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

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

Melanie Renee Beck

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

Advisor:
M. Claudia Scarlata

August, 2017

Acknowledgements

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

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

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

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

My family for years and years of support.

Dedication

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

Abstract

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

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

Contents

List of Tables

List of Figures

Chapter 1

Introduction

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

Astronomers have made use of visual galaxy morphologies to understand the dynamical structure of these systems for nearly ninety years (e.g., Hubble, 1936; de Vaucouleurs, 1959; Sandage, 1961; van den Bergh, 1976; Nair & Abraham, 2010; Baillard et al., 2011). The division between early-type and late-type systems corresponds, for example, to a wide range of parameters from mass and luminosity, to environment, colour, and star formation history (e.g., Kormendy, 1977; Dressler, 1980; Strateva et al., 2001; Blanton et al., 2003; 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., review by Kormendy & Kennicutt, 2004; Elmegreen et al., 2008; Sheth et al., 2008; Masters et al., 2011; Simmons et al., 2014). Modern studies of morphology divide systems into broad classes (e.g., Conselice, 2006; Lintott et al., 2008; Kartaltepe et al., 2015; Peth et al., 2016), but a wealth of information can be gained from identifying new and often rare classes, such as low redshift clumpy galaxies (e.g., Elmegreen et al., 2013), polar-ring galaxies (e.g., Whitmore et al., 1990), and the green peas (Cardamone et al., 2009).

Obtaining these morphologies has traditionally been a time-consuming visual endeavor and only in the past twenty years have automated morphological assignment been possible. Even with the varied automated approaches currently being exploited, an era of Even Bigger Data looms for the field of astronomy. The next decade will herald the first light of more powerful ground- and spaced-based telescopes such as the Large Synaptic 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 all of Everything? I don't know. Traditional techniques will not be sufficient for extracting galaxy morpholgoies on a pertinent timescale.

This thesis details a solution to the scalability of galaxy morphology designations by examinging classification obtained as part of the Galaxy Zoo project, a crowd-sourcing initiative that has obtained morphological classifications from over a million volunteers for over a million galaxies from several surveys. Though innovative, even crowd-sourcing will be unable to sustain the classification load for future surveys. Instead, these classifications are combined with supervised machine learning algorithms that train on automated measurements of galactic morphological structural indicators. This thesis begins with a detailed account of the methodology used to obtain these morphological structural indicators (Chapter 2). Chapter 3 demonstrates how crowd-sourcing techniques can be optimized by applying SWAP [come up with generic words] to Galaxy Zoo classifications, while Chapter 4 explores the combination of visual classifications with machine learning algorithms. Also included is a prelimary analysis of a rare sample of "clumpy" galaxies in the local universe discovered during the course of the analysis of Galaxy Zoo classifications [NOT GOOD WORDS] (Chapter 5). While common in the distant universe, these potential local counterparts could shed light on star-formation things and stuff. This Introduction provides a brief overview of galaxy formation, the sciece of morphology, and galaxy classification techniques, as well as an overview of Galaxy Zoo: Express, an integrated framework of human and machine galaxy classifiers.

1.1 Galaxy formation

Brief overview of how galaxies form and the dizzying variety of morphologies.

Standard morphological categorizations briefly discuss the Hubble tuning fork. Explain the difference between elliptical, spiral and in-between galaxies. As well as irregulars/peculiars.

1.2 Morphology as a tracer of galaxy evolution

then segue into how those different morphologies provide detailed insights into formation and evolution. Give examples of the science that can be gleaned from broad classifications (early- vs late-type galaxies), and how rare classifications can provide unique viewpoints (clumpy galaxies at low redshift, or green peas, i.e.)

Brief review of why we care about morphology – i.e., it correlates with all kinds of sciency bits so we should keep doing it.

1.2.1 Mass and luminosity

1.2.2 Morphology-density relation

A relationship between morphology and the density in which the galaxy resides has been known for some time. It has been found that in the richest and densest clusters of galaxies, the dominant morphology is elliptical, while for field galaxies, the dominant morphology is spiral. Furthermore, it's been determined that this relation evolves over time, at least for the densest environments. All of this points to scenarios in which the environment of a galaxy dictates, or influences its morphology and that the processes contributing to this effect have evolved with time. (e.g., Fasano et al., 2000; Shen et al., 2003; Smith et al., 2005; Peng et al., 2010)

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

1.2.3 Color

1.2.4 Star formation history

1.2.5 Insights from rare morophologies

What have we learned from the green peas? Clumpy galaxies at low redshift? Ring galaxies?

– insights concerning dynamical disks; star formation; blah.

1.3 Obtaining morphological designations

Now that we all believe we should get these morphologies – how should we do it?

1.3.1 Visual classifications

The history of galaxy morphology assignment is rife with historical precendents due in large part to Edwin Hubble's original "tuning fork." Seeing as Hubble originally thought these 'nebulae' were features residing in our own galaxy, we end up with goofy shit where we call elliptcals "early-type" and disks "late-type" though subsequent studies all confirm that the ages of ellipticals are far older than those of spiral galaxies.

Visual classification, though highly accurate due to the human mind's unique pattern recognition capability, is, however, entirely slow. Assignment of morphological type to galaxies thus resulted in small samples lacking statistical signifigance for decades (until the use of cartels of graduate students wherein morphologies for galaxy samples reached a few thousands). With surveys such as the Sloan Digital Sky Survey looming on the horizon, a new approach would be necessary in order to take advantage of the unique insights provided by galaxy morphologies.

This necessitated the birth of the Galaxy Zoo project, the first effort to crowd-source the task of galaxy morphology assignment to the general public. Blah blah blah Galaxy Zoo blah.

While the Galaxy Zoo project has provided a solution that scales visual classification for current surveys by harnessing the combined power of thousands of volunteers (Lintott et al., 2008, 2011; Willett et al., 2013, 2017; Simmons et al., 2017), 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); upcoming surveys such as *LSST* and *Euclid* will require a different approach, imaging more than a billion new galaxies (LSST Science Collaboration et al., 2009; Laureijs et al., 2011). If detailed morphologies can be extracted for just 0.1% of this imaging, we will have millions of images to contend with. A project of this magnitude would take more than sixty years to classify at Galaxy Zoo's current rate and configuration. Standard visual morphology methods will thus be unable to cope with the scale of data.

1.3.2 Automated classifications

Another approach has been the automated extraction of galaxy morpholgies with the development of both parametric and non-parametric structural indicators.

It is well known that a galaxy's light profile can be modelled according to the function [insert function here] where the Sersic index, n, has been shown to correlate strongly with a distinction between early- and late-type galaxies. In particular, a Sersic index of $n = 4$ corresponds to a de Vaucouleurs' profile which well describes the surface brightness of an elliptical galaxy as a function of its apparent radius from the galaxy's center. On the other hand, a Sersic index of $n = 1$ describes an exponential profile which is a good description for spiral galaxies.

A drawback to the parametric approach is the need to assume the underlying distribution and while this works technique works well for galaxies that are obviously elliptical or spiral, it produces mixed results for other morphological types, i.e., irregulars or peculiars, which have low central concentration resulting in a low Sersic index, but which do not have disks or spiral arms.

Non-paramteric structural indicators require fewer assumptions on the data and are instead derived observationally. Several of these diagnostics have been developed over the past couple decades, each probing a different part of the galaxy's light profile and thus its overall dynamical distribution. The most common diagnostics are described below.

Closely related to the Sersic index is the non-paramatric diagnostic of concentration. A galaxy's concentration aims to identify how dense a galaxy's central surface brightness profile is. Concentration, originally conceived by Abraham(?) has had several definitions over the years. The most common consider the ratio of the aggregated light within two concentric apertures about the galaxy's center. Typically, these apertures contain 50 and 90% of the galaxy's light (another common approach is to use 50% and 80%). Because

Peng(?) [was that the first?] developed the first model to decompose a galaxy's light profile into

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).

1.3.3 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. The location in this 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; Scarlata et al., 2007). Another approach is through deep learning, a machine learning technique that attempts to model high level abstractions. Algorithms like convolutional and artificial neural networks (CNNs, ANNs) have been used for galaxy morphology classification with impressive accuracy (Ball et al., 2004; Banerji et al., 2010; Dieleman et al., 2015; Huertas-Company et al., 2015). A drawback to all machine learning classification techniques is the need for standardized training data, with more complex algorithms requiring more data. Furthermore, these data must be consistent for each survey: differences in resolution and depth can be implicitly learned by the algorithm making their application to disparate surveys challenging.

1.4 Overview of Galaxy Zoo: Express

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

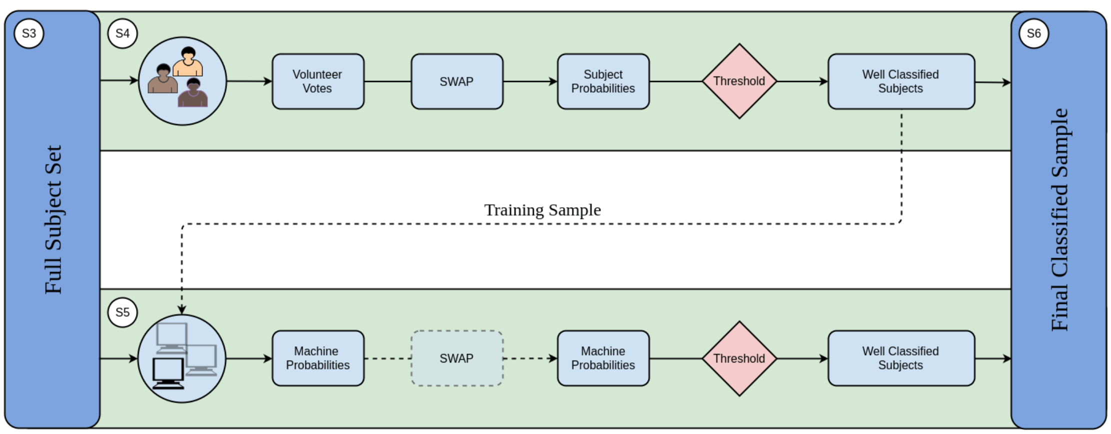

Figure 1.1 Schematic of our hybrid system. Humans provide classifications of galaxy images via a web interface. We simulate this with the Galaxy Zoo 2 classification data described in Chapter 2. Human classifications are processed with an algorithm described in Chapter 3. Subjects that pass a set of thresholds are considered human-retired (fully classified) and provide the training sample for the machine classifier as described in Chapter 4. The trained machine is applied to all subjects not yet retired. Those that pass an analogous set of machine-specific thresholds are considered machine-retired. The rest remain in the system to be classified by either human or machine. This procedure is repeated nightly.

The Galaxy Zoo Express (GZX) framework combines human and machine to increase morphological classification efficiency, both in terms of the classification rate and required human effort. Figure 1.1 presents a schematic of GZX including section numbers as a shortcut for the reader. We note that transparent portions of the schematic represent areas of future work which we explore in Chapter 6. Any system combining human and machine classifications will have a set of generic features: a group of human classifiers, at least one machine classifier, and a decision engine which determines how these classifications should be combined.

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

In Chapter 4, we incorporate a machine classifier. We have developed a Random Forest algorithm that trains on measured morphology indicators such as Concentration, Asymmetry, Gini coefficient and M_{20} well-suited for the top-level question of the GZ2 decision tree, discussed in Chapter 2. After a sufficient number of subjects have been classified by humans, the machine is trained and its performance assessed through cross-validation. This procedure is repeated nightly and the machine's performance increases with size of the training sample, albeit with a performance limit. Once the machine reaches an acceptable level of performance it is applied to the remaining galaxy sample.

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

Chapter 2

Data: visual and automated 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 classification procurement. 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 innitiative to visually classify large numbers of galaxies by asking members of the general public. It began in 2007 with the classification of 893,212 galaxy images from the Sloan Digital Sky Survey (SDSS) Data Release 6 with $r < 17.77$ Petrosian AB magnitudes Strauss et al. (2002); Adelman-McCarthy et al. (2008). The first iteration of the project was simplistic, inviting volunteers to determine whether a galaxy was elliptical, spiral, or a star / artifact. The project was an immediate success both in terms of the public interest and the resulting science. Following the completion of GZ1 Lintott et al. (2008), over a dozen peer-reviewed articles utilzing GZ1 classifications[1] were published. In addition to explorations of galaxy morphology and its dependence on color and environment Skibba et al. (2009); Bamford et al. (2009), significant results

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

also included substantial populations of red disks Masters et al. (2010), blue ellipticals Schawinski et al. (2009), and discoveries of rare objects such as the "green peas" Cardamone et al. (2009), and Hanny's Voorwerp Lintott et al. (2009), the first obvservation of a AGN ionization echo.

The early sucess of GZ1 led to several subsequent and progressively more complex projects. To date, Galaxy Zoo has provided morphologies for over a million galaxies from multiple imaging surveys of various wavebands and redshifts using classifications provided from over a million volunteers. A summary of Galaxy Zoo projects is given in Table XXX. The current research uses data specifically from the Galaxy Zoo 2 (GZ2) project Willett et al. (2013), the immediate successor of GZ1. The following sections provide an overview of the project specifics including the galaxy sample, the decision tree structure, and a brief description of how volunteer "clicks" are converted into useful classifications.

2.1.1 Galaxy sample selection

The original GZ1 project sought broad classifications for nearly one million galaxies in SDSS. Due to the sheer number of galaxies to quantify, the morphologies collected were broad, seeking to determine between early-type, late-type and mergers. However, much can be gained by probing more detailed morpholgies such as the existance of bars, bulges, dust lanes, rings, etc. Galaxy Zoo 2 thus selected the nearest and brightest 25% of galaxies from the original GZ1 sample, galaxies for which fine morphological structure could be resolved and classified. Pulled from Data Release 7 Legacy catalog Abazajian et al. (2009) which imaged the the North Galactic Cap, this galaxy sample required the Petrosian half-light magnitude be brighter than 17.0 in the r-band, along with a size limit such that 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 multipy-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 sets from this region: a set of 21,522 single-exposure images (though only about half conformed to the shallower Legacy magnitude cut specified above), and two sets of $\sim$30K co-added images from multiple exposures resulting in an object detection limit approximately two magnitudes deeper than the normal depth imaging. The research presented in this thesis utilizes the final GZ2 single-depth sample consisting of 295,305 galaxies of which 11,334 have the deeper magnitude limit.

2.1.2 GZ2 decision tree and project history

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*. With the exception of the first question, subsequent tasks depended on the response to the current question. 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'. 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.

