophysics and Space Science Library, Vol. 281, The IGM/Galaxy Connection. The Distribution of Baryons at z=0, ed. J. L. Rosenberg & M. E. Putman, 185

Kauffmann, G., Li, C., Zhang, W., & Weinmann, S. 2013, MNRAS, 430, 1447

Kauffmann, G., Heckman, T. M., White, S. D. M., et al. 2003, MNRAS, 341, 54

Kawata, D., & Mulchaey, J. S. 2008, ApJ, 672, L103

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

Kennicutt, Jr., R. C. 1998a, ARA&A, 36, 189

—. 1998b, ApJ, 498, 541

Kereš, D., Katz, N., Davé, R., Fardal, M., & Weinberg, D. H. 2009, MNRAS, 396, 2332

Kereš, D., Katz, N., Weinberg, D. H., & Davé, R. 2005, MNRAS, 363, 2

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., & Ho, L. C. 2013, ARA&A, 51, 511

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

Lee, K.-S., Ferguson, H. C., Wiklind, T., et al. 2012, ApJ, 752, 66

Lehnert, M. D., Nesvadba, N. P. H., Le Tiran, L., et al. 2009, ApJ, 699, 1660

Lilly, S. J., Carollo, C. M., Pipino, A., Renzini, A., & Peng, Y. 2013, ApJ, 772, 119

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., Jonsson, P., Cox, T. J., & Primack, J. R. 2008, MNRAS, 391, 1137

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

Marchesini, D., van Dokkum, P. G., Förster Schreiber, N. M., et al. 2009, ApJ, 701, 1765

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

Martin, D. C., Matuszewski, M., Morrissey, P., et al. 2015, Nature, 524, 192

—. 2016, ApJ, 824, L5

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

Mihos, J. C., & Hernquist, L. 1996, ApJ, 464, 641

Mo, H. J., Mao, S., & White, S. D. M. 1998, MNRAS, 295, 319

Monreal-Ibero, A., Colina, L., Arribas, S., & García-Marín, M. 2007, A&A, 472, 421

Moore, B., Katz, N., Lake, G., Dressler, A., & Oemler, A. 1996, Nature, 379, 613

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

Moster, B. P., Macciò, A. V., Somerville, R. S., Johansson, P. H., & Naab, T. 2010, MNRAS, 403, 1009

Muratov, A. L., Kereš, D., Faucher-Giguère, C.-A., et al. 2015, MNRAS, 454, 2691

Muzzin, A., Marchesini, D., Stefanon, M., et al. 2013, ApJ, 777, 18

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

Newman, S. F., Genzel, R., Förster-Schreiber, N. M., et al. 2012, ApJ, 761, 43

Nilsson, K. K., Fynbo, J. P. U., Møller, P., Sommer-Larsen, J., & Ledoux, C. 2006, A&A, 452, L23

Noeske, K. G., Weiner, B. J., Faber, S. M., et al. 2007, ApJ, 660, L43

Odewahn, S. C., Cohen, S. H., Windhorst, R. A., & Philip, N. S. 2002, ApJ, 568, 539

Oklopčić, A., Hopkins, P. F., Feldmann, R., et al. 2017, MNRAS, 465, 952

Oppenheimer, B. D., Davé, R., Kereš, D., et al. 2010, MNRAS, 406, 2325

Ouchi, M., Shimasaku, K., Okamura, S., et al. 2004, ApJ, 611, 660

Pahwa, I., & Paranjape, A. 2017, MNRAS, 470, 1298

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

Penzias, A. A., & Wilson, R. W. 1965, ApJ, 142, 419

Perlmutter, S., Aldering, G., della Valle, M., et al. 1998, Nature, 391, 51

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

Planck Collaboration, Ade, P. A. R., Aghanim, N., et al. 2016, A&A, 594, A20

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

Rees, M. J., & Ostriker, J. P. 1977, MNRAS, 179, 541

Riess, A. G., Filippenko, A. V., Challis, P., et al. 1998, AJ, 116, 1009

Robertson, B., Bullock, J. S., Cox, T. J., et al. 2006, ApJ, 645, 986

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., Rich, R. M., Charlot, S., et al. 2007, ApJS, 173, 267

