# Extragalactic machine learning: in theory and in practice

Sebastian Turner

A thesis submitted in partial
fulfillment of the requirements of
Liverpool John Moores University
for the degree of

Doctor of Philosophy

October 2020

# Abstract

Galaxy evolution is complicated. Throughout their lifetimes, galaxies are subject to an amalgamation of astrophysical and cosmological processes that direct the growth of their stellar masses, the transformation of their morphologies, and the cessation of their star formation. The variable action of these processes begets a diverse population of galaxies, which exhibit a variety of brightnesses, colours, shapes, and sizes, among myriad other features. Many of these features are bimodally distributed, which has led to the general acceptance of a simple empirical paradigm of galaxy evolution. However, connecting this diversity among galaxies with the array of processes that are involved in their evolution, and constraining the relative influences of each of these processes, requires that several features are analysed simultaneously. This has been enabled by the recent advent of machine learning techniques, which are capable of extracting scientifically useful information from complicated, multi-dimensional datasets, to astronomy and astrophysics. Unsupervised machine learning techniques, free from the requirement for pre-labelled training data, are especially well suited to the exploration of the data structures of galaxy samples in multi-dimensional feature spaces. This thesis assesses the use of clustering, an unsupervised machine learning technique, for the research of galaxy evolution.

Clustering is first tested on a well-characterised sample of galaxies from the GAMA survey. Galaxies are represented in five dimensions by a set of intrinsic astrophysical features. Use of a unique cluster evaluation framework enables the robust identification of reproducible and astrophysically meaningful clustering structures via the $k$-means method. Outcomes consisting of two, three, five, and six clusters are deemed stable, and form a hierarchical structure that agrees well with established notions of the galaxy bimodality. The two- and three-cluster outcomes are dominated in their structures by the stellar masses, colours, and star formation activity of galaxies, with Sérsic indices and half-light radii becoming important for the five- and six-cluster outcomes. Clusters also exhibit broad correspondence with detailed morphological classifications, and it is suggested that the inclusion of additional morphological features might improve this correspondence further. The five- and six-cluster outcomes indicate the differential role of environment in the evolution of galaxies with intermediate colours. This cluster evaluation framework is then applied for the validation of the cosmological, hydrodynamical EAGLE simulations against the GAMA survey.

Outcomes consisting of seven and five clusters respectively, determined using the same five features for both samples, are selected for analysis. These outcomes produce an agreement score of $V_a = 0.76$, indicating broad, overall agreement, but differences in their substructures. These differences include discrepancies in the growth of the central bulges of galaxies along the star-forming main sequence, an over-abundance of low-mass, bulge-dominated, star-forming galaxies in the EAGLE sample, and a subpopulation of high-mass, disc-dominated, star-forming galaxies in the EAGLE sample that is not present in the GAMA sample. These differences are attributed to the resolution of EAGLE, and to an active galactic nucleus feedback prescription that is not sufficiently effective in EAGLE. Finally, clustering is used to compare samples of galaxies at low ($z \sim 0.06$; GSWLC-2) and intermediate ($z \sim 0.67$; VIPERS) redshifts, in order to examine the evolution of subpopulations of galaxies. Galaxies are clustered in a nine-dimensional feature space defined by a series of ultraviolet-through-near-infrared colours using the Subspace Expectation-Maximisation algorithm, which includes iterative dimensionality reduction. The algorithm models both samples using seven clusters: four containing mostly star-forming galaxies, and three containing mostly passive galaxies. Both sets of star-forming clusters form clear morphological sequences, capturing the gradual internally-driven growth of galaxy bulges at both epochs. At high stellar masses, this growth is linked with quenching. However, it is only at low redshifts that additional, environmental processes appear to be involved in the evolution of low-mass passive galaxies.

The results of this thesis demonstrate the utility of clustering as a method with which to analyse the large galaxy samples that are anticipated from next-generation surveys, and with which to facilitate the multi-dimensional comparison of cosmological galaxy simulations with observations. Clustering is robustly able to identify astrophysically meaningful substructures in complex, multi-dimensional feature spaces, and these substructures may readily be interpreted with respect to the evolutionary contexts of the galaxies that they encompass.

# Declaration

The work presented in this thesis was carried out at the Liverpool John Moores University Astrophysics Research Institute. Unless otherwise stated, it is the original work of the author. While registered as a candidate for the degree of Doctor of Philosophy, for which the submission is herein made, the author has not been registered as a candidate for any other award. This thesis has not been submitted in whole or in part for any other degree.

Sebastian Turner
Astrophysics Research Institute,
Liverpool John Moores University,
IC2, Liverpool Science Park,
146 Brownlow Hill,
Liverpool,
L3 5RF,
United Kingdom

# Publications

During the course of the completion of the work presented in this thesis, the following papers have been accepted or submitted for publication in a refereed journal, or are currently being prepared for submission:

*Reproducible k-means clustering in galaxy feature data from the GAMA survey*
**Sebastian Turner**, Lee S. Kelvin, Ivan K. Baldry, Paulo J. Lisboa, Steven N. Longmore, Chris A. Collins, Benne W. Holwerda, Andrew M. Hopkins, and Jochen Liske, 2019. Monthly Notices of the Royal Astronomical Society, 482, 126.
(The work presented in this paper is the basis of Chapter 3 and Appendix A.)

*Testing a cosmological galaxy simulation with unsupervised machine learning*
**Sebastian Turner**, Ivan K. Baldry, Robert A. Crain, Paulo J. Lisboa, Steven N. Longmore, and Chris A. Collins. In preparation for submission to Monthly Notices of the Royal Astronomical Society.
(The work presented in this paper is the basis of Chapter 4 and Appendix B.)

*Synergies between low- and intermediate-redshift galaxy populations revealed with unsupervised machine learning*
**Sebastian Turner**, Małgorzata Siudek, Samir Salim, Ivan K. Baldry, Agnieszka Pollo, Steven N. Longmore, Katarzyna Małek, Chris A. Collins, Paulo J. Lisboa, Janusz Krywult, Thibaud Moutard, Daniela Vergani, and Alexander Fritz. Submitted to Monthly Notices of the Royal Astronomical Society.
(The work presented in this paper is the basis of Chapter 5 and Appendix C.)

# Acknowledgements

Thanks, first of all, to my superb supervisory team: to Steve Longmore, for all of your invaluable advice, guidance, and mentorship; to Ivan Baldry, for sharing your seemingly endless knowledge and wisdom in all things extragalactic astrophysics (and for graciously accepting the many defeats I dealt you on the squash court); and to Paulo Lisboa, for introducing me to the fascinating world of machine learning. Thanks to Lee Kelvin, for the significant and positive influence you had upon my work, and to Chris Collins, for your interest in, and your input on, my research. Thanks to my collaborators – Gosia Siudek, Agnieszka Pollo, Samir Salim, Rob Crain, and Kasia Małek – for the opportunity to get involved with such engaging projects, and for the many insightful and stimulating scientific discussions. Thanks to Steven Bamford and to Marie Martig for examining this thesis, and for all of the helpful comments, feedback, and suggestions.

I've been at the Astrophysics Research Institute for a whole heap of years now, and I'd like to thank all of the students and staff (including administrative and support staff) that I've known during my time here for having made it a thoroughly pleasant stay throughout. The unity shown by everyone (and spearheaded by Phil James) in the face of this year's challenging circumstances is a testament to the calibre of the department. Special thanks to the ExGal research group – our weekly meetings were both great fun and scientifically enriching. Shout-out to the cohort that I came through with for being an all-round lovely bunch of people, and a massive thanks to Kirsty Taggart for being a fantastic friend and colleague.

Thanks to Vincent, for allowing me to set up camp in your wonderful cafe whenever I needed a change of scenery. Of all the desks I've worked at over the last few years, my spot by the window is a firm favourite. Thanks to Breeny, and to all of the squash teammates and rivals that I've met on Liverpool's courts over the years. You provided me with the perfect distraction from my studies when I needed it (and also, frankly, when I didn't). Thanks to my friends, both near and far, for all of the good times that we've shared in, which helped to keep me going. Halcyon days.

Ganz besonders möchte ich meiner Mutter danken für die bedingungslose Unterstützung und das stete Vertrauen in meine Fähigkeiten. Ein großer Dank geht auch an meinen Opa für die Ermutigung und das anhaltende Interesse an meiner Arbeit. Danke auch für die Mirácolis!

And finally, a massive thanks to my sister and best friend, Freddie, for all of the motivation and moral support you've given me while I've been studying for my doctorate. It's not too long now until you get yours – that'll make it two of us! To put it in the timelessly poignant words of Yilmaz, the great Turkish philosopher and orator: *"double up, boys"*.

The funding for this postgraduate studentship was provided by the United Kingdom Science and Technology Facilities Council (STFC).

The work presented in this thesis was conducted using the following software: the `scikit-learn` (Pedregosa et al., 2011), `matplotlib` (Hunter, 2007), `scipy` (Jones et al., 2001), and `numpy` (Oliphant, 2006; Harris et al., 2020) packages for the `Python 3` programming language (`https://www.python.org`); the `Fisher-EM` subspace clustering package (Bouveyron & Brunet 2012; known in this thesis as `SEM`) for the `R` statistical computing environment (R Core Team, 2019); and the Starlink Tool for OPerations on Catalogues And Tables (`TOPCAT`; Taylor 2005).

The Galaxy And Mass Assembly (GAMA) project is a joint European-Australasian project based around a spectroscopic campaign using the Anglo-Australian Telescope. GAMA is funded by the STFC, the ARC (Australia), the AAO, and the participating institutions. The GAMA website is `http://www.gama-survey.org/`.

The Virgo Consortium is acknowledged for making their simulation data publicly available. The Evolution and Assembly of GaLaxies and their Environments (EAGLE) simulations were performed using the DiRAC-2 facility at Durham, managed by the ICC, and the PRACE facility Curie based in France at TGCC, CEA, Bruyeres-le-Chatel.

The construction of the GALEX-SDSS-WISE Legacy Catalogue (GSWLC) was funded through NASA award NNX12AE06G.

The VIMOS Public Extragalactic Redshift Survey (VIPERS) was performed using the ESO Very Large Telescope, under the "Large Programme" 182.A-0886. The participating institutions and funding agencies are listed at `http://vipers.inaf.it`.

Funding for the Sloan Digital Sky Survey (SDSS) III was provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Science Foundation, and the U.S. Department of Energy Office of Science. The SDSS-III website is `http://www.sdss3.org`.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

The nature of the Milky Way was the subject of curiosity already in antiquity. The Greek philosophers Anaxagoras and Democritus correctly guessed that the diffuse band of light that spans the night sky is composed of multitudinous stars (Aristotle, 350BC)[1], but it wasn't until the telescope was invented, nearly 2,000 years later, that this could be proven. Leonard and Thomas Digges (1571; 1576) are thought to have been the first to aim a telescope skyward (Gribbin, 2003), but Galileo Galilei (1610) was the first to use one to definitively resolve stars within the diffuse Milky Way. The publishing of his book *Siderius Nuncius*, which reported this ground-breaking result alongside many others, marked the commencement of observational astronomy. Another 140 years later, Thomas Wright (1750) discerned the shape of the Milky Way as a whole to be a disc and, departing from the heliocentric view of the Universe that was still prevalent at the time, proposed that the Solar System is radially displaced from its centre.

It was then suggested that some of the various "nebulae" that pervade the night sky might constitute a class of celestial object analogous to the Milky Way (Wright, 1750; Kant, 1755). Evidence in support of this "island universe" hypothesis started to accumulate in the Nineteenth and early Twentieth Centuries. William Parsons (1850), wielding the 72-inch "Leviathan of Parsonstown" – the largest ever telescope upon the completion of its construction – identified the spiral structures of several nebulae for the first time. Huggins & Miller (1864) observed a continuum "crossed ... by lines of absorption" in the spectrum of the nebula M31, revealing that it is composed of (then unresolvable) stars. This enabled a distinction between gaseous and stellar nebulae, the latter of which included all of the spirals. Slipher (1915), measuring the spectral redshifts of spiral stellar nebulae, found their velocities to be an average of 25 times greater than those of gaseous nebulae and of individual stars in the Milky Way, indicating that they could not be gravitationally bound to it.

[1] Aristotle himself thought the notion "impossible".

The growing speculation prompted a "Great Debate" on the matter between Harlow Shapley and Heber Curtis, on April 26$^{th}$ 1920 at Washington D.C.'s National Museum of Natural History. Shapley believed that the Milky Way encompassed the entire Universe, including all of the nebulae. He reasoned that if some were as large as the Milky Way, whose size he had calculated using a calibration which related the periods and the luminosities of Cepheid variable stars (Leavitt & Pickering, 1912), then they would have to be at "excessive" distances in order to be seen at their apparent angular diameters. In addition, he used a measurement of proper motion in the nebula M101 (van Maanen 1916; later discovered to be erroneous; Hubble 1935; van Maanen 1935) to impose an upper limit upon their sizes. Curtis, however, argued that some nebulae were external to, and analagous to, the Milky Way. He used the rates and the brightnesses of novae in spiral nebulae (Ritchey, 1917) to infer Milky-Way-like star counts and sizes (via distances) for them. He also cited the velocities that Slipher had measured as further proof for his argument.

There was no immediate victor of the Great Debate (Shapley & Curtis, 1921); it would ultimately be resolved conclusively by Edwin Hubble over the following few years. Using the new 100-inch Hooker telescope at Mount Wilson Observatory – the successor to the Leviathan as the world's largest telescope – he observed Cepheids in the nebulae M31 and M33 and, with Shapley's own calibration, calculated distances that definitively placed them outside of (and hence gave them physical sizes comparable to that of) the Milky Way[2] (Hubble, 1925). It wasn't long before other stellar nebulae were also confirmed as being beyond the extent of the Milky Way (Hubble, 1926a,b). His work on the topic culminated in the discovery of a relationship between the distances and the recession velocities of extragalactic nebulae (Hubble, 1929), which revealed the true scale of the Universe and paved the way for observational cosmology. The island universe hypothesis having been substantiated, the study of galaxies – systems of stars, gas, dust, and dark matter like our own Milky Way – could begin in earnest.

The subsequent century has seen a great deal of progress in the field of extragalactic astrophysics. The Universe is now known to contain trillions of galaxies (Gott III et al., 2005; Conselice et al., 2016), arranged along the threads of a "cosmic web" and exhibiting a variety of shapes, sizes, masses, colours, and more. Thorough inspection of this variety has revealed a series of dichotomies or bimodalities among the galaxy population, and has ultimately begotten a simple observational paradigm of galaxy evolution. This paradigm has been met with an expansive theoretical literature dedicated to identifying and describing the various astrophysical processes that drive it. However, the precise balance of these processes, and of the array of evolutionary pathways that they beget, is yet to be fully constrained. It is clear that a more detailed view of the galaxy population, achieved by combining several features at once, is needed to better understand galaxy evolution and the interplay of the astrophysical and cosmological processes involved.

[2]Oepik (1922) had previously also calculated a similarly large distance to M31, based instead on measurements of its internal dynamics.

Meanwhile, the last two decades have seen the burgeoning use of machine learning techniques within astronomy and astrophysics (Ball & Brunner, 2010; Baron, 2019). The uptake of these techniques has come primarily in response to the enormous data volumes anticipated from forthcoming surveys (e.g. Laureijs et al. 2011; Ivezić et al. 2019), which demand the use of fast, automated data analysis methods. Supplementary to these practical advantages is the ability of these techniques to distill complex, multi-dimensional input data into interpretable models, which invites a renewed examination of our understanding of astrophysics and especially (in the context of this thesis) of galaxy evolution. Clustering, an unsupervised machine learning technique which groups observations by their intrinsic similarity to one another, demonstrates substantial promise for exploration and discovery as it expresses the "natural" data structure of input observations.

Thus, the main questions that I aim to address in this thesis are:

- What place does clustering, and by extension unsupervised machine learning in general, have among the arsenal of methods used in future studies of galaxy evolution?
- Can clustering in feature spaces of high dimensionalities reveal substructures to the established dichotomies, or bimodalities, of galaxies?
- If so, can these substructures be used to constrain the balance of theoretical processes that have been proposed as driving galaxy evolution?

The remainder of this chapter, in which I introduce aspects of the current state of the field of galaxy evolution with a view to setting the scene for the chapters that follow, proceeds thusly. In Section 1.1, I briefly discuss the cosmological origins of galaxies and their resultant distribution among the cosmic web. In Section 1.2, I review observational progress in the study of the diversity of galaxies, including the convergence to a simple observational paradigm of galaxy evolution. In Section 1.3, I summarise theoretical processes that have been invoked to explain galaxy evolution and, in Section 1.4, I highlight the increasingly important role of cosmological simulations of galaxies in elucidating the balance of these processes. In Section 1.5, I summarise this chapter, and finally, in Section 1.6, I outline the structure of the remainder of this thesis.

## 1.1 Galaxies and cosmology

Though it was Hubble for whom the relationship between the distances and the recession velocities of galaxies was initially named, he did not consider its cosmological implications[3]. Instead, it was

[3]Hubble, at the time, referred to the recession velocities as "apparent" velocities, and merely viewed them as a "convenient" way of linking the redshifts of galaxies to their distances (Hubble, 1936).

Lemaître (1927, 1931a,b) who inferred that this relationship is a consequence of the expansion of a homogeneous and isotropic Universe[4]. Tracing this expansion back in time, he also inferred that the Universe originated as a singularity. This "Big Bang" model, as it would come to be known (Kragh, 2013), was validated by its accurate explanation of the abundances of the chemical elements (Alpher et al., 1948), and by the detection of Cosmic Microwave Background (CMB) radiation (Penzias & Wilson, 1965), whose blackbody spectrum was predicted by the model as a relic of the plasma-state of the early ($\sim 10^5$ yr) Universe (Dicke et al., 1965). The expansion of the Universe would ultimately be discovered, through the use of type 1a supernovae in distant galaxies as "standard candles", to be accelerating (Riess et al., 1998; Perlmutter et al., 1999). This acceleration is thought to be driven by "dark energy", which permeates the Universe and acts in opposition to gravity (Frieman et al., 2008). Its influence is represented in Einstein's equations of general relativity[5] (1917) by the cosmological constant $\Lambda$. Recent measurements ascribe $\sim 70$ per cent of the energy density of the Universe to dark energy (Planck Collaboration et al., 2018).

Parallel to these advancements in our understanding of the expansion of the Universe came advancements in our understanding of the origin of structures within it. The measurement of unexpectedly high velocities among galaxies in dense groups (Zwicky, 1933), and of unexpectedly high rotation velocities among stars in the outer discs of spiral galaxies (Rubin & Ford, 1970; Rubin et al., 1980), led observers to infer the presence of more mass within these systems than was implied by their luminous contents (Faber & Gallagher, 1979): gravitationally-dominant, electromagnetically-inert "dark matter", whose composition is unknown (Bertone & Hooper, 2018). Theorists, meanwhile, began using dark matter to develop an explanation for the emergence of the large-scale distribution of galaxies from predicted density fluctuations in the plasma of the early Universe (Sachs & Wolfe, 1967; Silk, 1967, 1968; Peebles, 1968, 1982; White & Rees, 1978; Blumenthal et al., 1984; Davis et al., 1985). The distribution of these fluctuations was fixed and enlarged to cosmic scales by a period of rapid expansion of the early Universe (called "inflation"; Guth 1981). While baryons were initially prevented from descending into the peaks of these fluctuations by their electromagnetic coupling to photons, dark matter – affected only by gravitational forces – was not, provided that it was "cold" (i.e. moving slowly). The cooling of the Universe as it continued to expand (then at a slower rate than during the period of inflation) eventually led to the decoupling of baryons and photons, thus allowing baryons to follow cold dark matter (CDM) into the overdensities that it had already started to establish and yielding the large-scale structure seen in the Universe today. The now freely propagating photons went on to produce the CMB radiation, and the discovery of small angular anisotropies in the CMB radiation by the Cosmic Background Explorer (Smoot et al., 1992), echoing these primordial density

[4]In order to acknowledge Lemaître's contribution, members of the International Astronomical Union voted in 2018 to rename "Hubble's Law" to "the Hubble-Lemaître Law".

[5]Einstein had initially included a similar constant in his equations erroneously, in order to model what was thought to be a static Universe at the time.

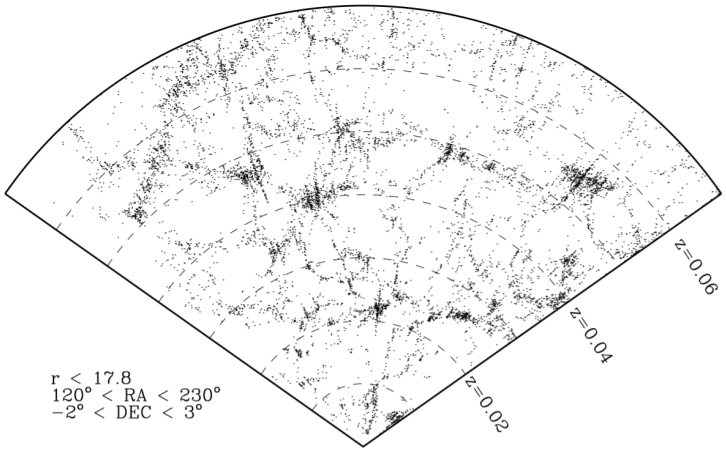


**Figure 1.1:** The large-scale structure of the Universe, shown using $r < 17.8$ galaxies from the Sloan Digital Sky Survey (SDSS; York et al. 2000), the 2-degree Field Galaxy Redshift Survey (2dFGRS; Colless et al. 2001), and the Galaxy And Mass Assembly project (GAMA; Driver et al. 2009). Each individual point represents a galaxy. This figure has been adapted from Baldry et al. (2012).

fluctuations, was a strong validation of this explanation (Wright et al., 1992). CDM contributes $\sim 25$ per cent of the energy density of the Universe, and $\sim 85$ per cent of its matter density (Planck Collaboration et al., 2018). Together, the cosmological constant $\Lambda$ and CDM constitute the twin pillars of the presently leading model of cosmology: $\Lambda$CDM (Bartelmann, 2010).

The continued dissipative collapse of gaseous baryons into the centres of gravitationally self-bound CDM "haloes" eventually led to the formation of stars and galaxies (Binney, 1977; Silk, 1977; White & Rees, 1978). Disc galaxies form from gas whose angular momentum is preserved during this collapse, and during its subsequent accretion into CDM haloes (Fall & Efstathiou, 1980; Mo et al., 1998). Massive elliptical galaxies result from the hierarchical merging of CDM haloes, along with their constituent galaxies (Toomre 1977; White & Rees 1978; see also Section 1.3.2). This hierarchical merging has dictated the growth of large-scale structure since decoupling; hence, the spatial distribution of galaxies in the Universe is directly contingent upon that of dark matter. While dark matter does not emit electromagnetic radiation and cannot be observed directly, galaxies do and can. Thus, galaxies constitute our best tracers of the structure and expansion of the Universe, providing crucial observational constraints for cosmological models. The spatial distribution of galaxies has been observed throughout the history of extragalactic astrophysics, with Shapley & Ames (1926) first identifying a "cloud" of galaxies, and extensive catalogues of similar such congregations following later (Zwicky, 1952; Abell, 1958; Abell et al., 1989). The redshift surveys of the last $\sim$ 30 years (e.g. Geller & Huchra 1989; York et al. 2000; Driver et al. 2009; Garilli et al. 2014) have catalysed progress, producing thorough maps of large-scale structure[6] (e.g. Gott III et al. 2005) which show that matter in the Universe forms a "cosmic web"

[6]The lensing of galaxy light by foreground dark matter represents another way in which the large-scale structure of the Universe may be discerned (Refregier, 2003).

made up of groups, filaments, sheets, and voids[7]. An example of one of these maps is shown in Figure 1.1.

The relative position of a galaxy among the cosmic web – whether it lies within a group, filament, etc. – is called its environment. The environment of a galaxy may be measured in several different ways. For example, one may determine the number density of other, nearby galaxies surrounding it (e.g. Baldry et al. 2006; Brough et al. 2013; Cucciati et al. 2017). Alternatively, one may use a "Friends-of-Friends" algorithm (Geller & Huchra, 1983; Yang et al., 2005, 2007; Robotham et al., 2011) to identify groups of galaxies via their proximity to one another. This facilitates the estimation of a galaxy's properties in the context of its group membership, such as its distance from the centre of the group (e.g. Blanton & Berlind 2007; Bamford et al. 2009; Woo et al. 2013) or its status as either a central (i.e. the most massive member) or a satellite (i.e. embedded within the CDM halo of a central). Measures such as these enable the investigation of how the environments of galaxies influence their evolution (see Sections 1.2.3 and 1.3.2). In turn, an understanding of the role of environmental processes in galaxy evolution grants further insight into cosmology.

## 1.2 The diversity of galaxies

In this section, I describe the variety of galaxies that are observed in the Universe. I begin with individual discussions of morphologies in Section 1.2.1, and of spectral energy distributions (including colours, star formation rates, stellar masses, and more) in Section 1.2.2. Then, in Section 1.2.3, I bring these features together, along with environment, to present an overview of several multi-feature distributions and scaling relations that have been observed among the galaxy population, and what it is that they reveal about galaxy evolution.

### 1.2.1 Morphologies

The morphologies of galaxies - their visual appearances, modulo inclination - have been the target of specific interest since before the genesis of extragalactic astrophysics (Sandage, 2005). Early observers took stock of the variety of what were still "nebulae" at the time by organising them into classes[8]. Herschel (1864) and Dreyer (1888), mustering the New General Catalogue (NGC) of nebulae, contrived an intricate system for describing its entries, including such characteristics as

[7]The most massive congregations of galaxies in the Universe are called "clusters". However, I reserve my use of the term "cluster" in this thesis for data structures that are modelled with unsupervised machine learning techniques (see Chapter 2) in order to avoid any possible confusion between these two different contexts. For the purposes of this thesis, the term "group" (as above) may be assumed to refer also to cosmological clusters.

[8]Initially also including nebulae internal to the Milky Way.

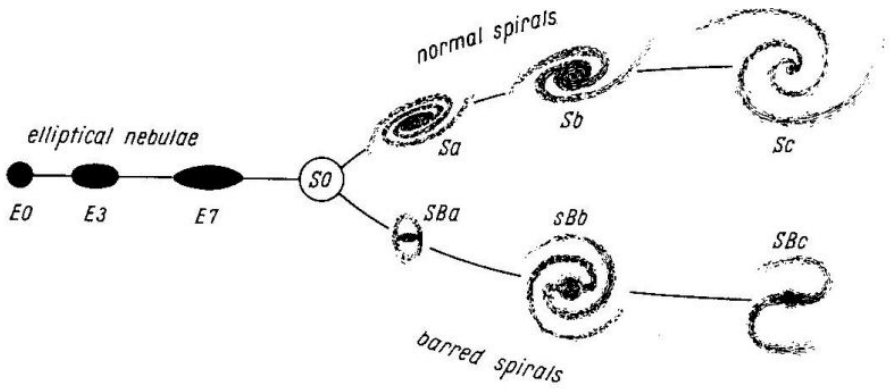


**Figure 1.2:** The Hubble tuning fork diagram, a schematic representation of the Hubble sequence for galaxy morphological classification. This figure has been reproduced from *The Realm of the Nebulae* (Hubble, 1936).

shape, apparent size, and brightness. Wolf (1908) defined 23 different types of nebula and listed NGC archetypes for several of them. Reynolds (1920) set out a sequence of seven grades of spiral nebulae, ordered by the prominence of their central nuclei (or bulges, as they're known today) and the amount of apparent structure (or "resolution") within them. Hubble (1922, 1926b) made the most lasting contribution, building on the sequence of Reynolds to develop a simple yet effective scheme for the classification of extragalactic nebulae (i.e. galaxies) that is still in use today (e.g. Bremer et al. 2018; Kelvin et al. 2018).

The Hubble sequence distinguishes primarily between elliptical and spiral galaxies. Ellipticals (denoted E) do not exhibit any internal structure; their light profiles are smooth, and they range in their shapes from spherical (E0) to flattened (E7). Spirals (S), which *do* exhibit internal structure, are ranked by the complexity of this structure (just as in Reynolds' sequence). The simplest spirals (Sa) have prominent central bulges, tightly wound spiral arms, and minimal structure within their discs, while the most complex spirals (Sc) have weak bulges, open arms, and flocculent structures. Borrowing from stellar spectroscopy, Hubble (1926b) used the terms "early" and "late" to describe this gradient from simple to complex morphologies[9]; nowadays, these terms tend to refer more generally to ellipticals and spirals (or "disc galaxies") respectively (e.g. Kelvin et al. 2014a,b). Spirals may also have a central bar connecting their spiral arms (SB). Galaxies that are neither elliptical nor spiral are classified as Irregular (Irr). Figure 1.2 shows the Hubble tuning fork diagram[10], which maps out Hubble's galaxy classes. Included in the diagram is an additional class, S0, which Hubble (1936) conjectured to be the then-unobserved missing link between ellipticals and spirals.

[9]Hubble was careful, though, to avoid any inadvertent implications about galaxy evolution through his use of these terms, intending them only express observed morphological complexity (Hubble, 1926b, 1927; Baldry, 2008).

[10]A similar diagram had previously been produced by Jeans (1928; who *did* propose that galaxies evolve along the sequence, spinning up from late- to early-type), although it is not known whether this was the inspiration for Hubble's diagram, which was published eight years later (Sandage, 2005).

Since its invention, the Hubble sequence has been subject to gradual amendments. Shapley & Paraskevopoulos (1940) proposed the addition of an Sd class for disc galaxies with especially late-type morphologies. Shapley (1943) and Holmberg (1958) used subclasses (e.g. $Sa^{-}$, $Sb^{+}$) for finer divisions among spirals. de Vaucouleurs (1959) split the Irr class (Sm, for irregular galaxies that have weak spiral structure; and Im), and included notation to account for the presence of outer rings (r). Sandage & Binggeli (1984) incorporated classes for dwarf galaxies (dE, dS0); these galaxies had been excluded from the original Hubble sequence by Malmquist (1922) bias. Kormendy & Bender (1996, 2012) removed the dependence of elliptical classifications on inclination by using isophotes to distinguish between boxy (Eb) and discy (Ed) types, and introduced a sequence of S0 classes (S0a, S0b, S0c; S0 galaxies having since been observed as smooth, disc-dominated galaxies) parallel to that of spirals (see also van den Bergh 1976). These amendments have reflected our increasingly detailed view of galaxies over time and revealed continuous variations in morphologies throughout the galaxy population. Nevertheless, the original elliptical-spiral dichotomy remains fundamental in the study of galaxy morphologies (e.g. Schawinski et al. 2014; Bremer et al. 2018; Kelvin et al. 2018).

The ongoing Galaxy Zoo project (Lintott et al., 2008, 2011; Willett et al., 2013) has enabled the description of galaxy morphologies with particularly high fidelity. Galaxy Zoo enlists citizen scientists to assign classifications, tallying their votes on the visual appearances of galaxies in order to build up comprehensive morphological profiles. As a result, the properties and evolution of galaxies that have specific morphological traits may be closely examined (e.g. Hart et al. 2017; Kruk et al. 2019; Newnham et al. 2020).

Galaxies' morphologies may also be characterised using quantitative features, which have the benefit of being more objective and scalable than Hubble-like classifications (e.g. Naim et al. 1995; Kelvin et al. 2014a)[11,12]. Reynolds (1920) had already included a quantitative component in his classification scheme, plotting the radial light profiles of spiral galaxies to establish the prominences of their bulges. Sérsic (1963, 1968), building on work by Hubble (1930) and de Vaucouleurs (1948), introduced a general equation with which to model galaxy light distributions, with the shape of the profile defined by a single parameter, $n_g$ (the "Sérsic index"). Measures of the sizes of galaxies, such as their effective/half-light radii or Kron radii, may be derived via the Sérsic equation as well (Graham & Driver, 2005). Furthermore, the fitting of a galaxy's light distribution with two Sérsic profiles enables its decomposition into separate contributions from the bulge and the disc (Freeman, 1970; Peng et al., 2002; Simard et al., 2011). Alternatively, the *CAS* system (Conselice 2003; see also references therein for predecessors) measures the concentration,

[11] Galaxy Zoo mitigates the subjectivity of individual classifications somewhat by exploiting "the wisdom of the crowd", and by down-weighting votes from unreliable classifiers (Willett et al., 2013).

[12] See Chapter 2 for a brief discussion on the relative scalability of quantitative morphologies, Galaxy Zoo classifications, Hubble-like classifications, and also of machine learning techniques.

asymmetry, and smoothness of a galaxy's light distribution non-parametrically. Other quantitative morphological features include the Gini coefficient ($G$; Abraham et al. 2003), the $M_{20}$ coefficient (Lotz et al., 2004), and $MID$ system (Freeman et al., 2013). A crucial strength of all of these features is their correlation with Hubble-like morphological classifications. Features like $n_g$, $C$, and $G$, which express how centrally concentrated a galaxy's light is, are sensitive to the presence of a bulge (or the absence of a disc in ellipticals). The amount of structure within a galaxy's disc may be captured with $S$ or $M_{20}$. In addition, several of these features have been combined to identify merging or irregular galaxies. Hence, these quantitative features incorporate key aspects of Hubble-like morphologies. As such, they are commonly used to morphologically classify galaxies[13], and to distinguish between early- and late-type galaxies (e.g. Shen et al. 2003; Scarlata et al. 2007; Lange et al. 2015).

Spectroscopic studies of the morphological components of galaxies reveal them to be, equivalently, dynamical components. Stars within the discs of late-type galaxies rotate around galactic centres in ordered, circular orbits (Slipher, 1914; Sofue & Rubin, 2001). Stars in early-type galaxies were initially all believed to have disordered, triaxial orbits (Binney, 1982), but integral field spectroscopy (Emsellem et al., 2004, 2011; Cappellari, 2016) has shown that early-type galaxies may be divided into slow rotators, which match the previous description, and fast rotators, which also contain an additional smooth disc component. The bulges of late-type of galaxies may be similarly divided on the basis of their dynamics (Kormendy, 1993; Kormendy & Kennicutt, 2004). "Classical" bulges are dynamically akin to slow-rotating elliptical galaxies, leading to the suggestion of a common origin for these structures and the introduction of the general term "spheroid" to unite them (Renzini, 1999). "Pseudobulges", on the other hand, more closely resemble discs in terms of their dynamics, being flatter than classical bulges and exhibiting rotation. These dynamical results have prompted another amendment to the Hubble sequence (Cappellari et al., 2011; Cappellari, 2016), which connects spirals and S0 galaxies of similar complexity and suggests evolutionary ties between them.

### 1.2.2 Spectral energy distributions

The ultraviolet-through-infrared (UV-through-IR) spectral energy distributions (SEDs) of galaxies are as multifarious as their morphologies. SEDs spanning these wavelength regimes are governed in their shapes chiefly by stellar emission, and the attenuation (in the UV and optical) and re-emission (in the IR) of stellar emission by interstellar dust. Hence, a galaxy's SED constitutes a precise inventory of its present contents, which, in turn, constitutes a record of its evolutionary history.

[13] See van der Wel (2008), though, for an example of how the differential use of quantitative morphological features can produce slightly different classifications for the same galaxies.

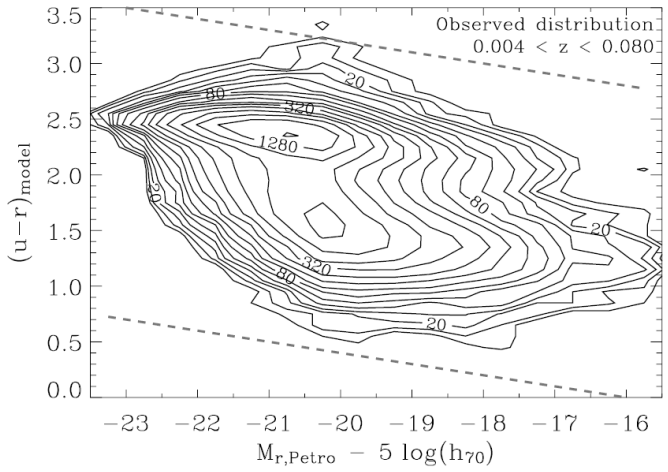


**Figure 1.3:** The $u - r$ colour bimodality, as seen in a sample of low-redshift galaxies derived from SDSS. Here, galaxies' $u - r$ colours have been calculated from "model", profile-based magnitudes (Stoughton et al., 2002), and are scaled against their absolute $r$-band Petrosian (1976) magnitudes. This figure has been reproduced from Baldry et al. (2004).

The brightness of a galaxy – its flux through a single broadband filter – is the simplest measurement that may be made of its SED, summing its stellar emission at the effective wavelength of the filter used. The colour of a galaxy – the ratio of its fluxes through two broadband filters at different effective wavelengths – can probe the contributions of different stellar populations to its total stellar emission. Optical colours have been used for this purpose since the infancy of extragalactic astrophysics, supported by what was already a mature understanding of stellar astrophysics (Roberts, 1963). Blue galaxies contain hot, massive stars with short lifetimes that can only have formed recently; hence, blue galaxies are actively star-forming. Red galaxies, lacking in these stars, are not actively star-forming, and are instead called "passive" or "quenched". The UV flux of a galaxy, which is particularly sensitive to the presence of hot, massive stars, may be used (following its correction for attenuation due to dust) to infer its star formation rate (SFR; Madau et al. 1998; Salim et al. 2007).

The onset of survey astronomy (e.g. York et al. 2000) facilitated the discovery that the optical colours of the overall galaxy population are bimodally distributed (Strateva et al., 2001; Baldry et al., 2004), with most galaxies occupying either the red peak or the blue peak of this distribution. The intermediate-colour region between these two peaks, containing fewer galaxies, is called the "green valley" (Wyder et al., 2007; Martin et al., 2007; Salim et al., 2007; Schiminovich et al., 2007). An example of the optical colour bimodality of galaxies is shown in Figure 1.3. Similar bimodalities of galaxies have also been observed in several other colours involving UV and near-IR (NIR) magnitudes (Williams et al., 2009; Arnouts et al., 2013). In general, these bimodalities represent a simple, useful framework with which to distinguish between star-forming (blue) and

passive (red) galaxies, and galaxies that are transitioning (green) relatively quickly between these two states (e.g. Schawinski et al. 2014; Smethurst et al. 2015; Bremer et al. 2018; Kelvin et al. 2018). However, different colours yield slightly different bimodalities: for example, galaxies occupying the blue peak of the $g - r$ colour distribution may instead occupy the green valley of the $NUV - r$ colour distribution (Salim, 2014). In addition, the red peaks of these bimodalities arise as a consequence of the saturation of the colours of particularly red galaxies while their star formation rates carry on decreasing (Salim, 2014; Eales et al., 2018). It is therefore clear that, for a complete and accurate description of the galaxy population in the context of the bimodality and the green valley, several colours spanning the UV-through-NIR wavelength regime must be considered simultaneously.

Spectroscopy offers another avenue by which to discern the contents of galaxies, with emission lines from the interstellar medium (ISM) proving particularly useful. The flux of the H$\alpha$ recombination line (Kennicutt et al., 1994; Charlot & Longhetti, 2001) or of the [OII] $\lambda$3727 forbidden doublet (Gallagher et al., 1989; Gilbank et al., 2010) of a galaxy may be used to infer its (Kennicutt, 1998), due to the photoionisation of HII regions by hot, massive stars at their centres. Estimates of the chemical abundance of the ISM have been calibrated to the fluxes of several emission lines (Kewley et al., 2019); stellar abundances, on the other hand, are calibrated to absorption line fluxes (e.g. Worthey 1994). The presence of an active galactic nucleus (AGN), i.e. of radiation driven by the accretion of material onto a galaxy's central supermassive black hole, may be diagnosed using emission-line diagrams (e.g. Baldwin et al. 1981; Kauffmann et al. 2003b; Cid Fernandes et al. 2010, 2011; Lamareille 2010). The size of the 4000 Å break in the spectra of galaxies (Balogh et al., 1999), like their colours, may be used to distinguish galaxies in their emission by young stars (small break) or old stars (large break), with the added benefit of being relatively unaffected by attenuation due to dust.

Continued progress in our understanding of stellar astrophysics has enabled the estimation of the full UV-through-IR SEDs of galaxies (Conroy, 2013; Ilbert et al., 2006; Da Cunha et al., 2008; Boquien et al., 2019). This estimation requires the validation of synthetically-constructed spectra against real observations (Walcher et al., 2011). As it is impractical to measure full galaxy spectra that span large wavelength ranges (e.g. UV-through-IR), especially for the large number of galaxies needed for a robust statistical study of galaxy evolution, SEDs are instead generally validated against curtailed, summary measurements, such as galaxy colours. An accurate estimate of the full SED of a galaxy enables the inference of its physical properties, including stellar mass, SFR, stellar metallicity, and more.

The construction of synthetic galaxy spectra requires two main ingredients: a library of template spectra of individual stars (either empirical, e.g. Pickles 1998; or theoretical, e.g. Martins et al. 2005), and a dust attenuation curve. Early studies simply matched sums of stellar template spectra

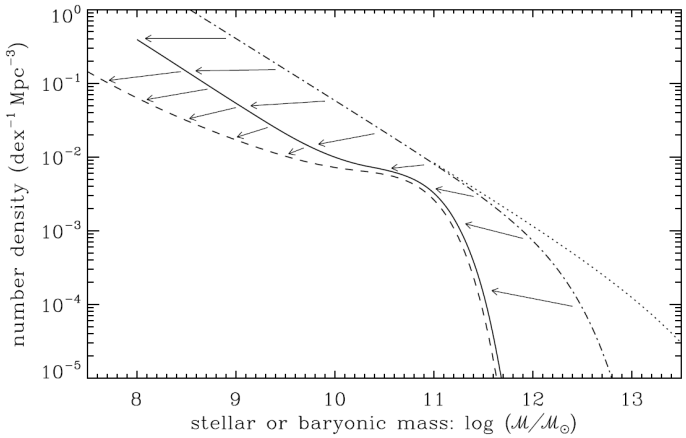


**Figure 1.4:** Baryonic mass functions (BMFs), including: a GSMF (dashed line; a fit to low-redshift galaxies), a galaxy BMF (solid line; including the ISM), a theoretical subhalo BMF (dot-dashed line; including the haloes of individual galaxies), and a theoretical halo BMF (dotted line; including group haloes). Deviations between the subhalo/halo BMFs and the GSMF, shown by the arrows, are largest at low and high masses. This figure has been reproduced from Baldry et al. (2008).

(i.e. synthetic composite spectra) to the observed optical colours of galaxies to discern their stellar contents (Spinrad, 1962; Spinrad & Taylor, 1971; Faber, 1972). This method, particularly prone to degeneracies, was superseded by stellar population synthesis (SPS; Tinsley 1968; Bruzual & Charlot 2003; Maraston 2005), which uses theories of stellar evolution to set astrophysical constraints upon these synthetic composite spectra. Individual stellar template spectra are aggregated using isochrones to build up template spectra of simple stellar populations (SSPs), the basic unit of SPS consisting of groups of stars that form at the same time and with the same metallicity, but with different masses given by an initial mass function (e.g. Salpeter 1955; Chabrier 2003). The template spectra of SSPs are built up for various stages of their evolution. These SSP spectral templates are then themselves combined via their convolution with models of the metallicity-evolution and star formation histories of galaxies to yield composite stellar spectra. Finally, the influence of dust, modelled with a wavelength-dependent attenuation curve (e.g. Calzetti et al. 2000; Charlot & Fall 2000) and with IR emission templates (e.g. Chary & Elbaz 2001), is added. A crucial aspect of accurate SED estimation is the disentangling of the degenerate influences of stellar ages, stellar metallicities, and attenuation due to dust, all of which may redden the spectrum of a galaxy (Worthey, 1994; Bell & de Jong, 2001; Papovich et al., 2001).

### 1.2.3 The observational view of galaxy evolution

Galaxies increase their stellar masses as they evolve, through star formation and/or through accretion and mergers. This evolution, across the population, begets the galaxy stellar mass function (GSMF), a low-redshift example of which is shown using the dashed line in Figure 1.4 (Baldry

et al., 2008). As a measure of the growth of structure in the Universe, the GSMF is an important constraint for models of cosmology. The GSMF has a two-component form, broken by a "knee" at $\sim 10^{10.6}$ $M_\odot$. The GSMF is offset to lower number densities from the subhalo and halo baryonic mass functions, which include the mass contribution of *all* baryons (i.e. not just stars) in individual galaxy haloes and in group haloes respectively. This indicates inefficiencies in the formation of stars and galaxies from the full supply of available baryons at low and high masses. Separate processes have been proposed to explain the inefficiencies at low and high masses (i.e. either side of the knee), where the deviation of the GSMF from the baryonic mass functions is largest (see Sections 1.3 and 1.4).

The decline of star formation at high masses has also been observed in distributions involving other features. The power-law correlation of the SFRs and stellar masses of actively star-forming galaxies – the "star-forming main sequence" (SFMS; Noeske et al. 2007; Salim et al. 2007) – ceases beyond $\sim 10^{10.5}$ $M_\odot$, where the rate of increase of star formation with stellar mass reduces (Whitaker et al., 2015; Eales et al., 2017; Popesso et al., 2019a). Furthermore, passive galaxies become more prevalent at higher steller masses. The colour-magnitude distribution of galaxies goes from being dominated by blue galaxies at faint magnitudes to red galaxies at bright magnitudes (Figure 1.3; Baldry et al. 2004). Galaxy colours trend similarly with stellar mass as well (Peng et al., 2010; Baldry et al., 2012; Taylor et al., 2015). Equivalent trends and divisions have also been observed at higher redshifts (Brammer et al., 2009; Popesso et al., 2019b). In addition, there are trends of the morphologies of galaxies with their brightnesses (reviewed in Binggeli et al. 1988) and with their stellar masses (Bundy et al., 2010; Kelvin et al., 2014b; Moffett et al., 2016), with massive, bright galaxies more likely to be early-type. Kauffmann et al. (2003a), by way of the 4000 Å break strengths and central concentrations of galaxies, divide the population into two clear subpopulations about a critical mass of $\sim 10^{10.5}$ $M_\odot$; Driver et al. (2006), swapping 4000 Å break strengths for colours, impose a similar division at an absolute magnitude of $M_B < -16$. Altogether, it is clear that stellar mass is a vital feature for the description of the evolutionary states of galaxies, particularly in terms of the quenching of their star formation. However, it is also clear that other features play an important role too.