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 drawn randomly with the exception that 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 charactrize the likelihood of the classification distribution. This resulted in a median of 44 classifications per galaxy with a wide spread (Show Kyle's figure?). The full project spanned just over 14 months with the final dataset consisting of over 16 million classifications from over 80 thousand volunteers.

2.1.3 Data reduction

The GZ2 catalog publishes several types of morphologies computed from volunteer classifications constisting of a vote fractrion for each response to every task, denoted $f_{\rm response}$. The most basic of these is computed simply as $f_r = n_r/n_t$, where n_r is the number of

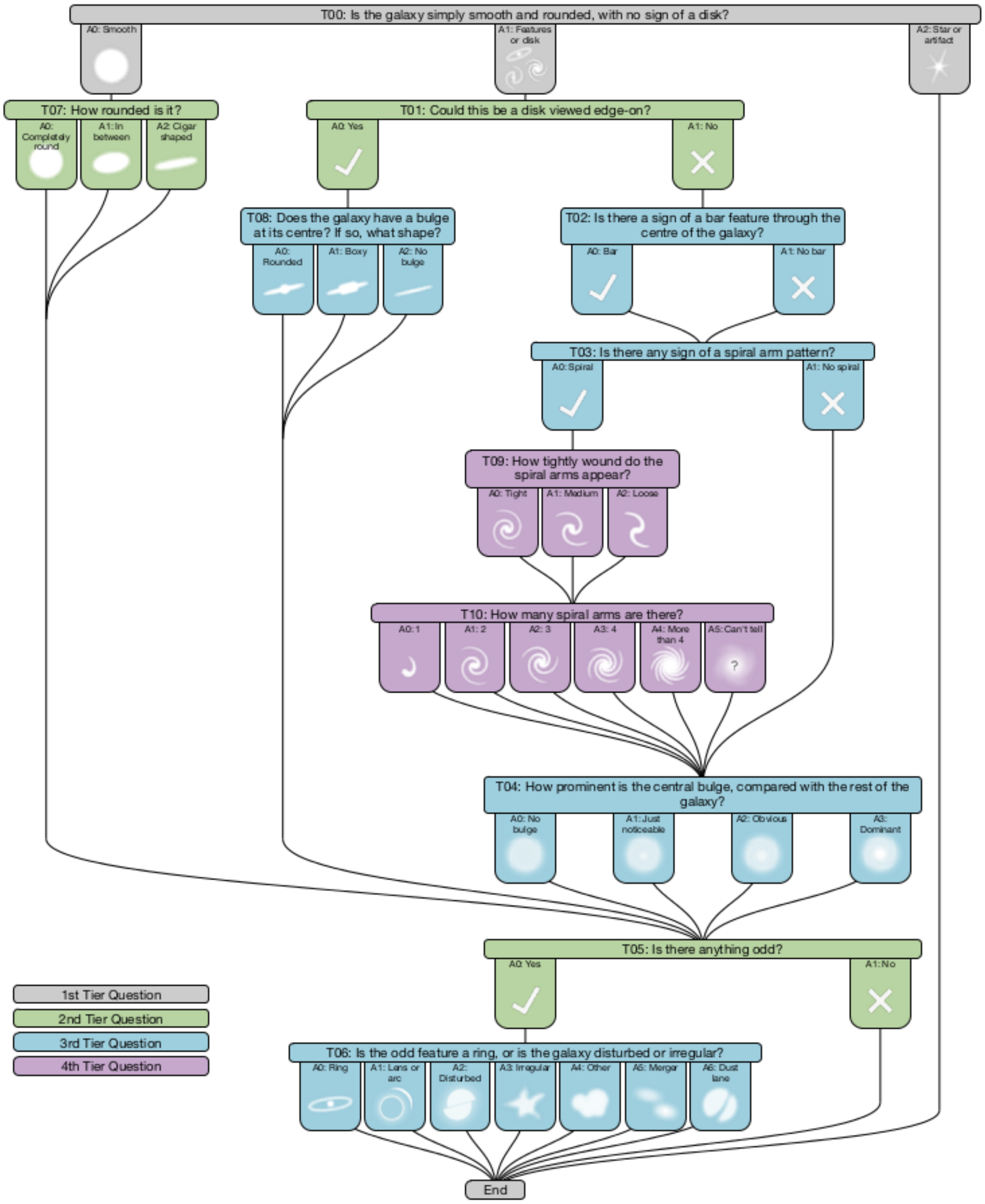

Figure 2.1 AHMAH MUTHAFUCKIN CAPTION BITCHES

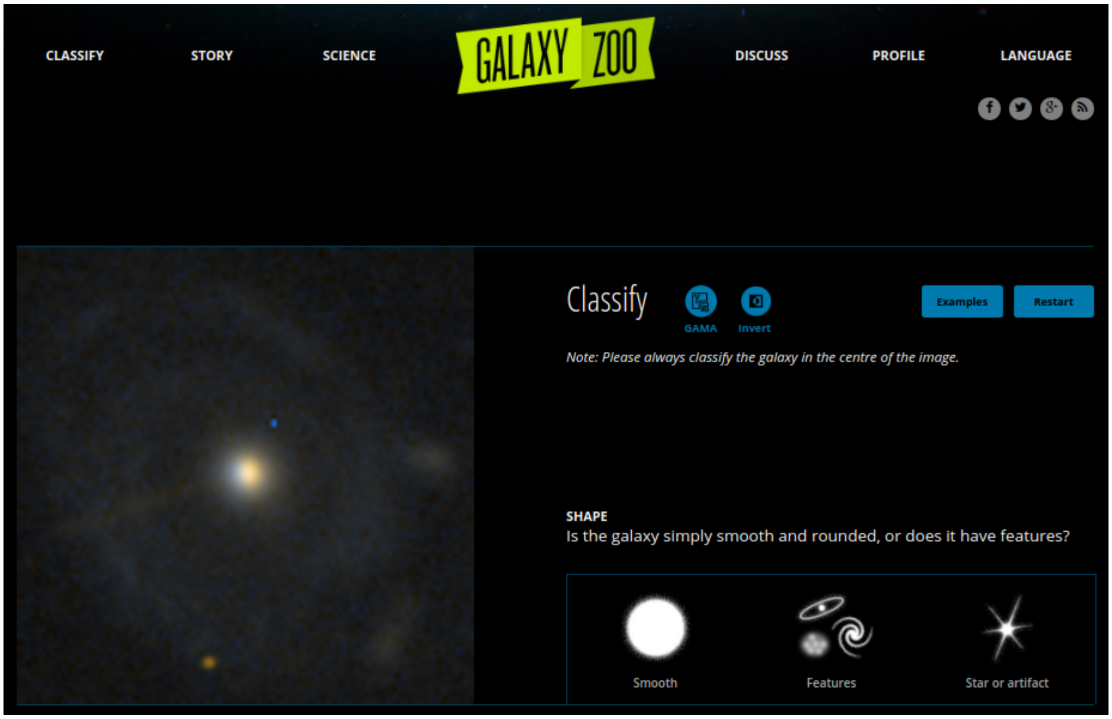

Figure 2.2 AHMAH MUTHAFUCKIN CAPTION BITCHES

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

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

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

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

This thesis draws on the *raw* vote fractions for reasons explained in Chapter 3. Focus now turns to the methodology of measuring automatic morphological indictors which are used in Chapter 4 to train a machine classifier.

2.2 Automatic morphology indicators

Any machine classifier must have a set of features from which to learn to differentiate between classes. These features can be anything: pixel values, spectra, anything that can correlates well enough with the class boundary that the machine can learn that boundary. Choosing which features are most appropriate for each task is not necessarily straight-forward or easy. Too few features, and the machine will not be able to learn the things that define each class. Too many features and machine training time could increase substantially or, worse, training will be subjected to the Curse of Dimensionality: with increasing parameter space dimensionality, the number of samples required to learn that parameter space increases exponetially!

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

2.2.1 Imaging data

The Galaxy Zoo 2 galaxy sample was originally composed of 295305 subjects. However, the catalog paper published by Willett et al. (2013) provided classifications for only 285XXX. In a private conversation with the author, the remaining ∼10K galaxies in the GZ2 sample were unfit for classification for one reason or another and were thus excluded [is this because they were the galaxies with m_r ¿ 17.0??. However, as we explain in the next chapter, we use volunteer classifications from the original GZ2 database which includes classifications of these ∼10K galaxies. Because it is possible that these galaxies could have their morphology more accurately characterized by machine (as opposed to human classifications during the GZ2 project), thus obtaining SDSS imaging for the full GZ2 sample is attempted.

i-band imaging (with central wavelength 7480Å) from SDSS Data Release 12 is obtained for 290,059 galaxies in the GZ2 project. These are drawn from the meta data associated with each subject in the catalog of GZ2 galaxies which includes the RUN, CAMCOL, FIELD, etc, location indicators specificaly for SDSS imaging data. Because the original GZ2 project obtained imaging from DR7, it is surmised that the route to some galaxies in DR12 no longer exist thus explaining the loss of 5246 galaxy images. Because this represents such a small percentage [XXX] of the total population these galaxies were not tracked down at this time.

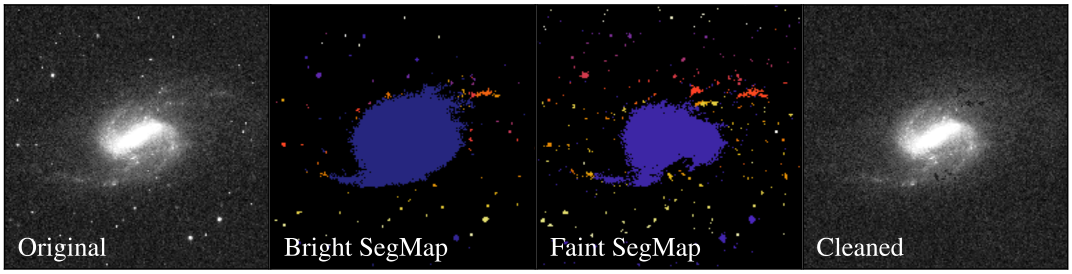

Figure 2.3 I'MMA FUCKING CAPTION BISH

151,987 SDSS fields were successfully downloaded for the remaining galaxies in the sample. Because each SDSS field is ∼130 square arcminutes, most fields contain more than one galaxy in the GZ2 sample. 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 mosaicing was not performed. This removed 7962 galaxies from our sample resulting in a final sample of 282,350 GZ2 galaxy postage stamps. A random sample of these postage stamps is shown in Figure .

2.2.2 Image cleaning

These postage stamps next undergo a cleaning process in order to remove the light from nearby sources so as not to contaminate the light profile of the galaxy of interest. Each stamp is processed through Source Extractor (ver. 2.8.6; Bertin & Arnouts, 1996), a software that automatically detects sources in CCD imaging based on a set of input parameters that control its sensitivity. Two sets of parameters are used: the first is designed to identify bright sources while the second is better optimized to detect fainter objects. This software produces segmentation maps that identify the boundaries of each detected object in an image. By design, the galaxy of interest is located at the center of the cutout. Extraneous sources are then identifed 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, while Figures XXX-XXX depict random samples of original and cleaned images for a variety

Full Galaxy Zoo 2 sample	295305
Successful cutouts	282350
Concentration	281945
M_{20}	282154
Asymmetry	282256
Gini coefficient	282252
Ellipticity $(1 - b/a)$	282337

of "difficulties". To zeroth order, 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

Once the postage stamps have been cleaned of nearby galaxies, measurements of the light profile can commence. The first stage in this process is the measurement of the Petrosian radius.

Compare my petro-rad to SDSS? –It doesn't look good.

We use the Petrosian radius to define the aperture size for some of the following morphology measurements. [why did I measure my own instead of using SDSS??] I have no idea.

We compute the following widely adopted nonparametric measurements of the galaxy light distribution on the cleaned postage stamps

Concentration

Concentration is computed as $C = 5\log(r_{80}/r_{20})$ where r_{80} and r_{20} are the radii containing 80% and 20% of the galaxy light respectively. Small values of this ratio tend to indicate disky galaxies, while larger values correlate with early-type ellipticals.

Asymmetry

Asymmetry quantifies the degree of rotational symmetry in the galaxy light distribution (not necessarily the physical shape of the galaxy as this parameter is not highly sensitive to low surface brightness features). A correction for background noise is applied (as in

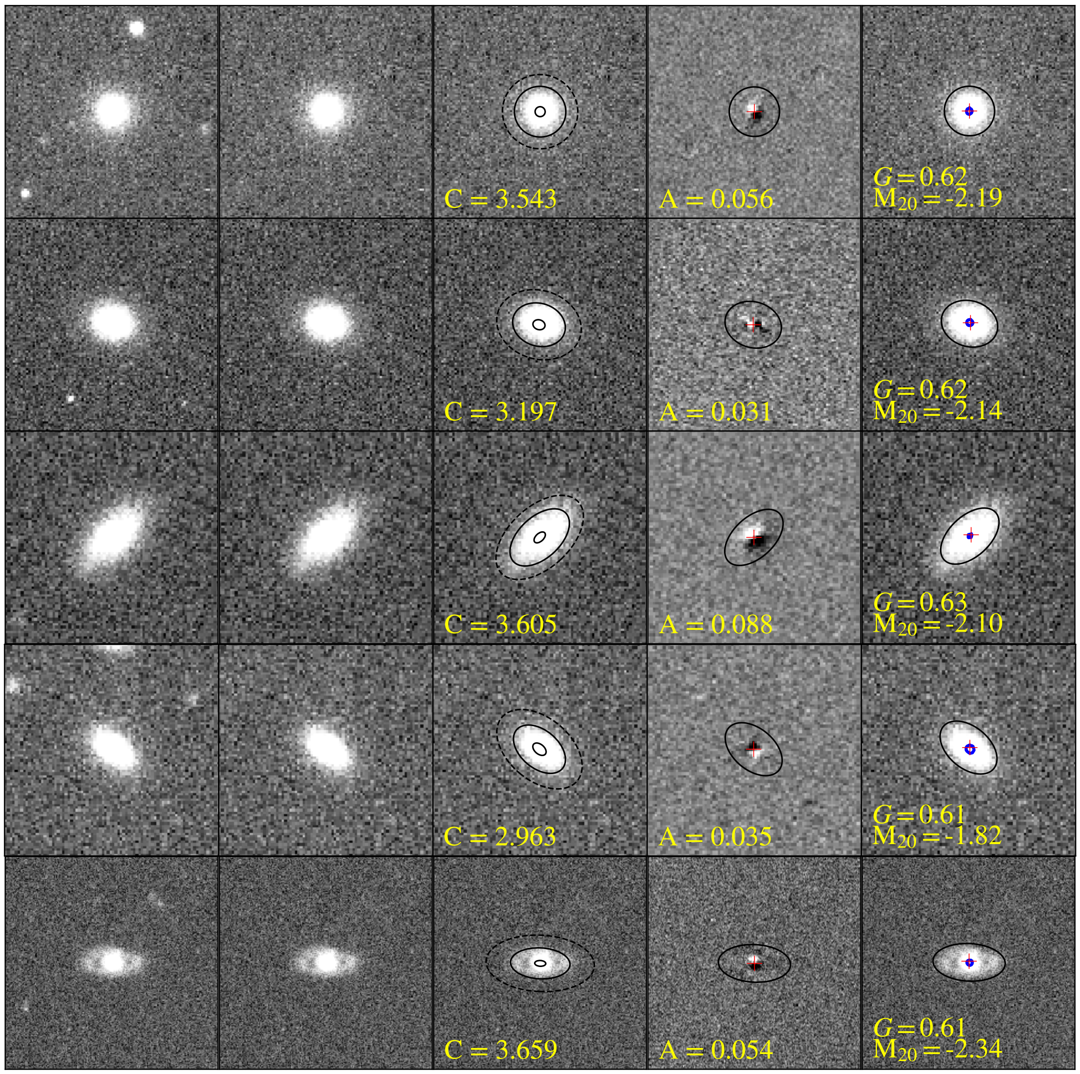

Figure 2.4 words words words I'm a fucking caption!