Salim, S., Lee, J. C., Janowiecki, S., et al. 2016, ApJS, 227, 2

Sánchez, S. F., Rosales-Ortega, F. F., Jungwiert, B., et al. 2013, A&A, 554, A58

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

Schiminovich, D., Wyder, T. K., Martin, D. C., et al. 2007, ApJS, 173, 315

Schmidt, M. 1959, ApJ, 129, 243

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

Somerville, R. S., & Davé, R. 2015, ARA&A, 53, 51

Soto, E., de Mello, D. F., Rafelski, M., et al. 2017, ApJ, 837, 6

Springel, V., White, S. D. M., Jenkins, A., et al. 2005, Nature, 435, 629

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., Giavalisco, M., Dickinson, M., & Pettini, M. 1999, ApJ, 519, 1

Steidel, C. C., Adelberger, K. L., Shapley, A. E., et al. 2000, ApJ, 532, 170

Steidel, C. C., Erb, D. K., Shapley, A. E., et al. 2010, ApJ, 717, 289

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

Swinbank, A. M., Webb, T. M., Richard, J., et al. 2009, MNRAS, 400, 1121

Tacconi, L. J., Genzel, R., Smail, I., et al. 2008, ApJ, 680, 246

Tacconi, L. J., Genzel, R., Neri, R., et al. 2010, Nature, 463, 781

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

The Dark Energy Survey Collaboration. 2005, ArXiv Astrophysics e-prints, astro-ph/0510346

Toomre, A. 1964, ApJ, 139, 1217

Toomre, A. 1977, in Evolution of Galaxies and Stellar Populations, ed. B. M. Tinsley & R. B. G. Larson, D. Campbell, 401

Tully, R. B., Mould, J. R., & Aaronson, M. 1982, ApJ, 257, 527

van de Voort, F., Schaye, J., Booth, C. M., Haas, M. R., & Dalla Vecchia, C. 2011, MNRAS, 414, 2458

van den Bergh, S. 1976, ApJ, 206, 883

van den Bosch, F. C., Burkert, A., & Swaters, R. A. 2001, MNRAS, 326, 1205

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

Weinmann, S. M., van den Bosch, F. C., Yang, X., & Mo, H. J. 2006, MNRAS, 366, 2

White, S. D. M., & Frenk, C. S. 1991, ApJ, 379, 52

White, S. D. M., & Rees, M. J. 1978, MNRAS, 183, 341

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

Woo, J., Dekel, A., Faber, S. M., et al. 2013, MNRAS, 428, 3306

Wright, D. E., Lintott, C. J., Smartt, S. J., et al. 2017, MNRAS, 472, 1315

Wuyts, S., Förster Schreiber, N. M., Genzel, R., et al. 2012, ApJ, 753, 114

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

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

Zanella, A., Daddi, E., Le Floc'h, E., et al. 2015, Nature, 521, 54

Appendix A

Spectro-polarimetry Confirms Central Powering in a Lyα 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 Lyα emission and detect integrated linear polarization of 9-13% $\pm$ 2-3% at a distance of approximately 15 kpc north and south of the peak of the Lyα 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 Lyα 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 polarizatin 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 suggests either that this region is not symmetric (which is likely) and exhibits variations in columns 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 Lyα 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 Lyα 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 simulataneously (Furlanetto et al., 2005). Theoretical studies have shown that polarization of Lyα 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 partiuclar, 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 Lyα surface brightness (SB) is more likely associated with an 850 μm 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 Lyα photons which interact with the surrounding medium. Deep integral-field spectroscopy of the Lyα 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 Lyα 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$\pm$2% 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 Lyα photons by HI.

Due to the observational expense involved, polarization measurements of spatially extended Lyα 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±4.6% in some parts of the nebula and conclude that at least a portion of the nebula is powered by the scattering of Lyα 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 Lyα 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 Lyα Polarization Basics

Lyα polarization requires photons be scattered imbuing them with a preferential direction or impact angle. Localized (*in situ*) production of Lyα 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 Lyα polarization signal.