The correlation between the colours and the morphologies of galaxies, with disc-dominated galaxies tending to be blue and spheroid-dominated galaxies tending to be red (Holmberg, 1958; Chester & Roberts, 1964; Bower et al., 1992a,b; Mignoli et al., 2009), suggests that these features are themselves evolutionarily linked. This link appears to be strongest among galaxies that have particularly concentrated morphologies, with several recent studies identifying the central density of a galaxy, which measures the prominence of the bulge, as a potent predictor of its being passive (Cheung et al., 2012; Wake et al., 2012; Fang et al., 2013; Bluck et al., 2014; Luo et al., 2020). This has been used to explain the different distributions of blue and red galaxies in the size-mass plane (van der Wel et al., 2009; van Dokkum et al., 2015; Haines et al., 2017). Furthermore, this has

prompted the suggestion that a massive or dense bulge is *necessary* for the permanent quenching of galaxies (Bell, 2008; van Dokkum et al., 2014). The general colour-morphology correspondence has led some observers to use the colours of galaxies as proxies for their morphologies when selecting samples (e.g. Bundy et al. 2006; Salimbeni et al. 2008; van der Wel et al. 2014). However, studies using Galaxy Zoo morphologies have shown that this correspondence does not always apply: some pure-disc galaxies have red colours (due to being genuinely passive, rather than due to attenuation by interstellar dust; Masters et al. 2010), and some spheroidal galaxies have blue colours, indicating recent or ongoing star formation (Schawinski et al., 2009a). As exceptions to the simple blue-and-discy versus red-and-spheroidal paradigm, an understanding of the origins of these galaxies is important for a complete understanding of galaxy evolution.

Several early studies noted a connection between the morphologies and the environments of galaxies (Hubble & Humason, 1931; Oemler, 1974; Davis & Geller, 1976). Dressler (1980a,b), combining Hubble sequence classifications with tenth-nearest neighbour surface densities, measured a "morphology-density" relation among galaxies in and around groups, showing that elliptical galaxies become more common, and spiral galaxies less common, with increasing local environmental density. Goto et al. (2003), studying galaxy concentrations, fifth-nearest neighbour surface densities, and groupocentric distances, and extending their analysis to the field (i.e. to lower environmental densities), further constrained this morphology-environment connection. Noting a drop in the fraction of intermediately-concentrated galaxies at the highest densities (following a gradual rise with density to that point), they also suggested that the morphological transformation of galaxies is influenced by different processes at different densities. The colours of galaxies are similarly connected to their environments, with blue, star-forming galaxies more common in low density environments than in high density environments, and vice versa for red, passive galaxies (Balogh et al., 2004; Baldry et al., 2006; von der Linden et al., 2010; Woo et al., 2013). Bamford et al. (2009) and Skibba et al. (2009), separating the influence of environment on colour and on morphology, found that the colour-environment connection is much stronger than, and may therefore even be the cause of, the morphology-environment connection.

Peng et al. (2010) disentangled the influences of stellar mass and local environmental overdensity on quenching by examining trends of these features with the changing fraction of red, passive galaxies. They discovered that low-mass galaxies are only quenched in high-density environments and that only high-mass galaxies are quenched in low-density environments (as shown in Figure 1.5). Thusly, they distinguished two phenomenological quenching pathways that act independently of one another: "mass quenching" and "environment quenching". Peng et al. (2012) and Wetzel et al. (2012, 2013) established that the evolution of satellite galaxies in groups, exhibiting little-to-no dependence on their stellar masses, is dominated by environment quenching. On the other hand, observations of the increasing incidence of AGN among high-mass and green valley galaxies (Kauffmann et al., 2003b; Schawinski et al., 2010), generally irrespective of their envi-

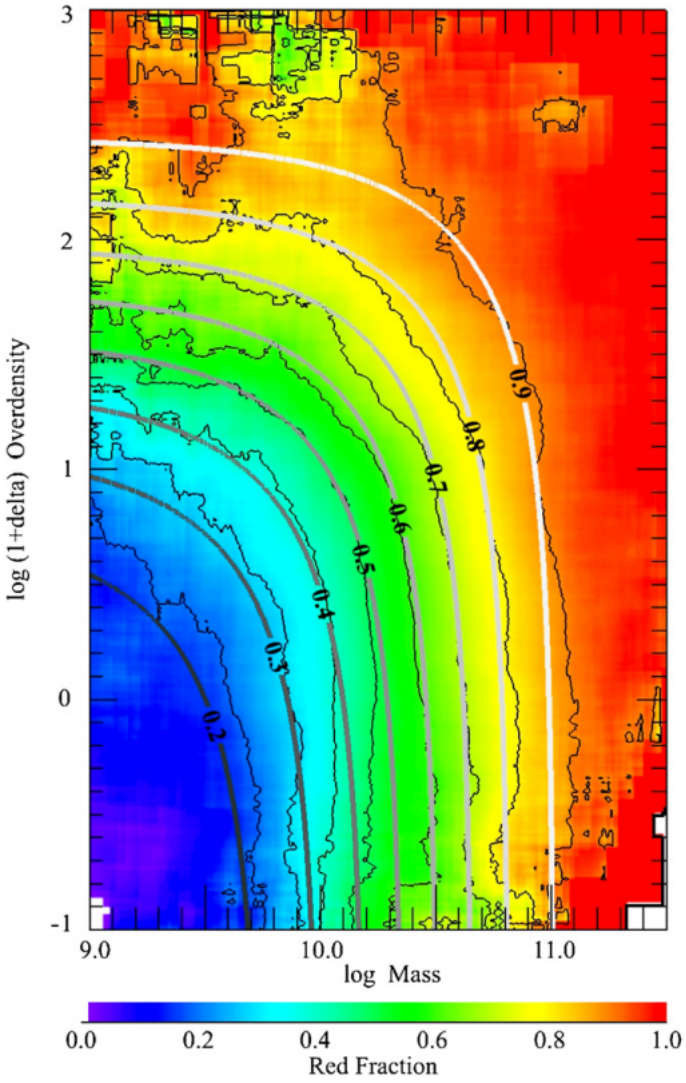


**Figure 1.5:** The disentangling of trends of the star formation activity of SDSS galaxies (via red, passive fraction) with their stellar masses and their local environmental overdensities. This figure has been reproduced from Peng et al. (2010).

ronments[14], provided empirical support for mass quenching. Schawinski et al. (2014), studying the morphologies of galaxies in the green valley, proposed two evolutionary pathways of their own. Most late-type galaxies quench slowly ($\gtrapprox$ 1 Gyr) once their halo (of which they are usually the central) exceeds a mass of $10^{12}$ $M_{\odot}$, and they retain their discs in doing so. Most early-type galaxies, meanwhile, quench quickly ($\sim$ 100 Myr) in a manner that is tied closely to their morphological transformation (e.g. via the merger of two late-type galaxies; see Section 1.3.2). Smethurst et al. (2015), modelling the star formation histories of galaxies, presented further evidence for the existence of multiple evolutionary pathways through the green valley.

[14]Kauffmann et al. (2004) find that only galaxies containing the brightest, most powerful AGN exhibit any significant dependence on environment, becoming less common with increasing local environmental density.

## 1.3 Processes of galaxy evolution

In this section, I list and briefly summarise astrophysical and cosmological processes that, together, influence the evolution of galaxies. In accordance with the observational and phenomenological work of Peng et al. (2010), Schawinski et al. (2014), and others, I distinguish between internal processes (Section 1.3.1), which apply to all galaxies, and external processes (Section 1.3.2), whose additional influence depends on the environment of the galaxy in question. This internal versus external dichotomy is most commonly asserted in the context of quenching processes (e.g. Smethurst et al. 2017); in this section, I extend it to include processes that engender morphological transformations in galaxies as well.

### 1.3.1 Internal processes

#### Bar inflows

Bars are present within the majority of low-redshift late-type galaxies (Eskridge et al., 2000; Nair & Abraham, 2010; Masters et al., 2011), and were recognised as an important morphological components already at the time of the inception of the Hubble sequence. They have been shown to form naturally from instabilities within dynamically-cold, thin stellar discs (Ostriker & Peebles, 1973; Sellwood & Wilkinson, 1993; Debattista et al., 2004, 2006). Once in place, bars draw gas into the inner regions of galaxies by redistributing its angular momentum (Hawarden et al., 1986; Bournaud & Combes, 2002; Sheth et al., 2005), thus promoting central star formation and the growth of pseudobulges (Kormendy & Kennicutt, 2004), and possibly fuelling AGN (Galloway et al. 2015, though this is disputed; Cheung et al. 2015). Bars are most common in disc galaxies with red colours and low levels of gas content (Masters et al., 2012; Cheung et al., 2013), which suggests that they are involved with quenching.

#### Morphological quenching

Martig et al. (2009), analysing zooms of cosmological simulations, discovered that the morphologies of galaxies may contribute directly to their quenching. The gaseous disc of an early-type galaxy may be stabilised against fragmentation and subsequent star formation by the gravitational potential of its central bulge, which azimuthally shears the gas, and by the lack of an accompanying stellar disc, which, if present, would contribute self-gravity to giant molecular clouds (i.e. the sites of star formation). This mechanism, not requiring the removal of gas from galaxies for their quenching (e.g. via feedback or stripping processes; see below and Section 1.3.2), naturally explains the observed lack of star formation activity in early-type galaxies that retain their gas

(e.g. Serra et al. 2012; Martig et al. 2013). The bulges of these galaxies may be grown via internal processes such as bar inflows (see above) or violent disc instabilities (see below), or via external processes (see Section 1.3.2). Morphological quenching offers an explanation for the link between spheroids and passiveness, but it is not believed to constitute a dominant quenching mechanism among the galaxy population as a whole (Bluck et al., 2014).

#### Stellar feedback

Recently-formed massive stars may supply energy to their surrounding ISM, inhibiting the continued formation of new stars (Hayward & Hopkins, 2017). Their radiation, stellar winds, and eventual explosions as supernovae can heat giant molecular clouds and induce turbulence within them, preventing their cooling and gravitational collapse, and drive the expulsion of gas from galaxies in powerful galactic outflows (Wada & Norman, 2001; Matzner, 2002; Murray et al., 2010, 2011; Hopkins et al., 2011, 2012). Evidence for the action of these mechanisms is provided by direct observations of galactic winds (Martin, 1999, 2005; Weiner et al., 2009) and of the chemical enrichment of the intergalactic medium (IGM; Aguirre et al. 2001; Songaila 2005, 2006). Stellar feedback is expected to be important for regulating star formation in galaxies with lower stellar masses; supernova winds are not anticipated to be able to escape the deep gravitational potentials of galaxies with higher stellar masses (Hopkins et al., 2014; Keller et al., 2016). Hence, stellar feedback offers an explanation for the inefficiency of star formation at low stellar masses (Figure 1.4; see also Section 1.4), but alternative processes are required to explain the inefficiency at high stellar masses (see below).

#### Supermassive black hole feedback

Supermassive black holes (SMBHs) at the centres of galaxies (Kormendy & Ho, 2013) grow via the accretion of gas (Lynden-Bell, 1969; Shakura & Sunyaev, 1973; Ichimaru, 1977). A corollary of this growth is the feedback of energy from the accreting gas to the wider gas supplies of galaxies. Galaxies whose inner regions contribute feedback in this manner are described as having an AGN. AGN are characterised by a unified model (Antonucci, 1993; Netzer, 2015) which invokes varied viewing angles to reconcile their disparate apparent properties. Two main modes of AGN feedback are distinguished, differing in how they originate and in how they deliver energy to the ISM (Churazov et al., 2005; Croton et al., 2006; Somerville et al., 2008; Heckman & Best, 2014). Kinetic mode feedback is driven by relativistic polar jets, which are believed to be generated by the magnetism and rotation of advective flows of hot gas around SMBHs. These jets imbue the ISM of galaxies with kinetic energy which can exceed the galaxies' gravitational binding energies and, as a result, lead to its expulsion. In massive ($> 10^{12}$ $M_{\odot}$) CDM haloes, kinetic mode feed-

back is expected to maintain the passivity of galaxies by continuously heating hot gas (see below). Radiative mode feedback, on the other hand, is associated with stochastic events like mergers (see Section 1.3.2) and violent disc instabilities (see above), or bar inflows (see above), all of which funnel cold gas into the centres of galaxies, where it encircles the SMBH in a thin accretion disc. This accretion disc then emits ionising radiation which launches powerful galactic winds that exert pressure on the ISM and, in the most extreme cases, carry star-forming gas out of galaxies.

AGN may be identified using emission-line diagnostics (Baldwin et al., 1981; Kauffmann et al., 2003b; Cid Fernandes et al., 2010, 2011; Lamareille, 2010). Direct tracers of AGN feedback (see Fabian 2012 for a review) include X-ray observations of jet-blown gas bubbles (McNamara et al., 2000; Hlavacek-Larrondo et al., 2012), and spectral features corresponding to fast outflows (Tremonti et al., 2007; Maiolino et al., 2012). AGN are most commonly hosted by green valley galaxies (Nandra et al., 2007; Schawinski et al., 2007, 2010; Hickox et al., 2009), substantiating their link with quenching. The close correlation between the masses of SMBHs and their surrounding stellar spheroids (i.e. bulges and/or whole elliptical galaxies; Silk & Rees 1998; Häring & Rix 2004; McConnell & Ma 2013) explains both the connection between the morphologies and the star formation activity of galaxies, and the sharp drop in the efficiency of star formation at high stellar masses (Figure 1.4; see also Section 1.4). In addition, AGN feedback can account for the observed downsizing of the galaxy population over cosmic time (i.e. the observation that the most massive galaxies formed earliest; Cowie et al. 1988; Cimatti et al. 2006; Cattaneo et al. 2008), which is in tension with its expected hierarchical assembly.

#### Violent disc instabilities

In comparison with those at low redshifts, star-forming galaxies at high redshifts are relatively rich in gas (Daddi et al., 2010; Tacconi et al., 2010). Their thick, turbulent, gaseous discs, fuelled by continuous cold inflows (Bournaud & Elmegreen 2009; Dekel et al. 2009; see also below), are particularly susceptible to fragmentation and gravitational collapse into large ($\sim$ 1 kpc), massive ($\sim 10^8$ $M_\odot$) stellar clumps due to "violent disc instabilities" (Bournaud et al., 2007b; Elmegreen et al., 2007; Genzel et al., 2011; Cacciato et al., 2012). These clumps can then migrate into the centres of galaxies via dynamical friction, where they contribute to spheroid growth such that there is a classical bulge in place at later times (Immeli et al., 2004; Elmegreen et al., 2008b), provided that the clumps survive any disruption to their star formation due to stellar feedback processes (see above). In addition, it has been proposed that the inward migration of these clumps may also facilitate the growth of central SMBHs (Elmegreen et al. 2008a; Bournaud et al. 2011, 2012; Gabor & Bournaud 2013; see also above). Chains of stellar clumps are a commonly observed morphological characteristic of star-forming galaxies at high redshifts (Cowie et al., 1996; van den Bergh et al., 1996; Conselice et al., 2004).

#### Virial shock heating

The manner in which galaxies acquire gas from the IGM is dependent upon the masses of their host CDM haloes (Kereš et al., 2005; Dekel & Birnboim, 2006). At low halo masses ($< 10^{12}$ $M_{\odot}$), gas descends into the discs of galaxies via cold, filamentary streams[15] (Katz et al., 2003; Kereš et al., 2009). Once in the disc, the cold gas can fuel star formation, and/or lead to violent disc instabilities (see above). At high halo masses ($> 10^{12}$ $M_{\odot}$), though, cooling becomes inefficient, and the incoming gas is instead heated upon entry to the halo by virial shocks such that it is not immediately available for star formation (Birnboim & Dekel, 2003; Cattaneo et al., 2006). Galaxies then remain quenched for as long as this gas remains hot. Kinetic mode AGN feedback has been proposed as a mechanism for maintaining the passivity of massive galaxies in this way (see above). It has also been argued that the continued accretion of gas can dynamically heat the halo (Dekel & Birnboim, 2008; Birnboim & Dekel, 2011). Supporting evidence for virial shock heating as a quenching mechanism comes from X-ray observations of the hot haloes of massive galaxies (Paolillo et al., 2002; Xia et al., 2002).

### 1.3.2 External processes

#### Gravitational processes: galaxy-galaxy tidal interactions and mergers

Galaxies in dense environments, such as groups, interact gravitationally with one another. These interactions are known as "tidal interactions" or "harassment" (Moore et al., 1996, 1998, 1999). Fly-by encounters of galaxies may remove gas (Combes et al., 1988; Mayer et al., 2006) and stars (Read et al., 2006; Chang et al., 2013) from their outer regions (especially if they are low-mass or diffuse), encourage central star formation within them (Ellison et al., 2008; Renaud et al., 2014), and dynamically heat their discs, all of which can cause a gradual transition from late- to early-type morphologies. The presence of stellar or gaseous tidal tails in the vicinity of galaxies is interpreted as evidence for prior galaxy-galaxy harassment (Arp, 1966; Combes et al., 1988; Kenney et al., 1995).

Mergers between galaxies, more common in environments of higher densities (Darg et al. 2010; Ellison et al. 2010; except in environments of the very highest densities; see below), are especially influential upon their evolution (Toomre, 1977; Hopkins et al., 2006). The precise outcome of a galaxy merger is contingent mostly upon the relative masses of the galaxies involved and upon their gas content (Lotz et al., 2010a,b). Major mergers (in which the masses of the progenitors – $M_1$ and $M_2$, where $M_1$ is the higher mass – satisfy $M_2/M_1 > 0.25$) can destroy the discs of

[15]This manner of gas acquisition also applies to higher halo masses at high redshifts (Kereš et al., 2005; Dekel & Birnboim, 2006; Kereš et al., 2009).

late-type galaxies and, via the violent relaxation of the merger remnant, produce an early-type galaxy (Barnes, 1988, 1992). The gravitational upheaval causes gas contained within the discs of the progenitors (if present) to lose its angular momentum and descend to the centre of the remnant (Barnes & Hernquist, 1996), where it may be consumed in a burst of star formation (Barnes & Hernquist, 1991; Mihos & Hernquist, 1994a,b, 1996; Cox et al., 2008) or trigger AGN activity by feeding the central SMBH (Di Matteo et al. 2005; Hopkins et al. 2005; Springel et al. 2005a,c; see also Section 1.3.1). If the progenitors are particularly rich in gas, then the gas may instead form a disc in the remnant via its reaccretion from tidal tails, and renew star formation (Barnes, 2002; Hopkins et al., 2009a,b, 2010). Furthermore, a major merger remnant may accrete *new* gas from the IGM with which to rejuvenate its star formation (Salim & Rich, 2010; Gabor et al., 2011). Minor mergers ($M_2/M_1 < 0.25$), which are much more common (Lotz et al., 2011), may foster the growth of a classical bulge within the more massive progenitor *without* the destruction of its surrounding disc (Walker et al. 1996; Bournaud et al. 2004, 2005, 2007a; though they have been shown to dynamically heat and thicken discs). Minor mergers may also induce starbursts, but they are not as productive as those induced by major mergers (Cox et al., 2008; Lambas et al., 2012; Kaviraj, 2014). Pre-merger systems are identified via close pairs of galaxies (Ellison et al., 2008; Silva et al., 2018), and ongoing mergers and post-merger systems via visual inspection of images (Arp, 1966; Lintott et al., 2008, 2011) or combinations of quantitative morphological features (Conselice, 2003; Lotz et al., 2004). In addition, the early-type morphologies of post-starburst galaxies (Yang et al., 2008; Almaini et al., 2017) are suggestive of major merger origins. Overall, the role of mergers in transforming galaxies from late- to early-type is clear. However, the effect of mergers on star formation activity varies case-by-case, such that mergers seemingly cannot be tied unequivocally with quenching (Weigel et al., 2017).

#### Hydrodynamical processes: removal of the ISM

Galaxy groups are permeated by hot ($\sim 10^7$ K), diffuse, gas (Voit, 2005). It may be detected via its bremmsstrahlung emission at X-ray wavelengths (Forman et al., 1972; Arnaud et al., 2010), or via the "Sunyaev-Zeldovich" effect (the inverse Compton scattering of CMB photons by energetic free electrons in the hot gas; Sunyaev & Zeldovich 1972; Kukstas et al. 2020). This hot IGM may interact hydrodynamically with the cold ISM of infalling galaxies in different ways, with the general consequence of the removal of the ISM and the quenching of star formation[16]. These hydrodynamical processes are invoked to explain the observed decrease in the cold gas content of galaxies with increasing local environmental density (Giovanelli & Haynes, 1983; Brown et al., 2017).

[16]Gabor & Davé (2015) combine external processes like gas-stripping and thermal evaporation with internal processes like virial shock heating and AGN feedback (see Section 1.3.1) to propose a unified model in which hot gas dictates the quenching of both centrals *and* satellites.

The motion of galaxies through the IGM can lead to the ram-pressure stripping of their cold gas (Gunn & Gott, 1972; Hester, 2006). The pressure exerted by the IGM on the ISM of a galaxy scales with the square of the velocity of the galaxy; hence, ram-pressure stripping is most efficient in the inner regions of groups (Cayatte et al., 1990; Roediger & Hensler, 2005). It has also been shown that ram-pressure stripping leads to a temporary increase in star formation, due to compression of the ISM, before quenching (Tonnesen & Bryan, 2012; Vulcani et al., 2018). While face-on stripping of gas from disc galaxies is expected to be dominated by ram-pressure, edge-on stripping is instead expected to be dominated by the viscosity of the IGM (Nulsen, 1982; Marcolini et al., 2003). Jellyfish galaxies, exhibiting extended trails of stars and/or gas (Gavazzi et al., 1995; Ebeling et al., 2014; Poggianti et al., 2016, 2017), constitute prototypical examples of both ram-pressure stripping and viscous stripping. Finally, the hot IGM may directly heat the cold ISM, causing it to evaporate from the gravitational grip of galaxies Cowie & Songaila (1977); Nipoti & Binney (2007). In general, these hydrodynamical processes are effective at quenching galaxies, but they are not anticipated to cause significant morphological changes (Boselli & Gavazzi, 2006).

**Pre-processing**

While the morphologies of galaxies in environments of the highest densities are uniformly early-type, their high orbital velocities mean that they are *unlikely* to merge with one another (Fujita, 1998; Darg et al., 2010; Ellison et al., 2010). This has prompted the suggestion that galaxies are "pre-processed" in small groups (Fujita, 2004; Mihos, 2004), where mergers and morphological transformations are more likely. These small groups of pre-processed galaxies are then, in line with the predictions of the ΛCDM cosmological model of the hierarchical assembly of structure in the Universe, accreted into large groups. Hence, pre-processing explains the uniformity of the properties of satellite galaxies in large groups (McGee et al., 2009; Balogh & McGee, 2010; Wetzel et al., 2013, 2015). Gravitationally-bound substructures have been observed within large groups (Kodama et al., 2005) and are interpreted as a signature of pre-processing.

**Starvation**

In addition to being subject to the gravitational influence of other, nearby galaxies in groups (see above), galaxies are also subject to the gravitational influence of groups as a whole. Group tides can remove the warm, gaseous envelopes of galaxies (Larson et al. 1980; also known as the circumgalactic medium). This prevents its accretion into the inner regions of galaxies; as a result, galaxies then quench either via the consumption of any remaining cold gas contained in their discs through star formation, or via the subsequent stripping and/or heating of this cold gas (see above). This overall process – the removal of a galaxy's gaseous envelope followed by its slow or delayed

quenching – is known as "starvation" or "strangulation" (Bekki et al., 2002; Peng et al., 2015). Starvation provides an explanation for the inhibition of star formation at large groupocentric radii and for apparent delays to the full quenching of galaxies in groups (Balogh et al., 2000; Wetzel et al., 2012, 2013; Schawinski et al., 2014). Group tides have also been linked with enhancements in the central star formation of infalling galaxies (Merritt, 1984; Byrd & Valtonen, 1990).

## 1.4 Cosmological galaxy simulations

Observations of galaxies can provide powerful constraints for theories of galaxy evolution. However, an exploration of the interplay of the aforementioned astrophysical and cosmological processes necessitates an experimental approach, which may be achieved through the use of numerical simulations which model the evolution of populations of galaxies in their cosmological contexts by including prescriptions with which to model the influence of these processes (Baugh, 2006; Somerville & Davé, 2015; Vogelsberger et al., 2020).

Two main methods are employed for this purpose: semi-analytic models and hydrodynamical simulations[17]. Semi-analytic models (e.g. Kauffmann et al. 1993; Somerville & Primack 1999; Somerville et al. 2008; Cole et al. 2000; Gonzalez-Perez et al. 2014; Henriques et al. 2015, 2020) separate their treatment of dark matter and of baryons. The growth of large-scale structure, dictated by the gravitation of dark matter, is traced in terms of the hierarchical merging of CDM haloes and subhaloes, based either on collisionless $N$-body simulations (e.g. Springel et al. 2005b; Boylan-Kolchin et al. 2009; Klypin et al. 2011) or on statistical considerations (e.g. Kauffmann & White 1993; Lacey & Cole 1993; Somerville & Kolatt 1999). Analytical prescriptions which capture the effects of astrophysical and cosmological processes like gas accretion, star formation, and feedback are then used to infer the evolution of the integrated baryonic properties of galaxies in the context of these merger trees. These prescriptions, which may have either a theoretical or an empirical basis, are validated by their combined ability to reproduce the observed galaxy population. Semi-analytic models are relatively efficient, given that only the dark matter components of galaxies are simulated numerically. In addition, model variations may readily be generated by applying different analytical prescriptions to the same underlying merger trees.

Recent advances in computing power have enabled the application of hydrodynamical simulations (e.g. Dubois et al. 2014; Vogelsberger et al. 2014; Schaye et al. 2015; Davé et al. 2016, 2019; Pillepich et al. 2018), which fully and self-consistently model the evolution of the baryonic com-

[17] "Zoom" simulations are a subclass of hydrodynamical simulations which remodel small regions within cosmological simulations (containing e.g. groups or individual galaxies) at higher resolution in order to facilitate a closer examination of the astrophysical processes at play (e.g. Katz & White 1993; Martig et al. 2009, 2012; Hopkins et al. 2014, 2018; Bahé et al. 2017; Barnes et al. 2017).

ponents of galaxies (including their distributions *within* galaxies) alongside that of their dark matter components. The addition of hydrodynamics comes at considerable computational expense, which, along with the requirement of modelling a cosmologically representative population of galaxies, limits the resolution of hydrodynamical simulations. Hence, "sub-resolution" prescriptions (e.g. Springel & Hernquist 2003; often similar in their design to the analytical prescriptions used in semi-analytic models) are added to implement astrophysical processes that operate at scales that are not resolved, such as (again) star formation and feedback. The influence of these sub-resolution prescriptions is commonly calibrated to reproduce key observational results (e.g. the $z = 0$ GSMF in Schaye et al. 2015); their subsequent ability to reproduce *other* observational results to which they are *not* calibrated (e.g. galaxy morphologies) constitutes a prediction, and may be used to validate hydrodynamical simulations. Differences between the outputs of hydrodynamical simulations are driven mostly by differences between their sub-resolution prescriptions; for example, while some simulations distinguish between the two modes of AGN feedback (see Section 1.3.1; Pillepich et al. 2018), others do not (Schaye et al., 2015), leading to differences between their resultant galaxies in terms of gas retention and star formation activity (Davies et al., 2020b). The focus of the remainder of this section will be on hydrodynamical simulations.

Cosmological simulations have played an increasingly significant role in the study of galaxy evolution over time. The widespread acceptance of the ΛCDM cosmological model was driven by the successes of dark-matter-only simulations in recreating the large-scale structure of the Universe (Springel et al., 2006). More recently, hydrodynamical simulations have highlighted the prominent role of feedback processes in regulating star formation. Early such simulations were subject to the spurious "overcooling" of their gas (Balogh et al., 2001), leading to star formation that was too efficient, and to present-day galaxies whose stellar masses were too high and whose morphologies were too spheroidal (Navarro & Steinmetz, 2000; Scannapieco et al., 2009). The development of prescriptions that implement stellar and AGN feedback (e.g. Dalla Vecchia & Schaye 2012; Weinberger et al. 2017) has been instrumental in bringing the stellar masses and the morphologies of galaxies in hydrodynamical simulations closer to those of observed galaxies in the real Universe (e.g. Schaye et al. 2015; Pillepich et al. 2018).

Previous studies have tended to validate hydrodynamical simulations in the context of one or two features at a time. Examples include stellar masses, colours, star formation rates, dust content, sizes, morphologies, and kinematics (Trayford et al., 2015; Furlong et al., 2015; Camps et al., 2016; Kaviraj et al., 2017; Trayford et al., 2017; Nelson et al., 2018; Genel et al., 2018; Donnari et al., 2019; Rosito et al., 2019; Rodriguez-Gomez et al., 2019; van de Sande et al., 2019; Bignone et al., 2020). However, these features are all intricately interrelated. Disentangling the influence of the astrophysical processes that drive the coevolution of these features requires that several of them are examined simultaneously. This multi-dimensional validation may be enabled through the use of machine learning techniques.

## 1.5 Summary

This chapter has offered an overview of the field of galaxy evolution. Astronomical observations over the last century have revealed a rich diversity among the galaxy population. Survey astronomy has, more recently, exposed bimodalities and dichotomies in the distributions of galaxy features, and this has begotten a simple empirical paradigm of galaxy evolution. Galaxies, as they evolve from low to high stellar masses, generally go from being actively star-forming and disc-dominated to being passive and bulge-dominated, and the manner in which they do so is linked with their environments. An array of theoretical processes have been proposed as drivers of this diversity and of this empirical paradigm. Large-scale cosmological processes are responsible for the formation of galaxies, and continue to have a significant external impact upon their evolution into their later lives. Galaxies are also subject to internal processes that operate on smaller scales. All of these theoretical processes act in concert, affecting the growth, star formation activity, and morphologies of galaxies, and constraining their individual influence is an open problem in extragalactic astrophysics. Cosmological, hydrodynamical simulations of galaxies have recently highlighted the importance of feedback mechanisms in the regulation of star formation. However, disentangling the interplay of evolutionary processes requires that many galaxy features are examined at once. Such multi-feature analysis is enabled by machine learning.

## 1.6 This thesis

The remainder of this thesis proceeds thusly. In Chapter 2, I introduce the application of machine learning techniques to astronomical and astrophysics contexts, focusing on clustering and dimensionality reduction – the techniques that are explored in subsequent chapters – and on their prior uses in studies of galaxy evolution. I also motivate my selection of particular algorithms for the work presented in this thesis. In Chapter 3 (supplemented by Appendix A), I trial the use of the $k$-means clustering method on a pilot sample of galaxies from the GAMA survey. Galaxies are characterised using five intrinsic astrophysical features, with a view to establishing the interpretability of the clustering structures of the pilot sample in terms the present understanding of galaxy evolution. I build on the work of Chapter 3 in Chapter 4 (supplemented by Appendix B), applying the same clustering approach to facilitate a comparison of simulated galaxies from the cosmological, hydrodynamical EAGLE models with observed galaxies from the GAMA survey. The aims are two-fold: to discern the utility of clustering as a tool for the multi-dimensional validation of cosmological galaxy simulations, and to use the simulated galaxies to make inferences about the evolution of observed galaxies (if they have been accurately recovered by the simulations). In Chapter 5 (supplemented by Appendix C), I use clustering to compare samples of galaxies at low and intermediate redshifts, in order to examine the cosmic evolution of subpopulations of galaxies.

Galaxies are characterised for the clustering in terms of their rest-frame UV-through-NIR colours only, to establish the extent to which the SEDs of galaxies encode their evolutionary states and to assess clustering for the analysis of galaxy catalogues from deep photometric cosmological surveys. Finally, in Chapter 6, I summarise the work presented in this thesis, offer conclusions, and discuss prospects for future work.

# Chapter 2

# An overview of machine learning

Machine learning techniques are computer algorithms that aim to discover patterns in input data, and to use these patterns to construct models with which to make predictions about yet-unseen data. While many such techniques are closely related to statistical methods that have been established for hundreds of years already (e.g. Bayes 1763; Legendre 1805), the emergence of machine learning as an analytical paradigm in its own right began with the development of artificial neural networks in the mid-to-late Twentieth Century (McCulloch & Pitts, 1943; Rosenblatt, 1957; Werbos, 1975; Fukushima, 1980). The general rise in interest in machine learning techniques since the turn of the Millennium has been driven by advances in computing power, and by the publishing of large catalogues of labelled observations which enable the validation of supervised machine learning models (e.g. LeCun et al. 1998; Deng et al. 2009). The continued increase of astrophysical data volumes into unprecedented regimes (e.g. $\sim$ 20 TB per night from the Legacy Survey of Space and Time; Ivezić et al. 2019) demands a scalable approach to data analysis, and machine learning techniques represent a promising solution; their application to the morphological classification of galaxies constitutes a pertinent example of their utility.

Historically, the morphologies of galaxies have been classified by expert observers, working either individually or in small teams of $\lesssim 10$, in accordance with schemes like the Hubble sequence. Examples of catalogues that have been compiled in this manner include those published by Sandage (1961), de Vaucouleurs et al. (1991), Fukugita et al. (2007), Nair & Abraham (2010), Oh et al. (2013), Kelvin et al. (2014a, 2018), and Moffett et al. (2016), which each list between $10^2$ and $10^4$ galaxies. Hubble-like morphological classifications are particularly scientifically useful, but are time-consuming to assign, involving the close examination of images of galaxies. Hence, the assembly of catalogues of classifications for more than $\sim 10^4$ galaxies by small teams of experts is impractical. The Galaxy Zoo project (Lintott et al., 2008) innovated upon this approach by amassing morphological classifications from up to $10^5$ citizen scientists at a time. Galaxy Zoo

catalogues, listing morphological classifications for galaxies from various surveys (Lintott et al., 2011; Willett et al., 2013, 2017; Simmons et al., 2017) and from simulations (Dickinson et al., 2018), contain between $10^4$ and $10^6$ galaxies. Crowd-sourcing, however, also has its limits; $\sim$ 14 months were required to reach the necessary number of independent classifications for each of the $\sim 200,000$ galaxies in Galaxy Zoo 2 (Willett et al., 2013). Next-generation galaxy surveys, such as those that will be delivered by Euclid (Laureijs et al., 2011) and the Vera C. Rubin Observatory (Ivezić et al., 2019), will observe yet more galaxies ($\sim 10^9$). In order to scale up to this regime, the assignment of morphological classifications will need to be quicker still, and automated. While simple, quantitative morphologies meet these criteria (and have previously been measured for up to $\sim 10^6$ galaxies at a time; e.g. Simard et al. 2011), it is machine learning techniques that have the potential to produce morphological descriptions of galaxies with a comparable fidelity to Hubble-like morphological classifications (Lahav et al., 1995; Ball et al., 2004; Huertas-Company et al., 2008, 2011, 2015, 2019; Banerji et al., 2010; Dieleman et al., 2015; Walmsley et al., 2020). Furthermore, while fulfilling the demand for scalable approaches to data analysis, machine learning techniques also invite a renewed examination of our understanding of astrophysics due to their ability to distill complex, multi-dimensional input data into interpretable models.

Two main types of machine learning techniques are distinguished: supervised techniques and unsupervised techniques[1]. Supervised techniques are useful for mapping existing domain knowledge onto new data. A supervised classification algorithm, for example, may assign labels to previously unseen observations after being trained on pre-labelled observations. Unsupervised techniques, on the other hand, demonstrate substantial promise for exploration and discovery because they are less reliant on prior knowledge than supervised techniques. An unsupervised clustering algorithm, for example, assigns labels to observations in accordance with their similarity to one another (i.e. the distances between observations in terms of the features used to represent them). Unsupervised techniques, then, construct models that are driven purely by the structure of input data, and require no training. They may therefore be said to express the "natural" structure of the input data, rather than expressing structures imposed upon it by assumptions that are explicitly built into the use of supervised techniques. The use of unsupervised techniques does, though, incorporate implicit assumptions, and the precise definition of similarity can vary between techniques. Ensuring the astrophysical utility of these models therefore requires carefully considered choices of algorithm and features.

In the remainder of this chapter, I provide a brief description of concepts involved in clustering (Section 2.1) and in dimensionality reduction (Section 2.2), which are the two machine learning

[1]While supervised techniques are suitable for fully labelled input data, and unsupervised techniques for input data with no labels, *semi*-supervised techniques (Zhu, 2005) combine aspects of both to enable the analysis of partially labelled input data by assuming relationships between unlabelled and labelled observations (e.g. that unlabelled observations are likely to share the label of their nearest labelled observation).

techniques that are most relevant to the work presented in this thesis. The literature on previous applications of these techniques within astronomy and astrophysics is vast; hence, my literature review (Section 2.3) focuses on studies of galaxy evolution. Various metrics exist with which to measure distances between observations in multi-dimensional feature spaces; throughout this thesis, I use Euclidean distances only. Further comments on concepts mentioned in this chapter are included throughout the remainder of this thesis (see Sections 3.1, 5.2, A.1, B.1, C.1, and C.3 in particular). Finally, I summarise this chapter in Section 2.4.

## 2.1 Clustering

Clustering is an unsupervised machine learning technique that aims to partition $N$ observations in a $D$-dimensional feature into $k$ clusters. Observations are partitioned in accordance with their intrinsic similarity to one another; their distances from one another in the $D$-dimensional feature space. Hence, clustering algorithms do not require any training on pre-labelled observations. Clustering algorithms vary in their definitions of a cluster and of similarity, and will find different clusters accordingly. I distinguish between three main methods of clustering in this thesis: prototype-based clustering, density-based clustering[2], and model-based clustering.

Clusters determined by prototype-based clustering methods are defined by singular, central points: their prototypes. Observations are assigned to their nearest prototype, with the assignments constituting cluster memberships. The prototype of a cluster is given by a measure of the central tendency of its members (e.g. mean, median). The positions of prototypes are optimised iteratively (via Expectation-Maximisation; Dempster et al. 1977) in order to minimise the distances (and maximise the similarities) between the observations and their prototypes. As a result, prototype-based clustering methods tend to produce spherical clusters with similar sizes to one another, and cannot effectively model *true* clusters (i.e. clusters that actually exist in the input data) that do not match this description. The outcomes of prototype-based clustering methods are only locally optimal, so repeated, randomised initialisations are commonly used to find the global optimum (Arthur & Vassilvitskii, 2007). An advantage of prototype-based clustering algorithms is their simplicity, which means that they may robustly scale to large samples and high dimensionalities. The number of prototypes and clusters, $k$, must usually be specified in advance of the use of these methods. Various strategies may guide the selection of optimal values of $k$, including compactness- and stability-based approaches (Liu et al., 2010; von Luxburg, 2010; Lisboa et al., 2013). Examples of prototype-based clustering methods include $k$-means (MacQueen, 1967; Lloyd, 1982), $k$-medoids (Kaufmann & Rousseeuw, 1987), and affinity propagation (Frey & Dueck 2007; which does *not* require that $k$ is specified in advance).

[2]This includes, in this thesis, what is also known elsewhere as connectivity-based clustering.

Density-based clustering methods determine clusters using thresholds in inter-observation distances[3]. Different approaches are used to connect observations and thereby construct clusters, from single-linkage (i.e. clusters include all observations that lie within a given distance of their nearest observation) to neighbourhood-based criteria (i.e. requiring that cluster members are surrounded by a given number of other cluster members, within a distance threshold). Density-based methods are versatile in that they can model clusters with arbitrary shapes and sizes. While not requiring that $k$ is set in advance (like most prototype- and model-based methods), they instead require the assertion of thresholds in inter-observation distances or densities. Density-based clusters are difficult to evaluate, due to the lack of prototypes with which to compare cluster members (though stability-based approaches may still be used for evaluation). The lack of prototypes (and of any estimated cluster parameters) also means that the identities of density-based clusters can be less clearly defined than those of clusters determined via other methods. Furthermore, they do not scale well to high dimensionalities, at which observations become uniformly sparse (the "curse of dimensionality"; Bellman et al. 1957). Examples of density-based clustering methods include hierarchical clustering (Ward Jr., 1963; Sibson, 1973), and the `DBSCAN` (Ester et al., 1996; Schubert et al., 2017) and `HDBSCAN` (Campello et al., 2013) algorithms.

Model-based clustering methods assume that the structure of input data may be described using a mixture of $k$ probability density functions. Observations are assigned probabilities of belonging to each of the $k$ functions, and clusters, corresponding to each of these functions, comprise observations that are most likely to belong to each function. Like prototype-based methods, model-based methods tend to require the specification of $k$ in advance of their use. Model parameters for each of the functions, such as means, standard deviations, and mixture proportions (in the case of Gaussian density functions), are then optimised iteratively using an Expectation-Maximisation approach. Model-based clustering methods are flexible; their iterative adaptation of the model parameters means that they can adapt to accommodate a variety of cluster shapes and sizes (although they cannot model concave shapes, like density-based clustering methods can). They exhibit poor scalability with increasing samples sizes and dimensionalities, due to the accompanying increase in the number of model parameters that must be estimated. Model-based clusters are relatively straightforward to evaluate, via measurement of the quality of fit of the model to the input data. However, the use of these methods incurs a risk of overfitting through the use of too many model components. Hence, evaluation of clusters determined via model-based methods also tends to penalise the number of model parameters in order to favour simpler models (e.g. Schwarz 1978; Biernacki et al. 2000). Examples of model-based clustering methods include Gaussian Mixture Models (McLachlan & Basford, 1988) and Subspace Expectation-Maximisation (Bouveyron & Brunet, 2012).

[3]Density-based clustering methods are similar in their design to group-finding methods (such as the "Friends-of-Friends" algorithm; Geller & Huchra 1983; see also Section 1.1) that are used to identify overdensities of galaxies in physical space.

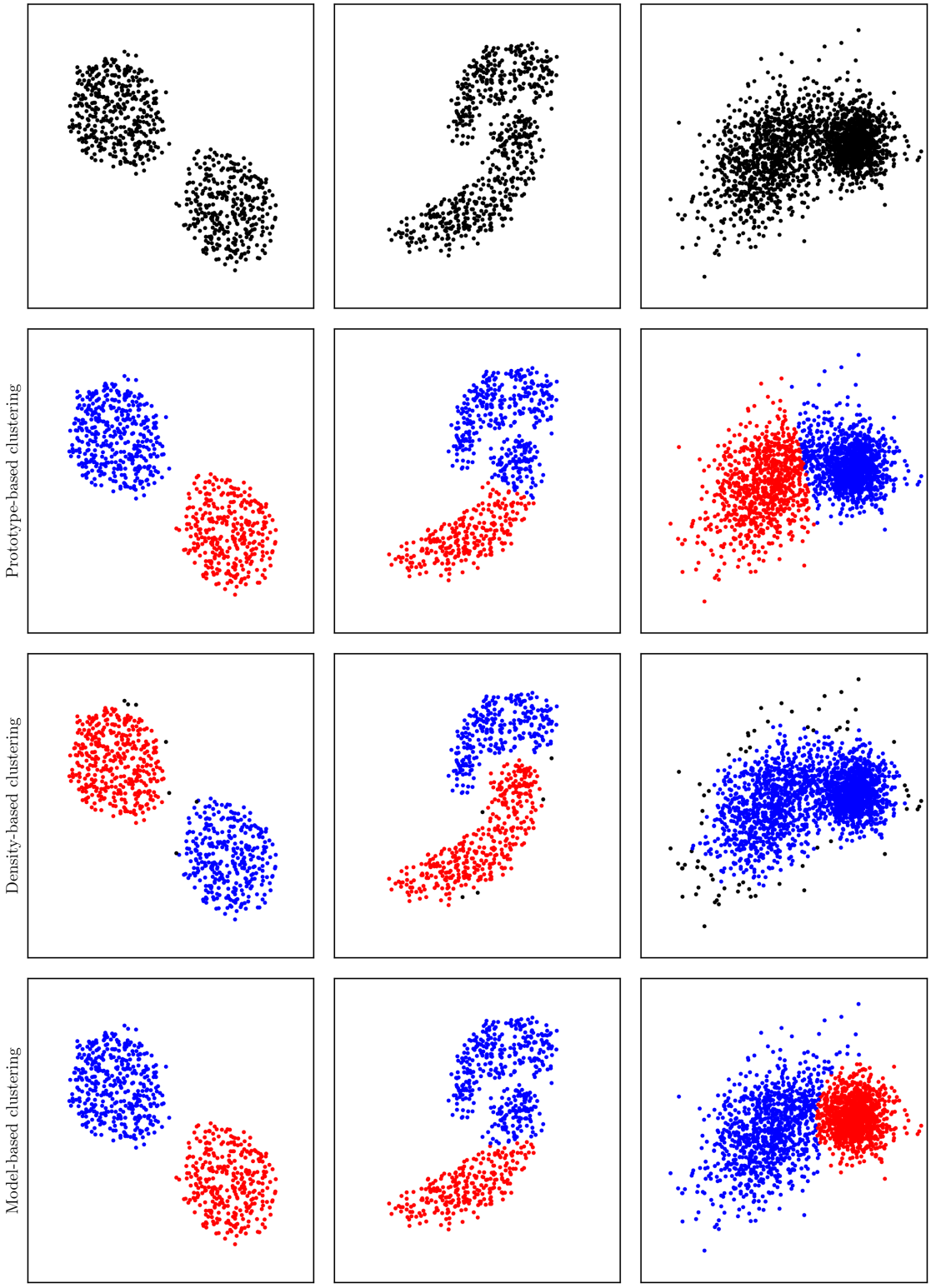


**Figure 2.1:** Examples of prototype-, density-, and model-based clustering outcomes in three simple two-dimensional data sets. Clustering is conducted using the *k*-means method, the `DBSCAN` algorithm, and a Gaussian Mixture Model respectively. Black points in panels showing density-based clustering outcomes are points that have been designated as "noise" by `DBSCAN`.