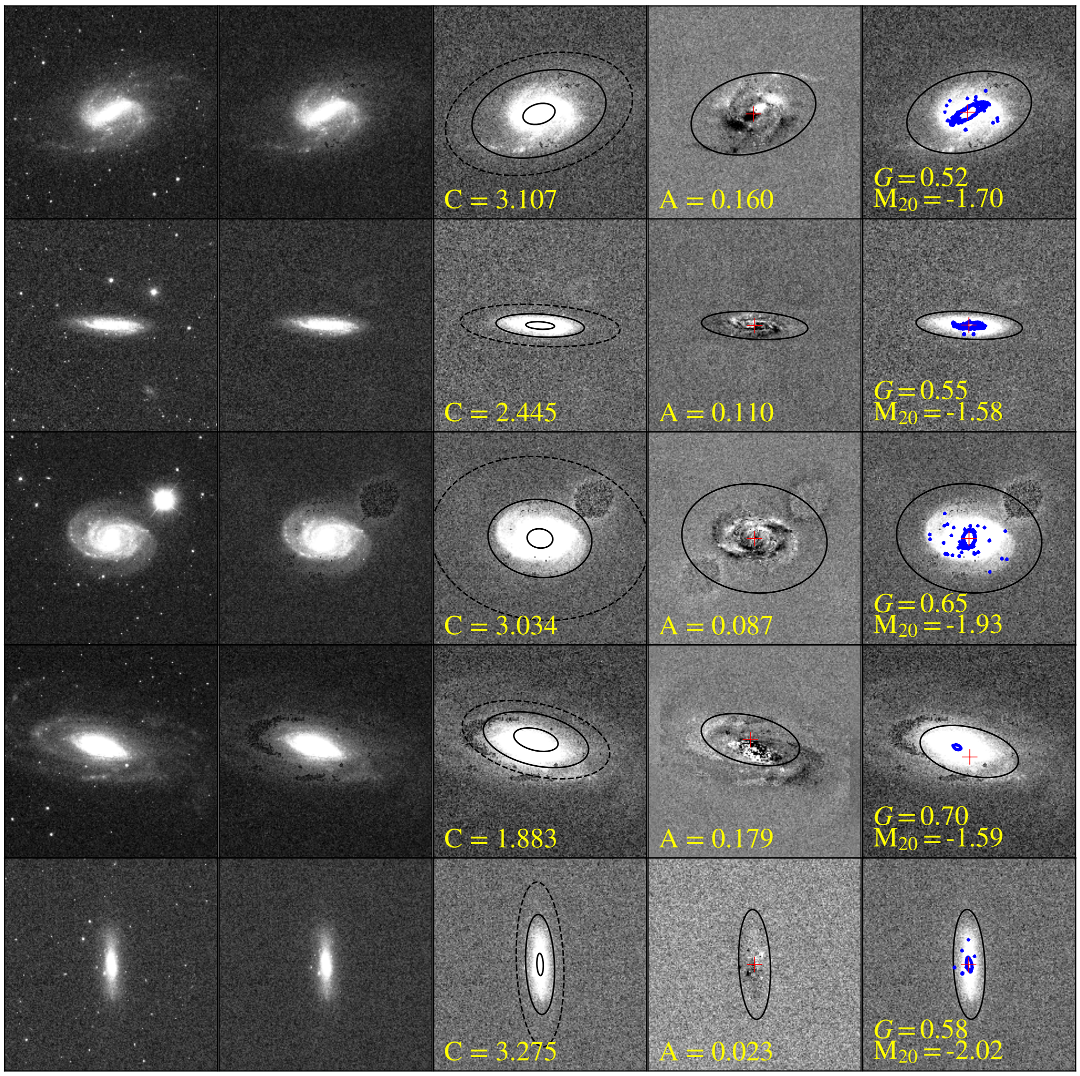

Figure 2.5 words words words I'm a fucking caption!

e.g. Conselice et al. (2000)), i.e.,

$$A = \frac{\sum_{x,y} |I - I_{180}|}{2\sum |I|} - B_{180} \tag{2.3}$$

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.

Gini coefficient

The Gini coefficient, G, (Glasser, 1962; Abraham et al., 2003) describes how uniformly distributed a galaxy's flux is. If G is 0, the flux is distributed homogeneously among all galaxy pixels; if G is 1, the light is contained within a single pixel. This term correlates with C, however, G does not require that the flux be in the central region of the galaxy. We follow Lotz et al. (2004) by first ordering the pixels by increasing flux value, and then computing

$$G = \frac{1}{|\bar{X}|n(n-1)} \sum_{i}^{n} (2i - n - 1)|X_i| \tag{2.4}$$

where n is the number of pixels assigned to the galaxy, and $\bar{X}$ is the mean pixel value.

$\mathbf{M}_{20}$

M_{20} (Lotz et al., 2004) is the second order moment of the brightest 20% of the galaxy flux. We compute it as

$$M_{tot} = \sum_{i}^{n} f_i[(x_i - x_c)^2 + (y_i - y_c)^2] \tag{2.5}$$

$$M_{20} = \log_{10}(\frac{\sum_i M_i}{M_{tot}}), \quad \text{while} \sum_i fi < 0.2 f_{tot} \tag{2.6}$$

where M$_{tot}$, the total moment, is computed first and f_{tot} is the total flux. For centrally concentrated objects, M_{20} correlates with C but is also sensitive to bright off-centre knots of light.

Ellipticity

Finally, 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.

In total, we measure morphological indicators for 282,350 SDSS galaxies. The relations between these diagnostics for the full sample is shown in the right panel of Figure 3.10. 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 Comparison with other samples

Compare sample against sersic index, SDSS concentration, split by morphology, etc.

2.3 Catalog of morphological indicators for 282350 SDSS galaxies

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

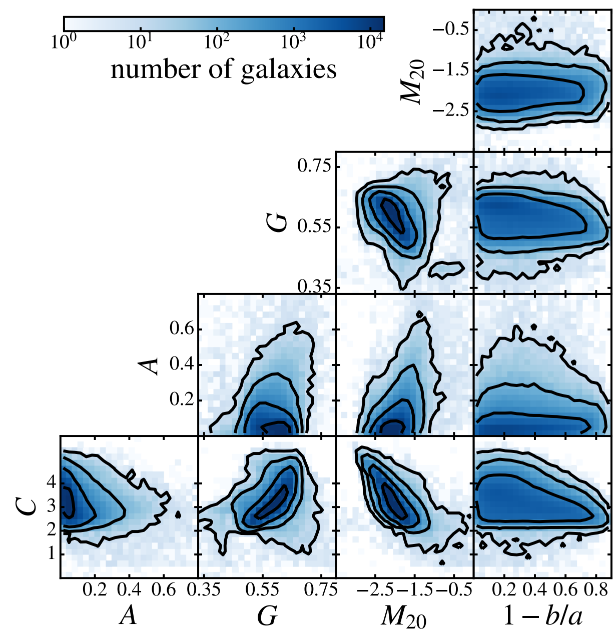

Figure 2.6 *Left.* Identifying 'Featured' subjects is independent of identifying 'Not' subjects. Both ROC curves use all subjects processed by SWAP where the score used to create the ROC curve is simply each subject's achieved posterior probability. The Featured curve demonstrates how well we identify '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. *Right.* Relation between measured morphology diagnostics for more than 280K SDSS galaxies. Most of these galaxies are processed through SWAP, receiving a posterior probability that estimates how likely each is to be 'Featured' or 'Not'.

Chapter 3

Intelligent management of visual classifications

3.1 Galaxy Zoo 2 Classification Data

Our simulations utilize original classifications made by volunteers during the GZ2 project. These data[1] are described in detail in Willett et al. (2013) and the preceding Chapter, though we provide a brief overview here. The GZ2 subject sample consists of 285,962 galaxies identified as the brightest 25% (r-band magnitude < 17) residing in the SDSS North Galactic Cap region from Data Release 7 and included subjects with both spectroscopic and photometric redshifts out to $z < 0.25$. Subjects were shown as colour composite images via a web-based 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

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

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

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

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

The labels we compute from GZ2 vote fractions are used solely to validate our classification method and are thus considered "ground truth," though this is, of course, subjective. Furthermore, we envision our framework being applied to never-before-classified image sets for which "ground truth" labels would not yet exist. Nevertheless, in Appendix **??** 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 labelled as type X, $\mathcal{N}_X$ is the number of subjects the volunteer has seen that were actually of type X, and $\mathbf{d}$ represents the history of the volunteer, i.e., all subjects they have seen. Therefore the confusion matrix for a single volunteer goes as

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

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

Each subject is assigned a prior probability that it exhibits feature A: $P(A) = p_0$. When a volunteer makes a classification, Bayes' theorem is used to compute how that subject's prior probability should be updated into a posterior using elements of the agent's confusion matrix. As the project progresses, each subject's posterior probability is updated after every volunteer classification, nudged higher or lower depending on volunteer input. Upper and lower probability thresholds can be set such that when a subject's posterior crosses the upper threshold it is highly likely to exhibit feature A; while if it crosses the lower threshold it is highly likely that feature A is absent. Subjects whose posteriors cross either of these thresholds are considered retired.

3.2.1 Gold-standard sample

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

We create a gold standard sample by selecting 3496 SDSS galaxies representative of the relative abundance of T-Types, a numerical index of a galaxy's stage along the Hubble sequence, at $z \sim 0$ by considering galaxies that overlap with the Nair & Abraham (2010) catalogue, a collection of ∼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

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

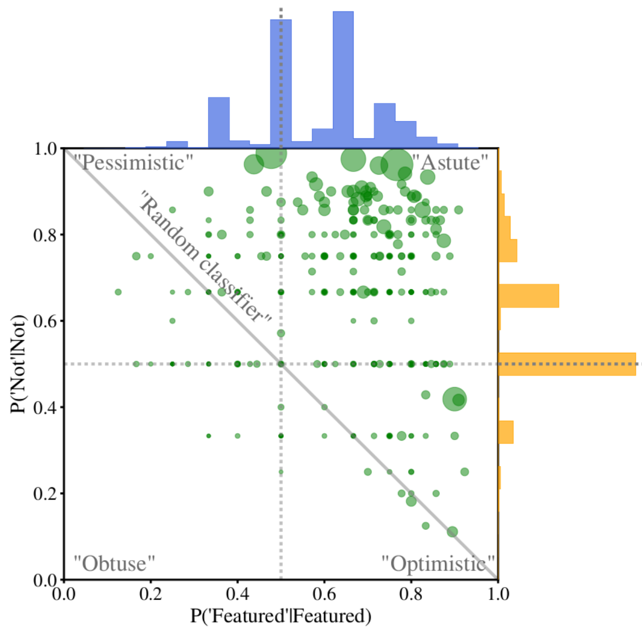

Figure 3.1 Confusion matrices for 1000 randomly selected GZ2 volunteers after fiducial SWAP assessment. Circle size is proportional to the number of gold standard subjects each volunteer classified. The histograms on top and right represent the distribution of each component of the confusion matrix for 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.

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

3.2.2 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 to be retired as 'Featured', to 0.99. Similarly, we set the 'Not' threshold, $t_N = 0.004$. In Section 3.3 we show that varying these parameters has only a small affect on the SWAP

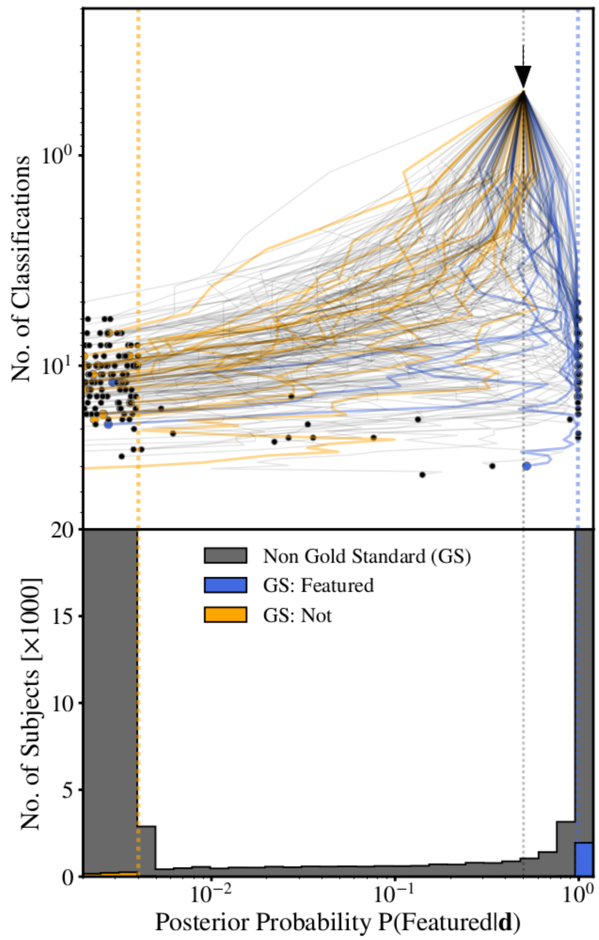

Figure 3.2 Posterior probabilities for GZ2 subjects. The top panel depicts the probability trajectories of 200 randomly selected GZ2 subjects. All subjects begin with a prior of 0.5 denoted by the arrow. Each subject's probability is nudged back and forth with each volunteer classification. From left to right the dotted vertical lines show the 'Not' threshold, prior probability, and 'Featured' threshold. Different colours denote different types of subjects. The bottom panel shows the distribution in probability for all GZ2 subjects by the end of our simulation, where the y axis is truncated to show detail.