The detection of signifiant polarization fraction of Lyα emission depends on two crucial factors: wing vs resonant scattering and Doppler boosting by thermal atoms in the surrounding medium. Lyα 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 scacttering 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 Lyα undergoes this type of scattering and the Lyα 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 Lyα 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 Lyα 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 Lyα 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 Lyα 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 Lyα 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 Lyα polarization is required. For the case of an outflowing thin shell, Dijkstra & Loeb (2008) predict that Lyα polarization will increase redwards of the line center. This is because the redder Lyα 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 Lyα 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 Lyα 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 Lyα 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 Lyα 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, arbritrarily 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 (∼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

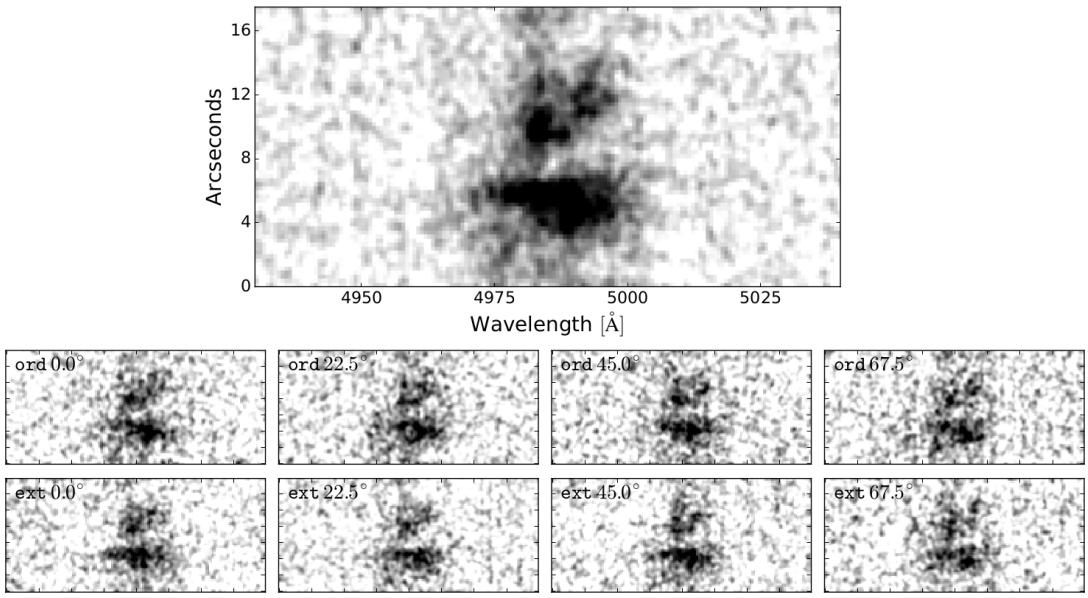

Figure A.1 *Top.* Spatial and wavelength distribution of the master total intensity Lyα spectrum. Co-added Lyα spectrum of 77 science spectra smoothed by a Gaussian with FWHM = $0\farcs5$. Due to light loss at the edges of the slit, we present here only the central $18''$. *Bottom.* Ordinary and extraordinary 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 $\sim 0.\!''75$. 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 intial 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 EsoRex[1] , an ESO recipe execution tool,

[1] `http://www.eso.org/sci/software/cpl/esorex.html`

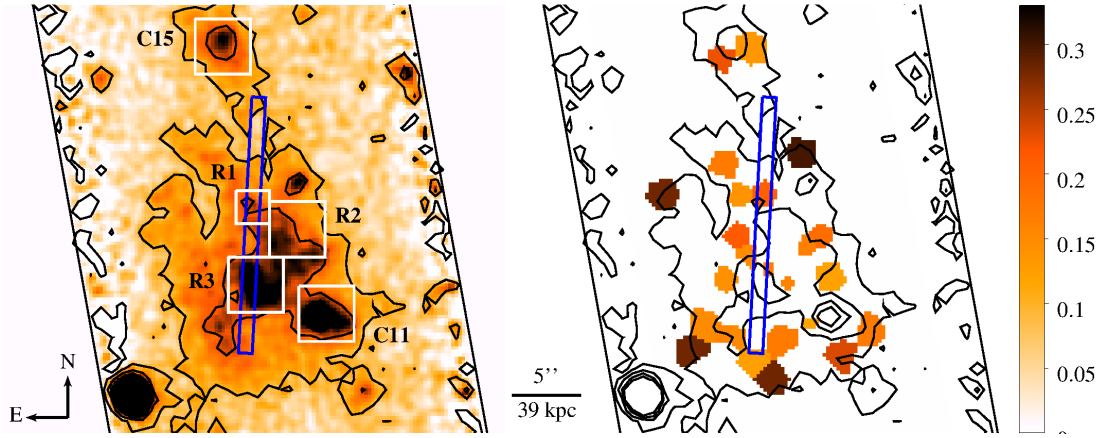

Figure A.2 Slit position over LAB1 as compared with results from H11. In the left panel we show the combined Lyα intensity frame from H11 adaptively smoothed to show detail including emission from LAB1 and nearby LBGs C11 and C15. Overlaid in white are boxes from Weijmans et al. (2010), regions in which they performed integral-field spectroscopy on the Lyα 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 Lyα 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 throroughly removed using L.A. Cosmic[2] (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/IRAF`[3] `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 Lyα 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 Lyα 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 Lyα 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:

$$F_\theta = \frac{f_\theta^{ord} - f_\theta^{ext}}{f_\theta^{ord} + f_\theta^{ext}} \tag{A.1}$$