Figure 2.1 shows examples of clustering outcomes determined in three simple two-dimensional data sets using each of these three main methods. Prototype-based clustering is conducted using the $k$-means method ($k = 2$; initialisation per Arthur & Vassilvitskii 2007), density-based clustering using the `DBSCAN` algorithm ($\epsilon = 0.04$, minimum cluster size of 20), and model-based clustering using a Gaussian Mixture Model ($k = 2$; $k$-means-based initialisation; full covariance freedom). All three methods are similarly successful in partitioning the trivial data set in the left column of Figure 2.1, which consists of two well-separated true clusters. Only density-based clustering is able to accurately partition the more complicated concave structure of the data set in the middle column. Prototype- and model-based clustering methods produce inaccurate partitions because the structure of the input data is inconsistent with the assumptions of these methods regarding the structures of clusters. It should, however, be noted that the ability of prototype- and model-based methods to partition complicated data structures (such as that of the middle column in Figure 2.1) improves as the number of clusters is increased. This is because the additional clusters may be used to segment such data structures. This effect is exploited throughout Chapters 3-5, and care is taken to ensure that the resultant clustering outcomes are robustly reproducible and astrophysically meaningful. Finally, only model-based clustering is able to partition the data set in the right column satisfactorily, giving a curved boundary between the two overlapping true clusters. Prototype-based clustering, due to its tendency to produce clusters of similar sizes, approximately bisects the data set. Density-based clustering, unable to define a boundary between the two overlapping true clusters, groups them together as one.

Clustering is appealing because it is descriptive, modelling the structure of input data directly without using any explicit prior notion of any classes that might exist. Choice of clustering method and of features will, however, implicitly influence the clusters that emerge. Features must offer sufficient discriminating information to the model while avoiding redundancies and overfitting, and, as outlined above, different clustering methods are suited to modelling different data structures. Hence, it is necessary to select an appropriate combination of features and algorithm for a given clustering situation. These choices may be informed by both domain-specific knowledge (pertinent scientific theory), and an awareness of the strengths and weaknesses of different clustering methods.

## 2.2 Dimensionality reduction

For $N$ observations in a $D$-dimensional feature space, dimensionality reduction aims to determine a $d$-dimensional subspace (where $d < D$, and usually with $1 \leq d \leq 3$ for visualisation purposes) within which to represent these observations, while incurring a minimal loss of information in doing so. Different methods vary in their definition of information, and in how they aim to preserve

it in the subspace that they determine. Both supervised and unsupervised methods exist. I distinguish between two main types of dimensionality reduction methods in this thesis: linear methods and non-linear methods.

The subspaces determined by linear methods of dimensionality reduction are linear projections of full, $D$-dimensional feature spaces. The $d$ subspace features are linear combinations of the $D$ original features, and so the relationship between the subspace and full space is clearly defined. This enables observations that were not used for the dimensionality reduction to be mapped onto the subspace afterwards. Principal component analysis (Pearson, 1901; Hotelling, 1936) is an unsupervised method which, by calculation of the eigenvectors of the covariance matrix of input data, yields $D$ perpendicular components which successively capture the maximum remaining variance in the input data. The projection of observations onto the first $d$ of these components yields a subspace which encompasses largest share possible (in $d$ dimensions) of their overall variance. Linear discriminant analysis, on the other hand, is a supervised method which maximises the variance between the prototypes of labelled observations which occupy $k$ classes. Because the positions of $k$ class prototypes in a $D$-dimensional full space may be described using $k-1$ vectors (assuming that one prototype is taken as the origin), $d \leq k-1$ for subspaces calculated using linear discriminant analysis.

Various non-linear methods exist. The architectures of both self-organising maps (Kohonen, 1982) and auto-encoders (LeCun, 1987; Bourlard & Kamp, 1988; Kramer, 1991) are based on artificial neural networks, which enables the preservation of global structures within the input data when mapping it onto a subspace, and makes these methods useful for pre-processing. Methods like Sammon (1969) projection, $t$-distributed stochastic neighbour embedding (van der Maaten & Hinton, 2008) and uniform manifold approximation and projection (McInnes et al., 2018) are instead based directly on inter-observation distances, and aim to preserve the local structures of observations with their transformation from $D$-dimensional full feature spaces to $d$-dimensional subspaces. While these methods do not produce models with which to map new observations (i.e. that were not used for the dimensionality reduction) onto subspaces[4], they can produce powerful visualisations which assist in the interpretation of input data (see Section 6.1).

## 2.3 Literature review

A common aspect among previous studies that have examined the use of clustering methods on samples of galaxies has been the comparison of clustering outcomes with the established early- versus late-type morphological dichotomy. Ellis et al. (2005), hierarchically clustering 350 galaxies from the Millennium Galaxy Catalogue (Liske et al., 2003) using 10 photometric and morpho-

[4] Although a neural network, for example, may be trained to *learn* this mapping.

logical features (including colours and central concentrations), found that a two-cluster outcome that correlated with Hubble-like classifications offered the most natural description of their sample. Barchi et al. (2016), specifically examining two-cluster outcomes given by the $k$-means and hierarchical clustering methods, determined a similarly broad morphological distinction for a sample of $1,962$ galaxies that were characterised by a set of five morphological features. They found that their clustering outcomes could predict simple Galaxy Zoo morphologies with $\sim 90$ per cent accuracy overall. Hocking et al. (2017, 2018) directly analysed the images of $\sim 60,000$ galaxies from the Cosmic Assembly Near-infrared Deep Extragalactic Legacy Survey (Grogin et al., 2011; Koekemoer et al., 2011; Skelton et al., 2014). Their clusters, given by the hierarchical clustering of a graph-based representation of their sample, also offered a clean early- versus late-type division and demonstrated good agreement with Galaxy Zoo classifications (Simmons et al., 2017). In addition, the high number of clusters that were determined ($\sim 1,000$) enabled the discovery of rare objects. Martin et al. (2020), building upon this work, applied the same method to a sample of galaxies from the Hyper Suprime-Cam Subaru Strategic Programme (Aihara et al., 2018a,b), used the $k$-means method to derive clear morphological identities for their $\sim 160$ clusters, and evaluated these clusters using silhouette scores (Rousseeuw, 1987) and trends with ancillary features (e.g. colours, stellar masses). Spindler et al. (2020) used an auto-encoder to learn a summary set of 20 morphological features from SDSS images, and a Gaussian Mixture Model to discern 12 clusters on the basis of these features. Their outcomes were evaluated against Galaxy Zoo 2 classifications (Willett et al., 2013). Use of an auto-encoder also enabled the generation of synthetic images from parameters estimated by the Gaussian Mixture Model. Cheng et al. (2020), also working with SDSS pixel data, combined an auto-encoder with $k$-medoids and hierarchical clustering to determine 27 clusters which predicted Oh et al. (2013) morphological classifications with $\sim 87$ per cent accuracy.

Sánchez Almeida et al. (2010) adopted a particularly exploratory stance, using the $k$-means method to cluster the spectra (in a $1,637$-dimensional feature space) of $788,677$ galaxies from SDSS. They converged upon a 28-cluster outcome following a trial-and-error model selection search, with 99 per cent of their sample being contained within 17 "main" clusters. The strong imbalance in the sizes of their clusters (spanning three orders of magnitude in the number of galaxies that they contained, in base 10) likely resulted from the extremely high dimensionality of their input feature space. Their clusters correlated closely (as would be expected) with spectral features (e.g. colours, emission lines), and also exhibited a trend with Hubble sequence morphologies. Sánchez Almeida et al. (2010) proposed the use of their cluster prototypes as templates for the estimation of redshifts and the imputation of missing data. de Souza et al. (2017) applied a Gaussian Mixture Model to find four main clusters of galaxies in a three-dimensional feature space that incorporated the axes of two two-dimensional emission-line diagrams (Baldwin et al., 1981; Cid Fernandes et al., 2010, 2011). Their model differed slightly from the classical view of emission-line galaxies,

distinguishing two types of star-forming galaxies and grouping Seyfert galaxies with those that host low-ionisation nuclear emission line regions.

Siudek et al. (2018b) used clustering to partition galaxies observed by the VIMOS Public Extragalactic Redshift Survey (VIPERS; Scodeggio et al. 2018). They chose the Subspace Expectation-Maximisation, which implements a clustering approach called the "Discriminative Latent Mixture" model. The approach incorporates dimensionality reduction via linear discriminant analysis as it iterates rather than as a part of any preparation of input data ahead of clustering. This ensures that improvements to the estimated parameters of the model are adaptive, and that the clustering uses only the most important information encoded within the input features. They aimed to establish the ability of Subspace Expectation-Maximisation to find a naturally-defined, astrophysically meaningful partition in terms of their input features: spectroscopic redshifts and 12 rest-frame UV-through-NIR colours. In this sense, their approach amounted to an unsupervised machine learning manifestation of SED estimation. The 12 clusters that they found revealed substructure to the colour bimodality of galaxies, and correlated with a variety of astrophysical features including stellar masses and morphologies. Their final outcome was evaluated using a series of quality-of-fit criteria and the hierarchical structure of outcomes with different $k$. Their study was repeated using photometric redshifts in place of spectroscopic redshifts, demonstrating that a meaningful partition may be determined using photometrically-derived features alone Siudek et al. (2018a). Further comments on this work are made in Chapter 5 and Appendix C.

Principal component analysis has a rich history of application in studies of galaxy evolution. Its use on galaxy spectra, which have particularly high dimensionalities, was pioneered by Sodré & Cuevas (1994) and Connolly et al. (1995) on samples consisting of 24 and 70 galaxies respectively. From the first two principal components of the spectra of 2dFGRS galaxies, Madgwick et al. (2002, 2003) manufactured a feature ($\eta$) which, by way of its correlation with emission line and absorption line strengths, summarised their current star formation activity. Yip et al. (2004), analysing $\sim 170,000$ SDSS galaxies, and Marchetti et al. (2013), analysing $\sim 30,000$ VIPERS galaxies, derived template eigenspectra for spectral classification and imputation. Ellis et al. (2005), parallel to their clustering work (see above), revealed via principal component analysis that their 10 photometric and morphological features each generally captured an even share of the variance of their sample. They also demonstrated, using linear discriminant analysis, a clear separability of E and S0 galaxies from late-type galaxies. Principal component analysis has enabled the excision of sky emission from SDSS images (Wild & Hewett, 2005) and the robust identification of post-starburst galaxies from spectra and SEDs (Wild et al., 2007, 2009, 2014). Cochrane & Best (2018) used principal component analysis to distinguish the differential roles of mass quenching and environment quenching upon the evolution of centrals and satellites from a cosmological, hydrodynamical simulation in terms of their stellar masses, halo masses, and SFRs.

Self-organising maps have also proved popular with extragalactic astrophysicists. Molinari & Smareglia (1998) used them to isolate early-type group galaxies in a four-dimensional photometric feature space. Geach (2012), exploring their potential for the pre-processing of survey data, demonstrated their utility in source classification and photometric redshift estimation on the basis of multi-wavelength photometry. Masters et al. (2015) suggested refinements to photometric redshift calibration strategies for upcoming large scale cosmological surveys (like Euclid) by using self-organising maps to highlight regions of their photometric feature space that yielded less reliable photometric redshifts. Nolte et al. (2018) projected a 41-dimensional sample of $7,356$ galaxies from the GAMA survey onto a two-dimensional 20-by-20 self-organising map, and concluded that Kelvin et al. (2014a) and Moffett et al. (2016) morphological classifications were not cleanly separable in their map. Hemmati et al. (2019) and Davidzon et al. (2019), studying observations and simulations respectively, applied self-organising maps for the inference of the physical properties of galaxies from their colours and SEDs, with a view to bypassing the need for the full estimation of the SEDs of galaxies. Steinhardt et al. (2020), by way of the *t*-distributed stochastic neighbour embedding of 30 bands of photometry onto two dimensions, were able to make a finer distinction between passive and dusty galaxies than that which would be possible through the use of traditional colour-colour plots.

## 2.4 Summary

Much of the interest in machine learning techniques within astronomy and astrophysics has been targeted at supervised techniques (Ball & Brunner, 2010; Baron, 2019), with a view to scaling existing astrophysical knowledge up to next-generation sample sizes. However, the use of unsupervised techniques has also been explored, and, among these, clustering and dimensionality reduction have shown significant potential in terms of their scientific utility. As motivated in Chapter 1, the aims of this thesis are to assess the use of clustering for the study of galaxy evolution. Hence, I focus on prototype- and model-based clustering algorithms, whose clear cluster identities will assist in the astrophysical interpretation of clustering outcomes. The *k*-means method is a simple, robust, and versatile clustering approach, and it is employed for the exploratory work presented in Chapters 3 and 4. In Chapter 5, I switch to Subspace Expectation-Maximisation, whose model-based clusters are freer to vary in their shapes and sizes. In addition, the move to a feature space of a higher dimensionality in Chapter 5 (nine, up from five in preceding chapters) is addressed by the inclusion of iterative dimensionality reduction in Subspace Expectation-Maximisation.

# Chapter 3

# Reproducible $k$-means clustering in galaxy feature data from the GAMA survey

The work presented in this chapter and in Appendix A has been published in Turner et al. (2019). See also the Publications page of the front matter of this thesis.

In this chapter, I test the viability of the $k$-means method as a galaxy classification solution for the next generation of extragalactic surveys and as a tool with which to explore feature spaces of high dimensionalities using a redshift- and magnitude-limited pilot sample of $7,338$ galaxies from the GAMA survey. I represent the sample using a preliminary selection of five features. I comment on how this preliminary feature selection influences the clusters that I find in the pilot sample, and how the selection might be improved for future studies. Cluster identities are discussed in terms of the input clustering features, by comparison with Hubble-like morphological classification, and in relation to the local environmental densities of the galaxies that they contain.

This chapter proceeds as follows. In Section 3.1, I describe my $k$-means implementation and my cluster evaluation method. In Section 3.2, I outline the pilot sample and feature selection. In Section 3.3, I present and analyse clustering results, and in Section 3.4, I summarise and conclude this chapter. This chapter is supplemented by Appendix A, in which I use a simple two-dimensional simulation to explain my use of stability for cluster evaluation (Section A.1), examine the stability of the clustering results in the context of bootstrap resampling of the pilot sample (Section A.2), and present examples of images of the galaxies in each of the clusters comprising each of the outcomes that I examine in the chapter (Section A.3). Where required in this chapter, I assume a $(H_0, \Omega_m, \Omega_\Lambda) = (70 \text{ km s}^{-1} \text{ Mpc}^{-1}, 0.3, 0.7)$ cosmology.

## 3.1 $k$-means and cluster evaluation

The $k$-means method is an unsupervised clustering approach that aims to partition a sample of $N$ observations, represented in a $D$-dimensional feature space, into $k$ compact, spherical clusters. Each of the clusters (a set of observations $C$) is characterised by its centroid ($\bar{\mathbf{c}}$); its arithmetic mean in each of the $D$ features. The standard $k$-means implementation ("`k-means`"; MacQueen 1967; Lloyd 1982) is a simple, fast algorithm comprising three steps:

0. Initialise: $k$ initial centres are selected (e.g. uniformly at random) from the observations.
1. Assign: the observations are assigned to their nearest centre (by Euclidean distance); these assignments are clusters.
2. Update: the centroid of each assignment is calculated; these become the new, updated centres.

Steps 1 and 2 are iterated until the algorithm converges; until there are no further differences between subsequent iterations. The convergence of `k-means` to a clustering outcome is provably always finite (Selim & Ismail, 1984), with a complexity $O(NDki)$ and generally requiring far fewer iterations ($i$) than there are observations (Duda et al., 2000). The final assignment of the observations may be taken as a classification scheme. The resultant partition of the sample is a Voronoi tessellation based on the final centroids. The final centroids are cluster prototypes: a $k$-point characterisation of the sample.

By iteratively recalculating the centroids, `k-means` inherently minimises the Sum of SQuare residuals Within ($SSQW$; Equation 3.1; i.e. variance within) each of its clusters. The `k-means` definition of a cluster follows: clusters are data structures that are compact and separated (and therefore accurately characterised by their centroids), such that they have a lower $SSQW$. The total $SSQW$ of a set of `k-means` clusters, $\phi$ (Equation 3.2), may be applied as an overall measure of their clustering quality. A consequence of the minimisation of $\phi$ is that `k-means` tends to produce clusters of similar sizes in the feature space. This is called the "uniform effect" (Liu et al., 2010), and it acts equally in all dimensions, leading also to spherical (rather than extended) clusters. It is common to normalise data to mitigate the influence of this effect on the results of `k-means`.

$$SSQW_j = \sum_{\mathbf{c}\in\mathbf{C_j}} \|\mathbf{c} - \bar{\mathbf{c}}_\mathbf{j}\|^2. \tag{3.1}$$

$$\phi = \sum_{j=1}^{k} SSQW_j = \sum_{j=1}^{k} \sum_{\mathbf{c}\in\mathbf{C_j}} \|\mathbf{c} - \bar{\mathbf{c}}_\mathbf{j}\|^2. \tag{3.2}$$

`k-means` is a local search heuristic: the behaviour of the centres as the algorithm iterates is dictated by those observations in their vicinities. Therefore, the outcome of `k-means` is dependent on the input initialisation. A common initialisation technique is to select $k$ observations from the sample uniformly at random. However, different such random initialisations may result in different, locally optimal clustering outcomes. It is computationally impractical to search for the global optimum in all of the $k^N$ clustering permutations of a large sample. Hence, when presented with different outcomes generated from different runs of `k-means` with the same $k$ on the same sample, it is standard practice to select as the optimal outcome that with the lowest $\phi$.

To mitigate the local dependency of `k-means`, I apply the random initialisation technique of Arthur & Vassilvitskii (2007) in *all* of my runs of the algorithm. It spreads out the initial centres, making the subsequent results of `k-means` more competitive with globally optimal outcomes. The first of the $k$ centres is selected from the sample with uniform probability. Subsequent centres are then selected with an increasing probability at larger distances from all preceding centres. This encourages optimisation to separated clusters. Whilst this initialisation is slower than a uniformly random initialisation, it generally yields a faster convergence over the iteration steps, resulting in a lower overall computation time for `k-means`.

### 3.1.1 Stability

A key consideration with applying `k-means` is the use of a suitable value of $k$; this is required as an input to the algorithm. `k-means` will always converge to an outcome, even in the absence of clustering structure in the sample. Assuming the sample has a clustering structure, it is generally the case that $k_{true}$, the *true* number of clusters in a given sample, is not known. It may even be that the true clusters in a sample have a hierarchical structure, such that there are several unknown values of $k_{true}$. Hence, it is common to trial clustering on a sample at several values of $k$ and to identify good values for modelling the true clustering structure of the sample post-clustering. Comparing these results necessitates an additional, alternative measure of clustering quality; $\phi$ decreases systematically as $k$ increases because more clusters occupy the same sample.

I identify good values of $k$ based on the stability of their clustering outcomes (von Luxburg, 2010; Lisboa et al., 2013). Specifically, I examine the stability of outcomes in spite of random initialisations, which may result in different clustering outcomes. Outcomes at some values of $k$ may be more or less different to one another than outcomes at other values of $k$. Those values of $k$ at which outcomes are more similar to one another are more stable; `k-means` consistently converges to similar outcomes, which implies a clustering structure in the sample at those values of $k$.

Stability may be understood by considering the behaviour of the `k-means` centres as the algorithm iterates. A key expectation is that if there is at least one centre in each of the $k_{true}$ clusters at

initialisation, then the centres will remain within those true clusters as `k-means` proceeds (Bubeck et al., 2012). For $k = k_{true}$, the centres will then settle to the centroids of the true clusters, in accordance with the algorithm's inherent minimisation of $\phi$. For $k > k_{true}$, this key expectation means that true clusters containing more than one centre at initialisation will be split. For $k < k_{true}$, where this key expectation does not hold, centres may move between true clusters and lead to mergers. The exact splits and mergers that occur are dependent on the locations of the centres at initialisation and will therefore change with different initialisations. It is important to note that when $k = k_{true}$, the Arthur & Vassilvitskii (2007) initialisation technique facilitates the ideal situation in which the $k_{true}$ clusters contain one centre each at initialisation. I demonstrate these concepts using a simple two-dimensional simulation in Section A.1.

To measure the difference between a pair of clustering outcomes at the same $k$, I use Cramér's $V$ index of association (Cramér, 1946). My use of the index in the context of stability is denoted with the symbol $V_s$:

$$V_s = \sqrt{\frac{\chi^2}{N \cdot (k-1)}}. \tag{3.3}$$

Here, $\chi^2$ is the chi-squared value for two clustering outcomes (categorical variables $A$ and $B$) on the same sample, each consisting of the same number ($k$) of unique labels. It is calculated (Equation 3.4) using a $k \times k$ contingency table (a.k.a. cross tabulation), comparing the observed frequency of observations ($o$) in each cell $(a, b)$ with its expected frequency ($e = N/k^2$; equal in every cell) given a null hypothesis of independence of the two outcomes. I provide examples of contingency tables and of the calculation of their corresponding $\chi^2$ and $V_s$ values in Section A.1.

$$\chi^2_{A,B} = \sum_{a,b} \frac{(o_{a,b} - e_{a,b})^2}{e_{a,b}}. \tag{3.4}$$

$V_s$ is normalised, reporting $\chi^2$ as a square-root-scaled fraction of its maximum possible value given $N$ (the number of observations in the sample) and $k$. It ranges from 0 for no agreement (i.e. the outcomes are independent; agreement is consistent with uniform random chance) to 1 for perfect agreement. In practice, `k-means` cannot produce outcomes that disagree to the extent that $V_s = 0$. I assess the stability of an individual clustering outcome by calculating its median $V_s$ with respect to other outcomes at the same $k$. I assess the stability of a set of clustering outcomes at the same $k$ by examining their distribution in median $V_s$ (see Figures 3.4, A.5, and A.6). Stability also enables the determination of whether there is *no* clustering structure in the sample (i.e. no $k_{true}$), as no particular value of $k$ will stand out from the others as being particularly stable.

## 3.2 The pilot GAMA sample

I use data from phase II of the Galaxy And Mass Assembly (GAMA) survey (Driver et al., 2009, 2011; Liske et al., 2015). The main aim of the survey is to study cosmic structure on scales of 1 kpc to 1 Mpc in the context of CDM models of the Universe. The survey is structured around its spectroscopic campaign, conducted at the Anglo-Australian Telescope using the AAOmega spectrograph (Sharp et al., 2006), and based on an input catalogue defined by Baldry et al. (2010). The spectroscopy provided reliable heliocentric redshifts for $\sim 238,000$ objects to a limiting $r$-band Petrosian magnitude of 19.8 and across five regions covering a total area of 286 deg$^2$. It has been supplemented with reprocessed imaging in 21 bands from a variety of other surveys (e.g. SDSS) that overlap with the GAMA spectroscopic campaign footprint (the Panchromatic Data Release; Driver et al. 2016). Data derived from these spectra and images are listed in tables hosted at `http://www.gama-survey.org/`.

I derive my pilot sample from the well-characterised sample of Moffett et al. (2016) (listed in the GAMA survey table `VisualMorphologyv03`) with a view to facilitating the interpretation of clustering results, particularly by comparison with results from other, previous GAMA survey studies. It is a flow-corrected-redshift- ($0.002 < z < 0.06$) and magnitude- ($r_{PETRO} < 19.8$) limited sample of $7,556$ local objects that have been morphologically classified using the method of Kelvin et al. (2014a). Note that while I intend to compare clustering outcomes with these visual, Hubble-like morphologies (see Section 3.3), I do not aim to reproduce them.

The Kelvin et al. (2014a) method assigns classifications by the consensus of three expert observers, whose visual inspection of optical three-colour images of galaxies is guided by a decision tree. The tree discriminates galaxies firstly as being either spheroid- or disc-dominated, and secondly as consisting either of a single component or of multiple components. The tree goes on to discern multiple component galaxies with bars from those without, but I ignore this distinction in this chapter due to the relatively low number of barred galaxies in my pilot sample ($\sim$ 4 per cent of all of the galaxies have bars). An examination of bars in the context of clustering is reserved for a future study, though, due to the significant role they play in the evolution of any barred galaxy (see Section 1.3.1).

The two levels of the tree that I do consider lead to four morphological types: E, S0-a, Sab-Scd, and Sd-Irr. A fifth type, "Little Blue Spheroid" (LBS), is identified separately at the top level of the tree. As star-forming, spheroid-dominated, blue dwarf galaxies, they have been likened to the blue early-type galaxies of Schawinski et al. (2009a) in that they defy expected galaxy trends with colour or morphology. Contaminants are also identified at the top level of the tree. There are 25 in the sample, which are mostly secondary galaxies, partial galaxy structures, or star-galaxy blends. These are removed from the sample, leaving $7,531$ galaxies.

**Table 3.1:** The features used to characterise galaxies, and the survey tables from which they are retrieved.

| Feature | Unit | Table | Version | Column | Reference |
|---|---|---|---|---|---|
| Stellar mass | $M_\odot$ | MagPhys | 6 | mass_stellar_best_fit | Driver et al. 2016 |
| u-r colour | mags | StellarMassesLambdar | 20 | uminusr | Taylor et al. 2011 |
| Sérsic index | - | SersicCatSDSS | 9 | GALINDEX_r | Kelvin et al. 2012 |
| Half-light radius | kpc | SersicCatSDSS | 9 | GALRE_r | Kelvin et al. 2012 |
| Specific star formation rate | $yr^{-1}$ | MagPhys | 6 | sSFR_0_1Gyr_best_fit | Driver et al. 2016 |

I then retrieve feature data for the galaxies in the pilot sample. There are hundreds of features available in the GAMA database with which to characterise the galaxies in the sample. One may be tempted to find clusters in the sample using all of them at once, in order to provide the algorithm with as much information as possible. However, as the dimensionality of the feature space containing the sample increases, the observations become more sparse, and `k-means` (or any clustering algorithm) will tend to overfit its clusters to the observations. Representing the sample using a smaller subset of features instead results in clusters that are more readily generalisable to the overall galaxy population.

Feature selection may involve both domain-specific knowledge and statistical considerations. I select features that capture intrinsic properties of galaxies, relating to their formation and evolution, and aim for a selection that expresses most known aspects of these processes. The redundancy of features - the extent to which they provide the same information as other features - may be assessed statistically, such as by calculation of their Spearman rank-order correlation coefficients. A higher correlation between features implies greater redundancy. Redundant features exert a higher influence over the clusters that `k-means` finds in that they lead to a projection of the feature space in which the sample is highly extended in one direction over others (e.g. Figure 3.3). `k-means` will therefore tend to split the data along the extended direction of the data due to the uniform effect. While it is common to discard such features to avoid this bias, retaining them could instead serve to strengthen a desirable pattern in the data.

Table 3.1 lists the feature data I retrieve and the main GAMA survey tables I access to do so. Stellar masses ($M_*$) and Gigayear-timescale specific star formation rates ($sSFR$) are taken from `MagPhysv06`, generated by Driver et al. (2016) from a run of `MAGPHYS` (Da Cunha et al., 2008) on all 21 bands of foreground extinction-corrected photometry listed in `LambdarCatv01` (Wright et al., 2016). `MAGPHYS` estimates SEDs from input redshifts, fluxes, and flux errors using star- and dust-emission template spectra, and corrects for light-attenuation by dust within galaxies. Rest-frame $u-r$ colours come from `StellarMassesLambdarv20`, derived from the same photometry (Taylor et al., 2011). I select $u-r$ colours for their ability to express the galaxy bimodality in the colour versus stellar mass plane. Unlike $M_*$ and $sSFR$, it does not include corrections for dust attenuation, meaning clustering outcomes will be influenced in some capacity by the presence of dust in some of the galaxies in the pilot sample. I take $r$-band Sérsic indices ($n_g$) and half-light

**Table 3.2:** The limits for truncation that have been imposed on each of the features. The truncated histograms are viewable in Figure 3.1. Limits marked with an asterisk do not actually exclude any galaxies.

| Feature | Units | Lower | Upper |
|---|---|---|---|
| $M_*$ | $\log_{10}(\mathrm{M}_\odot)$ | 6 | *12 |
| $u-r$ | mags | 0.3 | 2.7 |
| $n_g$ | $\log_{10}(\mathrm{n})$ | −0.6 | 1.2 |
| $R_{1/2}$ | $\log_{10}(\mathrm{kpc})$ | −1.0 | *1.5 |
| $sSFR$ | $\log_{10}(\mathrm{yr}^{-1})$ | −14 | −8 |

radii ($R_{1/2}$) from `SersicCatSDSSv09`, whose derivation is described in Kelvin et al. (2012) and is based on reprocessed SDSS imaging (Hill et al., 2011). The accuracy of Sérsic indices is expected to be consistent throughout the pilot sample due to the low redshifts of the galaxies therein (Vika et al., 2013).

Matching for data in all features leaves 7,516 galaxies in the sample, with 15 lost due to incompleteness. My use of features derived primarily from broad-band photometry facilitates a comparison of clustering results with a wide range of other surveys. This feature selection is preliminary, and I comment on the consequences of it for clustering later in this section and in Section 3.3, and on the potential for optimising feature selection in Section 3.4.

The half-light radii listed in `SersicCatSDSSv09` are presented in units of arcseconds, which are a function of the distances to the galaxies as well as of their intrinsic sizes. Flow-corrected redshifts (`DistanceFramesv12`; Baldry et al. 2012) are used to convert these angular radii to intrinsic kiloparsec radii.

A series of transforms are applied to standardise the data, with the intention of avoiding unintuitive partitions due to the uniform effect, and of granting equal weight to all of the features. The distributions of all of the features except $u-r$ colour are strongly skewed in linear units. The centroids that `k-means` iteratively recalculates are sensitive to the uneven tails of skewed distributions. I therefore ensure that all features are represented in logarithmic units (see Table 3.2), as it typical in the astrophysics literature.

Outliers in the sample are more readily apparent when examining the distributions of each of the features in logarithmic units. In order to mitigate the influence of outliers on the calculation of centroids by `k-means`, the sample is truncated in each of the features, removing galaxies that lie outside given limits. The limits that are imposed are listed in Table 3.2. While some of these limits have been set simply by inspection of histograms of the sample in each of the features, others have also involved astrophysical considerations. For example, the limits in Sérsic index have been set to eliminate subcomponent fits, or fits affected by light from sources near the galaxy that was fitted. Removing 198 outliers from the sample leaves 7,338 galaxies.

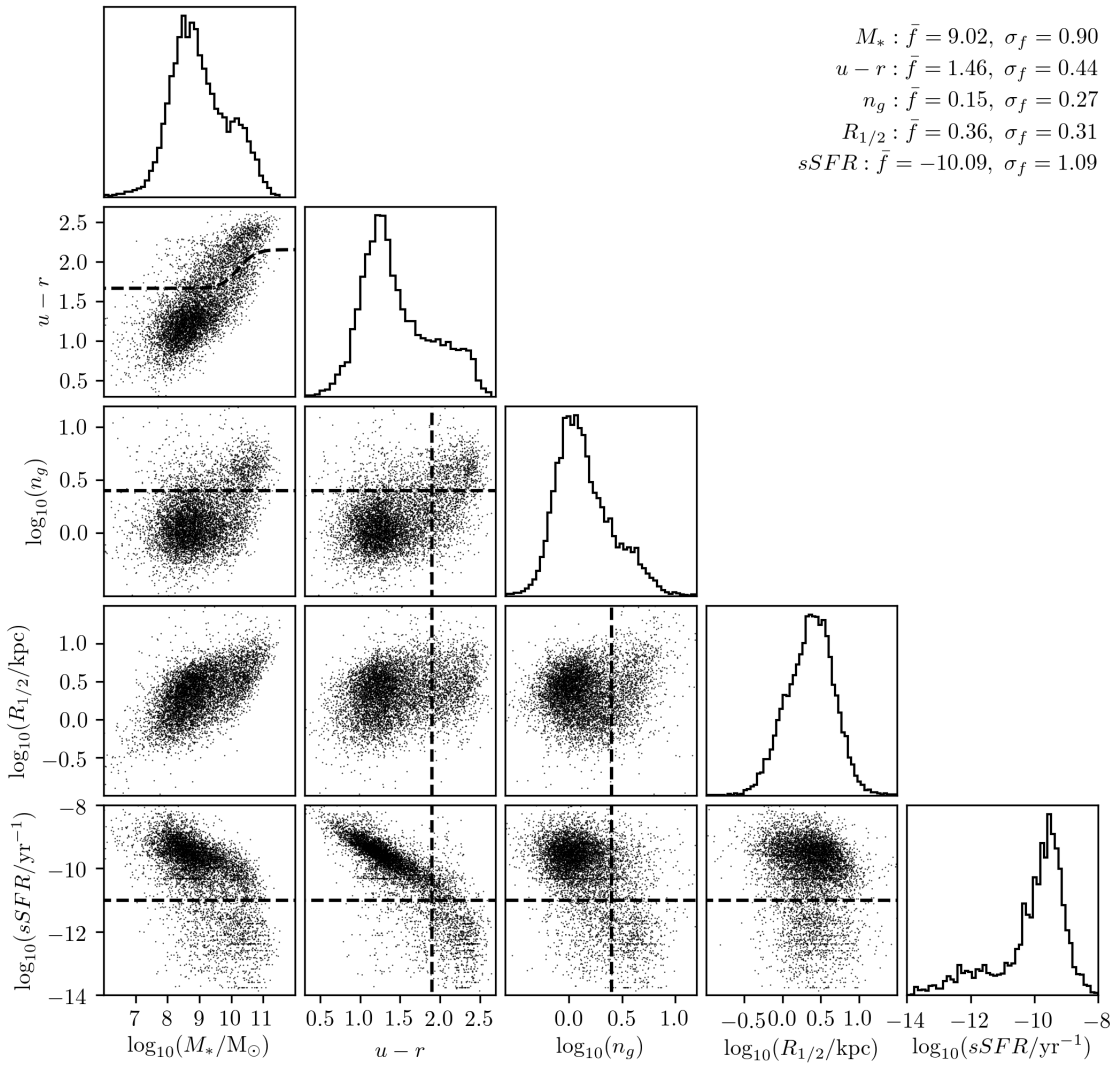


**Figure 3.1:** A profile of the pilot sample. It is represented using histograms and scatter plot projections. The dashed lines mark "classical" distinctions between the two main populations of galaxy (see text). The mean ($\bar{f}$) and standard deviation ($\bar{\sigma}$) of the sample in each feature is also listed in the upper right of the figure, in the units shown on the axes.

This now constitutes the final pilot sample, which I profile in Figure 3.1. Histograms and scatter plots are used to show the distribution of the sample in one- and two-dimensional projections. These distributions reveal a dominance of low-mass, blue, star-forming galaxies with low Sérsic indices in the sample. Most features exhibit significant secondary components to their distributions. The bimodality of galaxies is visible in several of the scatter plot projection panels.

I include dashed lines in Figure 3.1 which mark "classical" distinctions from the literature that have been made between the two main populations of galaxies. The line in the colour-mass panel is based on equation 11 of Baldry et al. (2004). I use a solar $r$-band absolute magnitude of 4.71 and equation 12 from the same paper to adapt it from applying to magnitudes to applying to masses. While the original line was calculated using SDSS "model" magnitudes, the `LAMBDAR` apertures were set using `Source Extractor` (Bertin & Arnouts 1996; see also section 6.1 of Wright et al. 2016). Model magnitudes report redder colours than magnitudes derived using top-hat apertures: I calculate an approximate mean offset of $-0.15$ over the range of colours in the sample and adjust the line accordingly. The lines in the scatter panels involving Sérsic indices and $u - r$ come from Lange et al. (2015). I apply a similar colour offset of +0.4 as the $u - r$ colours in Lange et al. (2015) are corrected for dust attenuation. The line for *sSFR* is taken from Pozzetti et al. (2010); specifically, I take their distinction between passive and non-passive galaxies.

Figure 3.2 shows the distribution of Kelvin et al. (2014a) morphologies in the pilot sample. The histogram reveals a dominance of late-type morphologies in the sample. The morphological types are ordered from highest (E) to lowest (LBS) mean stellar mass.

Some features span larger numerical ranges in logarithmic units than the others. For example, $M_*$ spans six orders of magnitude (base 10) while $n_g$ spans 1.8. Given that `k-means` minimises $\phi$ in all dimensions, it will tend to split the sample along any direction in which it is extended. To mitigate any bias of `k-means` for or against any of the features on this basis, the data is coded in each of the features using $Z$-scores. These are more strongly influenced by the centres of the feature distributions than their extremities[1]; hence, they are weighted towards the majority of galaxies near the feature distribution means, rather than the minority of outliers. Here (Equation 3.5), $f$ is the value of an observation in a given feature, $\bar{f}$ is the mean value of that feature, $\sigma_f$ is its standard deviation, and $Z_f$ is the $Z$-score of $f$:

$$Z_f = \frac{f - \bar{f}}{\sigma_f}. \tag{3.5}$$

[1]Other normalisation techniques, such as the more common min-max normalisation (which maps features to a consistent numerical range – usually 0 to 1 – based on their minimum and maximum values), are alternatively influenced more by the extremities of feature distributions. Min-max normalisation is applied in Chapter 4 where, due to my use of truncation, it yields similar results to the $Z$-scores used in this chapter.

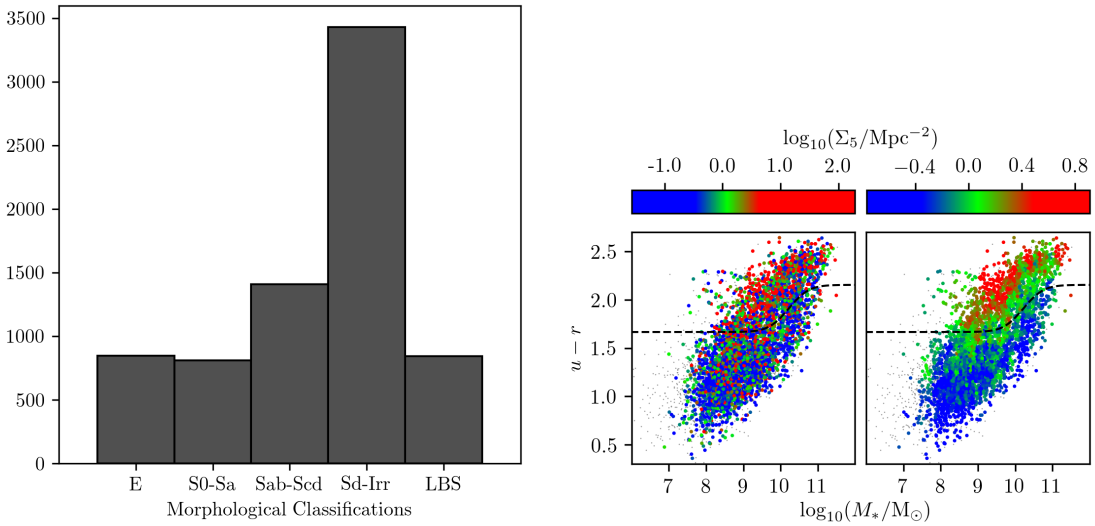


**Figure 3.2: Left:** Histogram showing the distribution of Kelvin et al. (2014a) and Moffett et al. (2016) morphologies in the sample. The sample is dominated by late-type morphologies. LBS stands for "Little Blue Spheroid". **Right:** The sample, projected onto the $u-r$ versus $M_*$ plane, with points coloured by their measured (left panel) and smoothed (right panel) local environmental densities ($\Sigma_5$). The small grey points represent those galaxies for which $\Sigma_5$ is not available; their bias toward lower masses and bluer colours is apparent. The dashed black line marks the Baldry et al. (2004) distinction between the blue and red sequences of galaxies.

**Table 3.3:** Spearman rank-order correlation coefficients for the features used to represent the sample.

| Feature | $M_*$ | $u-r$ | $n_g$ | $R_{1/2}$ | *sSFR* |
|---|---|---|---|---|---|
| $M_*$ | 1.00 | | | | |
| $u-r$ | 0.72 | 1.00 | | | |
| $n_g$ | 0.35 | 0.40 | 1.00 | | |
| $R_{1/2}$ | 0.55 | 0.25 | −0.04 | 1.00 | |
| *sSFR* | −0.61 | −0.83 | −0.38 | −0.19 | 1.00 |

Having been standardised, I assume that `k-means` will now be able to recover clustering structure in the sample that reflects the astrophysics involved in the formation and evolution of the galaxies therein. I now assess my feature selection pre-clustering. In Table 3.3, I show the Spearman rank-order correlation coefficients for pairs of the five features used to represent the sample. I note that $u-r$ colours are involved in the two strongest correlations of features (with $M_*$ and *sSFR*) for the sample, suggesting redundancy. I opt to retain it, however, to strengthen the bimodal structure of the data and because it includes information about the dust content of the galaxies in the sample, which *sSFR* does not. I expect that strengthening the bimodality will encourage `k-means` to search for more clusters *within* the two peaks of the bimodality at higher values of $k$. Other correlations that exist among my selection of features are weaker and do not suggest any further significant redundancies.

**Table 3.4:** Results of a principal component analysis of the sample.

| Feature | PC1 | PC2 | PC3 | PC4 | PC5 |
|---|---|---|---|---|---|
| $M_*$ | 0.51 | −0.28 | −0.02 | −0.63 | −0.51 |
| $u-r$ | 0.53 | 0.15 | −0.32 | −0.22 | 0.74 |
| $n_g$ | 0.37 | 0.40 | 0.84 | 0.07 | 0.04 |
| $R_{1/2}$ | 0.29 | −0.80 | 0.21 | 0.44 | 0.19 |
| *sSFR* | −0.49 | −0.30 | 0.39 | 0.18 | 0.40 |
| Relative Variance | 0.59 | 0.21 | 0.12 | 0.05 | 0.03 |

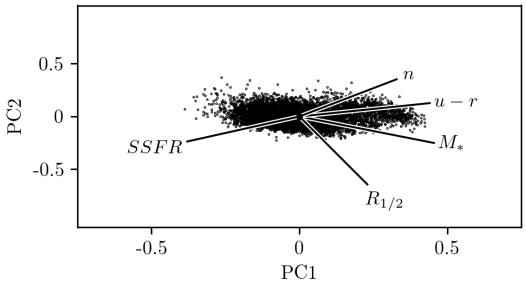


**Figure 3.3:** The features as functions of the first two principal components of the sample. The axes are scaled to show the relative variance that each principal component encompasses. I also represent the sample using scatter points. The size of the sample in this two-dimensional principal component space is normalised to fit within the area shown. $M_*$, $u-r$, and *sSFR* are most strongly associated with PC1, while $R_{1/2}$ is most strongly associated with PC2. $n_g$ is evenly balanced between the two, but is most strongly associated with PC3, which is not shown.

I also conduct a principal component analysis of the sample, to gain insight into its covariance structure and to anticipate clustering results. The results of this analysis are listed in Table 3.4. The results reveal that the structure of the sample is dominated by the first principal component (PC1), which encompasses 59 per cent of the sample's variance. This indicates that the sample has an elongated shape in the five-dimensional feature space. PC1 is defined mostly by those features that reflect aspects of the stellar populations within galaxies (i.e. $M_*$, $u-r$, and *sSFR*), so these features are expected to most strongly dictate the clusters that `k-means` finds. $R_{1/2}$ and $n_g$ are most strongly associated with PC2 (encompassing 21 per cent of the variance in the sample) and PC3 (12 per cent) respectively. Hence, these features are expected to play a role in dictating clusters at higher values of $k$, at which the use of additional centroids enables the algorithm to explore subtler, more local substructures within the sample. These relationships are clearly apparent in Figure 3.3, which shows the features and sample as functions of the first two principal components.

Finally, I also retrieve environmental data for the galaxies in the sample in order to probe the role of environment in dictating the clusters I find via its influence on the features. I choose the surface density $\Sigma_5$, defined using the projected comoving distance from a galaxy to its fifth-nearest neighbour, as a measure of local environmental density (via `EnvironmentMeasuresv05`; Brough

et al. 2013). This feature is only available for 4,195 of the 7,338 galaxies in the sample if I filter for a `SurfaceDensityFlag` of 0, which ensures that the fifth-nearest neighbour of a given galaxy lies within the GAMA survey footprint. Nearly all of the other 3,143 galaxies in the sample have a SurfaceDensityFlag of 2, indicating no neighbours within their distance from the survey edge, meaning they occupy particularly low-density environments.

These 3,143 galaxies are not evenly distributed in feature space. Most are blue, low-mass, and have low Sérsic indices, large radii, and high specific star formation rates, consistent with the Baldry et al. (2006), Bamford et al. (2009), and Peng et al. (2010) findings that such galaxies tend to occupy lower-density environments. I comment on the consequences of this incompleteness for my examination of local environmental densities within clusters where relevant in Section 3.3. Naturally, confidence in conclusions based on this data would be greater were this data available for the entirety of the sample.

Figure 3.2 shows the sample projected onto the $u - r$ versus $M_*$ plane. In the left panel, points are coloured by their measured $\Sigma_5$, taken directly from EnvironmentMeasuresv05. In the right panel, the local environmental densities have been smoothed. The smoothing is calculated by taking the average $\Sigma_5$ value of each galaxy and its seven nearest neighbours in the full five-dimensional feature space. This smoothing is applied to clarify the average trend of the five-dimensional clusters of galaxies in the two-dimensional $u - r$ versus $M_*$ plane. It is applied in all of the following environmental analyses in this chapter. This smoothing inhibits the range of $\Sigma_5$ for the sample in the right panel compared with the left. The colour bar levels have been set in order to distinguish galaxies in intermediate-density environments from those in low- and high-density environments. Both panels show that the sample is dominated by galaxies in low-density environments. The Baldry et al. (2004) dashed black line, intended as a separator of the blue and red sequences of galaxies in this feature plane, traces intermediate densities particularly closely.

## 3.3 Analysis of stable clustering outcomes

I combine stability and compactness (see Section 3.1) to evaluate `k-means` clusters in the sample, adapting the approach of Lisboa et al. (2013). I do not assume a value of $k_{true}$ (other than $k = 2$, corresponding to the bimodality of galaxies), so I trial `k-means` clustering at $k = 2$ through $k = 15$, initialising 200 times at each $k$ using the Arthur & Vassilvitskii (2007) technique. I first identify stable values of $k$, at which there appears to be a clustering structure in the sample, using $V_s$ (a measure of the strength of association between two clustering outcomes; Equation 3.3). I then select the optimal outcome at each of the stable values of $k$ by considering compactnesses.