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.

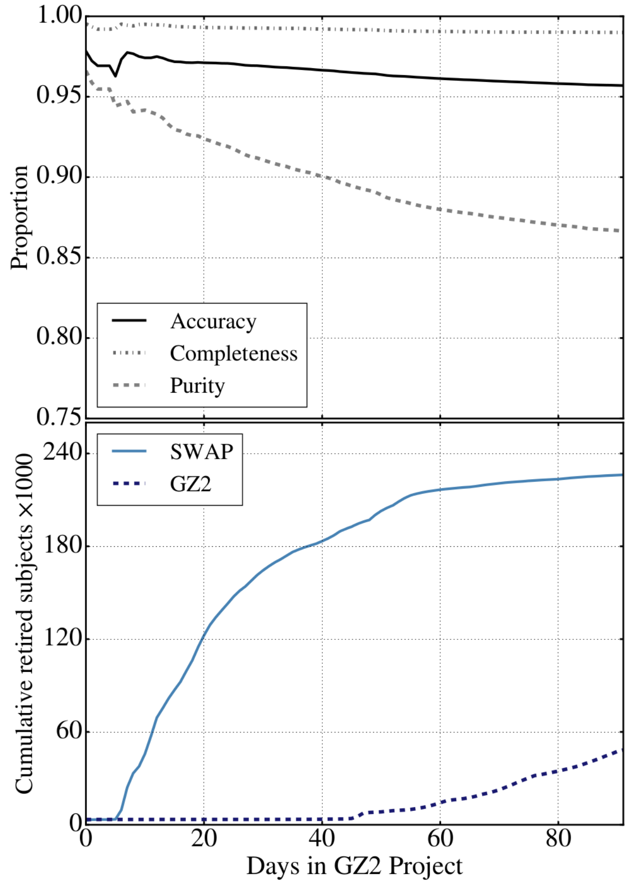

Figure 3.3 Fiducial SWAP simulation demonstrates a factor of 4.7 increase in the rate of subject retirement as a function of GZ2 project time (bottom panel, light blue) compared with the original GZ2 project (dashed dark blue). After 92 days, SWAP retires over 226K subjects, while GZ2 retires ∼48K. The top panel displays the quality metrics (greys). These are calculated by comparing labels predicted by SWAP to $\mathrm{GZ2}_{\mathrm{raw}}$ labels (Section A.3) for the subject sample retired by that day of the simulation. Thus, on the final day, SWAP retires 226,124 subjects with 95.7% accuracy, and with completeness and purity of 'Featured' subjects at 99% and 86.7% respectively. The decrease in purity as a function of time is due, in part, to the fact that more difficult to classify subjects are retired later in the simulation see Section 3.2.3

Figure 3.1 (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 3.2 (adapted from Figure 5 of Marshall et al. 2016) demonstrates how subject posterior probabilities are updated with each classification. The arrow in the top panel denotes the prior probability, $p_0 = 0.5$. With each classification, that prior is updated into a posterior probability creating a trajectory through probability space for each subject. The blue and orange lines show the trajectories of a random sample of 'Featured' and 'Not' subjects from our gold standard sample, while the black lines show the trajectories of a random sample of GZ2 subjects that were not part of the gold standard sample. The similarly coloured vertical dashed lines correspond to the retirement thresholds, t_F and t_N. The lower panel shows the full distribution of GZ2 subject 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 3.2.3). In contrast, a subject is considered SWAP-retired once its posterior probability crosses either of the retirement thresholds defined above.

However, it is important not to prioritize efficiency at the expense of quality. We thus also consider the metrics of accuracy, purity and completeness as a function of GZ2 project time. These are defined as follows: accuracy is the number of correctly identified subjects divided by the total number retired; completeness is the number of correctly identified 'Featured' subjects divided by the number of actual 'Featured' retired; and purity is the number of correctly identified 'Featured' subjects divided by

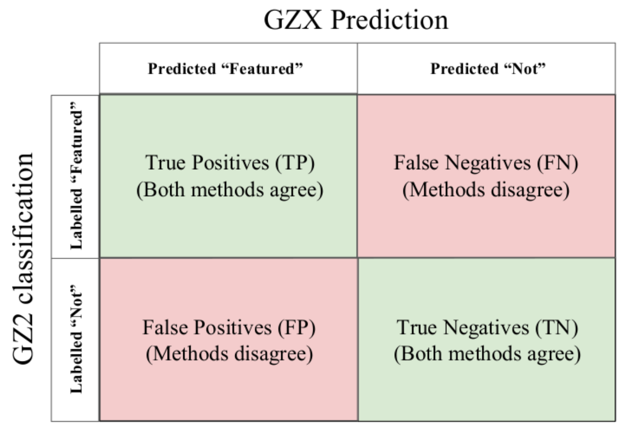

Figure 3.4 Confusion matrix for comparing GZ2 classifications to our method. True positives (TP) and true negatives (TN) indicate that the predictions from our method agree with GZ2 for subjects labelled 'Featured' and 'Not', respectively. When the two classification methods disagree, the result is a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

the number of subjects retired as 'Featured'.Because we have a binary classification we can construct a confusion matrix from which we can compute the quality metrics of accuracy, completeness and purity as a function of GZ2 project time by comparing our predicted labels to the GZ2$_{\mathrm{raw}}$ 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 has no false negatives whereas a pure sample has no false positives.Thus, a complete sample recovers *all* subjects labelled 'Featured' by GZ2, whereas a pure sample recovers *only* subjects labelled 'Featured' by GZ2. For example, by Day 20, SWAP retires 120K subjects with 96% accuracy, 99.7% completeness, and 92% purity.

Figure 3.3 and Table 4.1 detail the results of our fiducial SWAP simulation compared to the original GZ2 project. The bottom panel shows the cumulative number of retired subjects as a function of GZ2 project time. By the end of our simulation, GZ2 (dashed dark blue) retires ∼50K subjects while SWAP (solid light blue) retires 226,124 subjects. We thus classify 80% of the entire GZ2 sample in three months. Here we consider the number of classifications logged each day and not the length of time spent on a single classification. Under the assumption that collapsing the GZ2 decision tree to a single question would not have decreased the number of classifications collected each day during the GZ2 project Processing volunteer classifications through SWAP presents nearly a factor of 5 increase in classification efficiency. The top panel of Figure 3.3 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 3.2.5.

3.2.3 Intelligent subject retirement

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

We compute the raw vote fractions (f_{featured}, f_{smooth}, and f_{artifact}) for every subject in the GZ2 sample using only the first N classifications for $N \in [10, 15, 20, 25, 30, 35]$. From this, we compute descriptive labels as described in Section A.3. Because our SWAP simulation did not retire every subject in the GZ2 sample, we select 100 random subsamples each consisting of 226,124 subjects, and compute the average accuracy and the average number of total GZ2 classifications necessary retire that subsample. These results are shown in Figure 3.5 for each value of N along with the accuracy and total classifications for our SWAP simulation. We see that GZ2 needs at least 35 votes per subject in order to achieve consistent class labels 95% of the time. SWAP achieves the same accuracy with 30% fewer classifications. Furthermore, this justifies our choice of defining a subject as GZ2-retired once it reached at least 30 classifications.

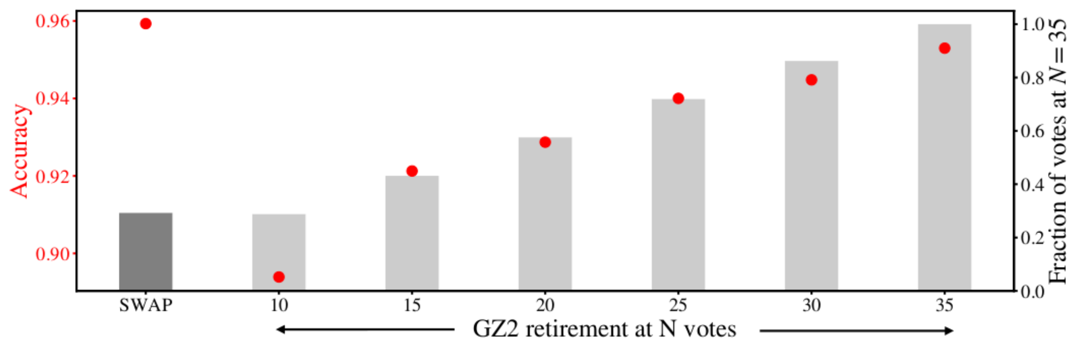

Figure 3.5 SWAP's intelligent retirement mechanism allows it to use 30% fewer classifications than GZ2 for the top-level question. To see this, we determine the minimum number of votes GZ2 needs in order to achieve consistent top-level classifications 95% of the time, allowing us to compare the total number of votes in our SWAP simulation to the total number required for GZ2. We compute the vote fractions f_{smooth}, f_{featured}, and f_{artifact} using only the first N votes in GZ2 and then compute the resulting class label. We compare these labels to those we originally computed in Section A.3 using the full GZ2 vote fractions and compute the accuracy and the total number of votes necessary for subject retirement. We do this for 100 randomly selected subsamples from GZ2 with the same number of subjects as retired during our SWAP simulation. The red dots show the average accuracy while the grey bars denote the average fraction of votes compared to that required for $N = 35$ (statistical error bars on these quantities are too small to be seen.) GZ2 needs at least 35 votes per subject in order to retain consistent subject classifications for 95% of the GZ2 sample. SWAP can surpasses this accuracy using only 30% as many votes.

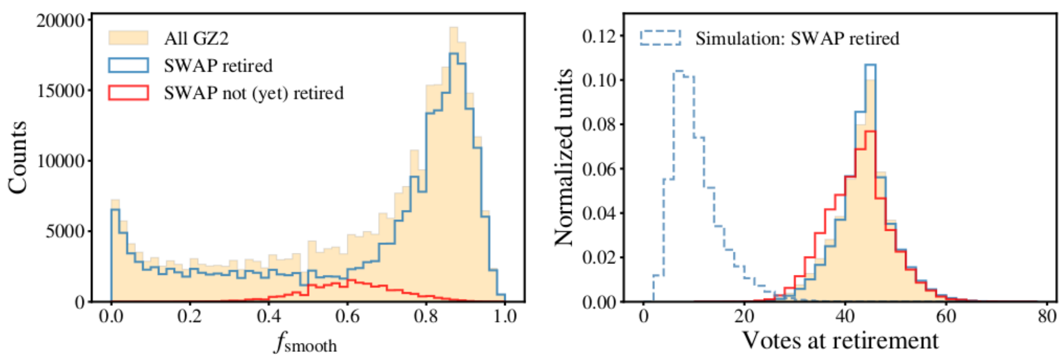

Figure 3.6 SWAP uses 30% fewer classifications to retire subjects due to its ability to retire easier subjects quickly, while more difficult subjects remain in the system to acquire additional classifications. The top left panel shows $f_{\rm smooth}$ for the entire GZ2 sample (orange), the subjects retired by SWAP (blue), and subjects that SWAP has not yet retired by the end of our simulation (red). The latter distribution peaks at $f_{\rm smooth} \sim 0.6$, which can intuitively be understood 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 top right panel provides additional evidence where here we show the number of votes at retirement for both the original GZ2 project (solid lines) and our SWAP simulation (dashed blue). The left-skew inherent in the red SWAP-not-yet-retired sample is due to difficult-to-classify subjects that received only 30-40 classifications during the GZ2 project. Even after processing all available classifications, SWAP cannot retire these subjects without additional volunteer input.

SWAP's performance can be explained through its retirement mechanism. GZ2 retired subjects randomly and required an arbitrary number of classifications per subject. In contrast, SWAP retires "easier" subjects first while harder subjects remain in the system for longer (requiring many more classifications to nudge that subject's posterior across a retirement threshold). Evidence for this can be seen in the top two panels of Figure 3.6. The top left panel shows the distribution of f_{smooth} for the entire GZ2 sample (orange), the SWAP-retired sample (blue), and the sample of subjects which SWAP has not yet retired, of which there are ~ 19K at the end of our simulation. The SWAP-retired sample generally follows the same distribution as GZ2-full except for the noticeable dip around $f_{\text{smooth}} = 0.6$. In contrast, the SWAP-not-yet-retired sample peaks at $f_{\text{smooth}} = 0.6$. These subjects can be interpreted as being the most difficult to classify which can be understood intuitively: galaxies with $f_{\text{smooth}} \leq 0.5$ are easily identified as having features, while galaxies with $f_{\text{smooth}} \geq 0.8$ are obviously elliptical.

This is further corroborated in the top right panel which shows the distribution of the number of classifications a subject had at the time of retirement. The solid lines show this distribution from the original GZ2 project for the same subsamples as the top left panel. For comparison, the dashed line shows the number of classifications at retirement realized during our SWAP simulation. Again, we see that the SWAP-retired sample is representative of GZ2 as a whole. However, the distribution for the SWAP-not-yet-retired sample is skewed toward fewer total classifications. To understand this, consider the following: GZ2 served subject images at random with the exception that, towards the end of the project, subjects with low numbers of classifications were shown at a higher rate to ensure that each galaxy had enough responses to accurately characterize the likelihood of the classification (Willett et al., 2013). The median number of classifications was 44 with the full distribution shown in orange in the top right panel of Figure 3.6. 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

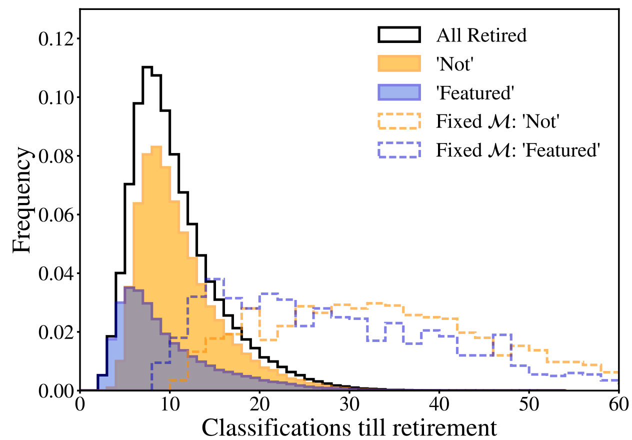

Figure 3.7 SWAP's volunteer-weighting mechanism provides a factor of three reduction in the human effort required to retire GZ2 subjects. The filled histograms show the number of volunteer classifications per subject achieved during our SWAP simulation broken down by class label, where the solid black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

to process the remaining GZ2 classifications, SWAP would eventually retire it, and 3) SWAP has seen most or all of the classifications for a subject but it is difficult and there are few or no remaining GZ2 classifications; without additional volunteer input, these subjects will never be retired by SWAP. It is this third category that skews the red distribution towards fewer GZ2 votes. These are difficult-to-classify subjects that have only 30 - 40 GZ2 classifications, all of which are processed by SWAP, but these subjects remain unretired.