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

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

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

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

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:

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

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 Modifed 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:

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

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

$$b_i^2 = \frac{q_i^2\sigma_u^2 + u_i^2\sigma_q^2}{q_i^2 + u_i^2} \tag{A.6}$$

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

$$SNR_p = \frac{p_{MAS}}{\sqrt{\frac{1}{2}(\sigma_q^2 + \sigma_u^2)}}, \tag{A.7}$$

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 $\mathrm{SNR}_p > 3.8$ 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 $\mathrm{SNR}_p > 0.5$.

A.4 Polarization of Lyα in LAB1

A.4.1 Polarization integrated over the line profile

Because our data have low signal-to-noise per pixel ($\mathrm{SNR}_p \lesssim 1$), binning of the science frames is a necessity. However, any Lyα 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 Lyα surface brightness (SB) as denoted by the set of arbitrary contours. Also shown in this figure are white boxes corresponding to Lyα 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 Lyα SB while **R1** is associated with a somewhat dimmer region. Between these two

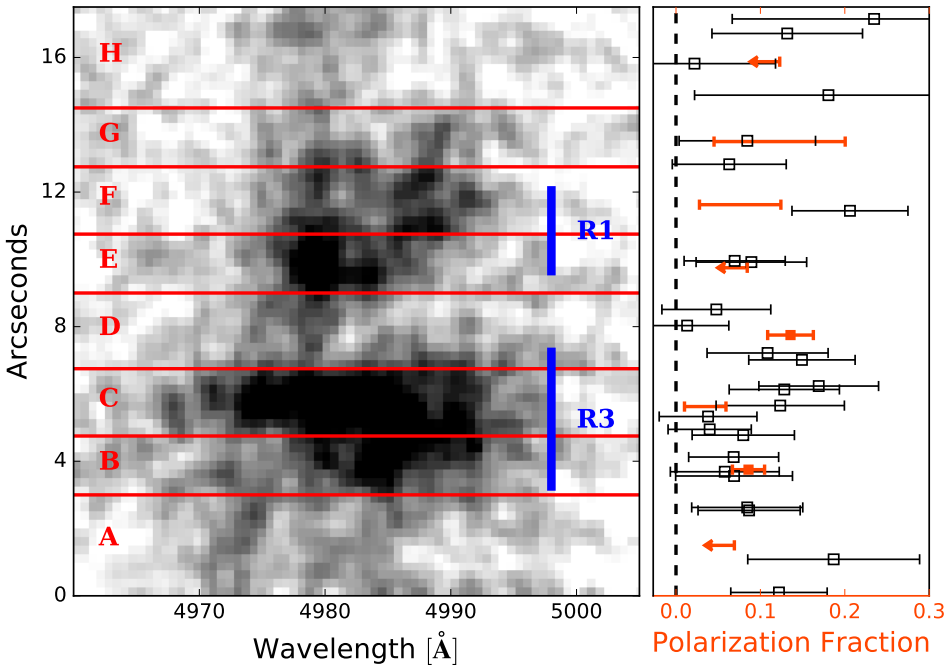

Figure A.3 Polarization of spectrally integrated Lyα. In the left panel we show the spatial binning of the Lyα emission spectrum with bins ranging from 2-3″ overlaid in red. Lyα emission is integrated over 4965-5000 Å. The blue lines indicate the spatial extent of Weijmans et al. (2010) IFU boxes. Depicted 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 corresonding $1\sigma_p$ 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 Lyα 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 Lyα 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 Lyα 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 $\mathrm{SNR}_p > 3.8$ and for these we report the measured polarization and $1\sigma_p$ error. The remaining bins have $\mathrm{SNR}_p < 3.8$ 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