In Figure 3.4, I map the stabilities of outcomes at different values of $k$. I calculate the median $V_s$

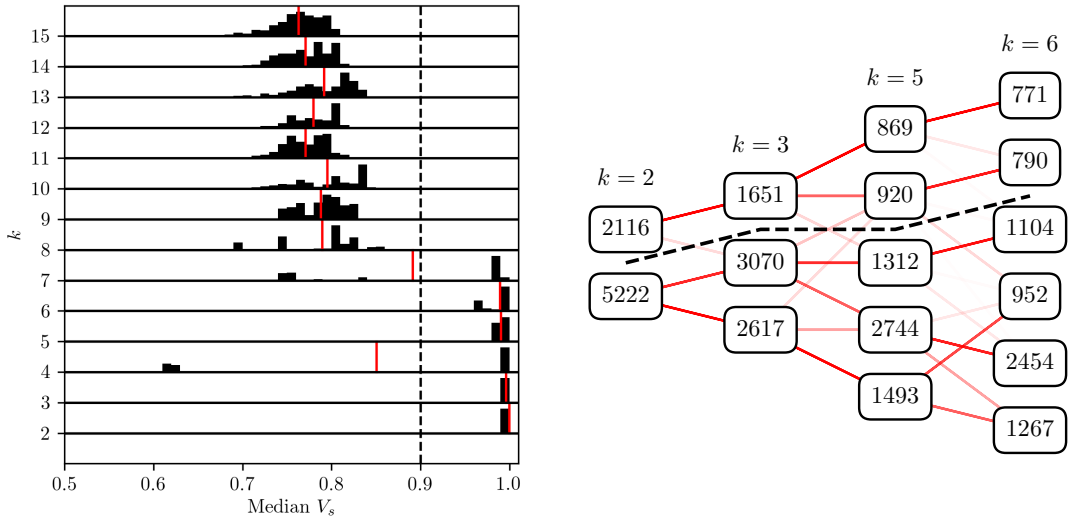


**Figure 3.4: Left:** Stability map of `k-means` clustering for the sample at $k = 2$ through $k = 15$. I calculate the median $V_s$ of each outcome with respect to all other outcomes at the same $k$. The distributions of all 200 medians at each $k$ are represented using histograms plotted along each of the horizontal black baselines. The heights of the histograms are normalised. Additionally, the means of these distributions are shown as vertical red lines. Outcomes at $k = 2, 3, 5$, and 6 are particularly stable. **Right:** Hierarchy tree showing the interrelation of $k = 2$, $k = 3$, $k = 5$, and $k = 6$. The text bubbles, representing the clusters, state the number of galaxies they contain and are ordered by the clusters' mean $u - r$ colours from reddest at the top to bluest at the bottom. The opacity of the red lines expresses how closely related connected clusters at different $k$ are. The dashed black line separates the basic structure of two "superclusters" that is found at all values of $k$.

of each individual outcome with respect to all other outcomes at the same $k$. The distributions of all 200 medians at each $k$ are represented using histograms plotted along each of the horizontal black baselines. The heights of the histograms are normalised. Additionally, the means of these distributions are shown as vertical red lines. There is a gap across all distributions, appearing to separate two distinct regimes of outcomes. These regimes are demarcated using the vertical dashed black line at median $V_s = 0.9$, but I emphasise that this "threshold" is a product of the sample and may differ for different samples. The key element for distinguishing between stable and unstable values of $k$ is the gap.

Values of $k$ at which the distributions are concentrated toward higher median $V_s$ (i.e. at which more outcomes are more consistent) are more stable. The outcomes at $k = 2, 3, 5$, and 6 stand out as being particularly stable. All of the outcomes at each of these values of $k$ have median $V_s > 0.9$, except for a single outcome at $k = 3$. The spread of outcomes at $k = 6$ corresponds to a maximum difference of $\sim 100$ galaxies ($\sim 1.5$ per cent of the sample) between outcomes.

The outcomes at $k = 4$ occupy a highly-peaked bimodal distribution, with a slight majority (123)

at median $V_s > 0.9$. The outcomes at $k = 7$ are similarly distributed with 113 at median $V_s > 0.9$, but with a larger spread in its secondary peak. While both distributions exhibit stable components, a significant number of outcomes in each are unstable. I focus presently on those values of $k$ that are most uniformly stable, and therefore exclude any outcomes from $k = 4$ and $k = 7$ from my analyses of clustering results.

The distributions of outcomes at higher values of $k$ are centred at lower median $V_s$ and have larger spreads, meaning they are unstable. The additional centroids used by `k-means` at these higher values of $k$ are more strongly influenced by the local structure within the sample at initialisation, such that the algorithm is more likely to converge to locally (rather than globally) optimal outcomes. The ability of the Arthur & Vassilvitskii (2007) initialisation approach to mitigate the local dependency of `k-means` becomes weaker as $k$ increases. The spreads of the outcomes for these unstable values of $k$ correspond to differences of thousands of galaxies between outcomes. The general trend of decreasing stability at higher $k$ continues beyond the outcomes at $k = 15$.

From each of the four values of $k$ that I have identified as being most stable, I select as my final, optimal outcomes for analysis those with the lowest $\phi$ (a measure of the compactness of the clusters in an outcome; Equation 3.2). I refer to these four "best" outcomes as simply $k = 2$, $k = 3$, $k = 5$, and $k = 6$. These outcomes at these values of $k$ retain their stability following application of the bootstrap method to the sample (Section A.2). The stability map that results from clustering for which $u - r$ colour is omitted as an input feature has the same general structure as Figure 3.4, but the distributions of outcomes are systematically offset to slightly lower median $V_s$ at each value of $k$.

A hierarchy tree, mapping the interrelation of the best outcomes, is shown in Figure 3.4. The clusters in each outcome are represented by text bubbles which state the number of galaxies that they contain. The red lines express how closely related clusters at different $k$ are. The opacities of the lines scale linearly with the fraction of galaxies in clusters at $k + 1$ that are also found in clusters at $k$. Clusters are ordered vertically at each $k$ by the mean $u - r$ colour of the galaxies they contain, with the reddest clusters at the top and the bluest at the bottom. It should be noted that outcomes at different values of $k$ are calculated independently of one another, so hierarchy is not imposed or assumed at any point in the clustering. Despite this, the best outcomes exhibit a broadly hierarchical structure. Considering them in sequence, clusters at higher values of $k$ generally emerge as splits of clusters at lower values of $k$. There is some mixing present, meaning some clusters at higher values of $k$ contain galaxies from multiple clusters at lower values at $k$. This is especially noticeable between $k = 3$ and $k = 5$, though it may be exaggerated by the omission of an outcome at $k = 4$ from the plot. The highly peaked bimodal distribution of outcomes in median $V_s$ at $k = 4$ (Figure 3.4) arises as `k-means` settles into one of the two splits that must occur between $k = 3$ and $k = 5$. $k = 5$ is stable and includes both of these splits, so no information is

**Table 3.5:** A summary of all of the clusters in outcomes $k = 2, 3, 5$, and 6. See the main text for an explanation of cluster names. The uncertainties on the centroids are estimated by application of the bootstrap method to the sample; this estimation is outlined in detail in Section A.2

| Cluster | $N_C$ | $\log_{10}(M_*/\mathrm{M}_\odot)$ | $u-r$ | $\log_{10}(n_g)$ | $\log_{10}(R_{1/2}/\mathrm{kpc})$ | $\log_{10}(sSFR/\mathrm{yr}^{-1})$ | $\log_{10}(\Sigma_5/\mathrm{Mpc}^{-2})$ | Loss % |
|---|---|---|---|---|---|---|---|---|
| $\mathrm{Ra}_2$ | 2,116 | $10.02^{+0.01}_{-0.01}$ | $2.01^{+0.00}_{-0.00}$ | $0.37^{+0.00}_{-0.00}$ | $0.51^{+0.00}_{-0.00}$ | $-11.35^{+0.01}_{-0.01}$ | 0.16 | 30 |
| $\mathrm{Ba}_2$ | 5,222 | $8.61^{+0.00}_{-0.00}$ | $1.24^{+0.00}_{-0.00}$ | $0.06^{+0.00}_{-0.00}$ | $0.30^{+0.00}_{-0.00}$ | $-9.58^{+0.00}_{-0.00}$ | −0.21 | 48 |
| $\mathrm{Ra}_3$ | 1,651 | $10.07^{+0.01}_{-0.01}$ | $2.10^{+0.01}_{-0.01}$ | $0.45^{+0.01}_{-0.01}$ | $0.47^{+0.00}_{-0.00}$ | $-11.64^{+0.04}_{-0.02}$ | 0.23 | 30 |
| $\mathrm{Bb}_3$ | 3,070 | $9.14^{+0.02}_{-0.04}$ | $1.39^{+0.01}_{-0.02}$ | $0.04^{+0.01}_{-0.01}$ | $0.53^{+0.01}_{-0.01}$ | $-9.80^{+0.03}_{-0.02}$ | −0.15 | 38 |
| $\mathrm{Ba}_3$ | 2,617 | $8.21^{+0.01}_{-0.01}$ | $1.15^{+0.01}_{-0.00}$ | $0.09^{+0.01}_{-0.01}$ | $0.09^{+0.01}_{-0.01}$ | $-9.44^{+0.00}_{-0.01}$ | −0.26 | 57 |
| $\mathrm{Rb}_5$ | 869 | $10.46^{+0.02}_{-0.04}$ | $2.23^{+0.01}_{-0.00}$ | $0.58^{+0.01}_{-0.01}$ | $0.60^{+0.02}_{-0.03}$ | $-11.91^{+0.04}_{-0.02}$ | 0.18 | 29 |
| $\mathrm{Ra}_5$ | 920 | $9.19^{+0.08}_{-0.08}$ | $1.86^{+0.04}_{-0.07}$ | $0.26^{+0.01}_{-0.01}$ | $0.18^{+0.01}_{-0.01}$ | $-11.30^{+0.20}_{-0.10}$ | 0.29 | 29 |
| $\mathrm{Bc}_5$ | 1,312 | $9.77^{+0.04}_{-0.06}$ | $1.61^{+0.03}_{-0.03}$ | $0.10^{+0.01}_{-0.01}$ | $0.66^{+0.01}_{-0.01}$ | $-10.03^{+0.04}_{-0.03}$ | −0.15 | 37 |
| $\mathrm{Bb}_5$ | 2,744 | $8.64^{+0.03}_{-0.03}$ | $1.20^{+0.01}_{-0.01}$ | $-0.01^{+0.01}_{-0.01}$ | $0.42^{+0.01}_{-0.01}$ | $-9.50^{+0.01}_{-0.01}$ | −0.22 | 45 |
| $\mathrm{Ba}_5$ | 1,493 | $8.12^{+0.02}_{-0.03}$ | $1.14^{+0.01}_{-0.02}$ | $0.16^{+0.02}_{-0.02}$ | $-0.04^{+0.01}_{-0.01}$ | $-9.41^{+0.03}_{-0.03}$ | −0.28 | 61 |
| $\mathrm{Rb}_6$ | 771 | $10.50^{+0.01}_{-0.05}$ | $2.24^{+0.01}_{-0.01}$ | $0.60^{+0.01}_{-0.01}$ | $0.63^{+0.00}_{-0.03}$ | $-11.91^{+0.03}_{-0.11}$ | 0.17 | 29 |
| $\mathrm{Ra}_6$ | 790 | $9.32^{+0.03}_{-0.11}$ | $1.94^{+0.01}_{-0.05}$ | $0.26^{+0.02}_{-0.03}$ | $0.20^{+0.00}_{-0.02}$ | $-11.69^{+0.19}_{-0.03}$ | 0.37 | 27 |
| $\mathrm{Bd}_6$ | 1,104 | $9.90^{+0.13}_{-0.06}$ | $1.68^{+0.07}_{-0.03}$ | $0.11^{+0.03}_{-0.02}$ | $0.66^{+0.01}_{-0.01}$ | $-10.11^{+0.03}_{-0.09}$ | −0.12 | 37 |
| $\mathrm{Bc}_6$ | 952 | $8.51^{+0.07}_{-0.28}$ | $1.30^{+0.02}_{-0.09}$ | $0.36^{-0.01}_{-0.10}$ | $0.04^{+0.03}_{-0.10}$ | $-9.67^{+0.15}_{-0.03}$ | −0.21 | 49 |
| $\mathrm{Bb}_6$ | 2,454 | $8.79^{+0.12}_{-0.06}$ | $1.24^{+0.04}_{-0.03}$ | $0.00^{+0.03}_{-0.01}$ | $0.47^{+0.04}_{-0.02}$ | $-9.58^{+0.04}_{-0.05}$ | −0.20 | 41 |
| $\mathrm{Ba}_6$ | 1,267 | $7.98^{+0.15}_{-0.03}$ | $1.06^{+0.03}_{-0.01}$ | $-0.04^{+0.04}_{-0.03}$ | $0.06^{+0.12}_{-0.06}$ | $-9.27^{+0.01}_{-0.05}$ | −0.32 | 65 |

lost by the exclusion of an outcome at $k = 4$ from my analyses.

Furthermore, there is a basic structure of two "superclusters" (separated by the dashed black line) at all values of $k$, including the simplest partition $k = 2$. This indicates the strength of the bimodality in the structure of the sample. As $k$ increases, `k-means` favours splitting the blue supercluster apart over the red supercluster, due mostly to its spread in the features and the higher number of galaxies in the blue supercluster, in conjunction with the "uniform effect" of `k-means`.

I introduce a preliminary naming scheme for the clusters based on their correlation with colour. The scheme is intended as a quick way to identify clusters in the various comparisons and analyses conducted in this section, rather than as a full description or explanation of cluster identities. My use of this scheme is not intended to imply that colour is entirely responsible for the clustering outcomes (though it clearly plays a strong role). Cluster names consist of three parts in the format "$\mathrm{Xy}_Z$". The first part, either "R" (for red) or "B" (for blue), corresponds to the supercluster (upper and lower respectively in the hierarchy tree in Figure 3.4) to which the cluster belongs. The second letter ranks the cluster by its mean $u-r$ colour in comparison with those of other clusters within the same supercluster at the same value of $k$. Rankings begin at "a" for the bluest cluster, and follow on alphabetically until all clusters within the supercluster are named. The third part, a number, indicates the outcome (i.e. the value of $k$) to which the cluster belongs.

`k-means` clusters are defined by their centroids. Table 3.5 summarises the clusters in the best outcomes ($k = 2, 3, 5$, and 6). Column $N_C$ lists the number of galaxies that each cluster contains. The next five columns contain the cluster centroids as coordinates in each of the input features, along with uncertainties estimated by application of the bootstrap method to the sample (see Section A.2). The final two columns list the "environment centroid" of the clusters (i.e. the mean smoothed $\Sigma_5$ of the galaxies they contain; *not* a feature used in the clustering), and the percentage of galaxies lost from each of the clusters due to incompleteness of environmental data (see Section 3.2). Sorting the clusters by their mean colours means they are also correlated with *sSFR* and $M_*$ in all four outcomes. This is consistent with the expectation that PC1 (with which these features are most strongly associated) dictates much of the clustering. Siudek et al. (2018b) also find a strong correlation of their colour-based galaxy classes (via an unsupervised method) with stellar masses and star formation activity. $n_g$ and $R_{1/2}$ do not correlate as strongly with the clusters (when sorted by colour), particularly at higher $k$, indicating that these features only play a role in dictating clusters as the number of clusters increases. The broad correlations of these features with the clusters at lower values of $k$ are due to their correlations with the PC1 features. I also find a correlation of environmental density with the clusters (when sorted by colour), indicative of the strong role of environment in quenching.

In order to understand cluster structures, I reprise the panels of Figure 3.1. The original histograms are omitted to avoid visual clutter, especially at higher values of $k$. Coloured histograms and coloured contours are used to show the distribution of the clusters in one- and two-dimensional projections. The contours are drawn to enclose 75 percent of the galaxies in each cluster. This level is chosen to strike a balance between generality and accuracy, given that clusters are best characterised by points near their centres. Cluster centroids are plotted as filled circles of the same colour. The classical dividers are retained for comparison with the clusters.

In the remainder of this section I describe each of the outcomes in detail.

### 3.3.1 $k = 2$

The hierarchy tree in Figure 3.4 shows that $k = 2$ forms the basic structure of two superclusters into which the clusters at higher values of $k$ may also be divided, indicating the influence of the bimodality on the clustering. Table 3.5 reveals that the clusters in $k = 2$ represent two distinct populations. $\mathrm{Ra}_2$, which contains fewer galaxies than $\mathrm{Ba}_2$, is made up of galaxies with higher masses, redder colours, higher Sérsic indices, larger radii, and lower specific star formation rates on average. This is consistent with established notions of an overall bimodality of galaxies. Cluster $\mathrm{Ra}_2$ has larger uncertainties on its centroid in all features in comparison with those of $\mathrm{Ba}_2$ because it less dense in feature space.

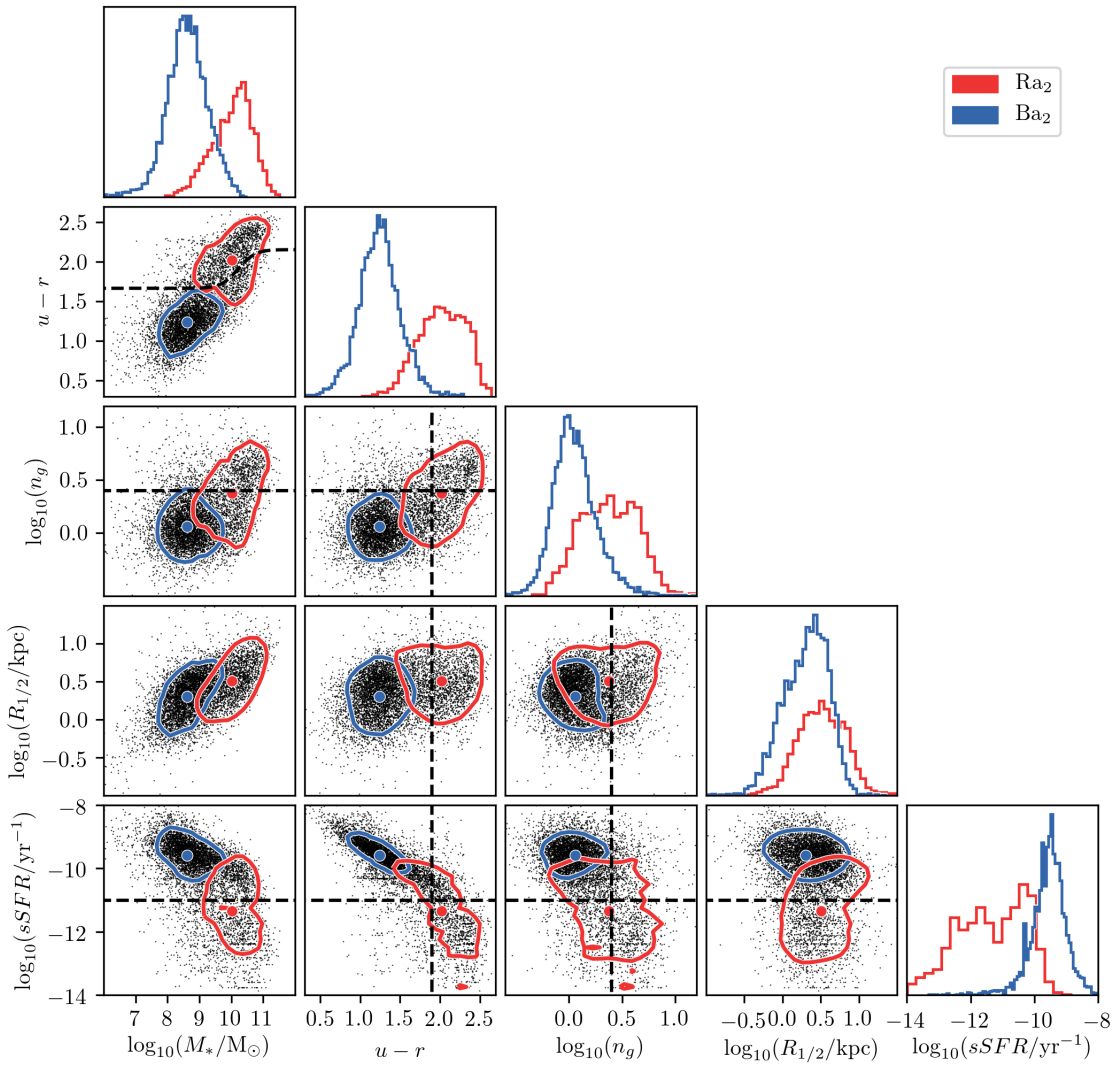


**Figure 3.5:** A profile of $k = 2$. Clusters are represented using coloured histograms and contours, and their centroids are marked using filled circles of the same colour.

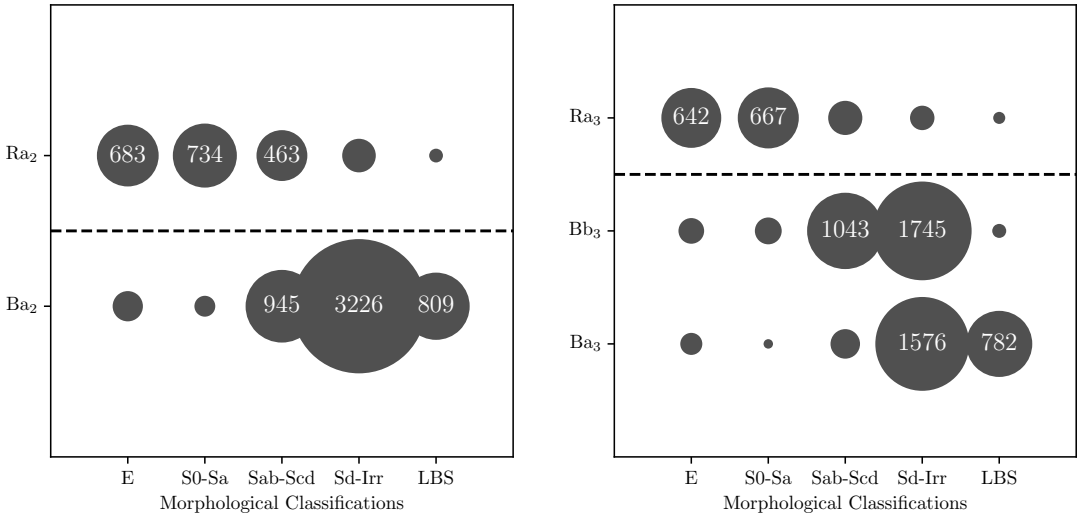


**Figure 3.6:** Bubble plot comparing $k = 2$ (**left**) and $k = 3$ (**right**) with the Kelvin et al. (2014a) and Moffett et al. (2016) morphological classifications. All bubbles containing more than 5 per cent of the galaxies in the sample are labelled with the number of galaxies that they contain. The dashed black line separates the two superclusters that `k-means` finds.

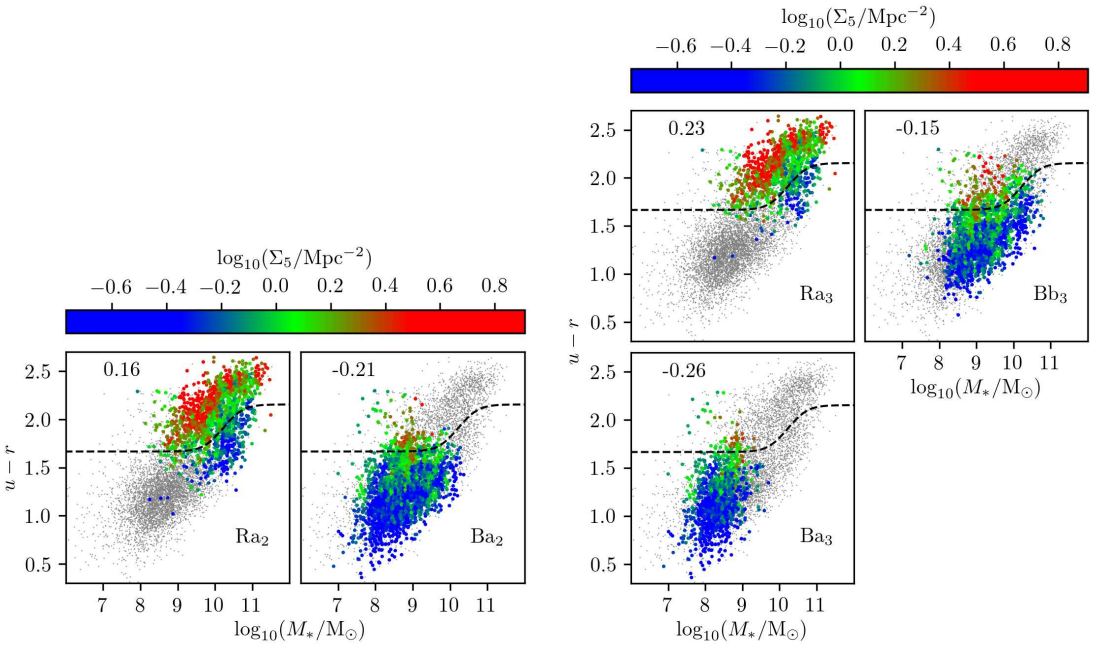


**Figure 3.7:** The $k = 2$ (**left**) and $k = 3$ (**right**) clusters, projected onto the $u - r$ versus $M_*$ plane, with points coloured by their smoothed local environmental densities ($\Sigma_5$). The small grey points represent the remainder of the sample, as well as those galaxies for which $\Sigma_5$ is not available (including those within the clusters highlighted in each panel). Cluster names are shown in the bottom right of each panel. The mean $\Sigma_5$ of each cluster is shown in the top left of each panel. The dashed black line marks the Baldry et al. (2004) distinction between the blue and red sequences of galaxies.

The $k = 2$ cluster projections in Figure 3.5 are best separated in panels involving $u - r$, $sSFR$, and $M_*$. These are the features that are most strongly associated with PC1, which dominates the covariance structure of the sample in feature space and hence dictates much of the clustering. The $k = 2$ cluster projections overlap more in panels involving $R_{1/2}$ and $n_g$, which are more strongly associated with PC2 and PC3 respectively. These features play a lesser role in the clustering in $k = 2$.

Cluster $\mathrm{Ra}_2$ spans the classical dividers (dashed black lines) in all panels of Figure 3.5. While appearing to represent red sequence galaxies (Baldry et al., 2004; Taylor et al., 2015) and passive (negligible $sSFR$) galaxies (Table 3.5), it extends well onto the blue sequence ($u-r$ versus $M_*$) and star-forming main sequence ($sSFR$ versus $M_*$). This is because the cluster boundary that `k-means` draws between its two centroids is a hyperplane, equidistant from them both and perpendicular to the line connecting them. The uniform effect of `k-means`, which produces clusters of similar sizes, essentially bisects the sample through PC1, along which the two centroids are evenly spaced. This gives a coarse partition of the sample. More clusters are needed to properly "resolve" the true structure, including the boundary, of the bimodality of galaxies.

In Figure 3.6, I use bubble plots to visualise agreement between the clusters and the Kelvin et al. (2014a) and Moffett et al. (2016) morphological classifications. The left-hand plot shows a considerable overlap of morphologies between the two clusters, verifying the relative weakness of $n_g$ and $R_{1/2}$ in the clustering in $k = 2$. This effect is also seen in Figure 3.5, which shows that cluster $\mathrm{Ra}_2$ is indiscriminate with respect to Sérsic indices in comparison with the classical divider. The broad correlation of clusters with morphological types (e.g. that earlier-type morphologies are more likely to be found in cluster $\mathrm{Ra}_2$; Figure 3.6) arises as a result of the correlation of morphology with the PC1 features. This effect is apparent in Figure A.7, which shows that the galaxies in $\mathrm{Ra}_2$ have smoother and more concentrated morphologies.

The mean local environmental densities (Table 3.5) of the $k = 2$ clusters reveal that the galaxies in $\mathrm{Ra}_2$ occupy denser environments on average than those in $\mathrm{Ba}_2$. This is mostly a reflection of the basic correlation of galaxy mass, colour, and star formation activity with environment, due to the coarse partition of the sample. A greater fraction of galaxies are lost from $\mathrm{Ba}_2$ than $\mathrm{Ra}_2$, such that this difference is likely to be underestimated. The panels in Figure 3.7 project the cluster onto the $u - r$ versus $M_*$ plane. Points are coloured by their smoothed local environmental densities ($\Sigma_5$; see Section 3.2). $\mathrm{Ba}_2$ consists mostly of galaxies in low-density environments. There is a gradual increase of $\Sigma_5$ with $u-r$ within $\mathrm{Ba}_2$. $\mathrm{Ra}_2$ exhibits more of a spread in $\Sigma_5$, but with a preference for higher-density environments. This suggests that environmental processes are the more common mechanism by which galaxies acquire redder colours.

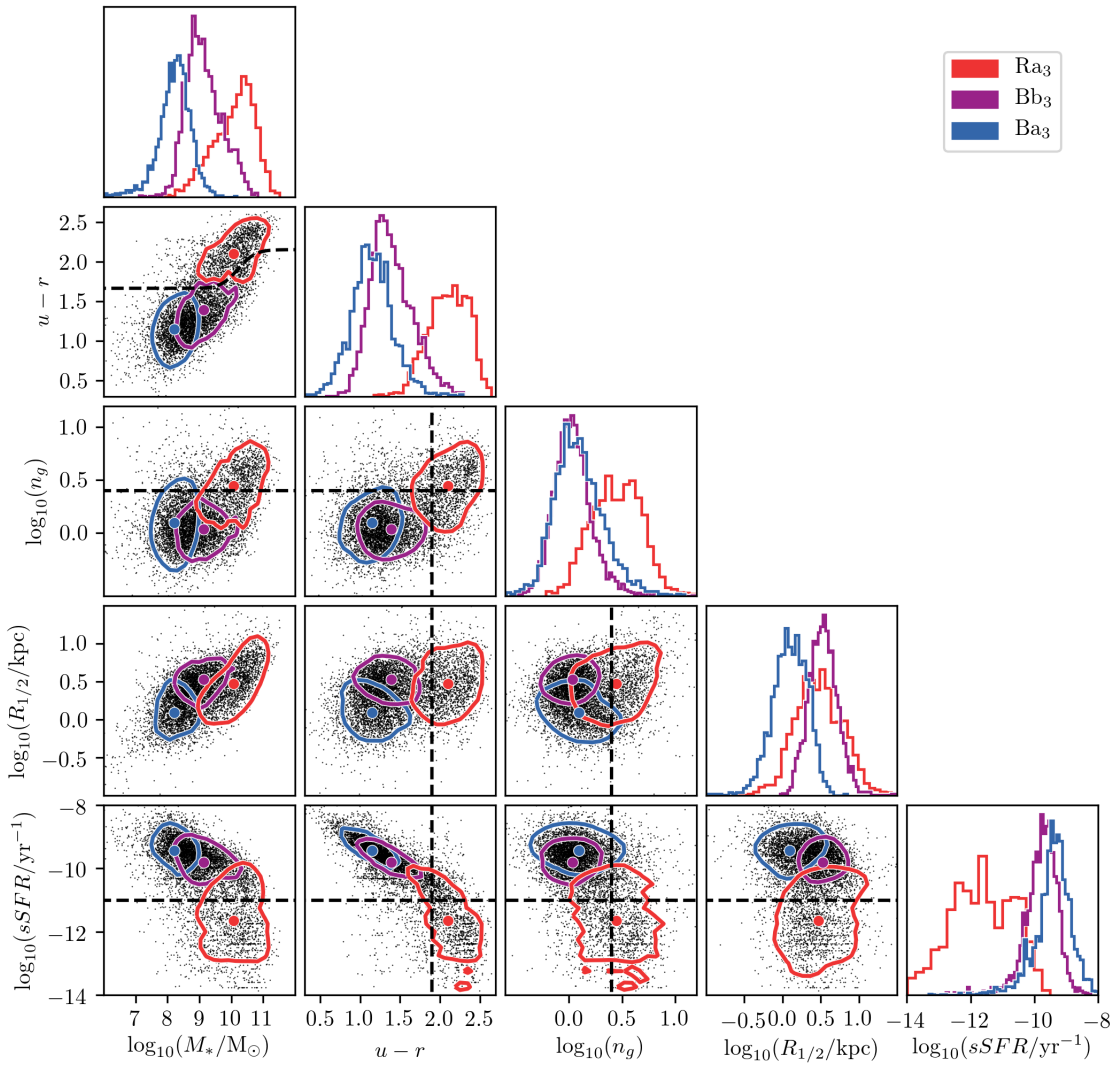


**Figure 3.8:** A profile of $k = 3$. Clusters are represented using coloured histograms and contours, and their centroids are marked using filled circles of the same colour.

### 3.3.2 $k = 3$

The right-hand plot of Figure 3.4 shows that $k = 3$ is hierarchical with respect to $k = 2$; only $\sim 6$ per cent of the galaxies in the sample do not follow a clean hierarchy between the two outcomes. Therefore, much of the clustering structure of $k = 3$ is derived from that of $k = 2$. The red supercluster remains relatively unchanged between the two outcomes. $\mathrm{Ra}_3$ contains $\sim 76$ per cent of galaxies that $\mathrm{Ra}_2$ contains, and Table 3.5 shows that both clusters share similar identities. The placement of $\mathrm{Ra}_3$ with respect to the classical dividers is improved in most panels, particularly those involving the PC1 features ($M_*$, $u-r$, and *sSFR*), indicating the strength with which they still dictate the clustering in $k = 3$. This improvement is enabled by the split of the blue supercluster, which has evened out the cluster sizes. The uncertainties on the centroid of $\mathrm{Ra}_3$ are generally less than or equal to those of $\mathrm{Ba}_3$ and $\mathrm{Bb}_3$ except for in *sSFR*, which is due to the spread in *sSFR* of the passive galaxies in $\mathrm{Ra}_3$ (see Figure 3.8).

The main change in $k = 3$ from $k = 2$ is that `k-means` splits the blue supercluster apart into two clusters: $\mathrm{Ba}_3$ and $\mathrm{Bb}_3$. The split happens due to the larger number of galaxies within the blue supercluster. The main features that distinguish these clusters are $R_{1/2}$ and $M_*$, in which the centroids (Table 3.5) differ by $1.42\sigma$ and $1.03\sigma$ respectively (as opposed to the $< 0.50\sigma$ differences in other features). Here, $\sigma$ is the standard deviation of the sample in a given feature. Figure 3.8 shows that the distributions of the clusters are most distinct in these features as well. $\mathrm{Ba}_3$ and $\mathrm{Bb}_3$ exhibit very similar distributions in $n_g$ and *sSFR*. While differing in mass and size, galaxies in both $\mathrm{Ba}_3$ and $\mathrm{Bb}_3$ are generally star-forming and have diffuse light profiles.

The use of an additional centroid has enabled `k-means` to explore the subtler variance in the sample (and in particular, in the blue supercluster) that PC2 encompasses. The particularly low masses and sizes of the galaxies in $\mathrm{Ba}_3$ suggest a distinction by `k-means` between dwarf galaxies and larger, more massive galaxies in the blue supercluster. Figure 3.6 appears to confirm this, showing that the more evolved spiral galaxies in the sample are more likely to be found in cluster $\mathrm{Bb}_3$. Figure A.8 shows that the galaxies in $\mathrm{Bb}_3$ have more prominent discs. The clusters still exhibit significant overlap in morphologies and cluster $\mathrm{Ra}_3$ is still indiscriminate with respect to $n_g$, indicating a continuing relative weakness of $n_g$ in dictating the clustering at $k = 3$.

The galaxies in $\mathrm{Ra}_3$ occupy denser environments on average than those in either $\mathrm{Ba}_3$ or $\mathrm{Bb}_3$ (see Table 3.5). Figure 3.7 shows that clusters $\mathrm{Ba}_3$ and $\mathrm{Bb}_3$ are similarly distributed in $\Sigma_5$. Both are dominated by low-density environments and exhibit tails towards higher-density environments. For $\mathrm{Bb}_3$, galaxies in low-density environments are found at all masses, suggesting that some galaxies are able to evolve to higher masses without any significant change in their morphology (see Figure 3.6) incurred by, for example, major mergers (Barnes, 1992).

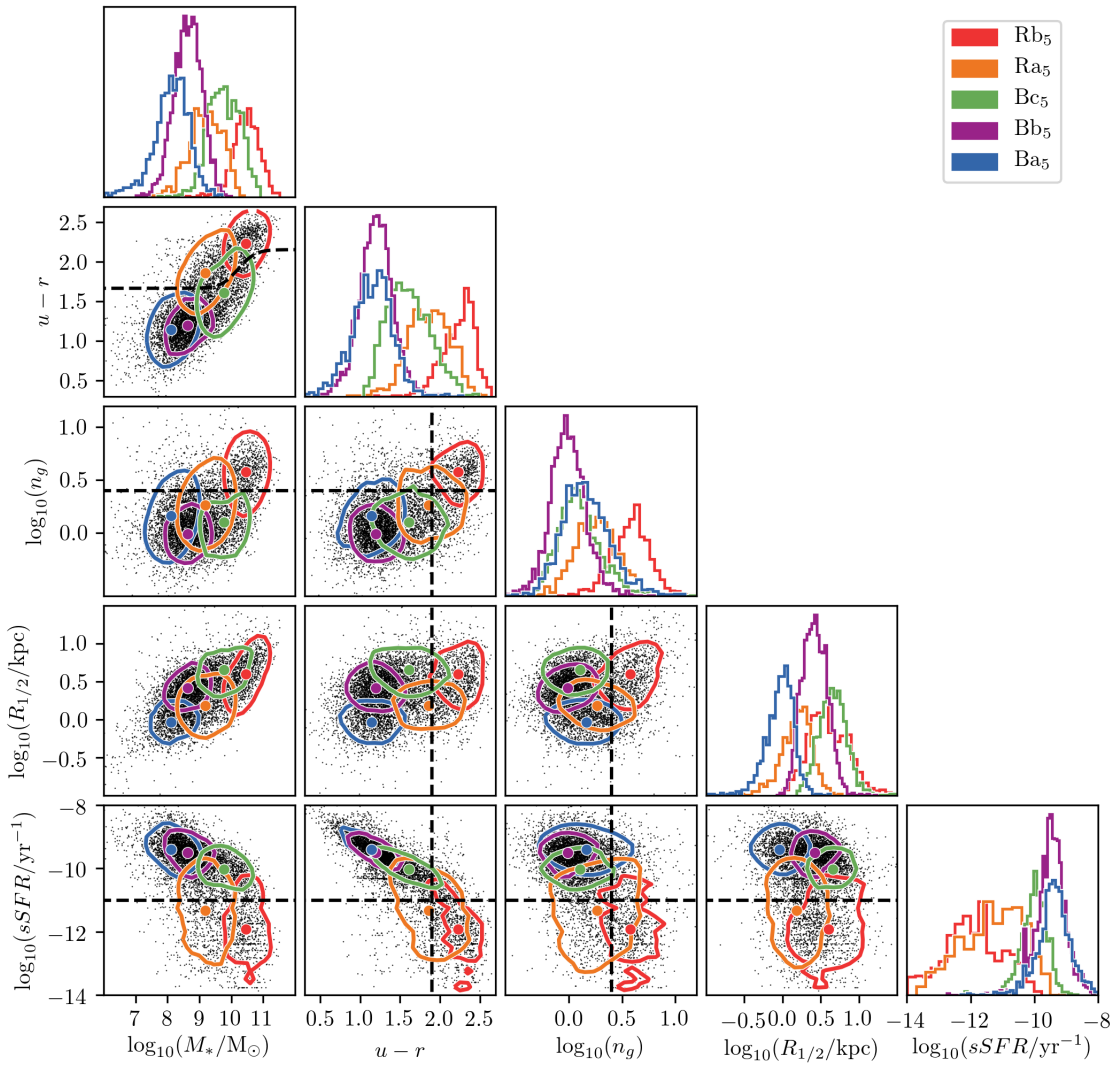


**Figure 3.9:** A profile of $k = 5$. Clusters are represented using coloured histograms and contours, and their centroids are marked using filled circles of the same colour.

### 3.3.3 $k = 5$

$k = 5$ is not as cleanly hierarchical with respect to $k = 3$ as $k = 3$ is with respect to $k = 2$ (Figure 3.4). There is mixing between the red and blue superclusters and within the blue supercluster, which involves $\sim$ 21 per cent of the galaxies in the sample. This indicates that `k-means` has probed an alternative structure of the sample in feature space to that which it finds in $k = 2$ and $k = 3$. My exclusion of an outcome from $k = 4$ due to instability (Figure 3.4) may exaggerate the apparent mixing. The bimodal stability of outcomes at $k = 4$ emerges as `k-means` settles into a split in *either* the red or the blue supercluster, rather than splits in both as in $k = 5$. Both of these splits must be made in order to achieve stability, suggesting genuine astrophysical differences between the galaxies occupying these clusters.

The red supercluster is split into two clusters: $\mathrm{Ra}_5$ and $\mathrm{Rb}_5$. The features that distinguish these clusters are $M_*$, $R_{1/2}$, and $n_g$, with differences of $1.41\sigma$, $1.35\sigma$, and $1.19\sigma$ respectively in their centroids (see Table 3.5). They are also separated in the *sSFR* versus $u - r$ plane. $\mathrm{Rb}_5$ consists mostly of evolved galaxies with the highest masses and reddest colours in the sample. Enabled by the use of additional centroids to probe subtler variances in the sample, `k-means` has also now made a morphological distinction between galaxies such that $\mathrm{Rb}_5$ contains galaxies with the highest Sérsic indices as well. This is also apparent in Figures 3.10 and A.9, which respectively show that $\mathrm{Rb}_5$ is made up mostly of early-type galaxies and of concentrated, smooth, spheroid-dominated galaxies.

$\mathrm{Ra}_5$ has a weaker cluster identity. While $\mathrm{Rb}_5$ mostly contains galaxies from only one cluster in $k = 3$, $\mathrm{Ra}_5$ contains galaxies from three. Its centroid has larger uncertainties in most features than those of the other $k = 5$ clusters (Table 3.5), and it exhibits a large spread in Figure 3.9, spanning the red and blue sequences and including both star-forming and passive galaxies. The galaxies $\mathrm{Ra}_5$ contains have red colours and low specific star formation rates like those in $\mathrm{Rb}_5$, but to a lesser extent, suggesting that they are not as evolved. Figure 3.10 reveals that it also contains a range of morphologies, with even E and Sd-Irr galaxies being grouped together. This suggests a lack of morphological information in my feature selection, despite the role $n_g$ is now playing in dictating cluster $\mathrm{Rb}_5$. I also note an apparent inability of `k-means` to properly distinguish E and S0-Sa galaxies. Nolte et al. (2018) find a similar inseparability of these morphological classes in the same sample using a self-organising map for dimensionality reduction.

The blue supercluster is split again in $k = 5$. The mixing between $k = 3$ and $k = 5$ within the blue supercluster means the clusters in each outcome do not correspond as strongly to one another. The galaxies in $\mathrm{Ba}_5$ and $\mathrm{Bb}_5$ have been distinguished from those in $\mathrm{Bc}_5$ by their masses, in which their centroids both differ from that of $\mathrm{Bc}_5$ by $> 1\sigma$. $\mathrm{Ba}_5$ and $\mathrm{Bb}_5$ are similar in terms of the colours and specific star formation rates of the galaxies they contain. The main distinction between them is in

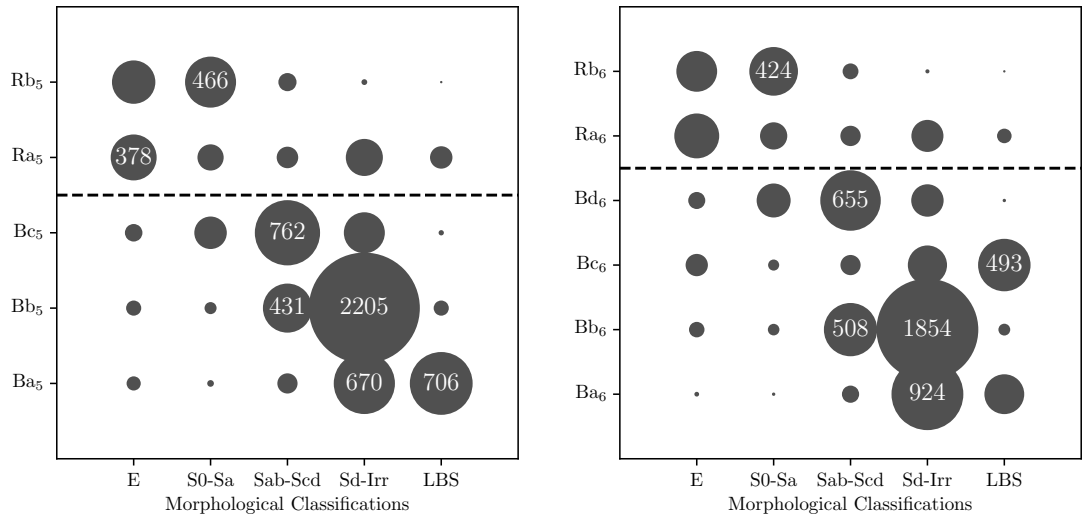


**Figure 3.10:** Bubble plot comparing $k = 5$ (**left**) and $k = 6$ (**right**) with the Kelvin et al. (2014a) and Moffett et al. (2016) morphological classifications. All bubbles containing more than 5 per cent of the galaxies in the sample are labelled with the number of galaxies that they contain. The dashed black line separates the two superclusters that `k-means` finds.

their sizes ($> 1\sigma$). Figure A.9 shows that $Bb_5$ contains more galaxies with extended discs, while $Ba_5$ contains more galaxies that are compact. $Ba_5$ contains the vast majority of LBS galaxies in the sample (Figure 3.10), while $Bb_5$ contains a significant number of more evolved spiral galaxies.

$Bc_5$ differs from the other two clusters in the blue supercluster, containing relatively massive and large galaxies with reduced star formation. The intermediate colours of the galaxies in $Bc_5$, and its location in the colour-mass plane in particular are consistent with previous descriptions of green valley galaxies (Salim et al., 2007; Schawinski et al., 2014). The low Sérsic indices of the galaxies in $Bc_5$ in comparison with those of the galaxies in the red supercluster are due to the presence of prominent discs (see Figures 3.10 and A.9). $Bc_5$ hence contains spiral galaxies at the later stages of their evolution.