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

3.2.4 Reducing human effort

The SWAP algorithm effectively weights volunteers according to how many gold standard subjects they correctly identify. This mechanism provides a reduction in the amount of human effort required to perform this classification task. 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 3.1. The dashed blue and orange distributions are the results from our toy model. When SWAP accounts for volunteer ability, most subjects are retired with between 6 and 15 votes, with a median of 9 votes. In contrast, when every volunteer is given equal weighting, subjects require 16 to 45 votes with a median of 30 votes before crossing one of the retirement thresholds. Thus the volunteer weighting scheme embedded in SWAP can reduce the amount of human effort required to retire subjects by a factor of three.

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

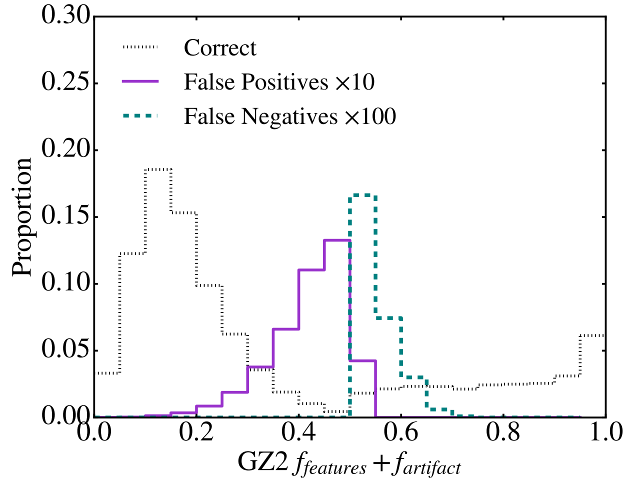

Figure 3.8 Distribution of GZ2 $f_{\text{featured}}+f_{\text{artifact}}$ vote fractions for subjects correctly identified by SWAP (dotted grey), along with those identified as false positives (solid purple), and false negatives (dashed teal). The false positives and false negatives are scaled by factors of 10 and 100 respectively for easier comparison. From Section A.3, subjects with values > 0.5 are defined as 'Featured', however, the teal distribution indicates that SWAP labels them as 'Not'. This is not a flaw of SWAP: 68.9% of incorrectly identified subjects have $0.4 \leq f_{\text{featured}}+f_{\text{artifact}} \leq 0.6$ 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 labelled 'Not'.

3.2.5 Disagreements between SWAP and GZ2

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

Instead we consider the stochastic nature of GZ2 vote fractions, which can be estimated as binomial. Let success be a response of "smooth" and failure be any other response. The 68% confidence interval on a subject with $f_{\mathrm{smooth}} = 0.5$ 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 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.

3.2.6 Summary

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

3.3 Exploring SWAP's Parameter Space

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

3.3.1 Initial agent confusion matrix.

In our fiducial simulation each volunteer was assigned an agent whose confusion matrix was initialized at $(0.5, 0.5)$, which presumes that volunteers are no better than random classifiers. We perform two simulations wherein we initialize agent confusion matrices as $(0.4, 0.4)$, slightly obtuse volunteers; and $(0.6, 0.6)$, slightly astute volunteers, with everything else remaining constant. Results of these simulations compared to the fiducial run are shown in the left panel of Figure 3.9. 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 3.9, 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 3.2.2. The dashed lines again denote the fiducial run. We maintain $\sim$95% accuracy, 99% completeness, and $\sim$84% purity; and no metric changes by $> 2\%$ regardless of initial confusion matrix values.

This spread is due to three effects: 1) subjects can receive an alternate SWAP label in different simulations, 2) subjects can be retired in a different order, and 3) the set of retired subjects is not guaranteed to be common to all runs. We find SWAP to be highly consistent: more than 99% of retired subjects are the same among all simulations, and, of these, 99% receive the same label. Instead we find that the order in which subjects

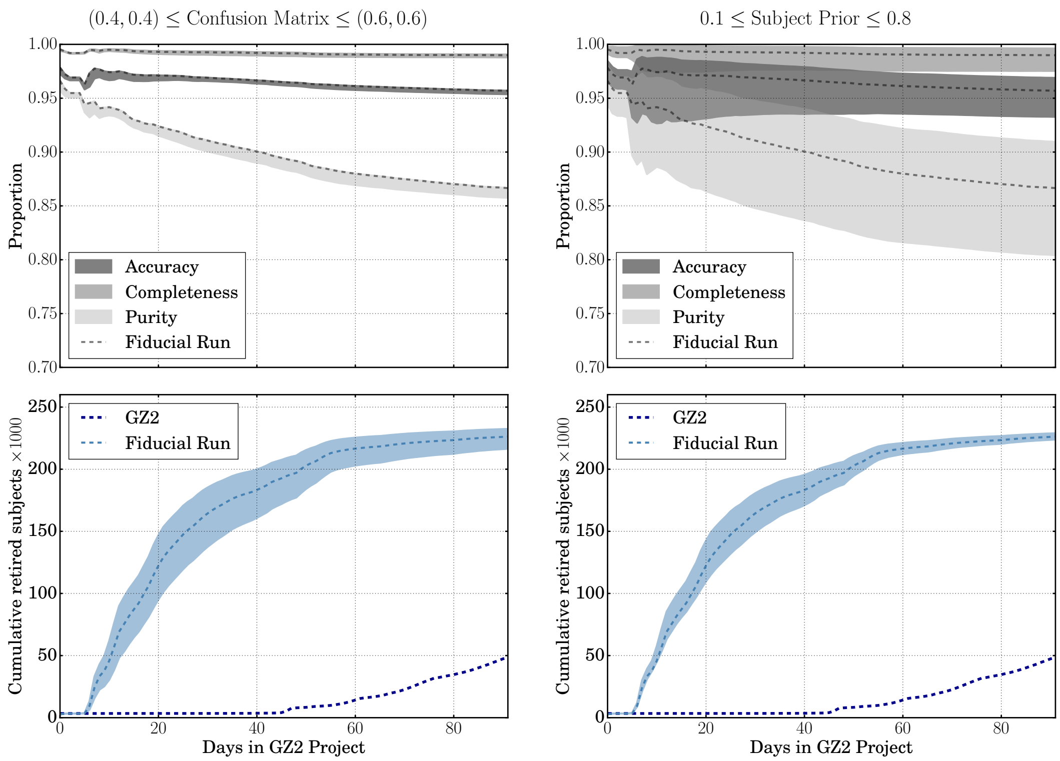

Figure 3.9 SWAP performance does not dramatically change even with a range of input parameters (shaded regions) as compared to the fiducial run of Section 3.2.2 (dashed lines). *Left.* The quality (top) and retirement rate (bottom) when the confusion matrix is initialized as (0.4, 0.4) and (0.6, 0.6), 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.

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.

3.3.2 Subject prior probability, p_0.

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

We run simulations allowing p_0 to take values 0.2, 0.35, and 0.8 and compare these to the fiducial run, with everything else remaining constant. The results are shown in the right panels of Figure 3.9. 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.

3.3.3 Retirement thresholds, t_F and t_N.

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

What thresholds should one choose? To answer this question, we consider the left panel of Figure 3.10, which depicts the receiver operating characteristic (ROC) curve for our fiducial simulation, an illustration of performance as a function of a threshold for a binary classifier. ROC curves display the true positive rate against the false positive

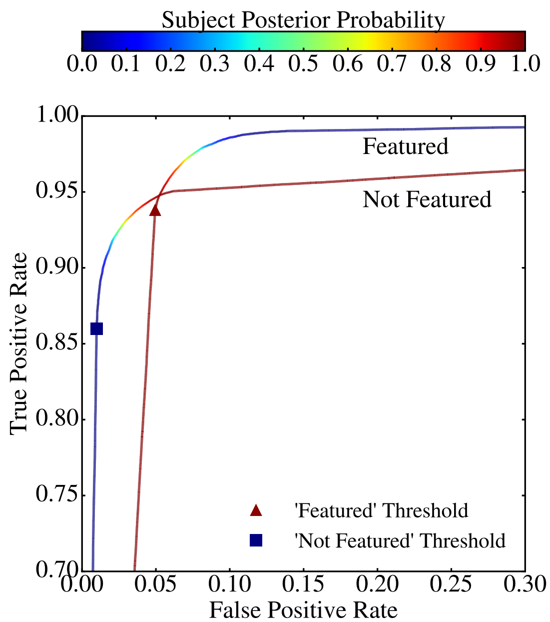

Figure 3.10 Identifying 'Featured' subjects is independent of identifying 'Not' subjects. Both ROC curves use all subjects processed by SWAP where the score used to create the ROC curve is simply each subject's achieved posterior probability. The Featured curve demonstrates how well we identify 'Featured' 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.

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

4.1 Efficiency through incorporation of machine classifiers

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

4.1.1 Random Forests

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

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

4.1.2 Grid Search and Cross-validation

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

4.1.3 Feature Representation and Pre-Processing

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

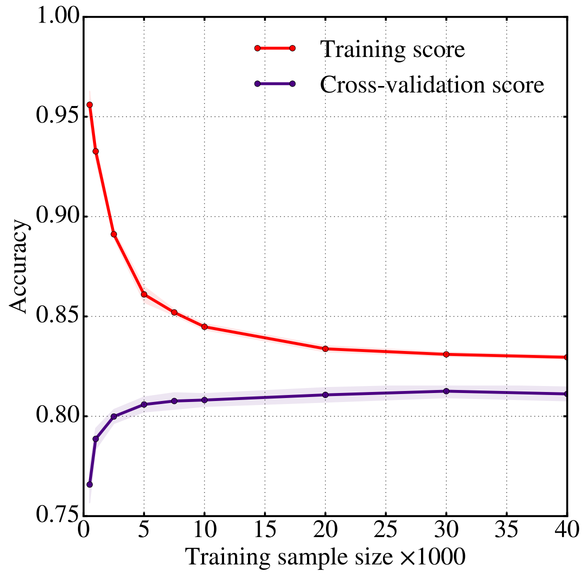

Figure 4.1 Learning curve for a Random Forest with fixed hyperparameters. These curves show the mean accuracy computed during cross-validation and on the training sample, where the shaded regions denote the standard deviation. When the training sample size is small, the machine accurately identifies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

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

4.1.4 Decision Engine

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

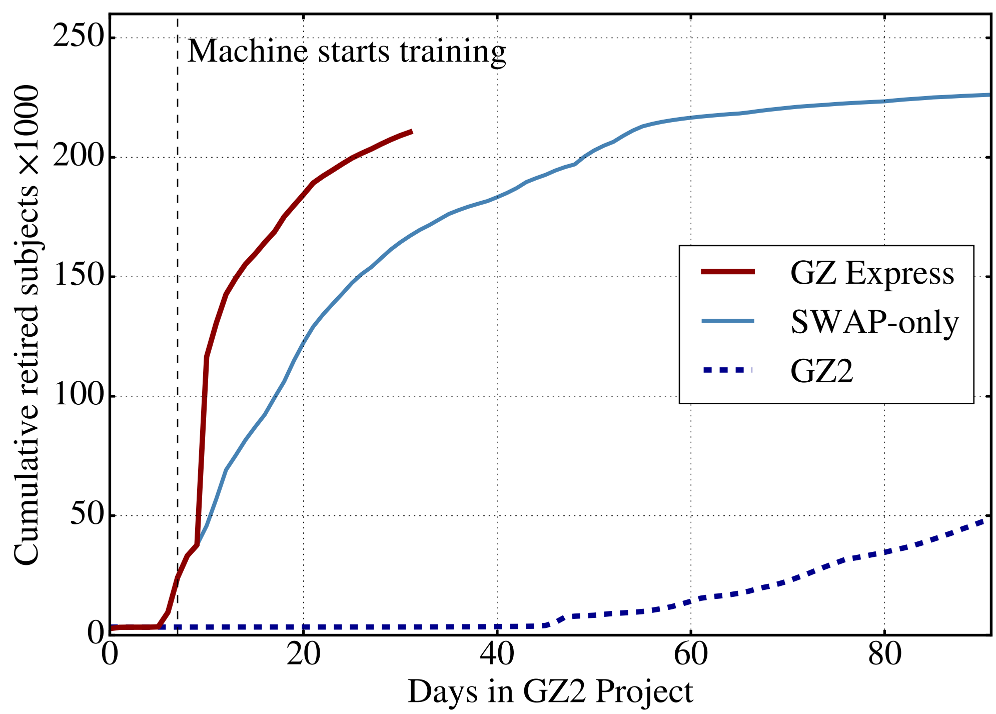

Figure 4.2 By incorporating a machine classifier, GZX (red) increases the classification rate by an order of magnitude compared to GZ2 (dashed dark blue) and out-performs the SWAP-only run (light blue), retiring more than 200K subjects in just 27 days of GZ2 project time. The dashed black line marks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

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

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

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

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

4.1.5 The Machine Shop

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

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

4.2 Results

We perform a full GZX simulation incorporating our RF with the fiducial SWAP run discussed in Section 3.2.2. 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 $\mathrm{GZ2}_{\mathrm{raw}}$ 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 93.5% and 84.2% 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 94.3%. We discuss this behaviour below.

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

4.2.1 Who retires what, when?

In the top panel of Figure 4.3 we explore the individual contributions to GZX subject retirement from the RF (dash-dotted teal) and SWAP (dashed orange). The solid black line shows the total GZX retirement (SWAP+RF), while the dotted grey line depicts the fiducial SWAP-only run from Section 3.2.2 for reference. Two things are immediately obvious. First, each component shoulders approximately half of the retirement burden with the machine and SWAP responsible for ∼100K and ∼110K 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 $GZ2_{\mathrm{raw}}$ labels as a function of GZ2 project time. Overall, GZX retires 73.6% of the GZ2 subject sample and this is evenly distributed between ‘Featured’ and ‘Not’ subjects as indicated by the solid black lines in both panels. However, SWAP retires more than 50% of all ‘Featured’ subjects while the machine retires only 18%. This divergence does not exist for ‘Not’ subjects

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	210543	932,017	93.5	84.2	94.3

where each component contributes 33-37%.