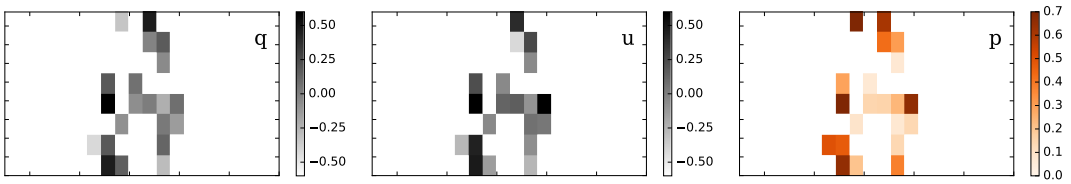

Figure A.4 2D Stokes parameters and polarization maps. The q and u maps are computed directly from the science frames shown in the bottom panel of Figure A.1 according to the prescription of Equation A.2 after binning as described in subsection A.4.2. The polarization map in the right panel is then computed 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 that 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.

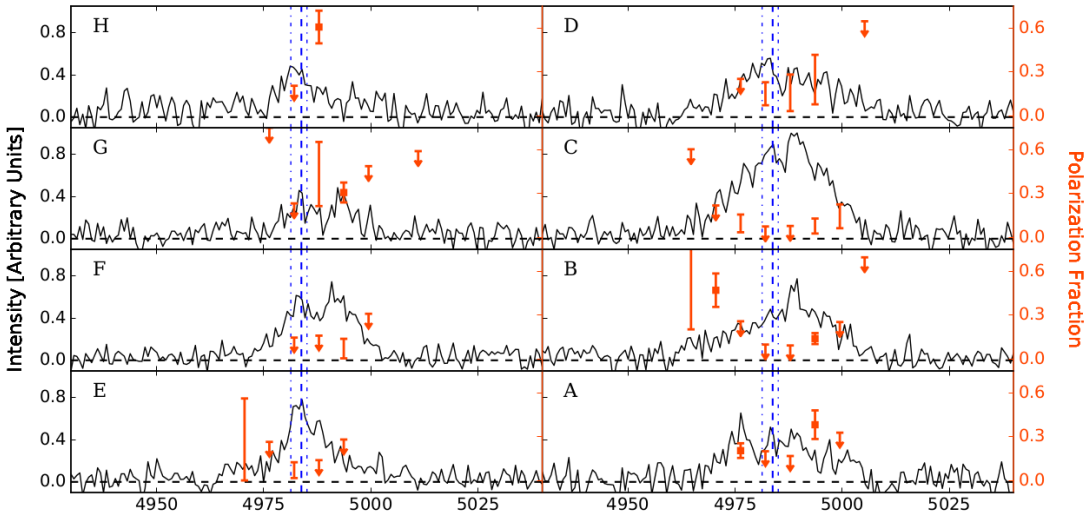

Figure A.5 Polarization of Lyα as a function of wavelength. Figures **A** through **H** contain the extracted 1D Lyα spectrum of the corresponding aperture from Figure A.3. All spectra have been scaled to show their relative intensity. Overplotted in orange we show the polarization fraction as a function 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 [OIII]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 Lyα 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 Lyα 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 Lyα 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 furher 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