Just as the use of additional centroids has enabled `k-means` to make finer distinctions between galaxies in terms of the features are used to describe them, it has also led to a finer view of the role of environment in influencing the clusters. The galaxies in clusters $Ba_5$, $Bb_5$, $Bc_5$ mostly occupy low-density environments (Table 3.5 and Figure 3.11). Meanwhile, clusters $Ra_5$ and $Rb_5$ generally contain galaxies in higher-density environments. Notably, the average local environmental densities of galaxies in clusters $Bc_5$ and $Ra_5$ differ, despite the fact that these clusters are adjacent in all panels of Figure 3.9 and span both the blue and red sequences in the left-hand plot in Figure 3.11.

I suggest that these "green valley" clusters, $Ra_5$ and $Bc_5$, each mostly contain galaxies on different evolutionary pathways (Peng et al., 2010; Schawinski et al., 2014). The evolution of galaxies

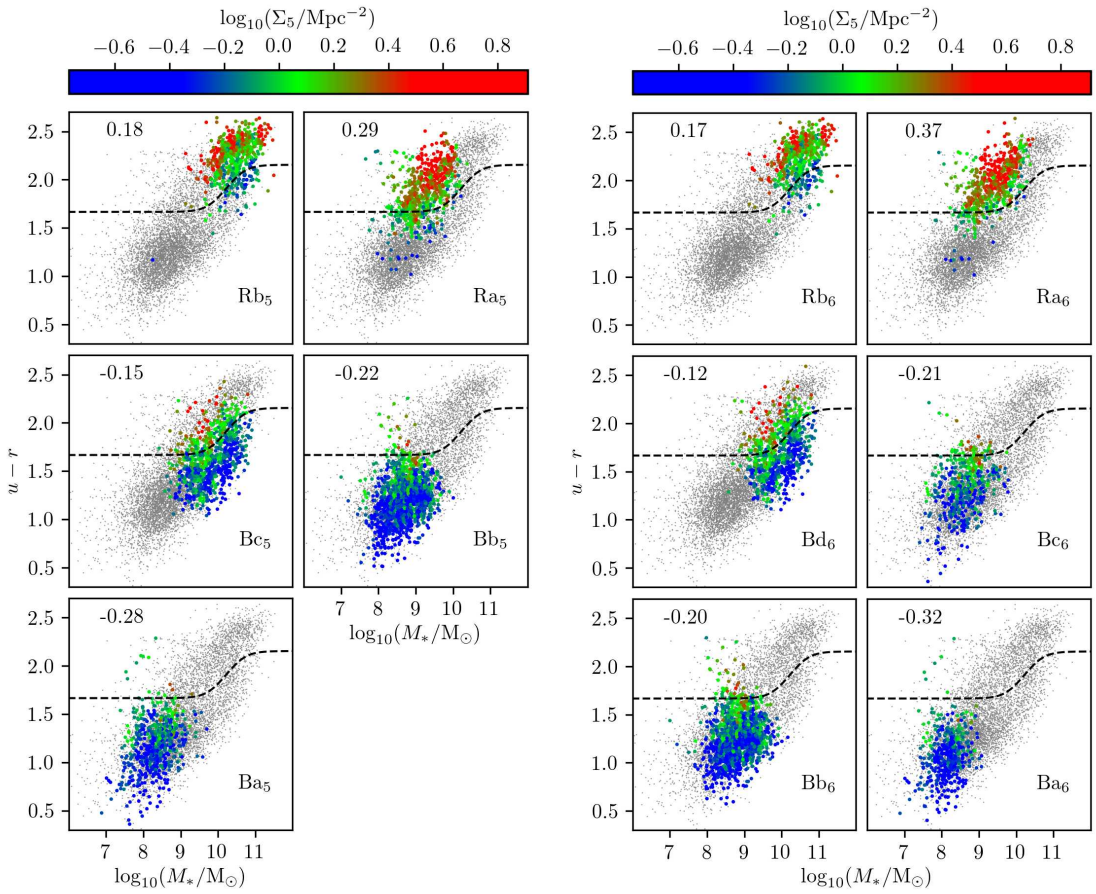


**Figure 3.11:** The $k = 5$ (**left**) and $k = 6$ (**right**) clusters, projected onto the $u - r$ versus $M_*$ plane, with points coloured by their smoothed local environmental densities ($\Sigma_5$). The small grey points represent the remainder of the sample, as well as those galaxies for which $\Sigma_5$ is not available (including those within the clusters highlighted in each panel). Cluster names are shown in the bottom right of each panel. The mean $\Sigma_5$ of each cluster is shown in the top left of each panel. The dashed black line marks the Baldry et al. (2004) distinction between the blue and red sequences of galaxies.

in $\mathrm{Ra}_5$, whose distribution in $\Sigma_5$ is skewed toward higher densities, is dominated by external processes (see Section 1.3.2), which both transform their morphologies and inhibit their star formation on short timescales. Examples of external processes include major (Barnes, 1992) and minor mergers (Toomre & Toomre, 1972), and ram-pressure stripping (Gunn & Gott, 1972). These processes are likely to be responsible for the early-type morphologies (Figures 3.10 and A.9), red colours, and inhibited star formation rates (Figure 3.9) in $\mathrm{Ra}_5$. In addition, these processes have previously been invoked to explain the build-up of the low-mass end of the red sequence by quenching satellites (Wetzel et al., 2012, 2013; Muzzin et al., 2013).

Cluster $\mathrm{Bc}_5$, which is dominated by galaxies in lower-density environments, appears to contain galaxies that are dominated in their evolution by internal processes (see Section 1.3.1). Internal processes include feedback from stars (Geach et al., 2014; Hayward & Hopkins, 2017) and AGN (Croton et al., 2006; Somerville et al., 2008) which drives star-forming gas out of galaxies, and morphological quenching (Martig et al., 2009), in which bulges at the centre of late-type galaxies stabilise their discs against collapse and thereby prevent further star formation. The high masses (Figure 3.9) and prominent bulges (Figure A.9) of the galaxies in $\mathrm{Bc}_5$ seem to confirm the dominance of these internal processes, particularly of AGN feedback, in their evolution.

The difference in the morphologies of the galaxies in $\mathrm{Ra}_5$ and $\mathrm{Bc}_5$ is consistent with Schawinski et al. (2014), who find a morphological dichotomy of galaxies in the green valley. The spread in morphologies in $\mathrm{Ra}_5$ may arise due to both its large spread in $\Sigma_5$, and the short timescales of morphological transformations. The dominance of galaxies with high $\Sigma_5$ in $\mathrm{Rb}_5$ suggests a preference of external processes for moving galaxies onto the red sequence over time. The additional presence of galaxies in low-density environments in $\mathrm{Rb}_5$ suggests that galaxies evolving mostly via internal processes will also converge on the red sequence at high masses, though.

### 3.3.4 $k = 6$

$k = 6$ is once again more strongly hierarchical with respect to $k = 5$ than $k = 5$ is with respect to $k = 3$; $\sim 15$ per cent of galaxies mix between $k = 5$ and $k = 6$. Clusters are more readily comparable with those in $k = 5$, from which most of their structure is derived. The clusters in the red supercluster retain their identities in terms of their centroids (Table 3.5), distributions (Figure 3.12), and morphologies (Figure 3.10) between $k = 5$ and $k = 6$ and remain mostly unchanged. Similarly, $\mathrm{Bd}_6$ matches well with $\mathrm{Bc}_5$, as does $\mathrm{Bb}_6$ with $\mathrm{Bb}_5$.

The main changes between $k = 5$ and $k = 6$ are at the blue end of the blue supercluster. $\mathrm{Ba}_6$ and $\mathrm{Bc}_6$ both have low masses and small radii. They differ most significantly in $n_g$ ($1.34\sigma$), appearing to indicate a morphological distinction between disc-dominated galaxies in $\mathrm{Ba}_6$ and spheroid-dominated galaxies in $\mathrm{Bc}_6$. While Figure 3.10 reveals a significant degeneracy of morphological

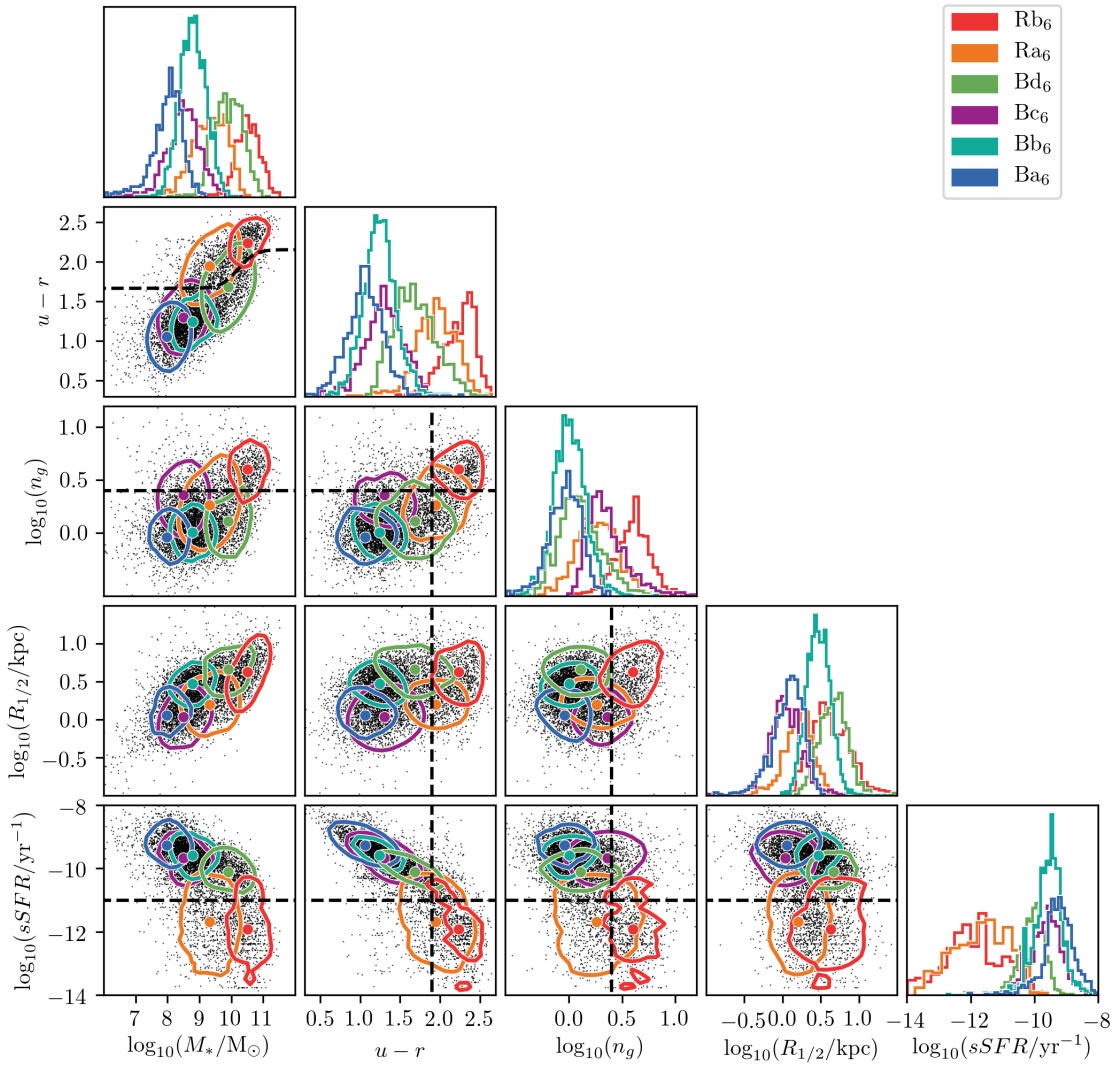


**Figure 3.12:** A profile of $k = 6$. Clusters are represented using coloured histograms and contours, and their centroids are marked using filled circles of the same colour.

classifications in these clusters, the distinction is more apparent in Figure A.10. The classification degeneracy may arise partially due to the difficulty incurred in visually classifying intrinsically faint objects. $Bc_6$ contains some potentially spurious spheroid-dominated E galaxies, which could explain some of the difference in $n_g$ between $Ba_6$ and $Bc_6$. Further morphological information in my feature selection may lead to a clearer distinction between galaxies at the blue end of the blue supercluster.

Clusters $Ba_6$, $Bb_6$, and $Bd_6$, which all contain discy galaxies (Figures 3.10 and A.10) in mostly low-density environments (Figure 3.11), appear to form a continuum of galaxy evolution along the blue sequence. This continuum appears to be dictated by internal processes, given the increase in mass and bulge prominence along the sequence of consistently low $\Sigma_5$. The environments of the galaxies in $Bc_6$ are also low-density, though they have an early-type morphology, suggesting that they may have formed differently. There is a significant tail of this cluster toward intermediate densities, suggesting that they may be in the early stages of morphological transformation due to environmental effects. Their origin and fate is unclear (Schawinski et al., 2009a).

The agreement of the clusters with the classical dividers has improved considerably as $k$ has increased. At $k = 6$, the clusters align particularly well with established notions of a fundamental bimodality of galaxies, which is expressed using the dashed black lines in Figure 3.12. Cluster $Ra_6$ still spans the dividers in some panels, though this may be due to the rapid timescales of the morphological transformations that the galaxies it contains are likely undergoing, such that they exhibit a larger spread in the features.

## 3.4 Summary and conclusions

I report the results of a test of the $k$-means clustering method as a galaxy classification solution for the unprecedentedly large surveys of the future and as a tool for exploring feature spaces of high dimensionalities. It is tested on a redshift- and magnitude-limited pilot sample of $7,338$ galaxies from the GAMA survey, which is represented using a preliminary selection of five features: stellar mass, $u - r$ colour, Sérsic index, half-light radius, and specific star formation rate. Analyses of correlations and covariances between features reveal that stellar mass, $u - r$ colour, and specific star formation rate dominate much of the structure of the sample in the feature space and hence dictate much of the clustering. I rescale, truncate, and normalise the sample ahead of clustering to mitigate a) the influence of outliers on results, and b) bias toward any of the features based on skewed distributions or large numerical ranges.

I apply the $k$-means method, which partitions data into $k$ clusters, in the context of a unique cluster evaluation approach that enables the robust identification of stable clustering structure in spite of

stochastic effects, including the random initialisations of `k-means` and bootstrap resampling. My clustering approach is highly scalable, taking just $\sim$ 3 minutes per values of $k$ for the present samples using a single core on a laptop computer. I find that the local galaxy population is stably divisible into 2, 3, 5, and a maximum of 6 clusters. I select optimal clustering outcomes from each value of $k$ for analysis, and reach the following conclusions:

1. Clusters in all four of the best outcomes agree with established notions of the bimodality of galaxies. Agreement improves as $k$ increases. The use of additional centroids to model the data structure of the sample in feature space enables a more detailed view of the bimodality via `k-means`. At higher values of $k$, there are distinct clusters that appear to follow different evolutionary pathways through the green valley. While $M_*$, $u - r$, and $sSFR$ dictate most of the clustering structure in all four of the best outcomes, $n_g$ and $R_{1/2}$ play an increasingly strong role at higher $k$ as `k-means` uses the additional centroids to explore subtler substructures in the sample.

2. Though I do not aim to reproduce any existing classification schemes with the clusters, there is a general agreement of the clusters with the Kelvin et al. (2014a) and Moffett et al. (2016) Hubble-like morphological classifications of the galaxies in the sample. At low $k$, this agreement is mostly due to the correlation of morphology with stellar mass, $u-r$ colour, and specific star formation rate, which dictate the majority of the clustering. This suggests a relative lack of morphological information among my feature selection. At higher $k$, though, it is found that `k-means` is able to explore subtler substructures in the sample and make genuine morphological distinctions between galaxies using Sérsic index. The addition of further morphological features to my selection is anticipated to improve these distinctions further.

3. Analysis of the local environmental densities of the galaxies in the clusters in outcomes $k = 5$ and $k = 6$ especially suggests the differential roles of internal and external processes in galaxy evolution. Those clusters containing more galaxies in high-density environments also contain more galaxies with early-type morphologies, with their spreads in morphologies suggesting rapid morphological transformation and their reduced $sSFRs$ indicating quenching. Clusters containing galaxies in low-density environments are found along the whole blue sequence, such that some galaxies are able to evolve to the highest masses while retaining a disc-dominated morphology. Clusters in the blue sequence appear to form an evolutionary continuum, whose galaxies are dominated in their growth by internal processes. There is an apparent preference for externally driven evolution of low-mass ($M_* < 10^{10}$ M$_\odot$) galaxies onto the red sequence. An availability of this environmental data for the entirety of the sample would significantly strengthen these (or alternative) conclusions regarding the role of environment in the evolution of galaxies.

I endorse $k = 6$ as being the most useful outcome for its ability to both capture the broad bimodal structure of the galaxy population in feature space and identify finer distinctions *within* this bimodality that highlight the differential role of environment in the evolution of galaxies.

While the concurrent use of too many features may lead to undesirable effects such as overfitting and redundancy, it is clear that my feature selection may be improved by the addition of further information. The inclusion of morphological features like asymmetry or distance-independent smoothness (Conselice, 2003), the Gini coefficient (Abraham et al., 2003), or those derived from two-component fits might yield stronger cluster identities and further disentangle the roles of internal and external evolutionary processes, particularly with respect to clusters $\mathrm{Ra}_5$ and $\mathrm{Ra}_6$ which both contain a spread of morphologies. Wijesinghe et al. (2010) show that morphological information from multiple bands of photometry also incorporates information about the distributions stellar populations within galaxies and how they have evolved. The inclusion of spectroscopic features would also improve clustering results and cluster interpretation. In particular, emission-line diagnostics (Baldwin et al., 1981; Lamareille, 2010) could highlight the role of AGN in galaxy evolution, and the strength of the 4000 Å break (Poggianti & Barbaro, 1997) could include galaxy stellar ages in cluster identities. In general, it appears that an optimal feature selection will consist of a combination of features derived from both photometry and spectroscopy.

My feature selection for the work in this chapter is preliminary, and based mostly on astrophysical domain knowledge. While some simple statistical consideration is applied (Spearman rank-order correlations, and a principal component analysis), a variety of other methods are also available (see e.g. chapter 2 of Aggarwal 2014) for feature selection and feature extraction (i.e. the manufacture of features) which may further improve clustering results.

The pilot sample is well-characterised by a number of previous studies, facilitating the interpretation of the clusters that are found, but it is small and limited to low redshifts. While the sample suffices for an initial test of `k-means` and my cluster evaluation approach, a more thorough test would be to apply the framework to a larger sample of galaxies, constituting a more complete representation of the diversity of galaxies in the local Universe. SDSS would be particularly suitable given the wealth of ancillary features available, and especially given, for example, its overlap with the Galaxy Zoo project which would enable more detailed study of morphologies within clusters. Furthermore, clustering a sample of galaxies spanning a greater range of redshifts, or a comparison of clusters in different redshift bins, invites the examination of the evolution of clusters themselves over cosmic time.

I conclude by emphasising that my cluster evaluation approach is malleable. It may readily be adapted for use with any algorithm and any sample to identify stable clustering structure. I test stability against random initialisations and application of the bootstrap method to the sample, but the approach may also be applied in the context of other Monte Carlo methods.

# Chapter 4

# Testing a cosmological galaxy simulation with unsupervised machine learning

The work presented in this chapter is the subject of a paper that is in preparation for submission to Monthly Notices of the Royal Astronomical Society. See also the Publications page of the front matter of this thesis.

In this chapter, I detail the results of a comparison of a sample of simulated galaxies from the Evolution and Assembly of GaLaxies and their Environments (EAGLE) project (Schaye et al., 2015; Crain et al., 2015) with a sample of observed galaxies from the GAMA survey (Driver et al., 2009). Galaxies in both samples are characterised using five features that are relevant to all aspects of galaxy evolution: stellar mass, specific dust mass, specific star formation rate, size, and bulge-to-total ratio. Thus, I aim to probe the structures and distributions of the samples within this shared five-dimensional feature space, and to offer astrophysical explanations for the similarities and differences between them. Identifying and quantifying these similarities and differences constitutes a robust, multi-dimensional validation of the EAGLE project. I apply the same clustering approach as in Chapter 3 – it is described in full in Section 3.1).

The remainder of this chapter proceeds as follows. In Section, 4.1 I outline the collection and preparation of the data for both of the samples, taking measures to ensure a fair comparison. In Section 4.2, I present results and consider their implications. Finally, in Section 4.3, I summarise this chapter and offer conclusions. Where required in this chapter, I assume the ($H_0$, $\Omega_m$, $\Omega_\Lambda$) = (67.77 km s$^{-1}$ Mpc$^{-1}$, 0.307, 0.693) cosmology of the Planck Collaboration et al. (2014), as implemented within the EAGLE simulations.

## 4.1 Samples

In this section, I describe the collection and preparation of the data for my two samples. I take the observed sample as the reference sample, and intend the simulated sample to match it, though in practice, the samples influence the construction of one another. I represent both samples with the same (or as close to being the same as is practical) feature set: stellar mass, specific (fractional) dust mass, specific star formation rate, size, and bulge-to-total ratio.

### 4.1.1 The GAMA survey

I construct the observational sample of galaxies using data from the third data release (DR3) of the GAMA survey (Driver et al., 2009, 2011; Liske et al., 2015; Baldry et al., 2018). The survey as a whole is introduced in Section 3.2. I gather feature data for the GAMA survey sample from two DR3 data tables: `MAGPHYSv06` and `BDDecompv02`. `MAGPHYSv06` contains outputs from the application of the spectral energy distribution (SED) fitting code `MAGPHYS` (Da Cunha et al., 2008) to the self-consistent extinction-corrected matched-aperture photometry listed in `LAMBDARCatv01` (Driver et al., 2018). `LAMBDARCatv01` was generated by the application of the photometric code `LAMBDAR` (Wright et al., 2016) to the reprocessed imaging collated for the Panchromatic Data Release. `MAGPHYS` uses template spectra that model the emission of optical light by stars (Bruzual & Charlot, 2003) and far-infrared light by dust (i.e. attenuated and then reemitted; Charlot & Fall 2000) in a given galaxy to estimate its SED based on fits of those spectra to input photometry. From `MAGPHYSv06`, I take the median stellar masses ($M_*$), dust masses ($M_d$), and Gigayear-timescale specific star formation rates ($sSFR$) given by the probability density functions of fits to the `LAMBDARCatv01` photometry. The dust masses are converted to specific (fractional) dust masses ($sM_d$) by dividing by the stellar masses in order to eliminate any linear dependence on stellar mass.

`BDDecompv02` lists outputs from the application of the image analysis codes `ProFit` (Robotham et al., 2017) and `ProFound` (Robotham et al., 2018) to Kilo-Degree Survey (KiDS; de Jong et al. 2015, 2017) $r$-band images of $z < 0.08$ galaxies in the three equatorial regions of the GAMA survey. `ProFound`, whose functions include sky subtraction and source extraction, was used to reduce the KiDS images and provide initial estimates of the final `BDDecompv02` outputs. `ProFit` was then used to optimise these estimates, using both gradient descent and Markov Chain Monte Carlo samplers in the context of various fits. I take half-light radii ($R_{1/2,l}$) derived from single-Sérsic fits, and bulge-to-total light ratios ($B/T_l$) derived from two-component fits consisting of a Sérsic bulge and an exponential disc. The half-light radii are converted from arcseconds to kiloparsecs using flow-corrected redshifts (`DistanceFramesv12`; Baldry et al. 2012). The $z <$ 0.08 restriction set by the inclusion of feature data from `BDDecompv02` matches well with the

**Table 4.1:** The bounds that are imposed on the feature data for the inclusion of galaxies in both samples. They are intended to exclude outliers and long marginal tails from the sample distributions, and thereby mitigate their influence on the clustering outcomes.

| Feature | Units | Lower | Upper |
|---|---|---|---|
| $M_*$ | $\log_{10}(\mathrm{M}_\odot)$ | 9.5 | 12.0 |
| $sM_d$ | $\log_{10}(M_*/M_d)$ | −6 | −1 |
| $sSFR$ | $\log_{10}(\mathrm{yr}^{-1})$ | −13.00 | −8.75 |
| $R_{1/2,l}$ or $R_{1/2,m}$ | $\log_{10}(\mathrm{kpc})$ | −0.3 | 1.6 |
| $B/T_l$ or $B/T_m$ | - | 0 | 1 |

$z = 0$ restriction that is explicitly set for the EAGLE sample.

Table 4.1 lists bounds that are imposed upon the feature data for the inclusion of galaxies in both samples. The intent with these of bounds is to exclude outliers and long marginal tails from the feature distributions, and thereby mitigate their influence upon the clustering outcomes. This is more important for the GAMA sample, whose feature data is subject to the influence of observational noise. Hence, my choices of bounds are guided mostly by the GAMA sample. My choice of lower stellar mass bound, which is instead guided by the EAGLE sample, is addressed in Section 4.1.2. Imposing these bounds leads to a final GAMA sample of 3, 724 galaxies.

Differences in numerical ranges spanned by the feature distributions *within* these bounds would influence the clustering outcomes (see Section 3.1). To mediate these differences, I rescale the feature data such that each distribution spans the range 0 to 1. This rescaling is based on the minimum and maximum values of the distributions themselves, rather than the bounds that are imposed in Table 4.1 (some feature distribution do not extend all of the way to these bounds). Hence, it equalises the "volumes" of both samples within their feature spaces while also mitigating any bias of `k-means` towards any of the features based purely on a larger numerical range.

The environments of the galaxies in the GAMA sample are characterised using surface densities ($\Sigma_5$), which are listed in `EnvironmentMeasuresv05` (Brough et al., 2013). $\Sigma_5$ is determined with respect only to galaxies among what is called "the density defining population" (absolute $r$-band magnitudes $< -20$, with an additional redshift-dependent correction), and with respect only to galaxies within $1,000$ km s$^{-1}$ of the galaxy under consideration along the line of sight. It is calculated using the projected comoving distance to the galaxy's fifth-nearest neighbour. For those galaxies with fewer than five neighbours within their distance from the edge of the survey footprint ($d_f$), it is calculated using the distance to the farthest available neighbour. For those galaxies with zero neighbours within $d_f$, it is simply calculated as $1/(\pi d_f^2)$. In both of these latter two special cases, the calculated surface density should be taken as an upper limit.

3, 715 of the 3, 724 galaxies in the GAMA sample are listed in `EnvironmentMeasuresv05`. Five

of these galaxies lie within incomplete regions of the GAMA survey footprint; $\Sigma_5$ was not measured for these galaxies so they are omitted from the environmental analysis. $1,037$ of these galaxies have zero neighbours within their distance from the edge of the survey footprint, so their surface densities are calculated as outlined above and are taken as upper limits. For the remaining $2,673$ of these galaxies, $\Sigma_5$ is calculated exactly using their fifth-nearest neighbours.

### 4.1.2 The EAGLE simulations

The Evolution and Assembly of GaLaxies and their Environments (EAGLE) simulations (Schaye et al., 2015; Crain et al., 2015) are hydrodynamical, cosmological models of the formation and evolution of galaxies in a flat ΛCDM Universe based on fundametal parameters measured by the Planck Collaboration et al. (2014): $\Omega_\Lambda = 0.693$, $\Omega_m = 0.307$, $\Omega_b = 0.048$, and $H_0 = 67.77$ km s$^{-1}$ Mpc$^{-1}$. The outputs of the EAGLE simulations have been publicly released (McAlpine et al., 2016) and are available at `http://virgodb.dur.ac.uk/`. The simulations were run using the *N*-body Tree-PM smoothed particle hydrodynamics code `GADGET3`, an earlier version of which was described by Springel (2005). The updated version of the code that was used to run the EAGLE simulations featured changes to its hydrodynamics and time-stepping (Schaller et al., 2015), and its subgrid prescriptions, which model the effects of astrophysical processes (e.g. star formation, stellar feedback, active galactic nucleus feedback) that operate below the resolution limits of the simulations. Schaye et al. (2015) described these subgrid prescriptions in full in section 4 of their paper. The specific details of the subgrid prescriptions, and the resolution limits below which they were implemented, varied among the EAGLE simulations.

I construct the simulated sample of galaxies using the $z = 0$ outputs of the `RefL0100N1504` simulation. This was one of several fiducial EAGLE simulations (denoted by the prefix "`Ref`") which were run in various volumes using exactly the same subgrid prescriptions and at a standard resolution. The standard resolution EAGLE simulations were designed to just about resolve the Jeans mass and length in the warm component ($T \approx 10^4$ K) of the interstellar medium (ISM). `RefL0100N1504` had a cubic volume of side length 100 comoving Mpc, the largest volume modelled by the EAGLE simulations. It contained $1504^3$ dark matter particles, each with mass $9.70 \times 10^6$ M$_\odot$, and initialised with the same number of baryonic particles, each with initial mass $1.81 \times 10^6$ M$_\odot$. In addition, the Plummer-equivalent gravitational softening length was set to 2.66 comoving kpc until $z = 2.8$, and to 0.7 proper kpc (pkpc) after that.

I gather feature data for the EAGLE sample from several of the publicly released data tables. Stellar masses ($M_*$) are taken from `Aperture`. I opt for the stellar masses within 30 pkpc of the centres of galaxies due to their agreement with observed galaxy stellar mass functions (GSMFs) (Schaye et al., 2015), including the GAMA survey GSMF (Baldry et al., 2012). Spherical half-

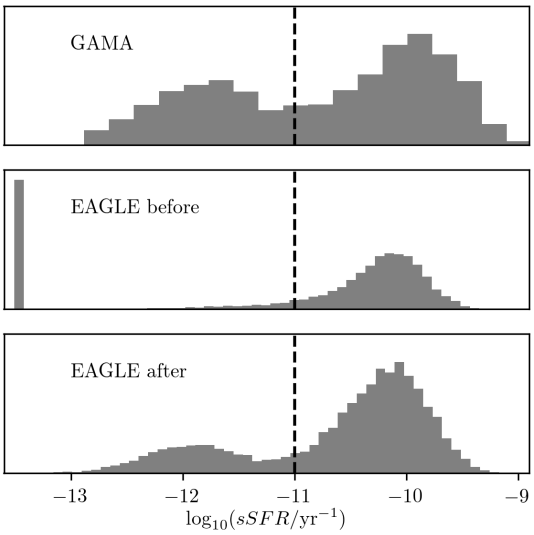


**Figure 4.1:** The effect of the addition of observational noise to measurements of *sSFR* for galaxies in the EAGLE sample. The upper panel shows the distribution of *sSFR* for observed galaxies in the GAMA sample (*with* Table 4.1 bounds applied). The middle panel shows the distribution of original EAGLE measurements of *sSFR*. Simulated galaxies with $sSFR = 0$ yr$^{-1}$ are plotted at $\log_{10}(sSFR/\mathrm{yr}^{-1}) = -13.5$. The lower panel shows the distribution of *sSFR* for simulated galaxies in the EAGLE sample following the addition of observational noise to the data (see main text). Galaxies with $sSFR = 0$ yr$^{-1}$ are set to having $\log_{10}(sSFR/\mathrm{yr}^{-1}) = -12$ before the addition of this noise. The addition of this noise makes the distribution of *sSFR* for the EAGLE sample much more readily comparable with that of the GAMA sample.

mass radii ($R_{1/2,m}$; Furlong et al. 2017), based on these 30 pkpc stellar masses, are taken from `Sizes`. Dust masses ($M_d$), derived from restframe submillimetre fluxes calculated with the radiative transfer code `SKIRT` (Baes et al., 2011; Camps & Baes, 2015), come from `DustFit` (Camps et al., 2018). Dust masses are divided by stellar masses to convert them to specific dust masses ($sM_d$). Bulge-to-total mass ratios ($B/T_m$) are given by inverting the disc-to-total mass ratios listed in `MorphoKinem` (Thob et al., 2019). These ratios are defined kinematically, by doubling the mass of counter-rotating stars in a given galaxy under the assumption that its bulge does not exhibit any net rotation (such that it is purely dispersion-supported; e.g. Sheth et al. 2003). A small fraction of galaxies ($\sim 1$ per cent) have $B/T_m > 1$; these galaxies are omitted from the final EAGLE sample. Finally, specific star formation rates (*sSFR*) are obtained from `Stars` by summing the initial masses of all star particles formed within the central 30 pkpc of galaxies within the last $10^9$ years, and then dividing those star formation rates by the aforementioned stellar masses.

The aforementioned resolution limits of the EAGLE simulations lead some galaxies in the EAGLE sample to have $sM_d = 0$ and/or $sSFR = 0$ yr$^{-1}$ (i.e. less than the minimum $sM_d$ or *sSFR* that the EAGLE simulations can resolve)[1]. Galaxies with $sSFR = 0$ yr$^{-1}$ are plotted at $\log_{10}(sSFR/\mathrm{yr}^{-1}) = -13.5$ in the middle panel of Figure 4.1. *Observational* measurements of

[1]Other features are unaffected due to the bounds that are later imposed on the samples (Table 4.1).

low $sM_d$ and low $sSFR$ are instead dominated by noise and exhibit a very different distribution, as shown for galaxies from the GAMA sample (*with* Table 4.1 bounds applied) in the upper panel of Figure 4.1. This difference in the distributions of galaxies at low $sSFR$ (and low $sM_d$; not shown), which arises simply due to how these features are measured from observations and simulations, complicates a comparison between the samples because it directly influences the clustering outcomes. Given that I wish to facilitate a fair comparison between the samples, this difference must be eliminated.

To do so, I add observational noise to EAGLE measurements of $sM_d$ and $sSFR$. Firstly, median uncertainties are determined in each of these features from the GAMA sample. Separate medians are determined at low and high values, either side of $\log_{10}(sM_d) = -3.5$ and $\log_{10}(sSFR/\mathrm{yr}^{-1}) = -11$ respectively. Separate medians are used as a simple, heuristic way of capturing the variation of the uncertainties with the features themselves. The low/high value cuts are chosen as the minima of the observed distributions of these features (for $sSFR$, see the dashed line in the upper panel of Figure 4.1). Gaussian scatter is then added to the measurements of $sM_d$ and $sSFR$ for galaxies in the EAGLE sample, with standard deviations given by the observational median uncertainties and depending on the values of the original simulation measurements relative to the aforementioned cuts. Galaxies with $sM_d = 0$ and $sSFR = 0$ yr$^{-1}$ are set to having $\log_{10}(sM_d) = -4$ and $\log_{10}(sSFR/\mathrm{yr}^{-1}) = -12$ (i.e. the low-value peaks of the observed feature distributions) respectively before the addition of this Gaussian scatter. For EAGLE $sSFR$, this process leads to the distribution seen in the lower panel of Figure 4.1, which is much more readily comparable with the observed distribution. The distribution of $sM_d$ (not shown) is similarly improved by the addition of observational noise.

The bounds that are imposed for the inclusion of galaxies in the EAGLE sample are listed in Table 4.1. As mentioned in Section 4.1.1, these bounds apply mostly to galaxies in the GAMA sample. The lower stellar mass bound of $10^{9.5}$ $\mathrm{M}_\odot$ excludes simulated galaxies with unreliable specific star formation rates and bulge-to-total ratios due to poor sampling of gas and star particles respectively (figure 11b of Schaye et al. 2015; section 2.2 of Thob et al. 2019). I impose the same lower stellar mass bound to the GAMA sample in order to ensure a fair comparison, especially as the EAGLE simulations were calibrated to reproduce observed stellar mass distributions (including that of the GAMA survey; Baldry et al. 2012; Schaye et al. 2015). As with the GAMA sample, I rescale the feature data for the EAGLE sample ahead of clustering such that each feature distribution spans the range 0 to 1. This is both to mitigate the bias of `k-means` towards features distributed over larger numerical ranges, and to equalise the volumes of both samples within their feature spaces. The final EAGLE sample consists of 7,117 galaxies.

Galaxies in the EAGLE simulations are hosted within structures called haloes. Haloes are analogous to groups of galaxies that are close to one another in the real Universe. The shapes and sizes

of haloes are defined in terms of the distances between dark matter particles only, but haloes also comprise all of the other particle types of the EAGLE simulations as well. Gas, star, and black hole particles are assigned to the same halo as their nearest dark matter particle. Galaxies are defined as self-bound substructures within haloes, and haloes may contain more than one galaxy. The galaxy containing the particle with the smallest gravitational potential energy in a given halo is defined as that halo's central galaxy. All other galaxies in that same halo are satellites. Halo assignments, halo masses, and central/satellite designations for the final EAGLE sample are taken from the public data tables tables `FoF` and `Subhalo`.

I acquire full merger trees (Lemson & Springel, 2006; Qu et al., 2017) for all of the galaxies in the EAGLE sample from the public database. Merger events in the EAGLE simulations are defined in terms of the exchange of particles between galaxies, between snapshots. Galaxy Y in snapshot 2 is the descendant of galaxy X in snapshot 1 if it, among all other galaxies in snapshot 2, contains the greatest fraction of particles that galaxy X contained in snapshot 1. Hence, every galaxy at every snapshot has one descendant in the next snapshot, but may have more than one progenitor in the previous snapshot; this, over several successive snapshots, yields the hierarchical branching structure of a galaxy's merger tree. The *main* progenitor of a galaxy at a given snapshot is the galaxy at the previous snapshot lying on the branch (i.e. across all snapshots) with the highest mass. The main progenitor of a galaxy may be considered as an "earlier version" of that same galaxy, such that the main branch traces its evolution with time.

Finally, I also gather initial masses and formation times of star particles from `Stars` so that the SFHs of EAGLE galaxies may be examined. I opt to consider only in-situ star formation (star particles formed in progenitors on the main branches of their merger trees), as opposed to ex-situ star formation (star particles formed in progenitors on other branches, which then went on to merge with main-branch progenitors).

## 4.2 Results and discussion

### 4.2.1 Identifying the best clustering outcomes

`k-means` is susceptible to yielding different, locally optimal clustering outcomes from varying, random initialisations. To ensure a full exploration of the feature spaces of each of the samples, I initialise `k-means` clustering 100 times each at $k = 2$ through to $k = 20$. I first determine the best value of $k$ for each of the samples using stability (measured with $V_s$). Specifically, I calculate the median $V_s$ for each of the 100 outcomes at a given value of $k$ with respect to the other 99 outcomes at the same value of $k$. Stabilities are illustrated using stability maps (Figure 4.2).

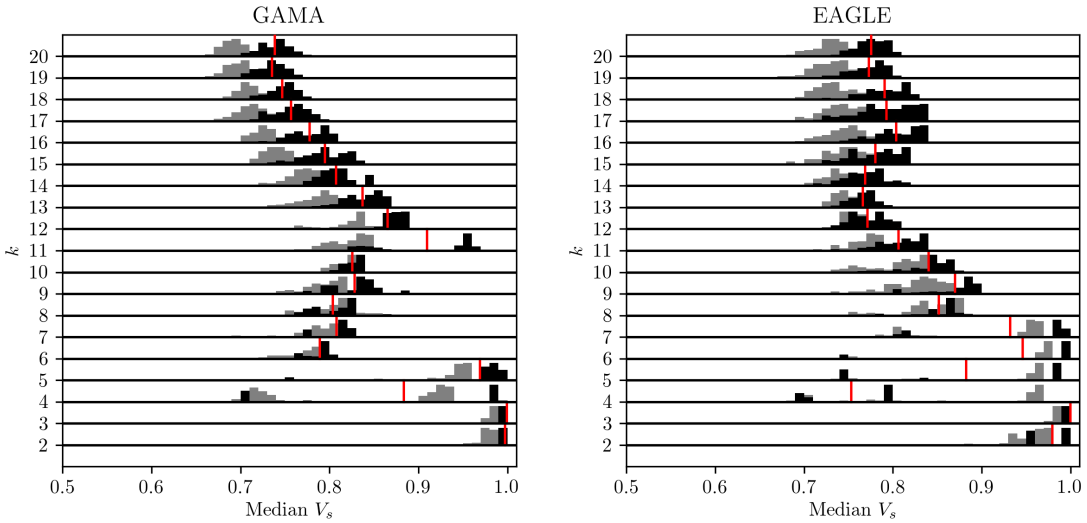


**Figure 4.2:** Stability maps, showing the stability of `k-means` clustering among 100 random initialisatons at each of $k = 2$ through to $k = 20$ for the GAMA sample of observed galaxies (**left**) and for the EAGLE sample of simulated galaxies (**right**). The stability of a set of outcomes at the same $k$ is illustrated by their distribution in median $V_s$, which is shown in black for clustering in the original samples (with means shown by the red markers) and in grey for partitions given by clustering in bootstrapped samples (see main text). Distributions skewed towards higher values of median $V$ indicate stable values of $k$, and I select the highest stable value for each sample as its best. The best outcomes are: $k = 5$ for the GAMA sample, and $k = 7$ for the EAGLE sample.

The black histograms show the distribution of median $V_s$ values at each $k$. The grey histograms are based on partitions given by `k-means` clustering in bootstrapped samples, generated thusly: the bootstrap resampling method is applied to the original sample (drawing a new sample, equal in size to the original sample, with replacement), conduct `k-means` clustering in the bootstrapped sample, and use the final centroids to partition the original sample. The black histograms illustrate stability in spite of different, random initialisations in the original samples. The grey histograms illustrate stability in spite of random variations of the samples themselves. Both sets generally exhibit similar behaviour.

Values of $k$ at which the histograms are concentrated toward higher values of median $V_s$ are more stable; outcomes are more consistently similar. I distinguish two types of histogram: stable histograms have strong peaks at $V_s \gtrsim 0.9$, and unstable histograms have multiple peaks of similar magnitude or a large spread over lower values of median $V_s$. The grey histograms are systematically offset from the black histograms towards lower median $V_s$, particularly at higher values[2] of $k$. For both samples, the black histograms appear to exhibit a limit in $k$ beyond which there are no further stable values of $k$. The grey histograms, with their offset, substantiate this effect.

I select the highest stable values of $k$ as the best because they offer the most detailed partitions while still capturing reproducible clustering structure. For the GAMA sample, $k = 5$ is the best. While the black histogram at $k = 10$ exhibits a strong peak at high median $V_s$, its corresponding grey histogram does not. Clustering at this value of $k$ is not robust to bootstrap resampling of the GAMA sample, so I consider it unstable overall. For the EAGLE sample, $k = 7$ is the best. Both the black and grey histograms exhibit a strong peak at high median $V_s$ this value of $k$, so clustering at this value of $k$ is stable and robust. I compare outcomes at two different values of $k$; identifying where and why the additional clusters have been determined within the EAGLE sample may reveal important differences in its structure from that of the GAMA sample.

The final, best *outcomes* of the 100 at each of these two best values of $k$ are those with the lowest compactness, which I measure with the sum of within-cluster variances $\phi$. I name the clusters that constitute these outcomes using two-part notation in the format "XY"; "X" (either "G" or "E") denotes the outcome to which the cluster belongs (in terms of the sample within with it was determined; GAMA or EAGLE), and "Y" (an integer with value $1 \leq Y \leq k$) denotes the cluster itself within that outcome. Cluster names are ranked by the average *sSFR* of the cluster to which they refer, such that X1 would denote the cluster containing galaxies with the highest average *sSFR* for outcome X.

[2]While the grey histogram at $k = 5$ for the EAGLE sample unexpectedly has a peak at high median $V_s$ where its corresponding black histogram does not, it has another peak at lower median $V_s$, such that both histograms agree that this value of $k$ is unstable.

**Table 4.2:** A summary of the clusters comprising the two best outcomes. The prefixes "G" or "E" in the cluster names denote the sample (GAMA or EAGLE) within which the cluster was determined. Column $N_C$ contains the numbers of galaxies belonging to each cluster. The remaining entries list the centroid of each cluster in each of the input features to the clustering.

| Cluster | $N_C$ | $\log_{10}(M_*/\mathrm{M}_\odot)$ | $\log_{10}(sM_d)$ | $\log_{10}(sSFR)$ | $\log_{10}[(R_{1/2,l}$ or $R_{1/2,m})/\mathrm{kpc}]$ | $B/T_l$ or $B/T_m$ |
|---|---|---|---|---|---|---|
| G1 | 947 | 9.86 | -2.82 | -9.95 | 0.74 | 0.11 |
| G2 | 781 | 9.91 | -2.89 | -10.10 | 0.59 | 0.68 |
| G3 | 567 | 10.54 | -3.23 | -10.66 | 0.95 | 0.25 |
| G4 | 695 | 9.97 | -3.88 | -11.70 | 0.46 | 0.55 |
| G5 | 734 | 10.72 | -4.24 | -12.06 | 0.86 | 0.73 |
| E1 | 1,223 | 9.83 | -2.70 | -10.07 | 0.74 | 0.37 |
| E2 | 1,433 | 9.86 | -2.90 | -10.24 | 0.62 | 0.59 |
| E3 | 1,106 | 9.79 | -3.02 | -10.28 | 0.55 | 0.84 |
| E4 | 1,148 | 10.48 | -3.00 | -10.35 | 0.74 | 0.33 |
| E5 | 603 | 10.81 | -3.83 | -11.20 | 0.77 | 0.83 |
| E6 | 801 | 10.24 | -4.03 | -11.78 | 0.56 | 0.47 |
| E7 | 803 | 9.82 | -4.03 | -11.81 | 0.53 | 0.82 |

### 4.2.2 Comparing the outcomes

Table 4.2 offers a summary of the clusters comprising both of the final outcomes in terms of the input features to the clustering. Figures 4.3 and 4.4 show the distributions of the clusters in two-dimensional projections of the input feature space. These figures reveal a broad similarity of the structures of the outcomes. The primary structural characteristic of both outcomes is the division of their clusters by the star formation activity and dust content of the galaxies they contain, which are themselves correlated. Clusters from both outcomes line up either along the SFMS, or along the parallel sequence of passive galaxies (except for E5, see below). This reflects the well-established bimodal nature of the galaxy population (Baldry et al., 2004; Schawinski et al., 2014). Subtler distinctions within these two sequences, which differ between the two outcomes, are driven by the remaining input features. These distinctions include the separation of disc- and bulge-dominated star-forming galaxies, and the grouping of passive galaxies with different stellar masses and morphologies.