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

4.2.2 Machine performance

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

We isolate the 6127 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 only 6.25% of all subjects the machine retires. We visually examine several hundred and assess that, to the expert eye, a majority are, in fact, 'Featured'. A random sample is shown in Figure 4.4.

That the machine strongly identifies these galaxies as 'Featured' ($p_{\text{machine}} \geq 0.9$) where humans instead classify them as 'Not' ($f_{\text{featured}} < 0.5$) has several contributing factors: 1) as discussed in Section 3.2.5, the threshold we chose carries with it a confidence interval such that subjects with $0.4 < f_{\text{featured}} + f_{\text{artifact}} < 0.6$ are most likely to receive disagreeing labels from other classifying agents, 2) the first task of the GZ2 decision tree asks a question that does not necessarily correlate with a split between early- and late-type galaxies, and 3) the machine learns on morphology diagnostics that are very different from visual inspection.

We find that 41.4% of these false positives have $0.4 \leq f_{\mathrm{featured}}+f_{\mathrm{artifact}} < 0.5$ 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 43.5% of false positives have $f_{\mathrm{featured}}+f_{\mathrm{artifact}} \leq 0.35$, and this discrepancy is not as easily explained. In Figure 4.4 we examine a random sample of false positives in this regime where, for clarity, we display only the f_{featured} value in the lower left corner. The majority of these subjects are discs lacking features such as spiral arms or strong bars. Whether this is the reason the majority of volunteers classify these objects as "smooth" is beyond the scope of this paper, however, this behaviour might be modified by providing actual training images and live feedback as performed in Marshall et al. (2016). We suggest that, at least for this particular question, if either human or machine identifies a subject as 'Featured', it is likely the subject is discy 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 3.10. 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 1.1, several machines could train simultaneously, their predictions aggregated through SWAP, creating an on-the-fly machine ensemble.

With the above upgrades implemented, we expect performance of both the classification rate and quality to further increase. However, even our current implementation can cope with upcoming data volumes from large surveys. By some estimates, *Euclid* is expected to obtain measurable morphology with its visual instrument (VIS) for approximately $10^6 - 10^7$ galaxies (Laureijs et al., 2011). Visual classification at the rate achieved with Galaxy Zoo today would require 12–120 years to classify.[1] If the *Euclid* sample is on the high end, GZX as currently implemented could classify the brightest 20% during the six years of its observing mission. As currently implemented, we obtain accuracy around 95% potentially leaving hundreds of thousands of galaxies with unreliable classifications. In a companion paper that seeks to identify supernovae, Wright et al. (submitted) demonstrate a dramatic increase in accuracy through an entirely different human-machine combination whereby the scores from human and machine are averaged together with the combined score yielding the most reliable classification. Again, a combination of both approaches will allow us to take full advantage of legacy output from large scale surveys.

4.3.1 Conclusions

In this paper we design and test Galaxy Zoo Express, an innovative 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 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

[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`

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*.

4.4 Acknowledgements

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

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

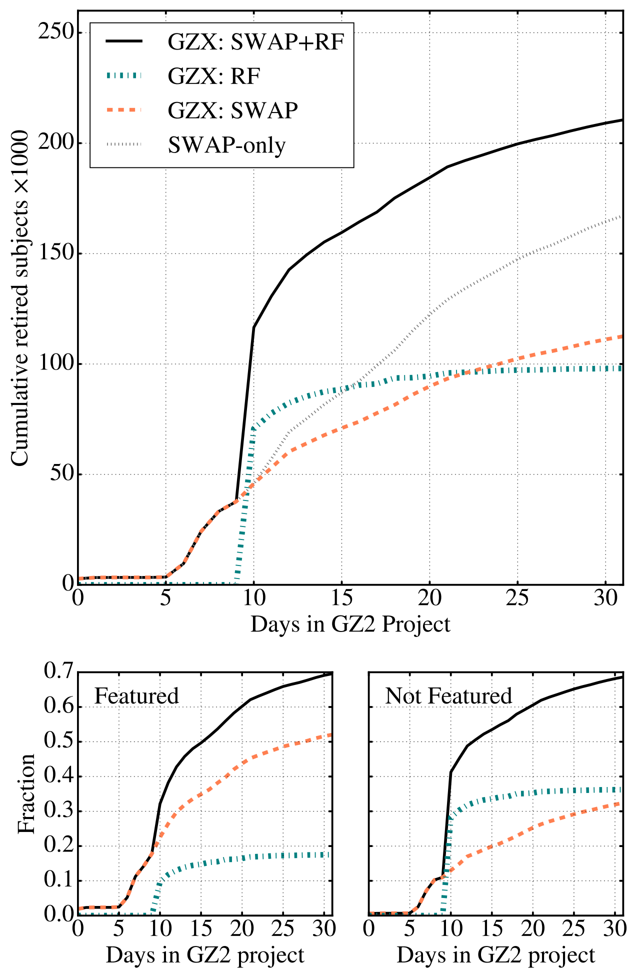

Figure 4.3 Contributions to subject retirement by both classifying agents of GZX: human (SWAP) and machine. The top panel shows cumulative subject retirement for GZX as a whole (solid black), along with that attributed to the RF (dash-dotted teal), and SWAP (dashed orange). The dotted grey line shows the fiducial SWAP-only run for comparison. Retirement totals for humans and machine are nearly equal over the course of the simulation but display different behaviours: SWAP's retirement rate is almost constant while the RF contributes substantially after its initial application and then plateaus. The bottom panels show what fraction of GZ2 subjects are retired, separated by class label. Overall, GZX retires 73.6% of the entire GZ2 sample in 32 days, retiring the same proportion of 'Featured' and 'Not' subjects as indicated by the black lines. However, humans retire 30% more 'Featured' subjects than the machine, while both components retire a similar proportion of 'Not' subjects.

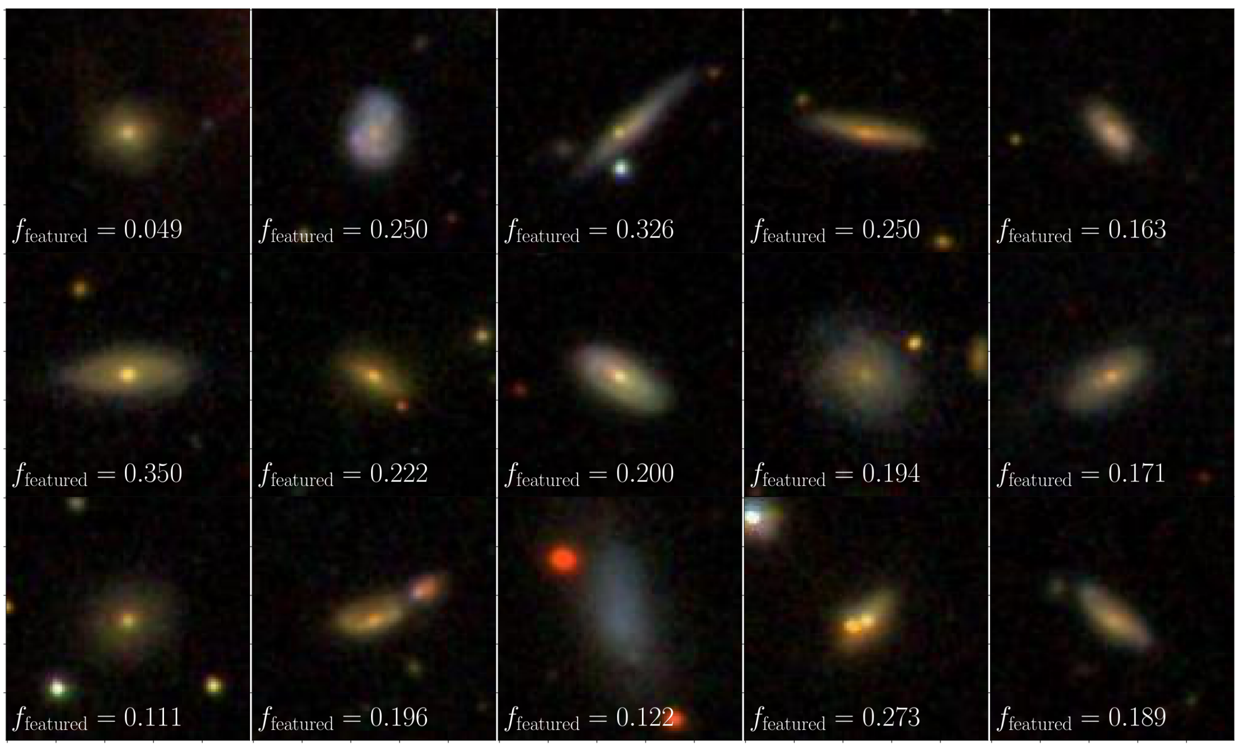

Figure 4.4 A random subsample of subjects identified as false positives: labelled by machine as 'Featured', but as 'Not' according to $\mathrm{GZ2}_{\mathrm{raw}}$. We display f_{smooth} in the lower left corner, that is, the fraction of volunteers who classified the subject as 'smooth' ('Not'). Values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

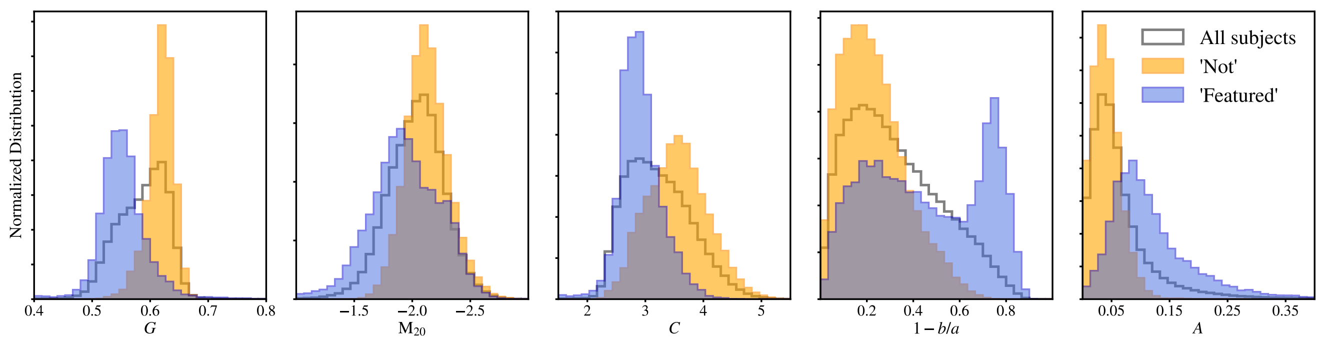

Figure 4.5 The RF is trained on a 5-dimensional morphology parameter space. We show the distribution of each morphology indicator for machine-retired 'Featured' (blue) and 'Not' (orange) subjects compared to the full GZ2 subject sample (black). The difference between 'Featured' and 'Not' subjects is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

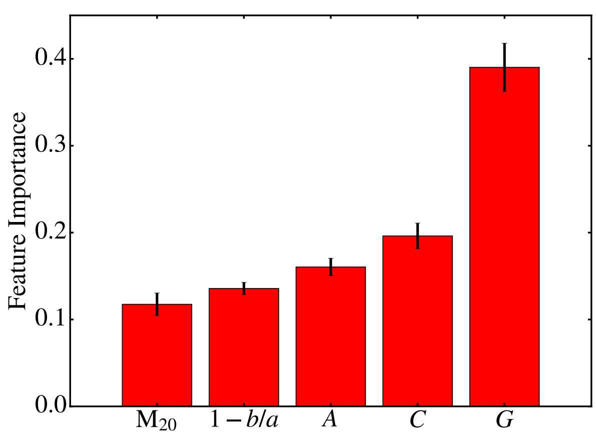

Figure 4.6 The RF's ranked feature importance averaged over all nights of training with black bars indicating the standard deviation. A larger value corresponds to higher importance. The machine computes feature importance according to how much each feature increases the purity of the resulting split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

Chapter 5

A population of "clumpy" galaxies in local Universe

All the things about clumps.

5.1 Sample Selection & Data

What was put into Galaxy Zoo: Hubble? All of Stripe 82 (but what does that mean?) Why were they put in?

From Willett+17:

Single-epoch images from SDSS Stripe 82 were selected using the criteria from Willett et al. 2013, which required limits of `petroR90_r` $> 3''$ and a magnitude brighter than $m_r < 17.77$. 21,522 galaxies in SDSS met these criteria. Co-added images from Stripe 82 were selected from the union of galaxies with co-added magnitudes brighter than 17.77 mag, and the galaxies detected in the stripe-82-single images and matched to a co-add source. This resulted in a total set of 30,339 images. Of the images in the co-added sample, 5144 (17%) were dimmer than the initial cut of 17.77 mag.

Delicious side-effect: Clumpies!

We consider the Stripe 82 subjects that were part of the GZH project. The GZH catalog doesn't provide debiased labels for the Stripe 82 subjects, most likely because this has already been done before in GZ2 and because the debiasing is completely different for these low redshift subjects. If we're now going to pick out subjects that are "featured" – shouldn't we use the debiased GZ2 values? How much does this change our sample? For that matter, how similar are the smooth and featured vote fractions between GZ2 and GZH? **Not going to answer this question in this paper.**

To select a sample of "clumpy" galaxies from the GZH Stripe 82 sample, we consider only those subjects with high featured and high clumpy vote fractions; the fraction of volunteers who classified each subject as being featured or clumpy. Specifically, we began with a selection criteria of $f_{\mathrm{featured}} \geq 0.5$ and $f_{\mathrm{clumpy}} \geq 0.5$ and $N_{\mathrm{votes}} \geq 20$, where N_{votes} is the number of volunteers who answerd the question "Does the galaxy have a mostly clumpy appearance?". This produced a a sample of 629 galaxies: 273 with single-epoch imaging and 356 with coadd imaging. After visual inspection we find that this is hardly a pure sample of traditional clumpy galaxies instead including many small groups of elliptical galaxies as well as galaxies in various merging states and possessing multiple nuclei. [**show example image?**] After excluding these and duplicate imaging, we retain 90 coadd-depth clumpy galaxies and 102 single-depth clumpy galaxies. Of these, 36 subjects have both single- and coadd-depth imaging.

[Note: technically ALL of these galaxies have single- and coadd-depth imaging because they are all in Stripe82. it's only a point of contention here because not all of them have the GZH morphologies assigned to them, for various reasons.]