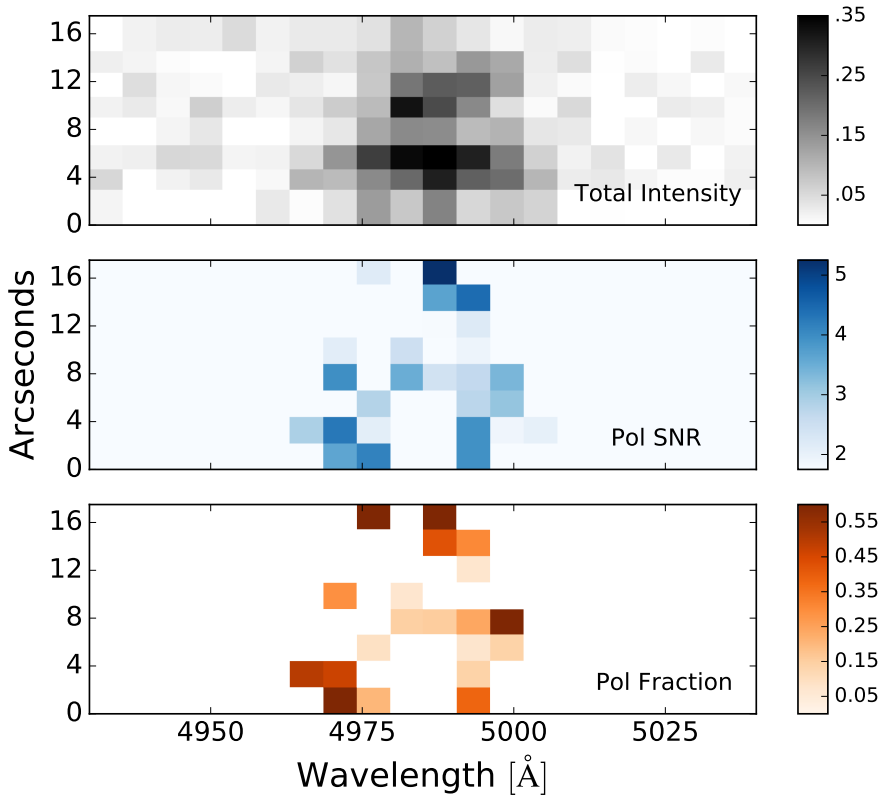

Figure A.6 *Top.* Total intensity of the 2D Lyα spectrum in bins of 5Å by 2-3″. *Middle.* SNR$_p$ map which demonstrates the relative quality of our polarization measurements. SNR$p > 2$ is generally enough for us to discriminate p statistically signicant from zero with 95% confidence and we show CIs for such bins in Figure A.5. 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 $\mathrm{SNR}_p > 2$. Errors on the polarization angles range from $\sim$6–15°. 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 Lyα SB.

A.5 Discussion and Conclusions

We have presented deep spectro-polarimetry of the LAB1 Lyα 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 Lyα SB would not be strongly polarized but the polarization would rise with increasing radius (Dijkstra & Loeb, 2008). The observed polarization of Lyα in the southern portion of the slit is consistent with Lyα 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 Lyα 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

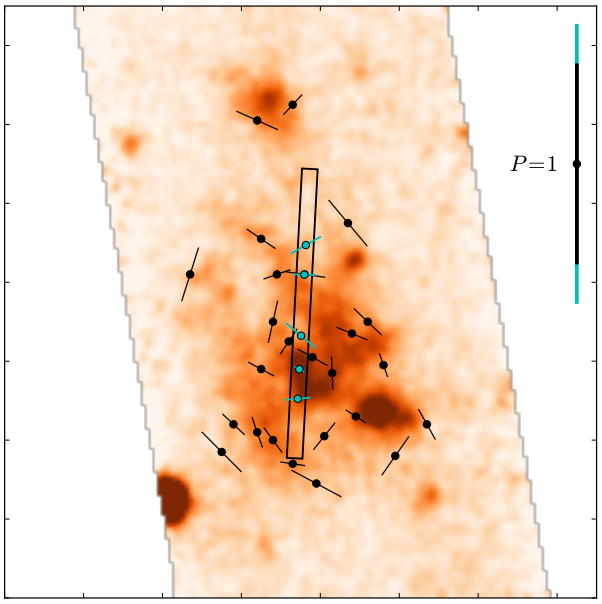

Figure A.7 Direction of Lyα polarization vectors. Smoothed total Lyα intensity from H11 overlaid with polarization vectors from H11 (black) and from this analysis (cyan) for those spatial bins from Figure A.3 which have $\mathrm{SNR}_p > 2$. Because we measure small polarization amplitude in our bins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

to the south (**B**), we find total $p = 8.59 \pm 1.9\%$. Thus we have increasing polarization away from the peak Lyα emission with a radius of $\sim$ 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 Lyα 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 Lyα 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 [OIII] 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 Lyα 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 Lyα 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 Lyα 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 consistant with low polarization in the core of the Lyα 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 Lyα 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 flourescence from ionizing radation eminating 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 souce 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 Lyα Polarization

With another successful detection of the spectral dependence of Lyα polarization the question arises: What does the future hold for Lyα 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 Lyα 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 by 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 Lyα 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 Lyα 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 Lyα 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 Lyα 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 Lyα scattering, it behooves us to test further what relationship, if any, exists between radio-loud and -quiet nebulae. Compact Lyα 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 Lyα 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 Lyα polarization in all Lyα radiative transport codes to generate predictions for various applications including clumpy and filimentary 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 Lyα line profiles. With the current limitation on instumentation 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 Jérémy Blaizot and Maxime Trebitsch for proofing the manuscript and providing feedback which helped to clarify the text.