In order to quantify the similarity of the two clustering outcomes ("G" and "E"), I compare how their centroids partition the *same* galaxies. Given that the GAMA sample comprises real galaxies in the real Universe, I take it as the reference sample with which to compare the partitions. This is enabled by my choice to cluster both samples within rescaled feature spaces of unit dimensions (Section 4.1.1). With both clustering outcomes having been determined within an equivalent feature space, the E centroids may be mapped onto the GAMA sample. The EAGLE-based partition of the GAMA sample, which I name $\mathrm{E_G}$, is then given by assigning GAMA galaxies to their near-

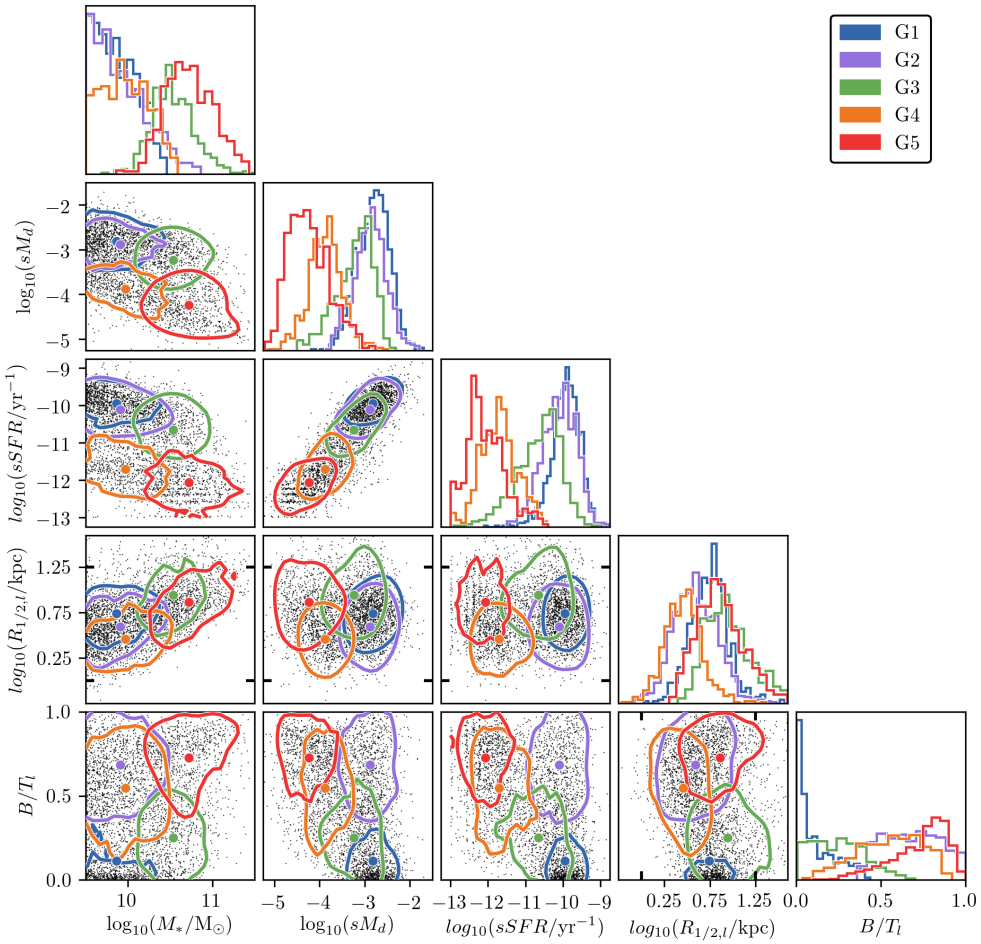


**Figure 4.3:** A profile of the best `k-means` clustering outcome for the GAMA sample of observed galaxies, consisting of five clusters. Cluster distributions are shown using coloured contours (drawn at a relative density of 0.25) and histograms, and their centroids using filled, coloured circles. The black markers on the inside of the $R_{1/2,l}$ axes show the $R_{1/2,m}$ axis limits in the profile of the best clustering outcome ($k = 7$) for the EAGLE sample (Figure 4.4).

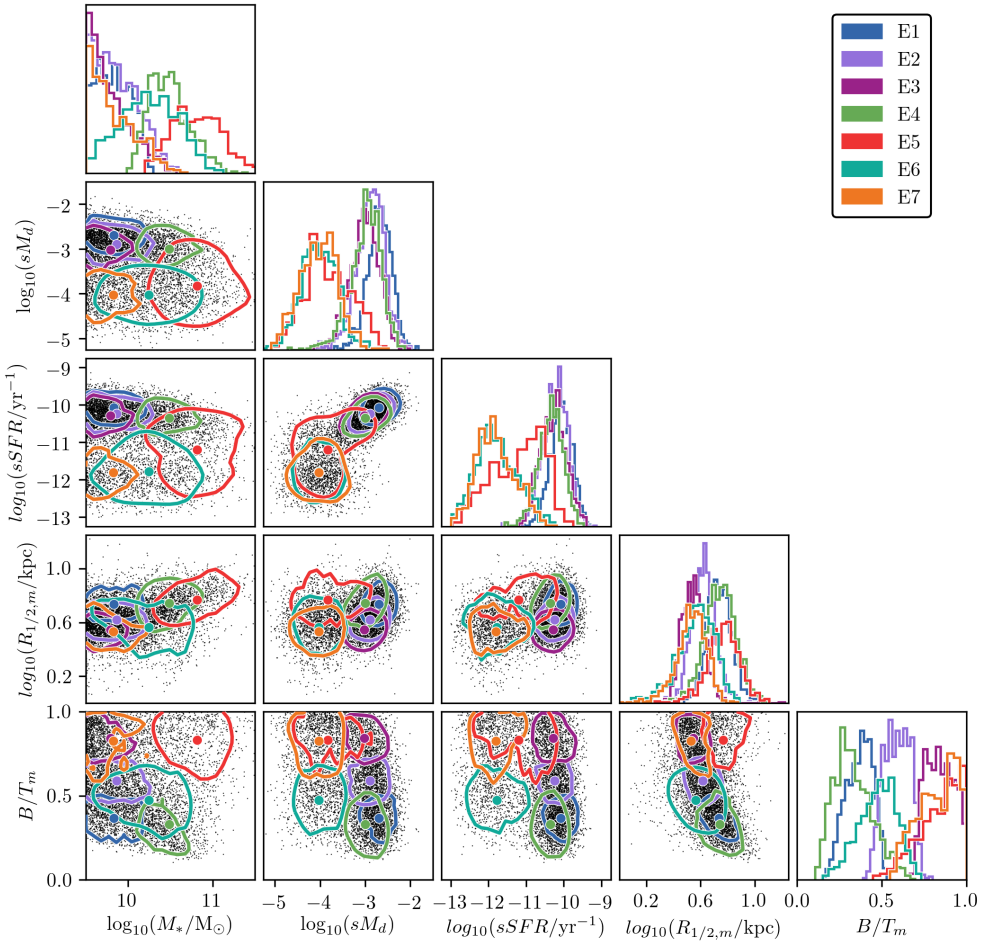


**Figure 4.4:** A profile of the best `k-means` clustering outcome for the EAGLE sample of simulated galaxies, consisting of seven clusters. Cluster distributions are shown using coloured contours (drawn at a relative density of 0.25) and histograms, and their centroids using filled, coloured circles. Note that the axis limits of $R_{1/2,m}$ differ from those of $R_{1/2,l}$ in the profile of the best clustering outcome ($k = 5$) for the GAMA sample (Figure 4.3), due to the tighter distribution of galaxy sizes in the EAGLE sample.

**Table 4.3:** A contingency table comparing two partitions of the GAMA sample: G (determined within the GAMA sample) and $E_G$ (determined within the EAGLE sample and mapped onto the GAMA sample; see main text for further explanation). Entries show the fraction of galaxies in the GAMA sample that are contained by each combination of G and $E_G$ cluster. Entries in bold show combinations which contain the majority of an individual $E_G$ cluster.

| | $E_G1$ | $E_G2$ | $E_G3$ | $E_G4$ | $E_G5$ | $E_G6$ | $E_G7$ |
|---|---|---|---|---|---|---|---|
| G1 | **0.17** | 0.01 | 0.00 | 0.07 | 0.00 | 0.01 | 0.00 |
| G2 | 0.00 | **0.09** | **0.09** | 0.00 | 0.02 | 0.00 | 0.01 |
| G3 | 0.00 | 0.00 | 0.00 | **0.11** | 0.01 | 0.03 | 0.00 |
| G4 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | **0.10** | **0.09** |
| G5 | 0.00 | 0.00 | 0.00 | 0.00 | **0.13** | 0.05 | 0.01 |

est $E_G$ centroid[3]. Differences between G and $E_G$ will arise both from global differences between the GAMA and EAGLE samples and from more local differences between G and E.

Table 4.3 directly compares the partitions, showing the fraction of GAMA galaxies shared by each combination of G and $E_G$ clusters. I use Cramer's V index to measure the overall agreement between the two partitions. My use of $V$ in the context of agreement (which is separate from my use of $V$ in the context of stability; see Sections 3.1 and A.1; Figure 4.2) is denoted with the symbol $V_a$. For the partitions G and $E_G$, I calculate $V_a = 0.76$. I provide a guide for the interpretation of $V_a$ in Appendix B.1. On the basis of this guide, and in agreement with my more qualitative assessment above, $V_a = 0.76$ indicates a broad similarity between the two outcomes, but the possibility of slight differences at the substructure level.

Where the majority of galaxies within an $E_G$ cluster is contained within a single G cluster, the corresponding Table 4.3 entry is highlighted with bold text. This tells us which of the G clusters each of the E clusters (via the $E_G$ clusters) is most strongly related to, in terms of their coverage of the shared feature space. Given that the E outcome comprises more clusters than the G outcome, G clusters may be related to more than one E cluster; indeed, there are two such cases in Table 4.3. The strength of these relationships is measured using the Jaccard (1901) index ($J$), dividing the number of GAMA galaxies shared by *all* of a related set G and $E_G$ clusters by the total number of GAMA galaxies contained by *any* of them[4]. I use these related sets to guide the following cluster-by-cluster comparison of the G and E outcomes, quoting $J$ values as a measure of how well each of the G clusters has been recovered, and discussing implications of the identities of E clusters for facilitating an understanding of the evolution of galaxies in their corresponding G clusters.

[3]Partitions E and $E_G$ have the same centroids and the same numerical naming scheme for their clusters, such that clusters E1 and $E_G1$ have the same exact centroid and span the same region of the shared feature space of both samples. However, E clusters contain EAGLE galaxies and $E_G$ clusters contain GAMA galaxies.

[4]Alternatively put, this is the intersection of the Venn diagram of a related set of clusters, divided by the union.

**G1, G3, E1, and E4: star-forming disc galaxies**

According to Table 4.3, E1 (via $E_G1$) is most closely related to G1, and E4 (via $E_G4$) to G3. These pairs produce Jaccard indices of $J$(G1,$E_G1$) = 0.65 and $J$(G3,$E_G4$) = 0.50 respectively, suggesting a limited recovery of the observed clusters by their corresponding simulated clusters. However, combining all four of these clusters for a unified comparison produces a much higher score of $J$(G1,G3,$E_G1$,$E_G4$) = 0.88.

My justification for unifying the comparison in this way is two-fold. Firstly, there is significant overlap between the two related pairs; the vast majority of G1 galaxies that are not contained by $E_G1$ are instead contained by $E_G4$ (Table 4.3). Secondly, the astrophysical identities of these clusters are largely similar. All four contain star-forming, dusty, disc-dominated galaxies (Table 4.2 and Figures 4.3 and 4.4). EAGLE has therefore succeeded in reproducing the SFMS at low and intermediate masses, accurately capturing the growth of the stellar masses of galaxies via star formation (see also Schaye et al. 2015; Furlong et al. 2015; Clauwens et al. 2018). Shortcomings at the high-mass end of the SFMS are addressed below.

These clusters divide the SFMSs of their respective samples by stellar mass: G1 and E1 contain low-mass galaxies, and G3 and E4 contain intermediate- to high-mass galaxies. However, the relatively low Jaccard indices given by each of these pairs of clusters are instead due to two main morphological differences between the disc galaxies in each sample. The first difference is the general offset of EAGLE disc galaxies towards bulgier morphologies: EAGLE galaxies have a minimum $B/T_m$ of $\sim$ 0.15 (Figure 4.4), while GAMA galaxies have $B/T_l$ values as low as 0 (Figure 4.3). Hence, EAGLE is unable to reproduce the pure disc galaxies that are observed in the real Universe. The influence of this offset upon my comparison of G and E is mitigated somewhat by my choice to rescale the feature distributions pre-clustering (see Section 4.1). Hence, in spite of this offset, I am still able to directly compare the disciest EAGLE galaxies with the disciest GAMA galaxies, which is a particular strength of my approach. However, the *cause* of this offset leads also to the second difference: while G3 galaxies have *more* prominent bulges than those in G1, E4 galaxies have bulges that are as promininent or slightly *less* prominent than those in E1. Hence, EAGLE does not accurately reproduce the observed evolution of the morphological components of disc galaxies with increasing stellar mass along the SFMS.

These morphological differences are caused by the combination of two resolution effects within EAGLE. Firstly, the spatial resolution of EAGLE ($\sim$ 1 pkpc), set by the Jeans length of the warm ISM ($T \approx 10^4$ K), leads the discs of EAGLE galaxies to have larger scale heights than the discs of observed galaxies (Schaye et al., 2015; Trayford et al., 2017; Thob et al., 2019). Hence, the dynamics of even the disciest EAGLE galaxies are fractionally supported by dispersion, resulting in their minimum $B/T_m$ of $\sim$ 0.15. This is also likely to be the cause of the relatively small sizes

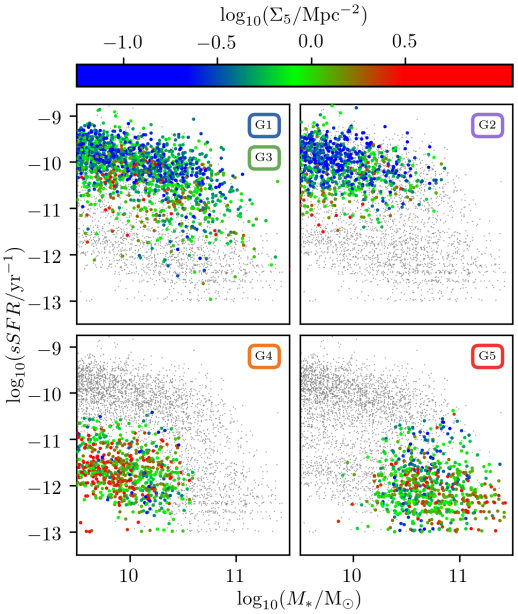


**Figure 4.5:** The local environmental densities of galaxies in each of the clusters determined within the GAMA sample. Galaxies are plotted as scatter points in the *sSFR* versus $M_*$ plane. The smaller grey points represent the full GAMA sample and the larger coloured points represent galaxies belonging to particular clusters, which are plotted in separate panels (see the panel labels; note that the upper-left panel highlights two clusters together for reasons given in the main text above). The colours of the points show the fifth-nearest-neighbour surface densities of galaxies, which have been smoothed using the surface densities of the seven galaxies that are nearest to them in the five-dimensional feature space.

of EAGLE's disc galaxies in comparison with those of GAMA (Table 4.2, Figures 4.3 and 4.4). Secondly, the coarse mass resolution of EAGLE's dark matter particles leads to their spurious scattering with better-resolved star particles, which also contributes dispersion-support (Ludlow et al., 2020). Hence, improvements to EAGLE's resolution are needed for more accurate morphologies among disc galaxies. These improvements, in turn, will beget a more accurate breakdown of EAGLE's SFMS. I note that my use of mass-based bulge-to-total ratios for EAGLE galaxies, which might be expected to understate the contribution of the disc component (containing recently formed, low mass-to-light ratio stars) in comparison with light-based bulge-to-total ratios, is unlikely to contribute significantly to this difference in disc galaxy morphologies between the samples; Scannapieco et al. (2010) find that while bulge-to-total ratios that are defined kinematically are systematically higher than those that are defined photometrically, differences in the mass-to-light ratios between the discs and the bulges of simulated galaxies are small.

The influence of these resolution limits upon the integrated morphologies of EAGLE disc galaxies diminishes slightly at higher stellar mass and larger half-mass radii, where the radial scales of their discs dominate over their vertical scales. As a consequence, in comparison with E1 galaxies, the morphologies of E4 galaxies more closely resemble those of G1 galaxies (i.e. more prominent

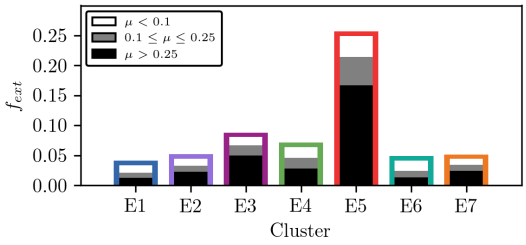


**Figure 4.6:** A stacked bar chart showing the cluster-average fractional contribution ($f_{ext}$) of accretion ($\mu < 0.1$), minor mergers ($0.1 \leq \mu \leq 0.25$), and major mergers ($\mu > 0.25$) since $z = 1$ to the $z = 0$ stellar masses of galaxies in the EAGLE sample.

discs). Hence, when the E centroids are mapped onto the GAMA sample, some G1 galaxies end up assigned to $E_G4$ rather than $E_G1$ (Table 4.3). I note that this trend between E1 and E4, of constant or decreasing bulge prominence, is contrary to what is seen in observations, where is has been shown that bulge prominence increases along the SFMS (Bluck et al., 2014; McPartland et al., 2019; Popesso et al., 2019a). Examining whether this trend is recovered with improvements to resolution will constitute a crucial test of the next generation of cosmological galaxy simulations. Furthermore, EAGLE fails to capture the observed turn-down of the SFMS at high stellar masses (see below), which has been linked with this rise in bulge prominence (Cheung et al., 2012; Fang et al., 2013; Bluck et al., 2014) as part of an internally-driven quenching pathway (Peng et al., 2010; Schawinski et al., 2014; Moutard et al., 2020).

#### G2, E2, and E3: bulge-dominated star-forming galaxies

For clusters G2, E2, and E3, I calculate $J(\mathrm{G2},\mathrm{E_G2},\mathrm{E_G3}) = 0.82$, which indicates good recovery of the observed cluster by the simulated clusters. Galaxies in these clusters have low stellar masses, high specific star formation rates and dust masses, and compact, bulge-dominated morphologies (Table 4.2, Figures 4.3 and 4.4). Of the few G2 galaxies that are not also contained within $E_G2$ and $E_G3$, most are instead contained within $E_G5$ or $E_G7$. This is due to morphological similarities between E2, E5, and E7 (Figure 4.4).

The matching of two EAGLE clusters to one GAMA cluster arises because of the strong continuum in the bulge-to-total ratios of low-mass star-forming galaxies in EAGLE (Figure 4.4). `k-means` tends to segment continua (see Section 3.1), and so it distinguishes between star-forming EAGLE galaxies with intermediate (E2) and high (E3) $B/T_m$ along this continuum. The equivalent distribution among star-forming GAMA galaxies is more diffuse (Figure 4.3), and as a result `k-means` models it using a single cluster (G2) that has a large spread in $B/T_l$. Hence, while the coverage of G2 within the shared feature space has been successfully reproduced by E2 and E3, the use by `k-means` of two clusters in EAGLE compared with one cluster in GAMA reveals that EAGLE

produces too many bulge-dominated galaxies at the mass regime of these clusters. Observational studies instead show that disc galaxies dominate the stellar mass function at these stellar masses (Kelvin et al., 2014b; Lange et al., 2015; Moffett et al., 2016). I note that this morphological difference partially arises from the resolution limitations of EAGLE that were discussed above, and so improvements to resolution in the next generation of cosmological simulations may mitigate this difference by yielding discier low-mass galaxies.

Clauwens et al. (2018) proposed three mass-based phases for the evolution of central galaxies and their morphological components. Their first phase ($\lesssim 10^{9.5}$ $M_\odot$) is characterised by in-situ star formation that is triggered by merger activity, and yields spheroidal galaxies. In their second phase ($10^{9.5}$ $M_\odot \lesssim M_* \lesssim 10^{10.5}$ $M_\odot$), continued in-situ star formation prompts the growth of a disc. Finally, their third phase ($\gtrsim 10^{10.5}$ $M_\odot$), in which stellar accretion and mergers dominate the growth of galaxies, results in bulge-dominated galaxies. Clusters E3 and E2 (and by extension E1) appear to correspond with the transition between the first two of these phases, given that their low-mass galaxies (the majority of which are centrals; Figure 4.7) exhibit a gradient in their morphologies (Figure 4.4) and in their merger activity (Figure 4.6). The spread of morphologies among the low-mass star-forming galaxies contained within G2 provide observational support for this suggestion.

Low-mass, star-forming, bulge-dominated galaxies, like those captured by E3, have been observed previously. Kelvin et al. (2014a) and Moffett et al. (2016) distinguished a class of "Little Blue Spheroids" (LBSs) among $z < 0.06$ galaxies in the equatorial regions of GAMA using Hubble classifications. Of the 291 G2 galaxies (of 781 total) for which Hubble classifications exist (due to the $z < 0.06$ restriction), only three are LBSs[5], while the rest are either ellipticals or early-type spirals. This mixture of Hubble classifications for G2 galaxies is consistent with their observed spread in $B/T_l$, and with the intermediate morphologies of E2 galaxies. The low number of LBSs in G2 is due in part to the lower stellar mass limit that is imposed upon the GAMA sample ($10^{9.5}$ $M_\odot$), which means that it is not inclusive to LBSs (Kelvin et al., 2014a; Moffett et al., 2016). Moffett et al. (2019), in an in-depth study of LBSs, find that most exhibit ordered rotation and suggest that this rotation may come from the accretion of star-forming gas (Graham et al., 2017). This is in contrast with E3 galaxies, whose high $B/T_m$ values come from their lack of any net rotation (Thob et al., 2019). Hence, altogether, the identities of E3 and G2 galaxies are not consistent with those of LBSs.

Schawinski et al. (2009a), by way of Galaxy Zoo morphologies (Lintott et al., 2008, 2011), identified a class of blue elliptical galaxies (i.e. another potential observational counterpart to E3 galaxies), and suggested starbursts via gas-rich mergers (Mihos & Hernquist, 1994a,b, 1996; Cox et al., 2008) as the mechanism for their formation. While E3 galaxies are subject to more merger

[5]The one other confirmed LBS in my sample is contained by cluster G1.

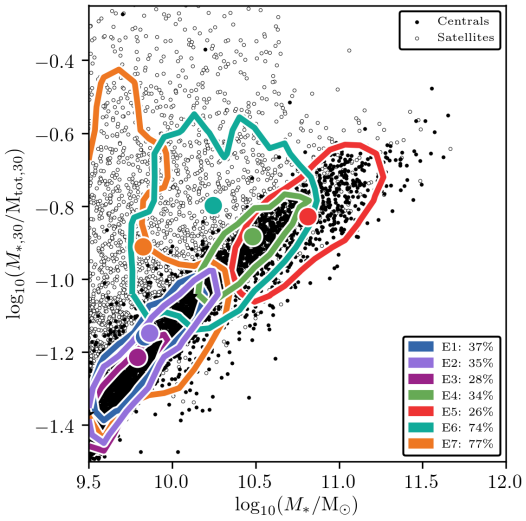


**Figure 4.7:** Stellar-to-total mass ratios of galaxies in the EAGLE sample, as a function of their stellar masses. These ratios are measured within 30 pkpc spherical apertures. Total masses are the sum of the masses of all of EAGLE's particle types within the aperture. Central and satellite galaxies are marked using filled and empty circles respectively. Cluster distributions are shown using coloured contours, and their centroids in this plane using filled, coloured circles. The percentages in the legend show the fraction of galaxies within each cluster that are satellites. The high values of the stellar-to-total mass ratios of E6 and E7 galaxies in Figure 4.7 arise from their accretion as satellites into the outer regions of massive group haloes.

activity than other galaxies with similar stellar masses (Figure 4.6), they exhibit typical SFMS specific star formation rates (Table 4.2, Figure 4.4) which indicate that they are generally not star-bursting. This is in agreement with Moffett et al. (2019), who suggest that the SFHs of LBSs are relatively constant. It is also consistent with my suggestion that E3 (along with E2 and E1) captures the transition from the first phase to the second phase of the Clauwens et al. (2018) evolutionary model, in that E3 galaxies appear to have specific star formation rates that are typical of the second phase.

### G4, E6, and E7: satellite galaxies

Table 4.3 shows that E6 and E7 are both most closely related to G4. These clusters contain galaxies that are passive, with little dust, have low-to-intermediate stellar masses, and exhibit a spread in their morphologies and sizes (Table 4.2, Figures 4.3 and 4.4). The relatively low Jaccard index calculated among these clusters, $J(\mathrm{G4,E_G6,E_G7}) = 0.63$, arises mostly from the inclusion of low-*sSFR* G3 galaxies and intermediate-$B/T_l$ G5 galaxies within $\mathrm{E_G6}$ (Figures 4.3 and 4.4). This is driven by the lack of coverage of this region of the feature space by other E clusters. E4 is confined to high specific star formation rates by EAGLE's flat and tight SFMS (where the GAMA

SFMS turns down), and E5 is confined to high bulge-to-total ratios by the high stellar masses of its galaxies and by their spread in star formation activity (where the GAMA SFMS does not extend to such high masses and where G5 is uniformly passive; see below). Hence, `k-means` uses E6 to model the distribution of EAGLE galaxies within this vacant region of the shared feature space and, as a result, E6 extends to slightly higher masses than G4. $E_G7$, meanwhile, is almost entirely contained within the extent of G4 in the shared feature space; hence, the region of the feature space occupied by *low*-mass, passive galaxies has been accurately recovered.

The majority of galaxies contained by E6 and E7 are satellites ($\sim$ 75 per cent; Figure 4.7), and they occupy haloes with a logarithmic mean total mass[6] of $10^{14.0}$ $M_\odot$. This is nearly a factor of ten greater than the logarithmic mean total halo mass of E1-E4 galaxies ($10^{13.2}$ $M_\odot$), which contain star-forming galaxies that span a similar range in stellar masses, and is in agreement with Cochrane & Best (2018), who link the quenching of EAGLE satellite galaxies to the total masses of their host haloes. Similarly, Figure 4.5 shows that the observational counterparts of these galaxies, in cluster G4, tend to inhabit environments of high densities. Overall, this highlights a prominent external influence upon the evolution of E6, E7, and G4 galaxies, and is consistent with studies which attribute the quenching of lower-mass galaxies to their environments (Peng et al., 2012; Wetzel et al., 2012, 2013; Trayford et al., 2016). The morphological diversity among E6, E7, and G4 galaxies (Figures 4.3 and 4.4) reflects the *variety* of environmental processes to which they are subjected (Correa et al., 2017; Smethurst et al., 2017). In addition, it indicates that the quenching of satellites galaxies is *not* always accompanied by a morphological transformation, such that some satellite galaxies retain their discs. Processes that meet this criterion include ram-pressure stripping (Gunn & Gott, 1972; McCarthy et al., 2008), which removes cold ISM from galaxies via their interaction with the hot intergalactic media of groups, or starvation (Larson et al., 1980; Peng et al., 2015), which inhibit the ability of galaxies to accrete new star-forming gas upon their accretion into group haloes. Furthermore, mergers and accretion, which would be expected to alter the morphologies of E6 and E7 galaxies (Toomre, 1977; Barnes, 1988, 1992; Walker et al., 1996), appear to have relatively little influence on their evolution (Figure 4.6) and are hence unlikely to be involved in their quenching (Weigel et al., 2017).

#### G5 and E5: high-mass central galaxies

Cluster E5 is most strongly related to cluster G5 (Table 4.3), with both clusters containing particularly massive and bulge-dominated galaxies (Table 4.2). G5 also includes galaxies with more intermediate morphologies (Figure 4.3), which E5 does not (Figure 4.4). As a result, when the E centroids are mapped onto the GAMA sample, some G5 galaxies end up included in cluster $E_G6$

[6]Measured within a radius from the central of that halo at which the mean enclosed density is 200 times the critical density of the Universe.

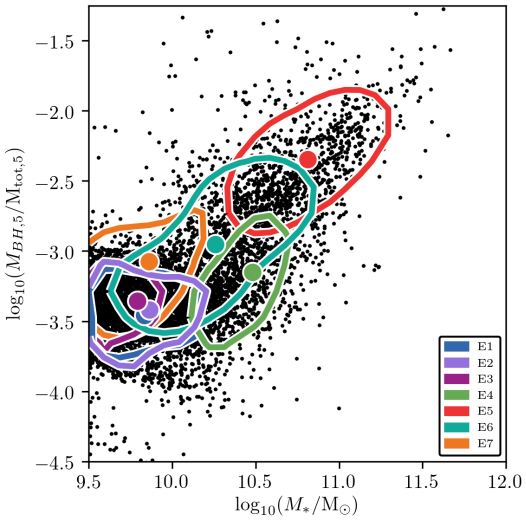


**Figure 4.8:** Black-hole-to-total mass ratios of galaxies in the EAGLE sample, as a function of their stellar masses. These ratios are measured within 5 pkpc spherical apertures. Total masses are the sum of the masses of all of EAGLE's particle types within the aperture. Cluster distributions are shown using coloured contours, and their centroids in this plane using filled, coloured circles.

instead (see also above), and I calculate a relatively low Jaccard index of $J(\mathrm{G5},\mathrm{E_G5}) = 0.59$.

Another important difference between G5 and E5 is that while G5 is made up purely of passive galaxies, 45 per cent of E5 galaxies are unexpectedly (in the context of their stellar masses and morphologies) star-forming ($sSFR > 10^{-11}$ yr$^{-1}$). These star-forming galaxies generally do not correspond with those in G3 (Table 4.3), which encompasses the high-mass end of the GAMA SFMS, because the E5 star-forming galaxies have more concentrated morphologies and higher stellar masses. Hence, EAGLE produces a subpopulation of high-mass, bulge-dominated, star-forming galaxies that are not observed in the real Universe. The same subpopulation has previously been identified through the use of non-parametric morphological features by Bignone et al. (2020) and a variety of morphological and kinematic diagnostics by Thob et al. (2019). In addition, Trayford et al. (2015) identified a similar subpopulation of galaxies on the basis of their colours and their stellar masses.

At the stellar masses of E5 galaxies, where the EAGLE SFMS remains flat, the GAMA SFMS turns down. This reduction in the star formation activity of observed galaxies at high stellar masses has been linked with an increase in their central densities (Cheung et al., 2012; Fang et al., 2013; Bluck et al., 2014) and is attributed to the growth of the central supermassive black holes (SMBHs) of galaxies (Häring & Rix, 2004) which can provide feedback that removes or heats star forming gas (Bower et al., 2006; Croton et al., 2006; Hopkins et al., 2006; Springel et al., 2006; Schawinski et al., 2006, 2007, 2009b). The lack of a turndown in the SFMS of the EAGLE sample, and the uniformly passive galaxies in G5, suggest that this feedback is insufficiently potent

in EAGLE (Trayford et al., 2015; Rosito et al., 2018). In addition, galaxies with stellar and halo masses typical of those galaxies in clusters G5 and E5 are expected to have gas that they accrete virially shocked to high temperatures, which would also act to maintain their passivity. However, the inability of cosmological simulations to reliably prevent this gas from cooling (Somerville & Davé, 2015) may also be partially responsible for the high-mass star-forming galaxies in EAGLE. I note that while improvements to simulation resolution (see above) may reproduce the observed growth of bulges within galaxies as they evolve along the SFMS, it is not expected that this will also recover their associated reduction in star formation.

The majority of E5 galaxies are centrals (74 per cent; Figure 4.7), and they occupy large haloes with a logarithmic average total mass of $10^{13.6}$ $M_{\odot}$. Similarly, G5 galaxies have high nearest-neighbour surface densities (Figure 4.5), which suggests that they are surrounded by satellites. Star-forming and passive E5 galaxies have similar halo masses, differing by less than 0.2 dex on average, and passive E5 galaxies are only 10 per cent more likely to be centrals than their star-forming counterparts, such that E5 galaxies in general seem to inhabit the same environments. Figure 4.8 shows that, in comparison with galaxies in other clusters, the inner masses of E5 galaxies are dominated by their SMBHs. E5 SMBHs also have high accretion rates, which drive the active galactic nucleus (AGN) feedback that they provide to their host galaxies (Schaye et al., 2015; Crain et al., 2015). Passive and star-forming E5 galaxies differ only slightly ($\lesssim$ 0.2 dex) in their SMBH masses and accretion rates. In accordance with the findings of Qu et al. (2017) and Clauwens et al. (2018) at high stellar masses, I find a significant external contribution to the growth of E5 galaxies ($\sim$ 25 per cent; Figure 4.6). Overall, star-forming and passive E5 galaxies are generally equivalent in terms of these ancillary features.

In Figure 4.9, I trace the evolution of the gas content and star formation activity of star-forming and passive E5 galaxies separately over the past 4 Gyr. Passive E5 galaxies resembled star-forming E5 galaxies in terms of both of these features $\sim$ 4 Gyr ago. The star-formation activity of E5 galaxies is directly dependent upon their gas content, with passive E5 galaxies quenching (i.e. passing below $10^{-11}$ yr$-1$) $\sim$ 2 Gyr ago. I attribute this quenching to the removal of the ISM of passive E5 galaxies by AGN feedback (Schawinski et al., 2009b; Davies et al., 2019, 2020b). Star-forming E5 galaxies, on the other hand, have remained consistent in terms of their gas content and star formation activity throughout the past 4 Gyr, and are generally neither star-bursting nor rejuvenating. I do not find a corresponding divide in the evolution of the halo masses and SMBH masses of star-forming and passive E5 galaxies, even when tracing this evolution back further to particularly early times ($\sim$ 12 Gyr ago). This is in contrast with Davies et al. (2020a), who show that present day star formation activity of central galaxies is linked with the assembly histories of their surrounding haloes, with the haloes of passive galaxies having assembled earlier. I *do* find that star-forming E5 galaxies are subject to their last major merger an average of 0.5 Gyr

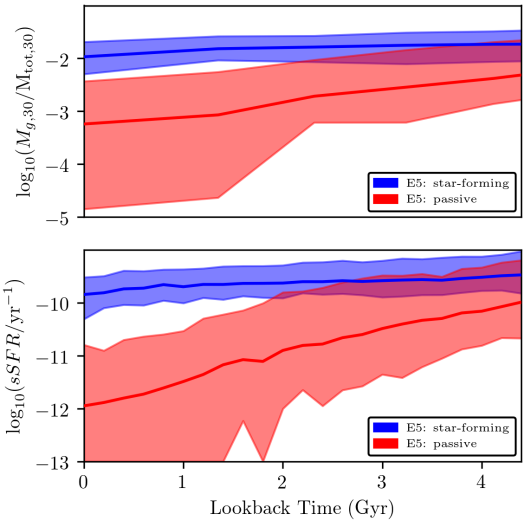


**Figure 4.9:** Mean gas-to-total mass ratios (within 30 pkpc apertures; upper panel) and specific star formation rates (lower panel) of star-forming (blue) and passive (red) E5 galaxies over the last 4 Gyr. Total masses are the sum of the masses of all of EAGLE's particle types within the aperture. The shaded regions shows the $16^{\mathrm{th}}$ and $84^{\mathrm{th}}$ percentiles of these evolving distributions. The star-forming and passive E5 galaxies shown (100 each) are those that have specific star formation rates closest to $10^{-10}$ yr$^{-1}$ and $10^{-12}$ yr$^{-1}$ respectively. The time resolution of the upper panel corresponds directly with EAGLE's snapshots; in the lower panel, it is 200 Myr.

more recently than passive E5 galaxies[7] (at average lookback times of $\sim$ 4.5 Gyr and $\sim$ 5.0 Gyr respectively), but it is clear that further analysis of the assembly histories of star-forming and passive E5 galaxies is needed to better understand differences in their evolution. The disagreement of my results with Davies et al. (2020a) may be partially influenced by two factors: while I examine the average evolution of many different galaxies in terms of their haloes and SMBHs, Davies et al. (2020a) examine modified resimulations of the same galaxy; and while Davies et al. (2020a) focus their analysis and discussion on central galaxies, my cluster E5 also includes some satellite galaxies. I note that Figure 4.9 also shows that morphological quenching (Martig et al., 2009) does not constitute a significant quenching pathway in EAGLE; while passive early-type galaxies that retain their gas supplies have been observed (Martig et al., 2013), passive E5 galaxies are distinguished from their star-forming counterparts *by* their lack of gas.

Figure 4.10 shows examples of forward-modelled optical images of galaxies in E5, split by their star formation activity. These images were generated by Trayford et al. (2017) using the radiative transfer code `SKIRT` (Baes et al., 2011; Camps & Baes, 2015). Both star-forming and passive E5 galaxies have apparent dense cores, which reflect their high $B/T_m$ values. However, some star-forming E5 galaxies also exhibit extended, clumpy discs, which are the sites of ongoing star

[7] I note, though, that this measurement is limited by the time-resolution of the merger trees, which corresponds directly with EAGLE's snapshots.

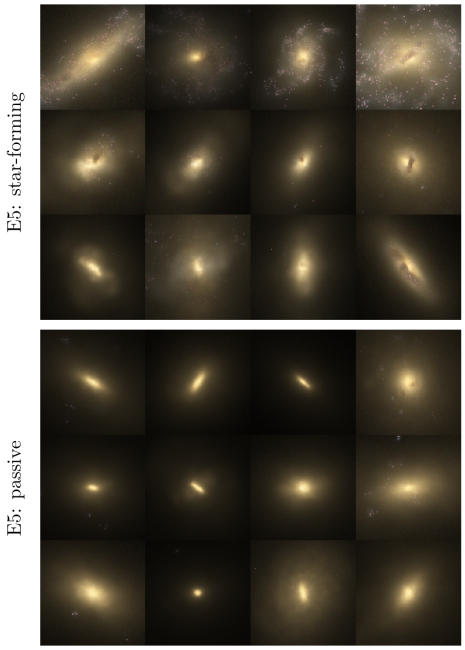


**Figure 4.10:** Example images of galaxies in E5, separated by their star formation activity. These mock three-colour optical images, cropped to a scale of 30 pkpc, were prepared by Trayford et al. (2017) using the radiative transfer code `SKIRT` (Baes et al., 2011; Camps & Baes, 2015). The star-forming and passive E5 galaxies shown are those that have specific star formation rates closest to $10^{-10}$ yr$^{-1}$ and $10^{-12}$ yr$^{-1}$ respectively. The SFHs in this figure have a

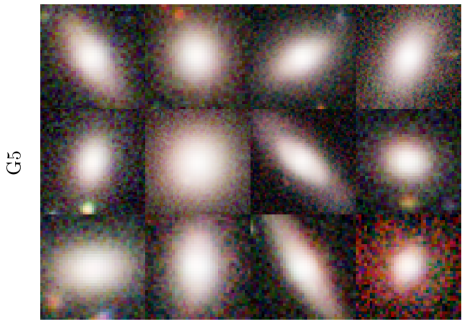


**Figure 4.11:** Example images of galaxies in G5. These three-colour *giH* images, with scales of $\sim$ 30 kpc, were prepared by Moffett et al. (2016) using photometry from SDSS and from the VISTA Kilo-Degree Infrared Galaxy Survey (Sutherland et al., 2015). The galaxies shown are those within G5 that are closest to its centroid.

formation, and dust lanes, demonstrating the retention of their ISM. Example images of G5 galaxies (Figure 4.11), meanwhile, show that while some have discs, they are uniformly smooth and do not exhibit any substructures. Thus, these images confirm that the presence of a dense core does not appear to be as strongly linked with internally-driven quenching in EAGLE as it is in observations (Peng et al., 2010; Cheung et al., 2012; Fang et al., 2013; Bluck et al., 2014; Schawinski et al., 2014). A preliminary investigation found that clear visual difference between star-forming and passive E5 galaxies is not captured by any of the available morphological and kinematic features calculated by Thob et al. (2019). I note, though, that even with the benefit of a morphological or kinematic feature that could distinguish between star-forming and passive E5 galaxies, star-forming E5 galaxies would still stand out as an unrealistic subpopulation.

## 4.3 Summary and conclusions

In this chapter, I present a novel method for the multi-dimensional validation of cosmological galaxy simulations against observations. The method is built around the use of clustering. Specifically, I use the `k-means` clustering algorithm to partition a sample of $3,724$ observed galaxies from the GAMA survey and a sample of $7,117$ simulated galaxies from the EAGLE models. Clustering is conducted within a five-dimensional feature space, common to both samples and defined by: stellar masses, specific (fractional) dust masses, specific star formation rates, half-light radii, and bulge-to-total ratios. The feature space is rescaled ahead of the clustering to ensure that the different numerical ranges of each of the features do not artificially bias the clustering. Clustering outcomes are validated with respect to their stabilities, to ensure that the outcomes are reproducible, and their compactnesses, which is a general aim of clustering. I find that the GAMA sample is best partitioned into five clusters, and the EAGLE sample into seven clusters (Figure 4.2). In order to compare these outcomes, I map the clusters determined within EAGLE onto the GAMA sample. By doing this, I can quantitatively examine how successful EAGLE has been in reproducing GAMA galaxies on a subpopulation level, and connect the properties of galaxies contained by equivalent clusters. Thusly, I make the following conclusions:

1. Comparing the GAMA clusters with a partition of the GAMA sample that is based on the EAGLE clusters (Table 4.3), I calculate $V_a = 0.76$. This, on the basis of Appendix B.1, indicates agreement between the two clustering outcomes in terms of their broader structures (i.e. the overall bimodality of galaxies), but differences between them at the substructure level that are expressed by individual clusters.

2. A particular strength of my approach is that features do not need to correspond to one another exactly in terms of their definitions. Though there is an offset between the morphological bulge-to-total ratios that I use for the GAMA sample and the kinematic bulge-to-total

ratios that I use for the EAGLE sample, a combination of my use of clustering with my choice to rescale the shared feature space of the samples means that I am still able to compare the disciest GAMA galaxies with the disciest EAGLE galaxies. The offset between the sizes of the largest galaxies in each sample is similarly mitigated in terms of its influence upon the clustering outcomes.

3. While EAGLE successfully captures the growth of the stellar masses of SFMS galaxies, it does not capture the observed growth of their morphological components. While the bulge-to-total ratios of GAMA galaxies increase along the SFMS (Figure 4.3), those of EAGLE galaxies remain the constant or decrease slightly (Figure 4.4). This is due to limitations of the resolution of EAGLE, which mean that low-mass galaxies exhibit excess dispersion support. Examining whether the observed trend is recovered with improvements to resolution will constitute a crucial test of the next generation of cosmological galaxy simulations.

4. EAGLE produces more low-mass, star-forming, bulge-dominated galaxies than are observed in the GAMA sample (Figures 4.3 and 4.4), such that they are modelled with two clusters rather than one. The resolution of EAGLE is likely to be at least partially responsible for this difference. Mergers appear to be significant in the formation of the galaxies in these clusters (Figure 4.6), consistent with previous results from the literature.

5. There are two clusters within the EAGLE sample in which the majority of galaxies are satellites in massive haloes (Figure 4.7). By comparison with a single cluster in the GAMA sample, they accurately capture the influence of external processes on galaxies in groups, particularly at lower stellar masses (Table 4.3). Discrepancies at higher stellar masses are driven by the lack of a turndown in EAGLE's SFMS, such that satellite quenching appears to extend to slightly higher stellar masses in EAGLE than is observed. The spread of morphologies among the galaxies in these clusters (Figs 4.3, and 4.4) indicates that their quenching is not always accompanied by a morphological transformation, and is therefore most likely to be due to processes like ram-pressure stripping and starvation.

6. While EAGLE recovers the high-mass, passive, spheroids that are observed in the GAMA sample, they are grouped in a cluster with a subpopulation of high-mass, star-forming, bulge-dominated galaxies that are not seen in the real Universe (Figures 4.3 and 4.4). These simulated star-forming and passive galaxies differ in their present day gas content (Figure 4.9) and in their visual appearances (Figure 4.10), and suggest that AGN feedback is not potent enough to remove the gas from those galaxies that are still star-forming. Otherwise, the assembly histories of the star-forming galaxies cannot be distinguished from those of the passive galaxies; it is clear that a specific study of differences their evolutionary pasts is required to explain the present day differences in their star formation activity.

This chapter has demonstrated the utility of clustering for the multi-dimensional validation of simulations against observations, and its ability to highlight particular features, and regions of the shared feature space, for analysis. EAGLE reproduces the global structure of the GAMA sample, but exhibits notable differences at more local scales. These differences are attributed to specific aspects of the EAGLE simulations (their resolution and their AGN feedback prescription), thus providing clear targets for improvement in the next generation of cosmological galaxy simulations. I emphasise that my approach may readily be adapted for use with any clustering algorithm, with prototype- and model-based algorithms being the most natural options.

My approach also has the potential to link observable features (such as those used in this chapter) with features that cannot be directly observed, and are hence only available for measurement in simulations. An example of such a feature is the gravitational binding energy of the halo of a central galaxy, which sets the threshold that AGN feedback must exceed in order to be able to expel gas from a galaxy (Davies et al., 2019, 2020b; Oppenheimer et al., 2020). Clustering within feature spaces that are defined by features like this, which are more directly linked with the processes involved in galaxy evolution, would be expected to produce outcomes that more closely trace the influence of these processes. The subsequent mapping of these outcomes onto observable features space would then facilitate and understanding of how these "hidden" features are connected to the observable features via evolutionary processes.

# Chapter 5

# Synergies between low- and intermediate-redshift galaxy populations revealed with unsupervised machine learning

The work presented in this chapter and in Appendix C is the subject of a paper that has been submitted to Monthly Notices of the Royal Astronomical Society. See also the Publications page of the front matter of this thesis.

In this chapter, I describe work that builds on that of Siudek et al. (2018b), which is described in Section 2.3. I adapt the approach of Siudek et al. (2018b) to compare samples of galaxies at two different redshifts. The aim is to uncover substructures within the established bimodality, to examine similarities and differences between these substructures at two cosmic epochs, and to interpret these similarities and differences in the context of galaxy evolution. My sample of galaxies at low redshift ($z \sim 0.06$) is drawn from the second edition of the GALEX-SDSS-WISE Legacy Catalogue (GSWLC-2; Salim et al. 2018), and my sample of galaxies at intermediate redshift ($z \sim 0.65$) is based on the VIPERS sample of Siudek et al. (2018b). I prepare the samples carefully to ensure a fair comparison of galaxies from different cosmic epochs and different surveys, and to mitigate methodological influences on the clustering outcomes. I also adjust the input features, defining nine neighbouring rest-frame colours that, together, represent the shapes of the UV-through-NIR SEDs of the galaxies in the samples.