We next select only those subjects which have spectroscopic redshift ¡ 0.06. Beyond this distance, the physical scale as observed with SDSS imaging is no longer similiar to Hubble's at z $\sim$ 3. The physical scale at z= 0.06 is 1"=1.1kpc. SDSS pixel scale is 0.396"/pixel. –¿ 0.43 kpc/pixel –¿ 2.3 pixels = 1kpc. Our final sample contains 105 unique galaxies, each with at least one SDSS spectrum. Half of our sample consist of objects with multiple spectra. We visually inspect these to verify that the SDSS fiber was indeed positioned over a star-forming region rather than the galactic bulge or other structure.

We obtain SDSS DR12 *ugriz* coadd imaging and spectra within 30" for every galaxy in our sample, discarding those spectra which belong to nearby sources or stars. We find that one "clumpy" galaxy is, in fact, a juxtoposition of three galaxies at disparate redshifts flagged as clumpy due to the low resolution of SDSS imaging. We exclude this subject(s) from our sample. Our final list includes 175 spectra. Approximately half of the galaxies in our sample have more than one SDSS spectrum with a handful having three or four spectra.

5.2 Analysis

During visual inspection we discover that several of the galaxies in our sample have very low surface brightness. Additionally, extremely bright clumps and nearby sources make SDSS photometry (and, subsequently, stellar masses) suspect. In this section we describe our analysis of basic galaxy parameters.

We create postage stamps of each galaxy from each field and all bands of our SDSS coadd imaging. The cutout sizes are determined to be 5 times the petrosian radius as computed by the SDSS pipeline. As mentioned above, however, this pipeline fails to properly account for galaxy extent and thus produces unrealistic measurements of the petrosian radius. We visually inspect all of our cutouts and choose an appropriate postage stamp size that fully encompasses each galaxy. The postage stamp radius is determine from the r-band petrosian radius and this value is used as the stamp size for each band.

We next process the r-band (maybe g-band cuz that's more blue?) postage stamp with Source Extractor (Bertin and whoever) using parameters designed to detect low surface brightness features. Important parameters are detailed in Table X that doesn't exist yet. Though these parameters adequately identify most of our sample, galaxies that fail are redone individually and SE parameters are tweaked for each particular case. These segmentation maps define the galaxy extent for all future analysis.

5.2.1 Measuring clump radial distance

5.2.2 Clump Halpha Hbeta ratios [Dust]

5.2.3 Clumpy Halpha equivalent widths

5.3 Discussion / Conclusions

our clumps are the best clumps. Everybody says so.

We thank all the people who contributed to Galaxy Zoo.

Chapter 6

Summary & Future Work

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

Our $2 - 5$ μm adaptive optics imaging of the cool hypergiant VY CMa penetrates deeper into its dusty nebula than in the optical, probing its recent mass-loss history. In Chapter 2 we analyzed the resolved images of its peculiar "Southwest" Clump, which has no obvious counterpart on the opposite side of the star. The distinct shape of the SW Clump is suggestive of a short-lived, localized event and may be analogous to a coronal mass ejection (CME) from a single location on the Sun's surface. A short-lived ejection event is consistent with the SW Clump appearing as a confined, coherent shape several hundred years after ejection. Using adaptive optics imaging polarimetry, in Chapter 3 we demonstrated the Clump is optically thick through at least 3.1 μm and reaffirmed the lower limit mass of 5×10^{-3} $M_{\odot}$ based on modeling its surface brightness as optically thick scattered light. Our 1.3 μm polarimetry detects several other prominent features of VY CMa's nebula include the NW Arc, Arc 2, S Knot, and S Arc. Using the polarized intensity as a lower limit on total scattered light intensity, we found each of these features to be optically thick as well. Their relatively high intrinsic polarizations are consistent with their high scattering optical depths since the depolarizing effect of multiple scatters is reduced for typical silicate grain albedos. In Chapter 4, we used 20 $-$ 37 μm infrared imaging with SOFIA/FORCAST to search for evidence of earlier mass loss. VY CMa's morphology at the longest wavelengths coincides with the general shape of the highly asymmetric nebulae seen in the visual, suggesting thermal emission from dust associated with the expanding arcs to the northwest and southwest. Modeling its SED we computed an average mass-loss rate of 6×10^{-4} $M_{\odot}$ yr^{-1} over the past $\sim$1200 years, with no clear evidence of mass loss much farther in its past.

Our study of the warm hypergiant IRC +10420 similarly traced its mass-loss over a range of angular scales. At the sub-arcsecond scale using adaptive optics, in Chapter 3 we used 2.2 μm polarimetry to reveal a relatively uniform nebula largely in the plane of the sky extending out to $2\farcs5$ from the star. This low-latitude ejecta is optically thick. Combining the polarimetry with 3 $-$ 5 μm imaging that shows extended emission, we modeled the flux of this nebula and found its emission is an order of magnitude brighter than can be explained by simple extrapolation of the scattered light seen at 2.2 μm. We hypothesized grains warmed to a temperature higher than the expected grain equilibrium temperature, but consistent with the local gas temperature in this region. In Chapter 4 we presented 8 $-$ 12 μm adaptive optics images that reveal spatially extended emission spanning nearly the same range as the 2.2 μm polarimetry. Applying radiative transfer modeling to the intensity profile of this extended emission and the SED, we found that IRC +10420's mass-loss history is divided into two distinct periods. Our best-fit model showed that it lost mass at a high average rate of 2×10^{-3} $M_{\odot}$ yr^{-1} from 6000 $-$ 2000 yr ago during its presumed RSG stage, followed by an order of magnitude decrease to an average rate of 1×10^{-4} $M_{\odot}$ yr^{-1} in the past 2000 yr.

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

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

In the future, infrared imaging over a range of angular scales will be used to expand the sample of RSGs and post-RSGs with resolved mass-loss. The SOFIA/FORCAST Cycle 3 program 03_0082 (PI: R. Humphreys) included three additional RSG targets (NML Cyg, VX Sgr and S Per) and contingent approval has been granted for observations of the yellow hypergiant HR 5171A during a Cycle 4 Southern deployment if flight scheduling permits. In Fall 2016, VY CMa is scheduled for observation with NOMIC, the $8-13$ μm imager on the Large Binocular Telescope. With NOMIC's $0\farcs27$ angular resolution (for single dish observing mode), VY CMa's SW Clump should be resolved. These observations will better sample the Clump's SED and could help resolve the puzzle of its non-detection by ALMA reported in **?**.

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

References

Abazajian, K. 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

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

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

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

Banerji, M., Lahav, O., Lintott, C. J., et al. 2010, MNRAS, 406, 342

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

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

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

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

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

Bunker, A., Smith, J., Spinrad, H., Stern, D., & Warren, S. 2003, Ap&SS, 284, 357

Cantalupo, S., Lilly, S. J., & Haehnelt, M. G. 2012, MNRAS, 425, 1992

Cardamone, C., Schawinski, K., Sarzi, M., et al. 2009, MNRAS, 399, 1191

Casteels, K. R. V., Conselice, C. J., Bamford, S. P., et al. 2014, MNRAS, 445, 1157

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

Colbert, J. W., Scarlata, C., Teplitz, H., et al. 2011, ApJ, 728, 59

Conselice, C. J. 2003, ApJS, 147, 1

—. 2006, MNRAS, 373, 1389

Conselice, C. J., Bershady, M. A., & Jangren, A. 2000, ApJ, 529, 886

Darg, D. W., Kaviraj, S., Lintott, C. J., et al. 2010, MNRAS, 401, 1552

de Vaucouleurs, G. 1959, Handbuch der Physik, 53, 275

Dieleman, S., Willett, K. W., & Dambre, J. 2015, MNRAS, 450, 1441

Dijkstra, M., & Kramer, R. 2012, MNRAS, 424, 1672

Dijkstra, M., & Loeb, A. 2008, MNRAS, 386, 492

—. 2009, MNRAS, 400, 1109

Dressler, A. 1980, ApJ, 236, 351

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

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

Fardal, M. A., Katz, N., Gardner, J. P., et al. 2001, ApJ, 562, 605

Fasano, G., Poggianti, B. M., Couch, W. J., et al. 2000, ApJ, 542, 673

Faucher-Giguère, C.-A., Kereš, D., Dijkstra, M., Hernquist, L., & Zaldarriaga, M. 2010, ApJ, 725, 633

Francis, P. J., Woodgate, B. E., Warren, S. J., et al. 1996, ApJ, 457, 490

Freeman, P. E., Izbicki, R., Lee, A. B., et al. 2013, MNRAS, 434, 282

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

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

Geach, J. E., Smail, I., Chapman, S. C., et al. 2007, ApJ, 655, L9

Geach, J. E., Matsuda, Y., Smail, I., et al. 2005, MNRAS, 363, 1398

Geach, J. E., Alexander, D. M., Lehmer, B. D., et al. 2009, ApJ, 700, 1

Geach, J. E., Bower, R. G., Alexander, D. M., et al. 2014, ApJ, 793, 22

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

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

Griffith, R. L., Cooper, M. C., Newman, J. A., et al. 2012, ApJS, 200, 9

Haiman, Z., & Rees, M. J. 2001, ApJ, 556, 87

Haiman, Z., Spaans, M., & Quataert, E. 2000, ApJ, 537, L5

Hayes, M., & Scarlata, C. 2011, in SF2A-2011: Proceedings of the Annual meeting of the French Society of Astronomy and Astrophysics, ed. G. Alecian, K. Belkacem, R. Samadi, & D. Valls-Gabaud, 129–133

Hayes, M., Scarlata, C., & Siana, B. 2011, Nature, 476, 304

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

Hubble, E. P. 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

Humphrey, A., Vernet, J., Villar-Martín, M., et al. 2013, ApJ, 768, L3

Jimenez, R., & Haiman, Z. 2006, Nature, 440, 501

Kartaltepe, J. S., Mozena, M., Kocevski, D., et al. 2015, ApJS, 221, 11

Kasper, M. E., Beuzit, J.-L., Verinaud, C., et al. 2008, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 7015, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, 1

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

Keller, C. U., Schmid, H. M., Venema, L. B., et al. 2010, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 7735, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, 6

Kormendy, J. 1977, ApJ, 217, 406

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, H.-W., & Ahn, S.-H. 1998, ApJ, 504, L61

Lintott, C., Schawinski, K., Bamford, S., et al. 2011, MNRAS, 410, 166

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

Lintott, C. J., Schawinski, K., Keel, W., et al. 2009, MNRAS, 399, 129

Loeb, A., & Rybicki, G. B. 1999, ApJ, 524, 527

Lotz, J. M., Primack, J., & Madau, P. 2004, AJ, 128, 163

LSST Science Collaboration, Abell, P. A., Allison, J., et al. 2009, ArXiv e-prints, arXiv:0912.0201

Marshall, P. J., Verma, A., More, A., et al. 2016, MNRAS, 455, 1171

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

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

Matsuda, Y., Yamada, T., Hayashino, T., et al. 2004, AJ, 128, 569

Matsuo, T., & Tamura, M. 2010, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 7735, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, 84

McLinden, E. M., Malhotra, S., Rhoads, J. E., et al. 2013, ApJ, 767, 48

Meert, A., Vikram, V., & Bernardi, M. 2016, MNRAS, 455, 2440

More, A., Verma, A., Marshall, P. J., et al. 2016, MNRAS, 455, 1191

Mori, M., & Umemura, M. 2006, Nature, 440, 644

Nair, P. B., & Abraham, R. G. 2010, ApJS, 186, 427

Nakamura, O., Fukugita, M., Yasuda, N., et al. 2003, AJ, 125, 1682

Neufeld, D. A. 1990, ApJ, 350, 216

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

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

Palunas, P., Teplitz, H. I., Francis, P. J., Williger, G. M., & Woodgate, B. E. 2004, ApJ, 602, 545

Patat, F., Moehler, S., O'Brien, K., et al. 2011, A&A, 527, A91

Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, Journal of Machine Learning Research, 12, 2825

Peng, C. Y., Ho, L. C., Impey, C. D., & Rix, H.-W. 2002, AJ, 124, 266

Peng, Y.-j., Lilly, S. J., Kovač, K., et al. 2010, ApJ, 721, 193

Pertenais, M., Neiner, C., Parès, L. P., et al. 2014, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 9144, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, 3

Peth, M. A., Lotz, J. M., Freeman, P. E., et al. 2016, MNRAS, 458, 963

Plaszczynski, S., Montier, L., Levrier, F., & Tristram, M. 2014, MNRAS, 439, 4048

Prescott, M. K. M., Kashikawa, N., Dey, A., & Matsuda, Y. 2008, ApJ, 678, L77

Prescott, M. K. M., Momcheva, I., Brammer, G. B., Fynbo, J. P. U., & Møller, P. 2015, ApJ, 802, 32

Prescott, M. K. M., Smith, P. S., Schmidt, G. D., & Dey, A. 2011, ApJ, 730, L25

Rosdahl, J., & Blaizot, J. 2012, MNRAS, 423, 344

Rybicki, G. B., & Loeb, A. 1999, ApJ, 520, L79

Saito, T., Shimasaku, K., Okamura, S., et al. 2006, ApJ, 648, 54

Sandage, A. 1961, The Hubble atlas of galaxies

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

Schawinski, K., Urry, C. M., Simmons, B. D., et al. 2014, MNRAS, 440, 889

Seager, S., Cash, W. C., Kasdin, N. J., et al. 2014, in American Astronomical Society Meeting Abstracts, Vol. 224, American Astronomical Society Meeting Abstracts 224, 311.06

Sersic, J. L. 1968, Atlas de galaxias australes

Shen, S., Mo, H. J., White, S. D. M., et al. 2003, MNRAS, 343, 978

Sheth, K., Elmegreen, D. M., Elmegreen, B. G., et al. 2008, ApJ, 675, 1141

Simard, L., Mendel, J. T., Patton, D. R., Ellison, S. L., & McConnachie, A. W. 2011, ApJS, 196, 11

Simmons, B. D., Melvin, T., Lintott, C., et al. 2014, MNRAS, 445, 3466

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

Simmons, J. F. L., & Stewart, B. G. 1985, A&A, 142, 100

Skibba, R. A., Bamford, S. P., Nichol, R. C., et al. 2009, MNRAS, 399, 966

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

Smith, D. J. B., & Jarvis, M. J. 2007, MNRAS, 378, L49

Smith, D. J. B., Jarvis, M. J., Simpson, C., & Martínez-Sansigre, A. 2009, MNRAS, 393, 309

Smith, G. P., Treu, T., Ellis, R. S., Moran, S. M., & Dressler, A. 2005, ApJ, 620, 78

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

Stapelfeldt, K. R., Brenner, M. P., Warfield, K. R., et al. 2014, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 9143, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, 2