This chapter proceeds as follows. In Section 5.1, I characterise the samples, the data I use to represent and analyse the galaxies that they contain (including the estimation of their SEDs), and

the measures that I take to ensure a fair comparison between them. In Section 5.2, I explain the Discriminative Latent Mixture (DLM) model and how Subspace Expectation-Maximisation (`SEM`) algorithm implements it, and I describe the feature space within which I cluster my samples. In Section 5.3, I present the outcomes of the clustering, and in Section 5.4, I offer my interpretation thereof. Finally, in Section 5.5, I summarise and conclude this chapter. This chapter is supplemented by Appendix C, in which I explore aspects of the use of `SEM` (Sections C.1 and C.3), detail my smoothing of GSWLC-2 galaxy colours (Section C.2), and highlight an AGN trend among GSWLC-2 clusters (Section C.4). Where required in this chapter, I assume a ($H_0$, $\Omega_m$, $\Omega_\Lambda$) = (70 km s$^{-1}$ Mpc$^{-1}$, 0.3, 0.7) cosmology.

## 5.1 Samples

### 5.1.1 GALEX-SDSS-WISE Legacy Catalogue 2

The second edition of the GALEX-SDSS-WISE Legacy Catalogue (GSWLC-2; Salim et al. 2016, 2018) was assembled using Data Release 10 (DR10; Ahn et al. 2014) of SDSS. GSWLC-2 aimed to characterise the star formation activity and dust content of galaxies in the local Universe. It contains all SDSS DR10 galaxies that meet the following criteria:

- apparent $r$-band petrosian magnitudes $< 18$,
- spectroscopic redshifts within the range $0.01 < z < 0.3$,
- lie within the Galaxy Evolution Explorer (GALEX) (Martin et al., 2005; Morrissey et al., 2007) observation footprint, whether they were detected by GALEX or not.

The lower redshift limit was imposed to exclude foreground stars, and particularly close galaxies with potentially unreliable photometry and/or distance estimates. Retaining galaxies that were not actually detected by GALEX itself preserves the optical selection of SDSS. In all, these criteria select 659, 229 SDSS DR10 galaxies.

$u$-, $g$-, $r$-, $i$-, and $z$-band optical photometry for galaxies in GSWLC-2 was drawn from SDSS. `modelMag` magnitudes, which are based on profile fits, were selected due to the accuracy of their colours. These `modelMag` magnitudes were corrected for extinction due to Milky Way dust using the empirical Yuan et al. (2013) coefficients.

The SDSS optical photometry was supplemented with near- ($NUV$) and far-UV ($FUV$) photometry from GALEX's final data release (GR6/7). GALEX conducted surveys at varying depths:

an All-sky Imaging Survey (which observed several targets per orbit), a Medium Imaging Survey (one target per orbit), and a Deep Imaging Survey (several orbits per target). These surveys were nested, such that it is possible for a galaxy to have been observed at more than one depth (although an observation of a galaxy at a given depth does not guarantee an observation of the same galaxy at shallower depths). Here, the UV photometry for galaxies in GSWLC-2 based on the deepest available observation of each galaxy (catalogue GSWLC-X2) is used. Salim et al. (2016) applied corrections to mitigate systematic offsets between the SDSS and GALEX photometry, which arose mostly due to the blending of sources in GALEX's low-resolution images. Peek & Schiminovich (2013) corrections for extinction due to Milky Way dust were applied to the UV photometry. UV photometry in at least one of GALEX's two bands (almost always $NUV$ if just one) is available for 65 per cent of GSWLC-2 galaxies, and for 80 per cent of the galaxies in the final GSWLC-2 sample (see below).

Wide-field Infrared Survey Explorer (WISE; Wright et al. 2010) observations at 12 and 22 $\mu$m (channels W3 and W4 respectively) were used to provide mid-IR (MIR) photometry for GSWLC-2 galaxies. Salim et al. (2018) opted for unWISE (Lang et al., 2016) forced photometry, which was based directly on SDSS source positions and profiles. MIR photometry in at least one of channels W3 and W4 is available for 78 per cent of GSWLC-2 galaxies, and for 87 per cent of the galaxies in the final GSWLC-2 sample (see below).

**GSWLC-2 rest-frame SEDs**

The rest-frame SEDs of GSWLC-2 galaxies were estimated using the Code Investigating GALaxy Emission (`CIGALE`; Noll et al. 2009; Boquien et al. 2019). Synthetic spectra generated by `CIGALE` were validated against the available observed UV-through-optical photometry in order to constrain the SEDs. Details of this fitting procedure are described at length in Salim et al. (2016, 2018); here, I offer a brief summary.

Synthetic spectra were generated using Bruzual & Charlot (2003) simple stellar population templates, based on a Chabrier (2003) initial mass function and with metallicities of $\log_{10}(Z) = -2.4$, $-2.1$, $-1.7$ ($\sim Z_{\odot}$), or $-1.3$. These templates were combined with Myr-resolution SFHs consisting of two exponentially declining episodes of star formation, which produce an old and a young population. Absorption of stellar emission by dust was implemented via a Noll et al. (2009) generalisation of the Calzetti et al. (2000) attenuation curve, modified to allow its slope to vary and to add a UV bump (see section 3.4 of Salim et al. 2018).

The SED estimation was additionally constrained by the galaxy's total IR luminosity (i.e. matching the energy absorbed by the dust in that galaxy with the energy it re-emits; see section 3.2 of Salim et al. 2018). Total IR luminosities were derived from the 22 $\mu$m WISE photometry (if avail-

able, 12 $\mu$m if not) using Chary & Elbaz (2001) templates, further corrected based on Herschel (Valiante et al., 2016) IR photometry (see section 3.1 of Salim et al. 2018). The overall quality of fit was measured by its reduced chi-squared value ($\chi^2_r$).

Astrophysical features including rest-frame absolute magnitudes, colour excesses [$E(B-V)$], stellar masses ($M_*$), stellar metallicities ($Z$), mass-weighted stellar ages ($MWSA$), and specific star formation rates [$sSFR$ (SED)] were derived from the full ensemble of possible synthetic spectra via a Bayesian approach (Salim et al., 2007). The likelihood of the fit of each synthetic spectrum to the photometry of each galaxy was used to generate a probability density function for each feature, with the likelihood-weighted means of the functions being quoted as the best estimates of the features, and the likelihood-weighted standard deviations as the errors.

**Final low-redshift sample**

The final GSWLC-2 sample is subject to the following selections. Firstly, only galaxies whose best-fitting `CIGALE` SEDs produce $\chi^2_r <= 11.07$ (i.e. the mean plus two standard deviations of the logarithmic GSWLC-2 distribution in $\chi^2_r$) are retained, in order to omit particularly poorly constrained fits. Spectroscopic redshifts are limited to the range $0.02 < z < 0.08$, and stellar masses (as estimated via Bayesian analysis of the synthetic `CIGALE` spectra) to $> 10^{9.5}$ $M_\odot$. These two restrictions ensure completeness above the imposed stellar mass limit. Finally, broad-line AGN are removed by asserting `flag_sed` = 0. The final GSWLC-2 sample has a median redshift of 0.06 and contains 177, 362 galaxies.

Brinchmann et al. (2004) specific star formation rates [$sSFR$ (ind.)] and 4000 Å break strengths [$D(4000)$] are invoked as additional, `CIGALE`-independent indicators of the stellar populations in GSWLC-2 galaxies. The SFRs sum two components: a spectroscopic fibre SFR, and a photometric SFR outside the fibre, given by an optical SED fit (Salim et al., 2007). The fibre SFR is given by either a H$\alpha$ calibration (Charlot & Longhetti, 2001) or, in the case of spectra that have a contribution from an AGN, a $D(4000)$-based estimate (itself calibrated on the emission lines of pure star-forming galaxies). These SFRs are then normalised by photometrically-determined stellar masses to give $sSFR$ (ind.). The timescale probed by $sSFR$ (ind.) lies between the 10 Myr timescale of the H$\alpha$-calibrated fibre SFRs, and the 1 Gyr timescale of optical SED-based SFRs (Salim et al., 2016). The $D(4000)$ measurements apply to fibre region only. Both of these features are available for 97 per cent of the galaxies in the GSWLC-2 sample.

Sérsic indices ($n_g$) and circularised half-light radii ($R_{1/2}$) for the galaxies in the GSWLC-2 sample are obtained from catalogues assembled by Simard et al. (2011). Both were derived from fits of singular Sérsic (1963, 1968) profiles to $r$-band images of galaxies in SDSS. The Sérsic indices have minimum and maximum allowed values of 0.5 and 8 respectively. Sérsic indices and half-

light radii are available for 96.2 per cent of the galaxies in the final GSWLC-2 sample. Simard et al. (2011) $r$-band bulge-to-total ratios ($B/T_r$) are also used for these galaxies, which were based on fits consisting of two components: a Sérsic bulge (fixed at an index of 4) and an exponential disc. Local environmental densities, available for 92.1 per cent of GSWLC-2 galaxies, come from Baldry et al. (2006). They averaged the surface densities of SDSS galaxies with respect to their fourth- and fifth-nearest density-defining neighbour within $1,000$ km s$^{-1}$ along the line of sight. Local overdensities ($\delta$) are calculated using $\delta = (\Sigma - \bar{\Sigma})/\bar{\Sigma}$, where $\Sigma$ is the local surface density and $\bar{\Sigma}$ the average surface density of the sample.

### 5.1.2 VIPERS

The VIMOS Public Extragalactic Redshift Survey (VIPERS; Guzzo et al. 2014; Garilli et al. 2014; Scodeggio et al. 2018) aimed to match the statistical fidelity of low-redshift surveys like SDSS, but at intermediate redshifts ($z \sim 0.7$). The survey was conducted using the VIMOS spectrograph (Le Fèvre et al., 2003) of the European Southern Observatory's Very Large Telescope. Its targeting was based on the Canada-France-Hawaii Telescope Legacy Survey Wide (CFHTLS-Wide) photometric catalogue[1], with objects qualifying for VIPERS if they had extinction-corrected $i$-band magnitudes $i_{AB} < 22.5$. An additional *ugri* colour cut was applied to remove low-redshift ($z \lesssim 0.5$) galaxies from the survey (Guzzo et al., 2014). PDR2, the second and final public data release of VIPERS, comprises spectroscopy for $97,414$ objects (Scodeggio et al., 2018). $52,114$ of these objects ($51,522$ galaxies and 592 broad-line AGN) have "secure" (> 99 % confidence) redshifts. This secure-redshift sample was the subject of the Siudek et al. (2018b) study, and is the basis of my present VIPERS sample.

Photometry for this sample was taken from a catalogue prepared by Moutard et al. (2016a). The CFHTLS-Wide photometric catalogue (i.e., the basis of the targeting for VIPERS) provided optical photometry for this sample in $u*$, $g$, $r$, $i$, and $z$ bands. Moutard et al. (2016a) derived total magnitudes for the galaxies in this sample by rescaling their isophotal magnitudes. These isophotal magnitudes were chosen for the accuracy of their colours with a view to photometric redshift estimation; this choice now benefits the SED estimation as well.

Like for the GSWLC-2 sample, UV photometry came from GALEX. Moutard et al. (2016a) supplemented existing Deep Imaging Survey observations of VIPERS galaxies with deep GALEX observations of their own in order to improve UV coverage within the VIPERS footprint. Coverage is complete in the W1 field of VIPERS, but not in the W4 field (see figure 1 of Moutard et al. 2016a). UV photometry was then measured using a Bayesian approach with the $u*$-band profiles of galaxies as priors (Conseil et al., 2011), which mitigated the confusion of sources due to their

[1] `http://www.cfht.hawaii.edu/Science/CFHLS`

blended UV profiles. UV photometry in at least one of GALEX's two bands (almost always $NUV$ if just one) is available for 52 per cent of galaxies in the Siudek et al. (2018b) sample and in the final VIPERS sample (see below).

Near-IR (NIR) $K_s$-band photometry came from a dedicated CFHT WIRCam (Puget et al., 2004) follow-up survey of VIPERS galaxies (Moutard et al., 2016a). This $K_s$-band photometry was validated against NIR photometry from the VISTA Deep Extragalactic Observations (VIDEO) survey (Jarvis et al., 2013), exhibiting good agreement. VIDEO survey $Z$, $Y$, $J$, $H$, and $K_s$ NIR photometry is also taken for the sample where available (11 per cent of the Siudek et al. 2018b sample, 10 per cent of the final VIPERS sample; see below). CFHT $K_s$-band photometry is available for 91 per cent of galaxies in the Siudek et al. (2018b) sample, and for 93 per cent of galaxies in the final VIPERS sample (see below).

#### VIPERS rest-frame SEDs

The SEDs of VIPERS galaxies are estimated via a full fit of synthetic `CIGALE` spectra to the available UV-through-NIR photometry. This differs slightly from the method used for the GSWLC-2, whose NIR SEDs were constrained not by their shapes but simply by their total IR luminosities (Section 5.1.1). While the same stellar templates (Bruzual & Charlot 2003, with Chabrier 2003 initial mass functions and metallicities of 0.004, 0.008, 0.02, or 0.05) are used for VIPERS as were used for GSWLC-2 , the SFHs are adjusted to reflect the change in cosmic epoch between samples and to account for the possibility of very recent bursts of star formation[2]. Astrophysical features are derived for VIPERS galaxies using the same Bayesian approach as for GSWLC-2 galaxies (see Section 5.1.1).

#### Final intermediate-redshift sample

The following selections are made to yield the final VIPERS sample: galaxies are kept if the $\chi^2_r$ of their best-fitting `CIGALE` SED has a value less than or equal to the mean plus two standard deviations (= 18.85) of the overall logarithmic VIPERS distribution. Spectroscopic redshifts are restricted to being within the range $0.5 < z < 0.8$, balancing my intent to define a co-eval population of galaxies against the need to keep the sample as large as possible. Like the GSWLC-2 sample, stellar masses are limited to $> 10^{9.5}$ $M_\odot$ with a view to mass completeness (though see Sections 5.3.4 and 5.4.2, where I discuss shortcomings). Broad-line AGN and serendipitous secondary spectral sources are removed using `zflag` $< 10$. Ultimately, this gives a final VIPERS

[2]Consequences of this adjustment are discussed in Section 5.3.4; the properties of most VIPERS galaxies appear accurate, except for those of a subpopulation of passive VIPERS galaxies.

sample consisting of 31,889 galaxies, with a median redshift of 0.65.

Emission-line SFRs, which are independent of the CIGALE SED estimation, were calculated from the [OII] $\lambda$3727 forbidden doublet fluxes of the galaxies in the VIPERS sample using the calibration (which includes empirical stellar-mass-based corrections) of Gilbank et al. (2010, 2011a,b). These [OII] $\lambda$3727 fluxes are available for 27,537 of the galaxies in the VIPERS sample, and they probe short timescales of star formation ($\sim$ 10 Myr). These [OII] SFRs are normalised by the CIGALE stellar masses to yield specific star formation rates[3] [*sSFR* (ind.)]. $D(4000)$ was measured from VIPERS spectra by Garilli et al. (2014), using the same Balogh et al. (1999) method as was used for SDSS (Brinchmann et al., 2004). Sérsic indices and circularised half-light radii for the galaxies in the VIPERS sample are given by Krywult et al. (2017), who fitted the *i*-band light distributions of galaxies with single Sérsic (1963, 1968) profiles. These features are available for 96.2 per cent of the galaxies in the final VIPERS sample. The Sérsic indices are winsorised to values of 0.5 and 8 in order to match the GSWLC-2 sample. The overdensities of 91.7 per cent VIPERS galaxies were derived by Cucciati et al. (2017), based on fifth-nearest neighbour surface densities.

## 5.2 Clustering method

I apply the Subspace Expectation-Maximisation algorithm, which estimates the parameters of the Discriminative Latent Mixture model. Bouveyron & Brunet (2012) offer full, rigorous, mathematical derivations of both the Discriminative Latent Mixture model and the Subspace Expectation-Maximisation algorithm in their paper (where it is called the Fisher-EM algorithm)[4]; here, I offer brief summaries of the model (Section 5.2.1), and of its implementation via the algorithm (Section 5.2.2). In Section 5.2.3, I discuss some additional relevant practicalities to the use of the model and algorithm, and in Section 5.2.4, I describe the shared feature space within which I cluster the two samples.

### 5.2.1 The Discriminative Latent Mixture model

The Discriminative Latent Mixture (DLM) model is a clustering approach that incorporates dimensionality reduction on the fly to determine a frugal fit to the structure of an input sample, which is assumed to consist of $k$ clusters. Selection of the value of $k$ is discussed in Section 5.2.3.

[3] My use of stellar masses given by CIGALE means that these *sSFR* (ind.) estimates are not entirely independent of CIGALE, however CIGALE's stellar masses are expected be consistent with those estimated via other methods, given that stellar mass estimates are generally quite robust (Bell & de Jong, 2001).

[4] Note that I have renamed their algorithm for this thesis to offer a clearer indication of how aspects of the algorithm work.

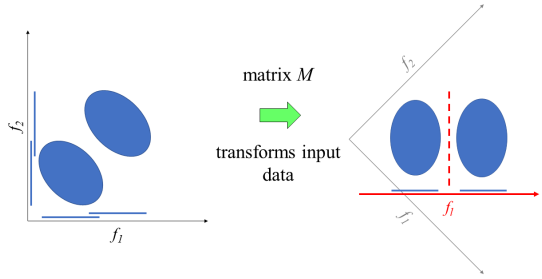


**Figure 5.1:** A simple demonstration of the principles behind subspace clustering. Here, a sample consisting of two clusters (represented by the two blue ellipses) is represented in a two-dimensional full space defined by features $f_1$ and $f_2$. Matrix $M$ enables the transformation of the sample to a one-dimensional subspace, defined by latent feature $f_l$, in which the two clusters are easily discriminated.

The key premise of the DLM model is thus: a sample represented in a $D$-dimensional space that is defined by observed features actually occupies an intrinsic $d$-dimensional subspace ($d < D$; the "empty space phenomenon"; Scott & Thompson 1983) that is defined by unobserved, latent features. Hence, the clustering structure of the sample should be fitted in this intrinsic subspace.

The subspace has two important properties in the context of the DLM model. Firstly, of all possible $d$-dimensional subspaces, it is the one that best discriminates the $k$ clusters in the sample. The model assumes $1 \leq d \leq k - 1$: that $k$ clusters may be distinguished in $k - 1$ dimensions or fewer (see Section 5.2.3 for further explanation). Secondly, the subspace is linearly related to the full $D$-dimensional space, such that the unobserved, latent features are linear combinations of the observed features. Hence there exists a matrix $M$, common to all of the $k$ clusters, that enables the transformation of the sample between the full space and the subspace. This transformation matrix is constrained by the condition that the basis vectors of the subspace must be orthonormal. Estimation of the transformation matrix $M$ is explained in Section 5.2.2. Selection of the value of $d$ is explained in Section 5.2.3. Figure 5.1 demonstrates these two important properties of the subspace.

The DLM model assumes that the sample is distributed among a mixture of $k$ Gaussian density functions within the discriminative latent subspace. The functions, each of which corresponds to a cluster, are defined by three parameters: a mean vector ($\mu_k$), a covariance matrix ($\Sigma_k$), and a scalar relative mixture proportion ($\pi_k$). The matrix $M$ enables the transformation of these parameters back to the full space. For the covariances, this includes the addition of Gaussian "noise" ($\delta_k$; unique to each of the clusters), which is defined as non-discriminative structure that exists in the full space but not in the subspace. While $\Sigma_k$ captures the cluster covariances inside the discriminative latent subspace, $\delta_k$ captures the cluster covariances outside the subspace. Full space covariances are the sum of both. Estimation of the cluster means, covariances, and noise terms is discussed in Section 5.2.2.

Implementation of the DLM model hence requires the estimation of the following parameters:

- $k - 1$ relative mixture proportions ($\pi_k$; given that one cluster has a proportion of 1);
- $kd$ parameters for the mean vectors ($\mu_k$) in the subspace;
- $kd(d + 1)/2$ parameters for the covariance matrices ($\Sigma_k$) in the subspace (fewer than $kd^2$ parameters because covariance matrices are symmetric);
- $d(D-(d+1)/2)$ parameters for the transformation matrix $M$ (the number of free parameters, given the constraint that the basis vectors of the subspace must be orthonormal);
- $k$ noise terms ($\delta_k$; given that this non-discriminative structure is Gaussian and spherical and may therefore by parametrised by a single value in reference to the Gaussian density function estimated for each cluster).

The total number of parameters ($q_{DLM}$) is most strongly influenced by the value of $d$. The maximum $q_{DLM}$ at a certain combination of $D$ and $k$ is given by setting $d$ to its maximum value of $k - 1$ (based on the aforementioned assumption that $k$ clusters may be distinguished in $k - 1$ dimensions or fewer). $q_{DLM}$ is smaller than the number of parameters that must be estimated for a Gaussian Mixture Model in the full space ($q_{GMM}$), especially if $d << D$. $q_{GMM}$ is given by the sum of $k - 1$ relative mixture proportions, $kD$ parameters for the mean vectors, and $kD(D + 1)/2$ parameters for the covariance matrices.

Parameter $q_{DLM}$ may be further reduced by imposing additional constraints upon the DLM model. For example, the covariance matrices ($\Sigma_k$) may be assumed to be the same for all Gaussians ($\Sigma$; the Gaussians all have the same shape). Alternatively, they may be assumed to be diagonal ($\alpha_{k,j}$, where the subscript $j$ indicates a different variance in each dimension of the subspace), meaning the latent features that define the subspace are uncorrelated. These diagonal covariance matrices may then also be assumed to be isotropic ($\alpha_k$; spherical Gaussians in the subspace), the same for all Gaussians ($\alpha_j$), or both ($\alpha$). The noise terms ($\delta_k$) may be assumed to be the same for all Gaussians ($\delta$) as well. Constraints like these may be imposed to speed up the clustering, in anticipation of a particular clustering structure, or (as in my case) to compare fits of models of varying complexities (see also Section 5.2.3). The various combinations of these constraints on the covariance matrices and noise terms yield 11 submodels of the full $\Sigma_k$, $\delta_k$ DLM model. They are listed in full in table 1 of Bouveyron & Brunet (2012) (and listed partially in Table 5.1 of this chapter).

### 5.2.2 The Subspace Expectation-Maximisation algorithm

The Subspace Expectation-Maximisation algorithm (`SEM`) estimates the parameters ($\pi_k$, $\mu_k$, $\Sigma_k$, $M$, $\delta_k$) of the DLM model, fitting a sample of $N$ observations, observed in a $D$-dimensional space (the "full" space, defined by $D$ observed features), with $k$ Gaussian density functions in a $d$-dimensional discriminative latent subspace ($1 \leq d \leq k-1$). `SEM` comprises the following steps:

0. Initialisation: $k$ starting points are selected from within the extent of the sample in the full space;
1. Expectation (E): transform the parameters of the mixture of Gaussians to the full space, and calculate the probability of each observation having originated from each Gaussian;
2. Subspace (S; based on discriminant analysis): using the observation probabilities, find the subspace that best separates the Gaussians;
3. Maximisation (M): update the parameters of the mixture of Gaussians (including non-discriminative structure, termed "noise") within the subspace.

The Expectation, Subspace, and Maximisation steps are iterated such that `SEM` improves its estimates of the DLM model parameters as it proceeds. `SEM` is slow to run on my large samples and, unlike traditional Expectation-Maximisation algorithms, does not always converge perfectly (such that there are no changes between successive iterations; due to the Subspace step). I therefore terminate `SEM` at the completion of 25 iterations; changes between iterations become negligible well before this number (see Section C.1). The final output of `SEM` is a series of $k$ probabilities for each of the observations: probabilities of each observation having originated from each of the $k$ Gaussians. Final cluster labels are given by assigning each observation to the Gaussian with the highest probability of having originated it.

While successive iterations of `SEM` improve its estimates of the DLM model parameters, these estimates improve only towards local maxima in their likelihood functions. `SEM` is hence run with varying initialisations, which may intuitively be considered as "exploring the surfaces" of the likelihood functions of the model parameters. This encourages optimisation towards different local maxima and, hopefully among these, the global maximum, which corresponds to the very best estimates of the DLM model parameters.

Initialisation techniques may be as simple as a uniform random selection of $k$ observations from the sample. I opt to use the `k-means` algorithm (MacQueen, 1967; Lloyd, 1982), which implements a simple centroid-based clustering approach, to generate initialisations for `SEM`. `k-means` is an Expectation-Maximixation algorithm and, like `SEM`, only optimises to local maxima. I therefore initialise `k-means` *itself* 100 times in the hope of encouraging optimisation towards the global

maximum of *its* objective function (which measures how separated the clusters are). Use of varying initialisations provided by a heuristic like `k-means` leads to "pre-optimisation" of `SEM` because the separated centroids are likely to span the full extent of the sample in its full space. This facilitates improvement of `SEM`'s estimates of the DLM model parameters towards the global maximum of their likelihood functions. Following this initialisation, `SEM` proceeds to the Subspace step, in which it finds the subspace that best separates the final `k-means` clusters, and to the Maximisation step, in which it fits the observations with a mixture of Gaussians within this subspace. `SEM` then loops back around to the Expectaton step and begins iterating proper.

The Expectation step uses the parameters estimated in the Maximisation step ($\pi_k$, $\mu_k$, $\Sigma_k$, $\delta_k$) to calculate the conditional probability of each observation having originated from each of the $k$ Gaussians. These parameters are transformed from the subspace, within which they are estimated in the Maximisation step, to the full space using matrix $M$, found in the Subspace step.

The Subspace step finds the $d$-dimensional discriminative latent subspace that best separates the new partition calculated in the Expectation step. Bouveyron & Brunet (2012) base this step on discriminant analysis, which finds the linear combination of the input features that maximises the ratio of the scatter *between* clusters to the scatter *within* clusters. Similar principles have been applied for the visualisation of multi-dimensional clusters as well (e.g. Lisboa et al. 2008). These scatters are weighted by the probabilities calculated in the Expectation step. A constraint of the DLM model is that the $d$ basis vectors that define the subspace must be orthonormal, which is not necessarily a property of the $d$ basis vectors that linear discriminant analysis provides. Bouveyron & Brunet (2012) assert this constraint by applying the orthonormal discriminant vector method (Okada & Tomita, 1985). The orthonormal discriminant vector method uses linear discriminant analysis to find the $d$ basis vectors in succession while also ensuring the orthonormality of each new basis vector with respect to all of those that have already been calculated. The first basis vector, which is free of this constraint, is given by the direct application of linear discriminant analysis to the sample in the full space. The $d$ orthonormal basis vectors constitute the columns of $M$, the matrix that enables the transformation of the sample between the full space and the subspace.

The Maximisation step updates the estimates of the means, covariances, and relative mixture proportions ($\pi_k$, $\mu_k$, $\Sigma_k$) of the $k$ Gaussians in order to maximise the likelihood of the fit. These estimates are measured within the subspace found in the Subspace step, and are weighted by the probabilities calculated in the Expectation step. This step also updates the estimates of the noise terms ($\delta_k$), which is given by the differences between the full-space variances (again weighted by the probabilities calculated in the Expectation step) and the newly updated subspace variances.

### 5.2.3 Practicalities

I do not presume a DLM submodel or value of $k$ with which to fit the samples. Instead, I conduct a search over all of the DLM submodels and over a range of values of $k$ to determine the best-fitting combination. Three of the DLM submodels ($\alpha_j$, $\delta_k$; $\alpha_j$, $\delta$; $\alpha$, $\delta_k$) are not available for use in the version of `SEM`[5] that I use for the fitting. This reduces the total number of available submodels from 12 (including the full $\Sigma_k$, $\delta_k$ model) to nine.

I identify the best-fitting combination of DLM submodel and value of $k$ by using the Integrated Completed Likelihood criterion (ICL; Biernacki et al. 2000):

$$\mathrm{ICL} = \ln(L) - \frac{q_{DLM}}{2}\ln(N) - [-\Sigma_{i=1}^{N}\,\Sigma_{l=1}^{k}\, z_{i,l}\ln(p_{i,l})], \tag{5.1}$$

where $L$ is the likelihood of the fit, $p_{i,l}$ is the probability of observation $i$ belonging to cluster $l$, and $z_{i,l}$ denotes cluster membership, taking a value of 1 when $p_{i,l} = \max(p_{i,:})$ and a value of 0 otherwise. The ICL is closely related to the popular Bayesian Information criterion (BIC; Schwarz 1978). While both the BIC and ICL criteria penalise the likelihood using the number of parameters used for the fit (to avoid over-fitting), the ICL criterion also rewards separated clusters (a general aim of clustering). The combination of submodel and $k$ that returns the highest ICL score is deemed the the best fit.

The dimensionality of the discriminative latent subspace is constrained by the number of clusters being fitted: $1 \leq d \leq k-1$. The maximal $d = k-1$ case may intuitively be understood as setting the origin of the subspace at one of the $k$ cluster centres so that the full-space vectors to each of the remaining $k-1$ cluster centres define the basis vectors of the subspace. If multiple clusters lie along the same direction in the full space, the number of basis vectors needed to define the subspace is reduced. In my application of `SEM`, I hold $d$ at its maximum value of $k-1$. This is recommended by Bouveyron & Brunet (2012) to avoid omitting any discriminative structure from the subspace and to ease convergence of `SEM` (which may become unstable or fail to converge if $d$ is too small in comparison with $k$ and/or $D$). Hence, the maximum value of $k$ in my model selection search is 9 (set by $d = 8$, given $D = 9$).

### 5.2.4 Input features to the clustering

The fitting of the clustering structures of both of the samples is conducted within a nine-dimensional feature space defined by nine colours. These colours are calculated not from the observed photometry that is used as input to the SED fitting, but from rest-frame magnitudes estimated by `CIGALE`.

[5] Version 1.5.1, for the `R` statistical computing environment.

This ensures homogeneity among the input features, and that the feature space is defined by rest-frame colours (which is more difficult to ensure using colours calculated directly from observed photometry). In addition, the SED estimation can infer the rest-frame magnitudes of galaxies in bands for which there is no observed photometry. The full list of rest-frame colours used for the clustering is: $FUV - NUV$, $NUV - u$, $u - g$, $g - i$, $i - r$, $r - z$, $z - J$, $J - H$, and $H - K_s$. These rest-frame colours are intended to represent the shape of each galaxy's UV-through-NIR SED, and to remove the influence of the intrinsic brightnesses of the galaxies on the clustering outcomes. The rest-frame magnitudes of GSWLC-2 galaxies (but not VIPERS galaxies) are subject to some smoothing (see Section C.2). In addition, the rest-frame NIR colours of GSWLC-2 galaxies were inferred from UV and optical photometry (given the lack of input NIR photometry). Use of the term "colour" from this point forward in this chapter is intended in reference to these rest-frame colours, as estimated by `CIGALE`.

These colours differ from those used by Siudek et al. (2018b); they used rest-frame colours defined with reference to the rest-frame $i$-band magnitudes of galaxies ($FUV - i$, $NUV - i$, etc.), also with the aim of removing the influence of galaxy intrinsic brightnesses on their clustering outcomes. However, their UV colours, defined across the largest distances in wavelength among their features, exhibited large spreads (up to a factor of 10 larger than the spreads of other colours) and dictated much of their clustering. Preliminary tests of clustering with these $i$-band based colours for my present, carefully prepared samples confirmed this. The $\alpha_{k,j}$ and $\alpha_{k,j}$ submodels achieved the highest ICL scores for these $i$-band colours, but gave only relatively crude segmentations of the samples (see also Section C.3). My colours, defined using magnitudes in filters at neighbouring effective wavelengths, mitigate this effect and encourage `SEM` to converge to more detailed partitions (although, as shown in Figure 5.2, bluer colours are still most important).

## 5.3 Results

### 5.3.1 `SEM` submodel selection

As outlined in Section 5.2.3, I conduct a search for the best-fitting `SEM` submodel and number of clusters for both of the samples. I identify the best-fitting combination using the ICL criterion, which penalises the likelihood using the number of parameters of the submodel while favouring separated clusters. Table 5.1 lists ICL scores reported for both samples. The uncertainties on these scores, which span the *full* variation over 100 initialisations, show that `SEM` is extremely stable and self-consistent. The best-fitting combinations for each sample are highlighted using bold text. I briefly describe patterns of behaviour of the various submodels, including the large spread in ICL scores, in Section C.3. Despite it registering the highest score for the GSWLC-2 sample, I reject

**Table 5.1:** Integrated Completed Likelihood (ICL) scores reported by my search over all possible combinations of submodel (see Section 5.2.1 for further explanation) and $k$ for both samples. The uncertainties span the full range of ICL scores registered over 100 initialisations for each combination. As mentioned in Section 5.2.3, only nine of the 12 submodels are available in the version of `SEM` that I use for my fitting. The score of the best-fitting combinations are highlighted using bold text. While submodel $\Sigma$, $\delta$ produces the highest score for the GSWLC-2 sample (at $k = 9$), I reject it for reasons given in Section C.3. Blank entries correspond to combinations for which `SEM` did not converge (see Section C.3). The entries listed in this table are subject to the multipliers at the right-hand side of each section. The ICL scores for the GSWLC-2 sample are systematically higher than those for the VIPERS sample because it contains more galaxies.

| | | Submodel | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | $\Sigma_k, \delta_k$ | $\Sigma_k, \delta$ | $\Sigma, \delta_k$ | $\Sigma, \delta$ | $\alpha_{k,j}, \delta_k$ | $\alpha_{k,j}, \delta$ | $\alpha_k, \delta_k$ | $\alpha_k, \delta$ | $\alpha, \delta$ | |
| GSWLC-2 | $k = 2$ | 1.8 ± 0.0 | 1.5 ± 0.0 | 0.4 ± 0.0 | -6.2 ± 0.0 | 1.8 ± 0.0 | -4.7 ± 0.0 | 1.8 ± 0.0 | -4.7 ± 0.0 | -6.2 ± 0.0 | $\times 10^5$ |
| | $k = 3$ | 8.2 ± 0.0 | 7.8 ± 0.0 | -141.0 ± 0.0 | 0.0 ± 0.0 | 3.8 ± 0.0 | -5.1 ± 0.0 | 3.7 ± 0.0 | -5.3 ± 0.0 | -4.5 ± 0.0 | |
| | $k = 4$ | 11.7 ± 0.0 | 11.3 ± 0.0 | 2.7 ± 0.0 | 5.2 ± 0.0 | | | 4.9 ± 0.0 | -4.2 ± 0.0 | -3.8 ± 0.0 | |
| | $k = 5$ | 13.4 ± 1.4 | 13.4 ± 0.2 | -46.2 ± 51.5 | 8.7 ± 0.4 | 6.7 ± 1.0 | | 6.0 ± 0.0 | -2.2 ± 0.0 | -5.7 ± 1.3 | |
| | $k = 6$ | 16.7 ± 0.0 | | 9.4 ± 0.0 | 13.0 ± 0.1 | | | 6.8 ± 0.0 | 0.7 ± 0.0 | -5.2 ± 0.0 | |
| | $k = 7$ | **17.9 ± 0.2** | | 11.8 ± 2.0 | 14.2 ± 1.6 | | | 7.2 ± 0.0 | 2.3 ± 0.0 | -5.6 ± 0.0 | |
| | $k = 8$ | | | | 16.3 ± 1.3 | | | 8.1 ± 0.0 | 1.1 ± 0.0 | -5.7 ± 0.0 | |
| | $k = 9$ | | | 17.1 ± 0.0 | ~~**18.1 ± 0.9**~~ | | | 8.0 ± 0.0 | 3.9 ± 0.0 | -6.0 ± 0.0 | |
| VIPERS | $k = 2$ | | 2.1 ± 0.0 | 4.3 ± 0.0 | -8.4 ± 0.0 | | -8.0 ± 0.0 | | -8.0 ± 0.0 | -8.4 ± 0.0 | $\times 10^4$ |
| | $k = 3$ | | 11.1 ± 0.0 | -421.0 ± 0.0 | -0.4 ± 0.0 | | -4.7 ± 0.0 | 6.8 ± 0.0 | -5.3 ± 0.0 | -7.9 ± 0.0 | |
| | $k = 4$ | | | -294.0 ± 0.0 | 8.3 ± 0.0 | | -6.7 ± 0.0 | 8.3 ± 0.0 | -7.4 ± 0.0 | -9.0 ± 0.0 | |
| | $k = 5$ | 32.9 ± 0.0 | 32.4 ± 0.0 | 15.6 ± 0.0 | 16.3 ± 0.2 | | | 10.5 ± 0.0 | -2.9 ± 0.0 | -7.0 ± 0.0 | |
| | $k = 6$ | | | 20.9 ± 1.5 | 23.6 ± 0.0 | | | 13.0 ± 0.0 | 2.9 ± 0.0 | -3.8 ± 0.1 | |
| | $k = 7$ | | **41.8 ± 0.1** | 26.4 ± 2.3 | | | | 15.9 ± 0.8 | 7.1 ± 0.4 | -6.7 ± 0.0 | |
| | $k = 8$ | | | | | | | 14.6 ± 0.0 | 3.2 ± 0.0 | -10.8 ± 0.0 | |
| | $k = 9$ | | | | | | | 12.5 ± 0.0 | | -10.8 ± 0.0 | |

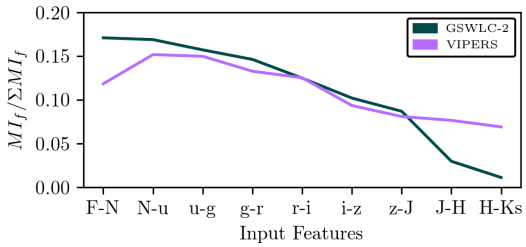


**Figure 5.2:** The relative importance of each of the input features to the clustering. "F" stands for FUV, and "N" for NUV. The mutual information (see Section 5.3.2 and Equation 5.2) of each of the input features with respect to the cluster labels has been normalised by the sum across all of the input features for each sample.

the $k = 9, \Sigma, \delta$ combination due to its inclusion of empty clusters (explained further also in Section C.3).

Both samples are best partitioned into seven clusters, within a six-dimensional discriminative latent subspace. The Gaussian density functions representing the clusters are each characterised by their own unique, full covariance matrices ($\Sigma_k$); the clusters each have different shapes, and the use of full covariance matrices indicates correlations (as expected) among the input features within the subspaces. While the best-fitting submodel for the GSWLC-2 sample uses unique noise terms for each cluster ($\delta_k$), the best-fitting submodel for the VIPERS sample does not ($\delta$), owing to the smoother distribution of the VIPERS sample in the feature space (see e.g. Figure 5.3). Submodels $\Sigma_k$, $\delta_k$ and $\Sigma_k$, $\delta$ report similar ICL scores and produce similar clustering structures in general and may therefore readily be compared with one another (see also Section C.3). That `SEM` has converged to highlighting these closely related submodels as being optimal for describing both samples is encouraging, and gives me confidence that I am conducting a fair comparison.

### 5.3.2 Feature importance

In Figure 5.2, I show the relative importance of each input feature to the clustering. Specifically, I calculate the mutual information ($MI$) between each input feature and the output cluster labels:

$$MI(f, l) = D_{KL}(p_{f,l} || p_f p_l). \tag{5.2}$$

Here, $D_{KL}$ is the Kullback-Leibler divergence (Kullback & Leibler 1951; also known as the relative entropy) between the joint probability distribution of input feature $f$ and output label $l$, and their independent distributions. For Figure 5.2, $MI_{f,l}$ is normalised by its sum across all input features to give a relative value.

The lines in Figure 5.2 are broadly similar, indicating that, on the whole, SEM uses the nine features in a similar way to determine its best partitions. This is further confirmed by noting that the subspaces within which SEM determined these best partitions have the same dimensionality (6) for both samples. The lines are especially consistent among the optical colours, which is expected given that optical photometry is ubiquitously available for galaxies in both samples. Altogether, the optical regime is the most important to the clustering. Individually, colours from the UV region of the SEDs of the galaxies in both samples are most strongly related to the output cluster labels. This highlights, as expected, the star formation activity and the dust content of galaxies as major influences on the shapes of their UV-through-NIR SEDs.

UV colours are slightly more important for the clustering in the GSWLC-2 sample, which reflects the increased UV coverage of its galaxies by GALEX (80 per cent, as opposed to 52 per cent for the VIPERS sample). NIR colours are less important for distinguishing clusters within the GSWLC-2 sample than within the VIPERS sample, which is likely due to their having been inferred purely from UV and optical input photometry[6]. This is in contrast with the galaxies in the VIPERS sample, whose NIR SEDs (more important to the clustering) were instead constrained by $K_s$-band photometry. For galaxies with incomplete photometry, the array of templates and synthetic spectra with which CIGALE may fit them is reduced, leading to reduced variation in the shapes of their SEDs. In addition, the rest-frame magnitudes (and hence, rest-frame colours) that CIGALE must infer from photometry at other wavelengths have larger uncertainties. Hence, availability of photometry with which to constrain the SEDs of galaxies is advantageous to the clustering. Nevertheless, Figure 5.2 shows that, for the most part, SEM uses the features similarly to model both samples despite slight differences in this availability, which is driven mostly by the ubiquitous availability of optical photometry for both samples.

### 5.3.3 Clustering structures

Table 5.2 profiles the clusters determined within each of the samples. Features are derived both from the same SEDs as the colours used for the clustering and from ancillary sources (see Sections 5.1.1 and 5.1.2). Clusters are named using two-part notation that will be used throughout the remainder of this chapter. The prefixes "G" or "V" denote clusters determined within the GSWLC-2 and VIPERS samples respectively. Clusters names have been ordered by their mean $NUV - r$ colours for ease of reference (see Table 5.2). I invoke the information in Table 5.2 when relevant throughout the following sections.

Figure 5.3 shows projections of the samples onto the two principal dimensions of their respec-

[6]While the Two Micron All-Sky Survey (Skrutskie et al., 2006) has NIR photometry for ∼ 50 per cent of GSWLC-2 galaxies, it is shallow and would not have provided strong constraints upon their NIR SEDs.

**Table 5.2:** Profiles, in terms of averages, of the clusters determined within each of the samples. See the main text for an explanation of the cluster naming scheme. Cluster means are listed in columns $NUV - r$ and $r - K_s$. For the remaining features, which are less directly linked to the clustering, I opt for medians to mitigate the potential influence of outliers on the cluster profiles. Column "%" lists the percentage of galaxies contained within each cluster for each sample. The data in the next seven columns [$NUV - r$ to $\log_{10}(sSFR/\mathrm{yr}^{-1})$ (SED)] originates from the same `CIGALE` SEDs as the rest-frame colours that were used as inputs to the clustering. Features listed in this table include colour excesses [$E(B-V)$], stellar masses ($M_*$), stellar metallicities ($Z$), mass-weighted stellar ages ($MWSA$), and specific star formation rates ($sSFR$). I list sSFRs both determined by `CIGALE` (SED; averaged over 100 Myr timescales) and determined from galaxy spectra (and hence independent of `CIGALE`; ind.; see Sections 5.1.1 and 5.1.2). Medians marked with asterisks have unexpected values given their corresponding $NUV - r$ colour and are discussed in Section 5.3.4.

| Cluster | % | $NUV-r$ | $r-K_s$ | $E(B-V)$ | $\log_{10}(M_*/\mathrm{M}_\odot)$ | $\log_{10}(Z)$ | $\log_{10}(MWSA/\mathrm{Myr})$ | $\log_{10}(sSFR/\mathrm{yr}^{-1})$ (SED) | $\log_{10}(sSFR/\mathrm{yr}^{-1})$ (ind.) |
|---|---|---|---|---|---|---|---|---|---|
| G1 | 24.0 | 2.39 | 0.42 | 0.11 | 9.90 | −2.22 | 3.80 | −9.87 | −9.87 |
| G2 | 15.2 | 3.29 | 0.91 | 0.20 | 10.26 | −1.81 | 3.85 | −10.02 | −10.19 |
| G3 | 17.3 | 3.51 | 0.78 | 0.14 | 10.37 | −2.11 | 3.89 | −10.38 | −10.47 |
| G4 | 8.5 | 4.31 | 1.16 | 0.13 | 10.70 | −1.75 | 3.92 | −10.87 | −11.22 |
| G5 | 9.7 | 5.07 | 0.67 | 0.22 | 10.35 | −2.30 | 3.90 | −10.78 | −11.97 |
| G6 | 11.3 | 5.24 | 0.78 | 0.08 | 10.57 | −2.11 | 3.93 | −11.92 | −11.93 |
| G7 | 14.0 | 5.27 | 0.73 | 0.11 | 10.54 | −2.20 | 3.93 | −11.85 | −12.02 |
| V1 | 26.8 | 1.86 | 0.25 | 0.01 | 9.87 | −2.12 | 3.52 | −9.34 | −9.25 |
| V2 | 18.4 | 2.17 | 0.60 | 0.02 | 10.14 | −1.90 | 3.55 | −9.22 | −9.34 |
| V3 | 9.3 | 2.62 | 0.75 | 0.05 | 10.10 | −1.40 | 3.52 | −8.99 | −9.35 |
| V4 | 18.5 | 3.26 | 1.05 | 0.12 | 10.67 | −1.80 | 3.58 | −9.71 | −9.92 |
| V5 | 5.2 | 4.75 | 0.91 | *0.15 | 10.61 | *−1.51 | *3.52 | *−9.43 | −10.09 |
| V6 | 10.3 | 4.81 | 0.90 | *0.15 | 10.69 | *−1.86 | *3.61 | *−9.90 | −10.29 |
| V7 | 11.5 | 4.86 | 0.96 | 0.02 | 10.91 | −2.05 | 3.74 | −11.27 | −10.42 |

tive six-dimensional discriminative subspaces. These projections, which offer direct views of the structures of the clustering outcomes, are determined uniquely for each sample by `SEM`: *the axes of the two plots do not correspond exactly to one another*. Nevertheless, these projections are broadly similar in terms of the shapes of the overall samples within them. Both samples exhibit a continuum, running from the lower right to the upper left of each plot, which has been segmented by `SEM`. In addition, both samples exhibit a cluster which extends into the sparser region to the upper right of each plot. This overall similarity gives me confidence in the success of the measures that were taken to ensure a fair comparison between samples at different redshifts and from different surveys (see Sections 5.1.1 and 5.1.2). In addition, it reinforces my conclusion that `SEM` has overall used the input features similarly for both samples in spite of slight differences in the availability of photometry between them (Section 5.3.2). The subtler differences between clusters in these projections are subject to the distributions of galaxies *within* the shapes of their respective samples. I comment on these differences where relevant in Section 5.3.4. Cluster colours in the plots in this chapter, like their names, are assigned based on their mean $NUV - r$ colours.