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

Stenflo, J. O. 1980, A&A, 84, 68

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

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

Taniguchi, Y., & Shioya, Y. 2000, ApJ, 532, L13

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

van Dokkum, P. G. 2001, PASP, 113, 1420

Vinokur, M. 1965, Annales d'Astrophysique, 28, 412

Wardle, J. F. C., & Kronberg, P. P. 1974, ApJ, 194, 249

Watanabe, M., Kodaira, K., & Okamura, S. 1985, ApJ, 292, 72

Weidinger, M., Møller, P., & Fynbo, J. P. U. 2004, Nature, 430, 999

Weijmans, A.-M., Bower, R. G., Geach, J. E., et al. 2010, MNRAS, 402, 2245

Whitmore, B. C., Lucas, R. A., McElroy, D. B., et al. 1990, AJ, 100, 1489

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

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

Yang, Y., Zabludoff, A., Tremonti, C., Eisenstein, D., & Davé, R. 2009, ApJ, 693, 1579

Yang, Y., Decarli, R., Dannerbauer, H., et al. 2012, ApJ, 744, 178

Appendix A

Spectro-polarimetry Confirms Central Powering in a 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, arbitrarily denoted the 'ordinary' (`ord`) and 'extraordinary' (`ext`) beams. Finally, the beams are passed through a grism dispersion element (`GRISM_1400V+18`) with a central wavelength of 5200 Å (∼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

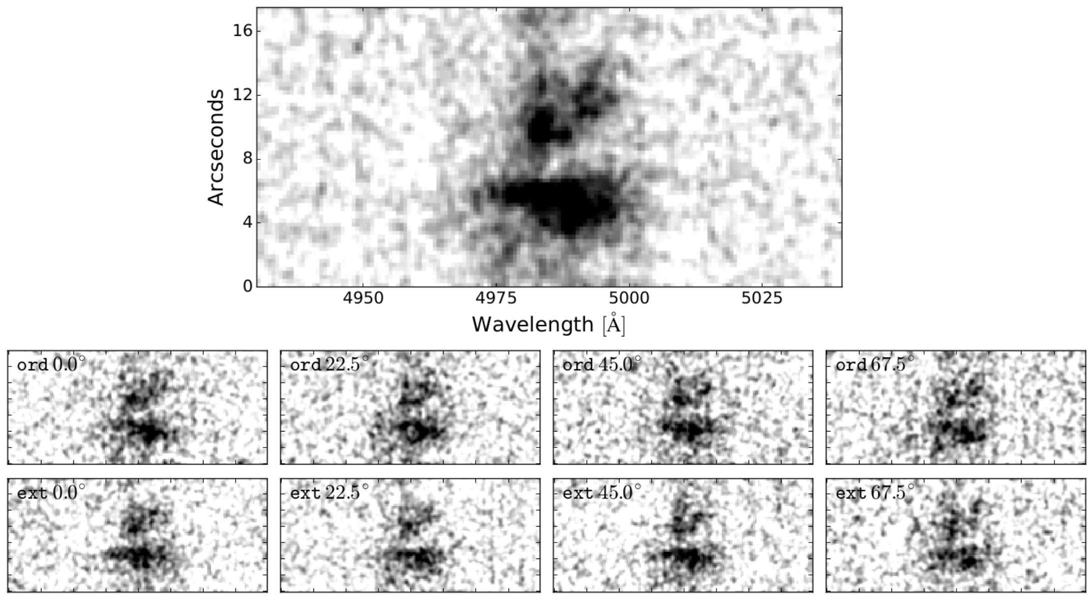

Figure A.1 *Top.* Spatial and wavelength distribution of the master total intensity Lyα spectrum. Co-added Lyα spectrum of 77 science spectra smoothed by a Gaussian with FWHM = $0.''5$. Due to light loss at the edges of the slit, we present here only the central $18''$. *Bottom.* Ordinary and extraordinary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

20 hours of observation time experienced astronomical seeing which varied between 0.5 and 1.0 arcsecond with median seeing at $\sim 0.''75$. 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`

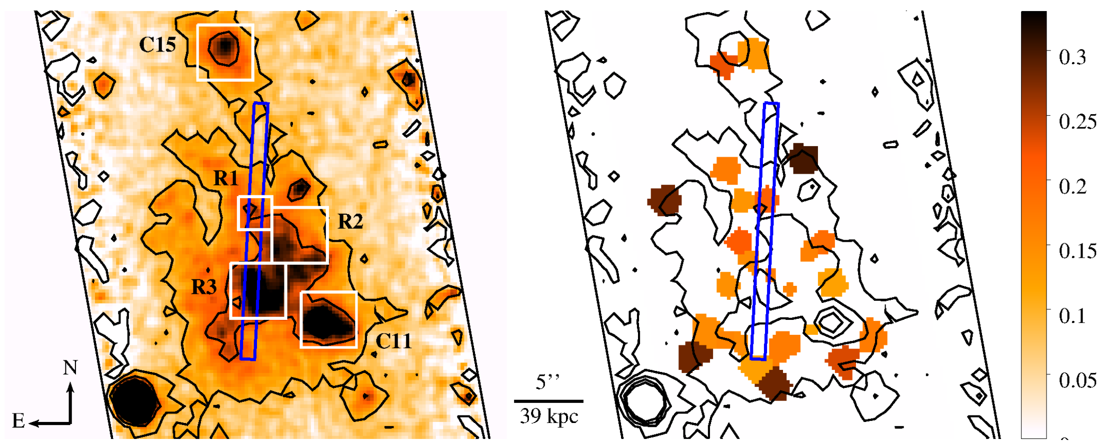

Figure A.2 Slit position over LAB1 as compared with results from H11. In the left panel we show the combined Lyα intensity frame from H11 adaptively smoothed to show detail including emission from LAB1 and nearby LBGs C11 and C15. Overlaid in white are boxes from Weijmans et al. (2010), regions in which 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:

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

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 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

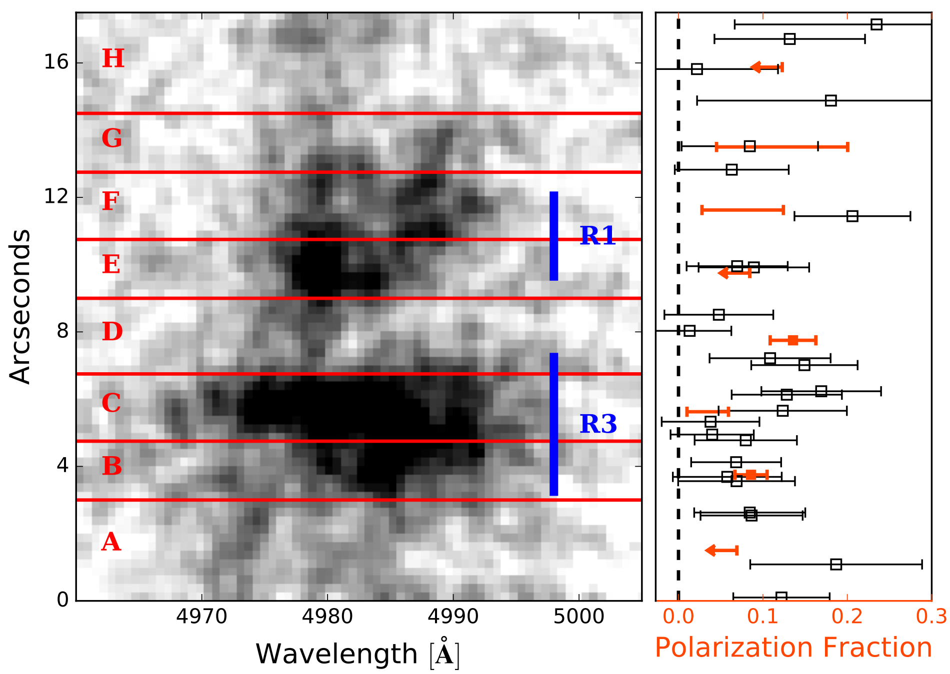

Figure A.3 Polarization of spectrally integrated Lyα. In the left panel we show the spatial binning of the Lyα emission spectrum with bins ranging from 2-3$''$ overlaid in red. Lyα emission is integrated over 4965-5000 Å. The blue lines indicate the spatial extent of Weijmans et al. (2010) IFU boxes. 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

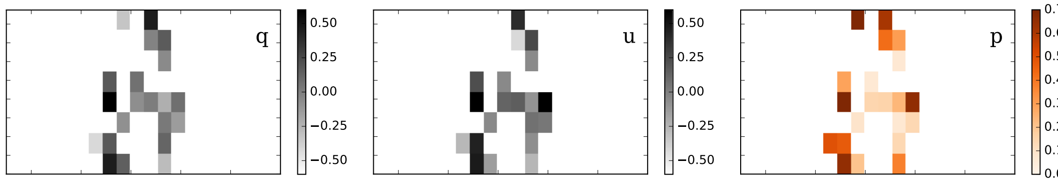

Figure A.4 2D Stokes parameters and polarization maps. The q and u maps are computed directly from the science frames shown in the bottom panel of Figure A.1 according to the prescription of Equation A.2 after binning as described in subsection A.4.2. The polarization map in the right panel is then computed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

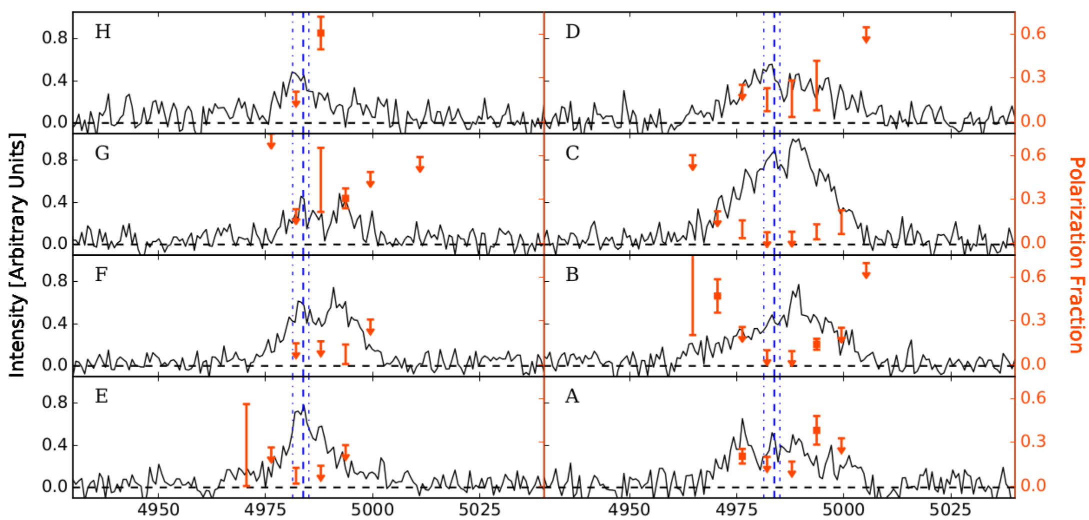

Figure A.5 Polarization of Lyα as a function of wavelength. Figures **A** through **H** contain the extracted 1D Lyα spectrum of the corresponding aperture from Figure A.3. All spectra have been scaled to show their relative intensity. Overplotted in orange we show the polarization fraction as a function 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 [O III]of four galaxies which are associated with LAB1. The dash-dotted lines are the minimum and maximum of those objects. In bins **B**-**E**, we see a trend of low polarization associated with the core of the 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

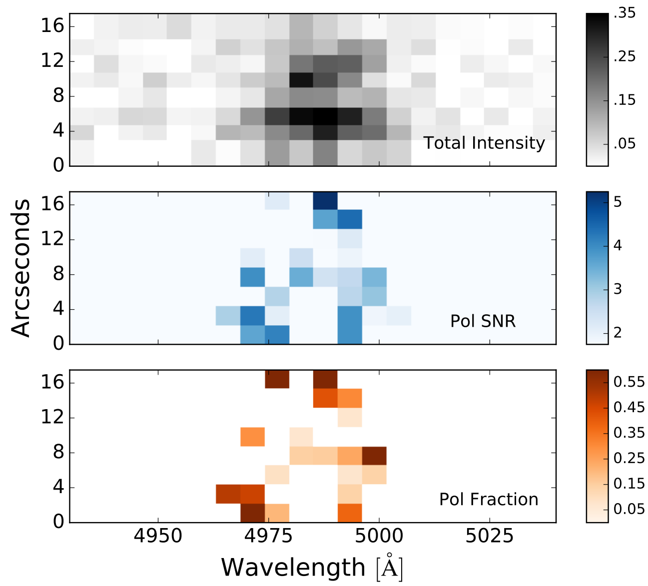

Figure A.6 *Top.* Total intensity of the 2D Lyα spectrum in bins of 5Å by 2-3″. *Middle.* SNR_p map which demonstrates the relative quality of our polarization measurements. SNR$p > 2$ is generally enough for us to discriminate p statistically signicant from zero with 95% confidence and we show CIs for such bins in Figure A.5. $\mathrm{SNR}_p > 3.8$ 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

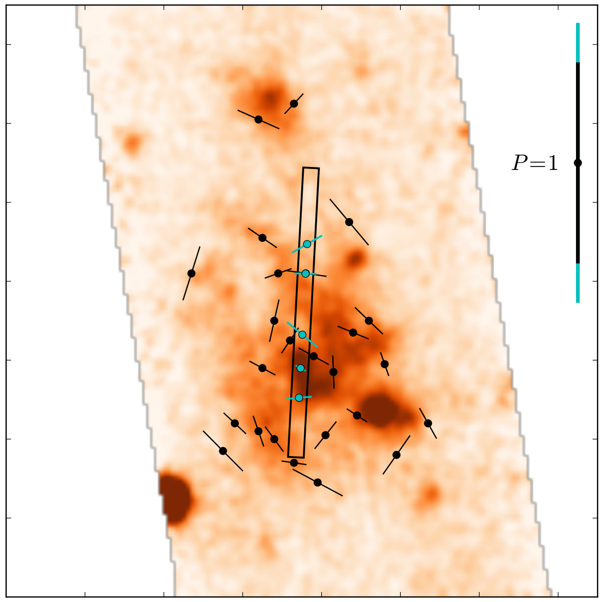

Figure A.7 Direction of Lyα polarization vectors. Smoothed total Lyα intensity from H11 overlaid with polarization vectors from H11 (black) and from this analysis (cyan) for those spatial bins from Figure A.3 which have $\mathrm{SNR}_p > 2$. Because we measure small polarization amplitude in our bins 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 [O III] and known to be components of LAB1 (Kubo et al., 2015; McLinden et al., 2013) (see Figure A.5). We note that these measurements are all within $\sim$ 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.