I break down the analysis of the clusters using the colour bimodality of galaxies. The colour bi-

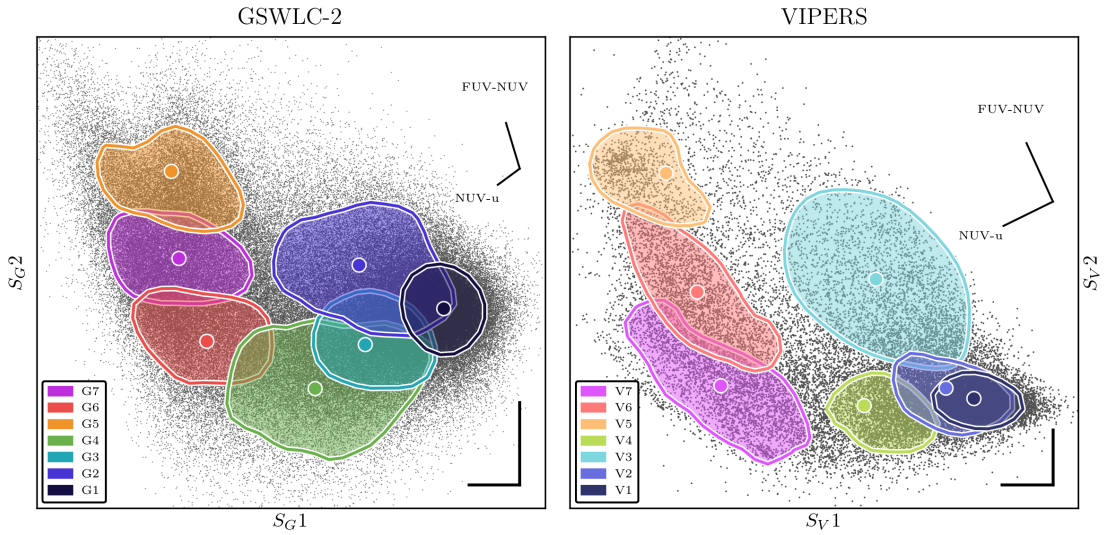


**Figure 5.3:** Projections of both samples onto the two dimensions that best separate their clusters. The axes of each plot are determined by `SEM` and are unique to each sample (as indicated by their labels; e.g. $S_G1$ represents the first axis of the subspace of the GSWLC-2 sample). but the resultant projections are mostly similar nonetheless. The distributions of clusters within this plane are shown using coloured, filled contours (drawn at a relative density of 0.4), and the coloured, circular markers show their means. The perpendicular black lines at the lower right of each plot show the extent to which the y-axis has been stretched relative to the x-axis to yield the projections as shown. The vectors at the upper right of each plot show the projections of the two input features that correlate most strongly with the axes of these projections.

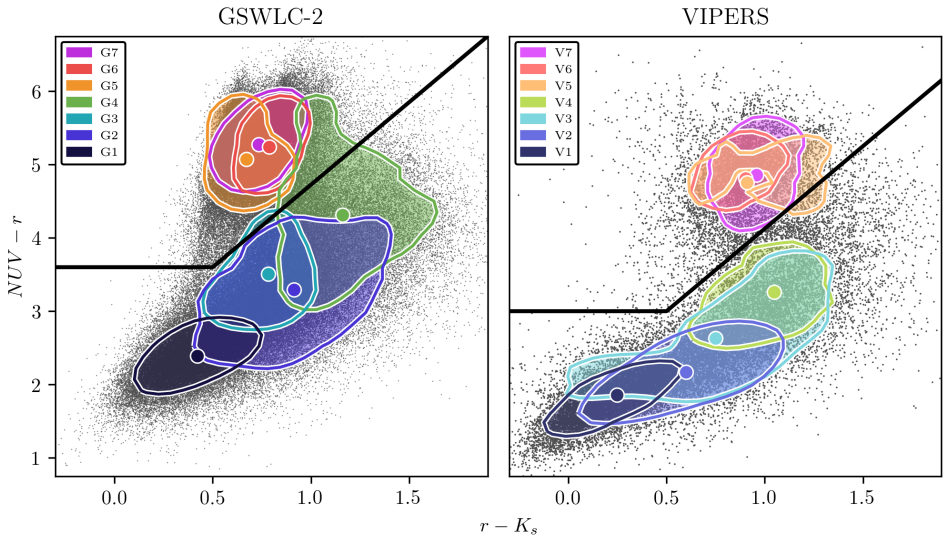


**Figure 5.4:** Colour-colour plots of the samples. Colours are derived from `CIGALE` SED estimation. The distributions of clusters are shown using coloured, filled contours (drawn at a relative density of 0.4), and the coloured, circular markers show their means. The black line in each plot (inspired by Moutard et al. 2016b; see main text) marks the boundary between star-forming galaxies (below the line) and passive galaxies (above the line).

modality is a steady property the galaxy population throughout cosmic time, having been observed among galaxies with redshifts as high as 4 (Wuyts et al., 2007; Williams et al., 2009; Ilbert et al., 2010, 2013). Hence, it may be used to separate clusters of star-forming galaxies (on the blue peak) from clusters of passive galaxies (on the red peak) in a way that is independent of redshift.

This separation is marked by the black lines in Figure 5.4. The $NUV - r - K_s$ colour-colour plane (Arnouts et al., 2013; Moutard et al., 2016b) is a useful tool with which to probe galaxy subpopulations due to its ability to separate star-forming (low $NUV - r$), passive (high $NUV - r$), and also dusty (high $r - K_s$) galaxies. It has been applied in several studies of galaxy evolution using data from VIPERS (e.g. Fritz et al. 2014; Davidzon et al. 2016; Moutard et al. 2016b; Siudek et al. 2017, 2018b; Vergani et al. 2018). The form of the black lines is inspired by Fritz et al. (2014) and Moutard et al. (2016b); they are placed independently in each panel, without reference to the positions of the clusters, to simply demarcate the star-forming and passive regions of the $NUV - r - K_s$ plane. Clusters whose means then lie below the black line in each plot are selected as blue, star-forming clusters, and clusters whose means then lie above the black lines are selected as red, passive clusters. As a result, both samples break down into four blue clusters and three red clusters. Deviations of the structures of the clusters from this simple blue/red (star-forming/passive) division that I enforce will highlight limitations of a purely binary view of the galaxy population.

None of the clusters determined within either of the samples are confined in their extent to just

the green valley (i.e. to just intermediate $NUV - r$ colours in Figure 5.4). This seems consistent with the notion that the green valley does not represent a single, unique, or excess subpopulation of galaxies, but instead encompasses a diverse array of galaxies that are transitioning from star-forming to quenched (Smethurst et al., 2015; Moutard et al., 2016b). By implication, this also suggests that there is more than one evolutionary pathway of galaxies through the green valley (Faber et al., 2007; Fritz et al., 2014; Schawinski et al., 2014). Previous studies (Baldry et al., 2004; Taylor et al., 2015) have suggested that the green valley comprises the overlapping tails of the red and blue peaks of the colour bimodality; Krywult et al. (in prep.), fitting the UV-through-optical bimodality with two Gaussian density functions in narrows stellar mass and redshift bins, show that this constitutes a particularly accurate description of the colour distribution of galaxies. Clusters G4 and V4, which are closest to the green valley, are most strongly associated with the high-mass end of the blue peak of the colour bimodality.

The blue peak of the bimodality corresponds closely with the SFMS, which is the tight correlation between the SFRs and the stellar masses of actively star-forming galaxies. The SFMS, like the bimodality, is ubiquitous throughout cosmic time (Speagle et al., 2014). It has a lower normalisation with decreasing redshift; this cosmological decline of star formation (Madau et al., 1996; Madau & Dickinson, 2014; Driver et al., 2018) is visible as a vertical offset between the samples in Figure 5.4. In this chapter, the terms "blue peak" and "SFMS" are synonymous, and I use them interchangeably.

The stronger $NUV-r$ split between star-forming and passive VIPERS clusters in comparison with those of GSWLC-2 (Figure 5.4, and also visible in Figure 5.3) is likely to result from two factors. First is the difference in the rest-frame wavelength coverage of GALEX photometry for the two samples; some rest-frame UV emission is redshifted out of the bandwidths of GALEX's filters at $z \sim 0.65$. Second is the difference in the completeness of UV photometry for each sample. GALEX observations exist for $\sim$ 80 per cent of galaxies in clusters G1-4. This proportion falls to $\sim$ 55 per cent in clusters G5-7, but this is expected given that these galaxies would be fainter in the UV regime. Meanwhile, $\sim$ 65 per cent of V1, V2, and V4 galaxies were observed by GALEX. Interestingly, only $\sim$ 20 per cent of galaxies in V3 have observed UV photometry, which may be a part of the reason for its separation from the other star-forming VIPERS clusters. Passive VIPERS clusters are $\sim$ 25 per cent complete in observed UV photometry. Together, these factors mean that low levels of UV emission from more evolved VIPERS galaxies with more intermediate colours are likely to be missed. On the other hand, Figure 5.2 shows that rest-frame $NUV - u$ colours are similarly important to the clustering structures of *both* samples, with $NUV$ emission expected to be a particularly accurate tracer of star formation (Salim, 2014).

**Table 5.3:** Profiles, in terms of averages of ancillary features, of the clusters determined within each of the samples. See the main text for an explanation of the cluster naming scheme. I list the median values of the galaxies that the clusters contain for each of the features. Column "%" lists the percentage of galaxies contained within each cluster for each sample. Features listed in this table include Sérsic indices ($n_g$), half-light radii ($R_{1/2}$), and environmental overdensities ($\delta$). The data is drawn from ancillary sources (see Sections 5.1.1 and 5.1.2).

| Cluster | % | $n_g$ | $\log_{10}(R_{1/2}/\mathrm{kpc})$ | $\log_{10}(1+\delta)$ |
|---|---|---|---|---|
| G1 | 24.0 | 1.04 | 0.57 | 0.40 |
| G2 | 15.2 | 1.34 | 0.50 | 0.51 |
| G3 | 17.3 | 1.57 | 0.55 | 0.55 |
| G4 | 8.5 | 2.38 | 0.61 | 0.59 |
| G5 | 9.7 | 4.09 | 0.40 | 0.85 |
| G6 | 11.3 | 4.18 | 0.45 | 0.80 |
| G7 | 14.0 | 4.25 | 0.44 | 0.83 |
| V1 | 26.8 | 0.92 | 0.49 | 0.29 |
| V2 | 18.4 | 0.95 | 0.48 | 0.29 |
| V3 | 9.3 | 1.11 | 0.50 | 0.36 |
| V4 | 18.5 | 1.53 | 0.55 | 0.35 |
| V5 | 5.2 | 3.31 | 0.42 | 0.40 |
| V6 | 10.3 | 3.29 | 0.40 | 0.40 |
| V7 | 11.5 | 3.40 | 0.43 | 0.43 |

### 5.3.4 Cluster identities

#### Clusters of star-forming galaxies

My $NUV-r-K_s$ cut (Section 5.3.3) yields the following blue, star-forming clusters: G1, G2, G3, and G4 for the GSWLC-2 sample; and V1, V2, V3, and V4 for the VIPERS sample. Though G4 also contains a significant number of galaxies with green or red $NUV-r$ colours (i.e. such that they are already quenching or quenched), I analyse it in this section due to apparent connections with other clusters on the SFMS (see also Section 5.4.1). Figure 5.5 shows that the SEDs of G4 galaxies tend to more closely resemble those of actively star-forming galaxies, being flatter in the UV regime (e.g. G3 galaxies) than those of typically passive galaxies (e.g. G5 galaxies). Hence, in terms of the influence of their evolution on the shapes of their SEDs, G4 galaxies are more closely related to G1-3 galaxies than G5-7 galaxies, despite some G4 galaxies being quenched.

Given that the SFMS is a smooth continuum, it is important where possible to establish why `SEM` has distinguished clusters within it, and to interpret the significance of these distinctions in terms of galaxy evolution. The position of a galaxy along the $NUV-r-K_s$ SFMS (Figure 5.4) is governed by a combination of its stellar mass and its dust content (Moutard et al., 2016a,b). The lobe at high $r-K_s$, which preferentially consists of edge on galaxies, is known to capture the

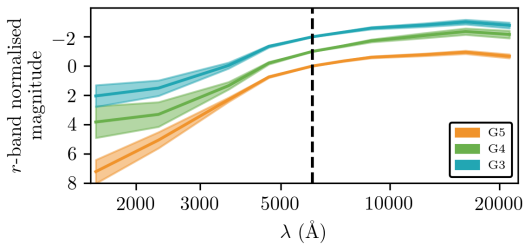


**Figure 5.5:** A comparison of the shapes of the mean (± standard deviation) estimated SEDs of galaxies in clusters G3, G4, and G5. Clusters G3 and G5 are chosen as they neighbour G4 in terms of their average $NUV - r$ colour. The estimated SEDs of individual galaxies are normalised by their $r$-band magnitudes (the effective wavelength of which is marked by a dashed black line) before the mean estimated SEDs are calculated. The y-axis applies to the mean SED of G5; those of G3 and G4 are vertically offset by $-1$ and $-2$ respectively to more clearly show the differences in their shapes.

excess reddening of high-mass star-forming galaxies (Arnouts et al., 2013), but it is more difficult to disentangle this combination of stellar mass and dust elsewhere within the SFMS. Hence, there is an overlap of star-forming clusters in Figure 5.4. In Figure 5.3, though, these clusters are more clearly separated.

G1 and V1 capture similar subpopulations of galaxies. Both clusters contain the galaxies with the bluest colours and the lowest masses (Figure 5.4, Table 5.2) within their respective samples; star-forming galaxies at relatively early stages of their evolution. The remaining star-forming clusters have higher masses and lie further along the SFMSs of each sample.

Clusters G2 and G3 overlap with one another in the left-panel of Figure 5.4, as do clusters V2 and V3 in the right-hand panel of the same figure. Figure 5.3 shows that G2 and V3 both extend away from the main continua within the subspace projections of their respective samples. The feature vector projections in Figure 5.3 show that the galaxies in these clusters have particularly red $FUV - NUV$ colours in comparison with other SFMS clusters. However, the astrophysical meaning behind this is unclear. `CIGALE` alternately attributes this reddening to high colour excesses for galaxies in G2 and to higher metallicities for galaxies in V3 (Table 5.2), suggesting that it has not fully resolved the degeneracy between the influences of dust and metallicity upon the colours of these galaxies. However, `CIGALE` *is* consistent in assigning G2 and V3 galaxies similar stellar masses and mass-weighted stellar ages to G3 and V2 galaxies (Table 5.2), which occupy similar regions of the $NUV - r - K_s$ plane. Stellar mass estimates are not strongly affected by an inability to resolve this degeneracy between the influences of dust and metallicity (e.g. Bell & de Jong 2001). Clusters G3 and V2, lying on the main continua in Figure 5.3, seem to be intermediate between clusters G1 and G4, and V1 and V4 respectively.

The star-forming clusters along the SFMS of the GSWLC-2 sample exhibit a gradient in their star

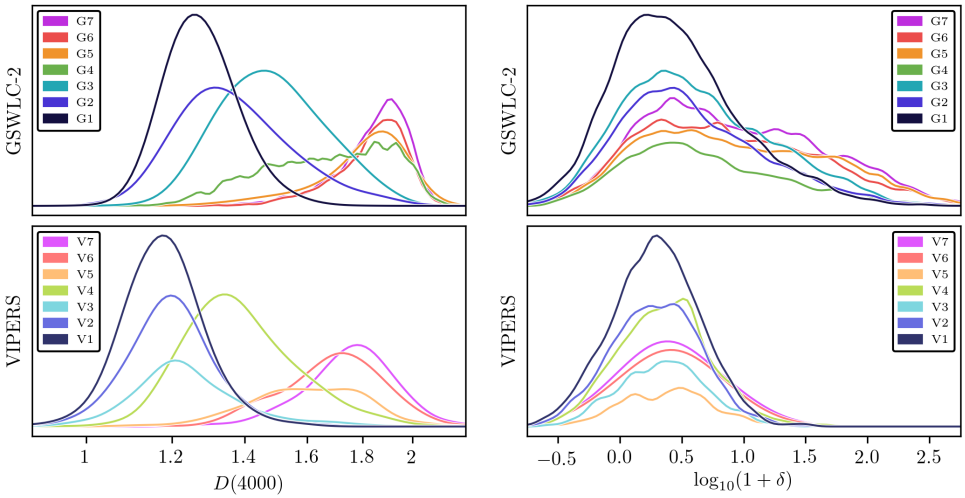


**Figure 5.6:** Smoothed kernel density estimates in $D(4000)$ (**left**, logarithmically distributed) and local environmental density (**right**) for each of the clusters from both outcomes. Here, $D(4000)$ was measured from the spectra of galaxies (Brinchmann et al., 2004; Garilli et al., 2014) using a method introduced by Balogh et al. (1999), and is hence independent of `CIGALE`'s estimated SEDs. For both samples, these overdensities are based on fifth-nearest neighbour surface densities (Baldry et al., 2006; Cucciati et al., 2017).

formation activity. Taking their increasing average stellar masses as a point of reference, clusters G1-4 exhibit a corresponding increase in their average $NUV - r$ colours (Table 5.2, Figure 5.4). decrease in their average sSFRs (both SED and ind.; Table 5.2), and increase in their average $D(4000)$ (Figure 5.6). High-mass galaxies in the GSWLC-2 sample do not form stars as readily as low-mass galaxies. This gradient is weaker for clusters V1-3 (particularly with regard to their median sSFRs; Table 5.2), though I note that clusters V2 and V3 have lower average stellar masses than G2 and G3. It is only in V4 that a rise in average stellar mass is accompanied by a decrease in average $sSFR$, and an increase in $D(4000)$.

The large median sizes and low-to-intermediate median Sérsic indices of star-forming clusters from both samples indicate that they are dominated by disc galaxies (Table 5.3). Clusters G1-4 exhibit a rise in their median $n_g$ to intermediate values along their SFMSs, indicating increasingly concentrated morphologies among their galaxies. In Figure 5.7, these clusters form morphological sequences that are separate from the distributions of passive clusters in the same plane. The sequence of V1-4 is not as strong as that of G1-4; again, it is only in V4 that a significant change is seen, with the higher stellar masses of its galaxies met with intermediate Sérsic indices.

While there are slight trends in the median local environmental overdensities of the star-forming clusters in both samples (Table 5.3), Figure 5.6 shows that their distributions thereof have very large spreads and exhibit a great deal of overlap with the distributions of other SFMS clusters from the same sample. Therefore, the reduction in the star formation activity of SFMS galaxies at

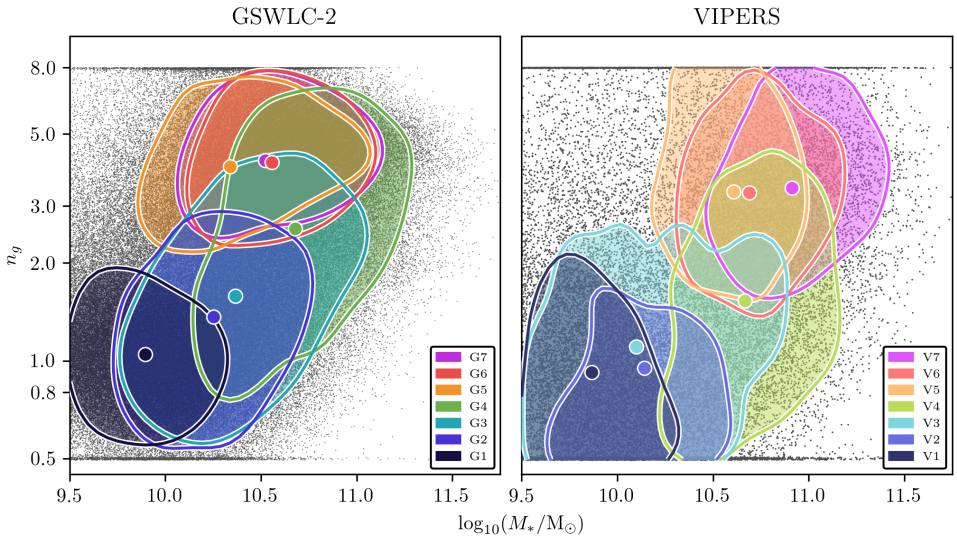


**Figure 5.7:** Sérsic index versus stellar mass for the galaxies in the samples. Sérsic indices were determined by Simard et al. (2011) for the GSWLC-2 sample, and Krywult et al. (2017) for the VIPERS sample. The distributions of clusters are shown using coloured, filled contours (drawn at a relative density of 0.4), and the coloured, circular markers show their medians. The Sérsic indices of the galaxies in the VIPERS sample have been winsorised to values of 0.5 and 8 in order to match the limits of the GSWLC-2 sample.

higher masses cannot be attributed to mainly environmental causes for either sample.

### Clusters of passive galaxies

The red, passive clusters, selected using the $NUV - r - K_s$ plots in Figure 5.4, are: G5, G6, and G7 for the GSWLC-2 sample, and V5, V6, and V7 for the VIPERS sample. The input colour that best separates the passive clusters in both samples is $FUV - NUV$. For G5-7, this separation appears to have captured the higher sSFRs and lower masses of G5 galaxies, and differences in the metallicities of G6 and G7 galaxies (Table 5.2). V7 has been distinguished due to the high masses and low sSFRs of its galaxies. However, `CIGALE`'s estimation of the astrophysical properties of V5 and V6 galaxies is less reliable (see below). In general, galaxies in the passive clusters are offset to redder $NUV - u$ colours than those in the SFMS clusters (see above).

Galaxies in clusters G6, G7, and V7 are alike with respect to most features. They share high stellar masses, low sSFRs, large $D(4000)$ (Figure 5.6), and early-type morphologies (Table 5.2), all of which are typical of canonically passive galaxies. `CIGALE` attributes the difference in the $FUV - NUV$ colours of G6 and G7 galaxies (i.e., the feature that best separates these clusters) to their metallicity distributions. While G6 peaks strongly at $Z \sim -2.1$, G7 is split evenly between peaks at $Z \sim 2.1$ and $Z \sim -2.4$. The metallicities of passive GSWLC-2 galaxies are discretised by the input Bruzual & Charlot (2003) grid, and due to a lack of any input NIR photometry

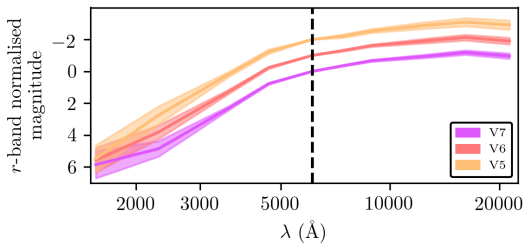


**Figure 5.8:** A comparison of the shapes of the mean (± standard deviation) estimated SEDs of galaxies in clusters V5, V6, and V7. The estimated SEDs of individual galaxies are normalised by their *r*-band magnitudes (the effective wavelength of which is marked by a dashed black line) before the mean estimated SEDs are calculated. The y-axis applies to the mean SED of V7; those of V5 and V6 are vertically offset by −1 and −2 respectively to more clearly show the differences in their shapes.

during their SED estimation (see Section C.2); with more precise metallicities, their distributions might overlap more. V7 also has low metallicities in comparison with other clusters determined in its sample. These sub-solar metallicities are unexpected for high-mass passive galaxies (e.g Gallazzi et al. 2006), indicating difficulties of breaking the age-dust-metallicity degeneracy with photometry alone, and suggesting that these metallicities are not entirely reliable. Altogether though, these clusters contain the oldest, most evolved galaxies among their respective samples: a subpopulation that is in place at the epoch of the VIPERS sample.

Galaxies in cluster G5, while also passive and early-type, have lower stellar masses than those in clusters G6 and G7. I also note a difference between their median sSFRs, as reported by `CIGALE` (SED) and by the independent Brinchmann et al. (2004) calibration (ind.; Table 5.2). I suggest that G5 is likely to contain post-starburst galaxies (PSBs; Wild et al. 2009), and that this difference in sSFRs is likely to arise due to the different timescales probed by these two measures (see Section 7 of Salim et al. 2016). While the fibre component of *sSFR* (ind.) is a more instantaneous measure of star formation activity (∼ 10 Myr, being based on H$\alpha$ emission), `CIGALE` averages star formation over a longer period of time (100 Myr, to match the timescale of UV emission as a tracer of star formation). Hence, even if the tail of a declining central burst of star formation activity is not captured by *sSFR* (ind.), it may still be captured by *sSFR* (SED). The spheroidal morphologies (Figure 5.7, Table 5.3) and the slight enhancement in the local environmental densities of G5 galaxies are suggestive of an external influence upon their evolution (see Section 5.4.2), which is consistent with previous studies which link PSBs with mergers (Zabludoff et al., 1996; Yang et al., 2008; Almaini et al., 2017).

Clusters V5 and V6 present conflicting identities in terms of features estimated by `CIGALE` (Table 5.2). While their galaxies have very similar stellar masses and morphologies to those in V7 (Table 5.3), they have unusually high colour excesses and metallicities and, in turn, high *sSFR* (SED). This is in contrast with *sSFR* (ind.) and observed $D(4000)$ of these galaxies (Table 5.2, Figure

5.6), which show that they are indeed passive. The large spread in $D(4000)$ of V5 may be due to some minor contamination of the cluster by star forming galaxies; its $NUV - r - K_s$ contour extends below the black line in Figure 5.4, into the region containing dusty star-forming galaxies. This may also have driven its median $E(B-V)$ to a higher value.

The inability of CIGALE to properly resolve the age-dust-metallicity degeneracy for V5 and V6 galaxies is due to the shapes of the UV regions of their SEDs. Figure 5.8 shows that galaxies in clusters V5 and V6 have steeper average UV SEDs than those in cluster V7. In order to explain the red UV colours (especially $FUV - NUV$) of these galaxies, CIGALE has invoked high colour excesses and metallicities rather than low *sSFR* (SED). This appears to be a consequence of CIGALE's two-burst SFHs, which may not be a realistic description of the SFHs of most passive VIPERS galaxies. These SFHs were adjusted for the epoch of the VIPERS sample by setting the formation time of the old population to 6.5 Gyr ago instead of 10 Gyr, and including the possibility of a particularly recent burst of star formation ($< 50$ Myr). However, a trial of the use of a gradual 1 Gyr quenching episode instead led to improvements in the quality of fit of passive SEDs, which produce low *sSFR*, to the photometry of the majority of V5 and V6 galaxies. Hence, it seems that further adjustments to CIGALE's SFH prescription are required when applying it at higher redshifts[7].

Galaxies contained within the passive clusters of the VIPERS sample tend to have higher stellar masses than those contained within the passive clusters of the GSWLC-2 sample (Table 5.2). The downsizing of the galaxy population (Cowie et al., 1988; Cimatti et al., 2006; Cattaneo et al., 2008) means that fewer passive galaxies with low stellar masses are expected to be observed at higher redshifts. However, stellar mass incompleteness of the VIPERS sample is likely to be a more significant factor in explaining this result. Davidzon et al. (2013) show that, even at its lower redshift limit of $z = 0.5$, the VIPERS sample is incomplete in passive galaxies below $\sim 10^{10}$ $M_\odot$. Furthermore, their completeness threshold increases with redshift to $10^{10.75}$ $M_\odot$ at my upper limit of $z = 0.8$, and thus skews the clusters of passive VIPERS galaxies towards higher stellar masses[8]. Hence, where the GSWLC-2 sample has two lobes of passive galaxies in Figure 5.3 (see also Section C.2), which differ in average stellar mass by $\sim 0.5$ dex, the VIPERS sample has only one. Though the VIPERS sample does contain some passive galaxies with low stellar masses (e.g. Figure 5.7), they are not substantial enough in number for SEM to model them with a dedicated cluster (i.e. like G5).

Passive clusters in both samples have high Sérsic indices and compact sizes (Table 5.3), indicating that their galaxies mostly have spheroid-dominated morphologies. As noted previously, they

[7] LePhare (Ilbert et al., 2006) SED estimation for the same galaxies (Moutard et al., 2016b; Siudek et al., 2018b), which used a single declining exponential for its SFHs, reported lower colour excesses and metallicities, and hence, lower *sSFR*.

[8] Star-forming galaxies and clusters are affected to a much lesser degree.

occupy separate regions of the plots in Figure 5.7 to their respective SFMS clusters. Figure 5.7 also shows that the $n_g$ distributions for passive clusters are highly consistent with one another, indicating a strong morphological homogeneity among their galaxies. While the passive clusters in the GSWLC-2 sample exhibit a slight offset to higher density environments in comparison with star-forming GSWLC-2 clusters, the environments of passive VIPERS clusters are consistent with those of star-forming VIPERS clusters. This difference between the two samples is, in part, expected, due to the emergence of environments of especially high densities over cosmic time (e.g. Marinoni et al. 2008; Kovač et al. 2010; Fossati et al. 2017). However, factors such as spectroscopic fibre collisions and the aforementioned incompleteness of passive VIPERS galaxies may also reduce the completeness of VIPERS at especially high densities. I note that this incompleteness does not appear to have strongly affected clusters elsewhere in the feature space (Figure 5.3).

## 5.4 Discussion

My clusters have been determined on the basis of the rest-frame colours of galaxies alone, which mostly express their star formation activity and dust content. In this section, I aim to discern what the trends of these purely colour-based clusters with other, ancillary features (revealed in Section 5.3) may say about the evolutionary histories of their constituent galaxies.

### 5.4.1 Internally driven evolution

Both sets of star-forming clusters – G1-4 and V1-4 – form clear morphological sequences in Figure 5.7. This implies an evolutionary pathway that connects the clusters in each set, and that acts at the epochs of both samples. In Figure 5.9, I examine the bulge-to-total ratios of GSWLC-2 galaxies (no such data exists for VIPERS). The G1-4 sequence is apparent here as well, capturing the rising prominences of the bulges of their galaxies. It does *not* extend to the highest $B/T_r$ values, despite G4 also containing some quenching and quenched galaxies. which indicates that G1-4 galaxies retain dominant disc components as they evolve, and that some G1-4 galaxies become passive without fully transforming their morphologies. The changing morphological bulge-disc balance appears to be captured also in the large spread in $D(4000)$ of G4 galaxies in particular (Figure 5.6). The overlapping environmental distributions of star-forming clusters in both samples (Figure 5.6), suggests that the evolutionary pathway that begets these morphological sequences of gradual bulge growth is more likely to be due to internal processes. I assume that the interpretation in this paragraph applies to galaxies in V1-4 as well.

Bar-driven inflows of star-forming gas (Sheth et al., 2005) are internal processes that act over long timescales, and are likely to be involved in this gradual evolution of galaxies along the SFMS.

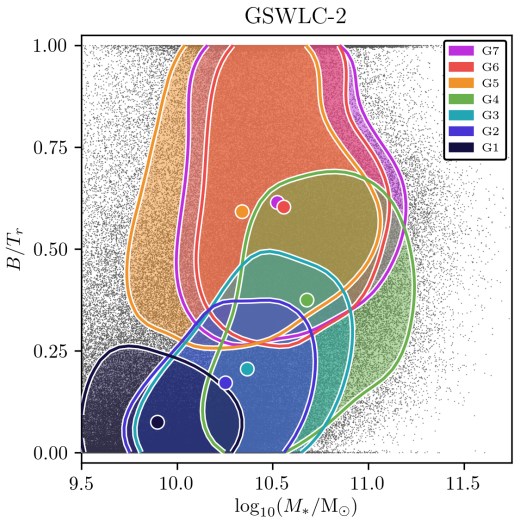


**Figure 5.9:** Bulge-to-total ratio ($B/T_r$) versus stellar mass for the galaxies in the GSWLC-2 sample. Here, the subscript "$r$" denotes the $r$-band photometry from which the ratios were derived (Simard et al. 2011; based on two-component fits). The distributions of clusters are shown using coloured, filled contours (drawn at a relative density of 0.4), and the coloured, circular markers show their means.

These inflows are commonly invoked to explain the formation of dynamically cold "pseudobulges" rather than the dynamically hot "classical" ($n_{cl} \gtrsim 2$) bulges that the Simard et al. (2011) two-component fits assume (Kormendy & Kennicutt, 2004; Fisher & Drory, 2008; Mishra et al., 2017). However, an increase in the prominence of pseudobulges would nonetheless be expected to be captured by the single-component fits which yield the Sérsic indices in Table 5.3 and in Figure 5.7. Internal processes such as the inward movement of clumps of newly formed stars from unstable discs during earlier epochs (such that the bulge is in place at later times; e.g. Elmegreen et al. 2008b; Bournaud et al. 2011; Tonini et al. 2016) may instead lead to the formation of a classical bulge, but these processes act over much shorter timescales and are less likely to lead to the gradual trend in bulge prominence along the SFMS. I do not rule out that SFMS galaxies may have undergone major and/or minor mergers or clump migration in their pasts; some have high total $n_g$ values, which may be capturing classical bulges formed as a result of merger activity or clump migration. Instead, I proffer that mergers do not contribute to their *growth* as they evolve *along* the SFMS (see also Section 5.4.2). It has been shown, for example, that the remnant of a gas-rich merger can reform a disc and continue to form stars, thus rejoining the SFMS (Hopkins et al., 2009a,b).

The decline in the sSFRs of galaxies along the sequences G1-4 and V1-4 suggests that their morphologies are also linked with an inhibition of their star formation. This could be due to the prevention of the collapse of gas clouds within the disc by the deep gravitational potential of the bulge ("morphological quenching'; Martig et al. 2009). It is more likely, though, that the prominence of the bulge among these galaxies is a marker of nuclear activity within them. More massive bulges

host more massive black holes at their centres (Häring & Rix, 2004), which in turn supply more radiative feedback to their surrounding galaxies. This feedback can heat or eject gas in the discs of the galaxies from which stars are formed and, as a result, inhibit star formation within these galaxies (Croton et al., 2006; McCarthy et al., 2010; Gabor et al., 2011; Vergani et al., 2018). In addition, the stronger inhibition of star formation in G4 and V4 galaxies (the former of which also contains some quenching and quenched galaxies) may be tied to their exceeding the "transition mass" ($\sim 10^{10.5}$ $M_\odot$ at $z \sim 0$; Kauffmann et al. 2003b), above which the inflow or subsequent cooling of star-forming gas is prevented by the heating of galaxy's halo by nuclear activity (Dekel & Birnboim, 2006; Kereš et al., 2009; Moutard et al., 2020). I explore the possibility of the influence of AGN among G1-4 in Section C.4, finding that galaxies in all four clusters, and especially those in G4, are likely to contain low-ionisation nuclear emission-line regions.

In all, this connection between the bulges and the SFRs of galaxies within the star-forming clusters of both samples implies that these galaxies evolve *along* the SFMS (in tandem with its downward movement over cosmic time), and that this evolutionary pathway is governed chiefly by internal processes that act over long timescales (Schawinski et al., 2014; Ilbert et al., 2015; Moutard et al., 2016b; Pacifici et al., 2016; Popesso et al., 2019a). At the highest masses, this pathway leads (as revealed by G4 in particular) *off* the tip of the SFMS, yielding a subpopulation of red, passive galaxies which retain a disc. I suggest that this is due to the prevention of the accretion of new gas with which to form stars by nuclear activity (Gabor et al., 2011; Moutard et al., 2020). This connection between the bulges and the star formation activity of SFMS galaxies has previously been established (Cheung et al., 2012; Fang et al., 2013; Bluck et al., 2014; Cano-Díaz et al., 2019; McPartland et al., 2019), but in my case it emerges purely from my clustering of galaxy colours, with morphologies invoked post-clustering for evaluation.

Clusters V1-3 all have low stellar masses and high sSFRs. There does not appear to be a strong trend between these two features for these clusters, which may be tied to their morphologies; all three also have very low median $n_g$, such that they appear to be very strongly disc-dominated. In the context of the internally-driven evolutionary pathway that I propose, this suggests that the bulges and/or supermassive black holes of V1-3 galaxies have not yet grown to the extent that they can affect star formation in the discs surrounding them. This would be consistent with Fang et al. (2013) and Bluck et al. (2014), who find that bulges must exceed a threshold in mass or central density before they become associated with quenching. For V4, a rise to higher median $M_*$ is met with a rise to intermediate median $n_g$ and a fall to lower sSFRs, suggesting that this threshold bulge mass has been achieved in some V4 galaxies. Hence, the rising prevalence of bulges grown by internal processes over cosmic time (e.g. Bruce et al. 2012; Gu et al. 2019) seems to be linked to the cosmic decline of cosmic star formation activity and the downsizing of the galaxy population. Given the long timescales over which these internal processes act, the gradual growth of V1-4 galaxies to higher masses and more prominent bulges would be expected to eventually lead to the

more evolved distribution of galaxies that is captured by clusters G1-4 by the present day, which are assumed to be their descendants.

### 5.4.2 Satellite quenching at low redshifts

The uniformly high Sérsic indices of galaxies in clusters G5-7 and V5-7 imply a strong link between their concentrated morphologies and their passiveness. At high masses, this link is likely to include a contribution from the internally-driven evolutionary pathway that I propose in Section 5.4.1, in which the growth of the bulges of galaxies ultimately leads to the quenching of star formation. Cluster V7 in particular, containing VIPERS galaxies with the highest masses, seems to align well with the sequence of clusters V1-4 in Figure 5.7, such that it could be an extension of this evolutionary pathway, consisting of the oldest galaxies with the most prominent bulges. This is in agreement with previous studies which find that the inner stellar density of galaxies is a successful predictor of its having been quenched (Bell, 2008; Cheung et al., 2012; Fang et al., 2013; Bluck et al., 2014), with the favoured explanation being the heating or ejection of star-forming gas by feedback from the supermassive black hole, whose mass scales with the inner stellar density.

However, other passive clusters are separated from their respective sequences of star-forming clusters in Figure 5.7. Clusters G7, G6, and in particular G5 (the latter containing the lowest-mass passive galaxies in the GSWLC-2 sample) have high median $n_g$ in comparison with other clusters centred at similar stellar masses (G2, G3). This separation, clearer towards lower stellar masses, invites the interpretation that their galaxies are subject to alternative and/or additional processes as they evolve. That these clusters contain those GSWLC-2 galaxies that occupy the highest-density environments (Figure 5.6) suggests an influence of external processes. Hence, I suspect that a significant proportion of galaxies among G5-7 are low-mass satellite galaxies (occupying the halos of more massive central galaxies), and are subject to external processes (Ilbert et al., 2010; Muzzin et al., 2013; Moutard et al., 2018). I note that no such separation or environmental offset is seen for V5-7, which I attribute to the incompleteness of low-mass passive galaxies in the VIPERS sample, which would also be expected to trace high-density environments. Hence, the following discussion on the influence of external processes upon satellite galaxies is conducted with reference to G5-7 only. However, I note that external processes may influence the evolution of $z \sim 0.65$ galaxies that have stellar masses below the minimum of $10^{9.5}$ $M_\odot$ that I impose.

Mergers and gravitational interactions, more common in environments of higher densities (Renzini, 1999; Tonini et al., 2016), are external processes which can increase the Sérsic indices of the galaxies involved (Naab & Trujillo, 2006; Aceves et al., 2006; Fisher & Drory, 2008). The precise morphology of a merger remnant, while generally early-type, is dependent upon the configuration of the merger (Toomre, 1977; Barnes, 1988, 1992; Walker et al., 1996). Major mergers are capable

of destroying the discs of late-type galaxies, while minor mergers can instead simply promote the formation of a classical bulge within a disc that remains mostly intact. In addition, gravitational interactions between galaxies as they pass by one another ("harassment") can gradually change their morphologies from disc- to spheroid-dominated (Moore et al., 1996; Smith et al., 2015). Figure 5.9 shows a range of bulge-to-total ratios among galaxies in G5-7, which may be capturing this varying degree to which mergers and gravitational interactions can disrupt the morphologies of their progenitors over time. While most of these galaxies are strongly spheroid-dominated, others (while still having high Sérsic indices which indicate the presence of bulge) retain a disc component (with $B/T_r$ values as low as $\sim 0.3$).

Whether these processes are also responsible for the quenching of G5-7 galaxies is unclear. Gravitational interactions between merging galaxies can induce central starbursts which rapidly exhaust their supplies of star-forming gas (i.e. like the PSB candidates that I suggest comprise G5), and/or can catalyse nuclear activity which inhibits any further star formation (Mihos & Hernquist, 1994a,b, 1996; Di Matteo et al., 2005; Springel et al., 2005a,c). Hence, a post-merger remnant is in fact quenched by internal processes (i.e. exhaustion of its fuel, or supermassive black hole feedback; see also above). However, a sufficiently gas-rich major merger may lead its remnant to form with a disc and continue forming stars (Barnes, 2002; Hopkins et al., 2009a,b, 2010). In addition, a merger remnant may eventually accrete new gas such that it can form a new disc begin a new episode of star formation (Salim & Rich, 2010; Gabor et al., 2011). Generally, mergers cannot be unequivocally linked with the quenching of galaxies (see also Weigel et al. 2017), and so it is more likely that galaxies are quenched mainly by other processes.

Several external processes have been proposed to explain the quenching of star-forming galaxies as they become satellites. Examples include ram-pressure stripping (Gunn & Gott, 1972; McCarthy et al., 2008), thermal evaporation (Cowie & Songaila, 1977; Nipoti & Binney, 2007), and viscous stripping (Nulsen, 1982; Kraft et al., 2017), all of which invoke the removal of the cold ISM of a galaxy via its hydrodynamical interaction with the hot IGM of high-density environments as the reason for quenching. These processes are correlated with the velocity of a galaxy as it travels through its environment, and generally quench galaxies quickly. Gas may also be removed from the extended halo of a galaxy at the outskirts of a dense environment, *by* the gravitational influence of that environment as a whole ("strangulation" or "starvation"; Larson et al. 1980; Peng et al. 2015). The galaxy then quenches slowly by exhausting any remaining gas in its disc. The balance of these processes is not yet known (Bahé & McCarthy, 2015; Peng et al., 2015; Smethurst et al., 2017). Recent studies, though, advocate for a general "delayed-then-rapid" quenching pathway (Wetzel et al., 2012, 2013; Muzzin et al., 2014; Moutard et al., 2018). Galaxies initially quench slowly at the outskirts of the environment, then quickly as they approach its core, where the conditions for the aforementioned hydrodynamical interactions are expected. This delay could also explain the large spreads in the environmental distributions among all of the clusters in Figure 5.6.

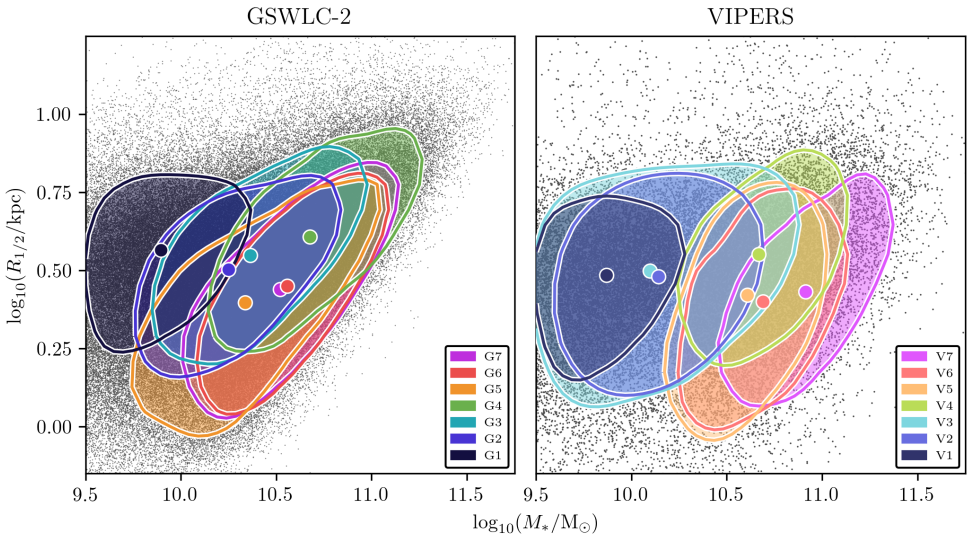


**Figure 5.10:** Half-light radius versus stellar mass for the galaxies in the samples. Circularised half-light radii are calculated from single Sérsic fits by Simard et al. (2011) for the GSWLC-2 sample, and Krywult et al. (2017) for the VIPERS sample. The distributions of clusters are shown using coloured, filled contours (drawn at a relative density of 0.4), and the coloured, circular markers show their medians.

These quenching processes are, in turn, unlikely to be responsible for the strongly spheroid-dominated morphologies of low-mass passive galaxies (Bekki et al., 2002; Boselli et al., 2009; Zinger et al., 2018). Hence, I suggest that the strong overlap between low-mass passive galaxies and spheroids appears to be a product of environment, which drives quenching and morphological transformation separately (Poggianti et al., 1999). In addition, it implies that the quenching of galaxies precedes, or at least be simultaneous to, their morphological transformation (Schawinski et al., 2014; Woo et al., 2017). While the merger of two gas-rich, star-forming galaxies may produce a rejuvenated remnant, mergers between passive progenitors will invariably produce passive remnants with increasingly spheroidal morphologies, ranging from lenticular galaxies with classical bulges (Mishra et al., 2017, 2018, 2019) through to pure spheroids.

### 5.4.3 Clusters in the size-mass plane

Figure 5.10 shows the size-mass distribution of the clusters in each of the samples. The stellar masses originate from the same `CIGALE` SEDs that were used to generate the colours with which the galaxies are represented for the clustering. The half-light radii of these galaxies are based on fits of single Sérsic profiles to their light distributions (see Sections 5.1.1 and 5.1.2). The size of a galaxy, in the context of its stellar mass and its morphology, is another important record of its assembly history. The positions and distributions of both sets of clusters in these plots match well with broader blue versus red, and early- versus late-type distinctions made in the same (or similar)

plane(s) by other studies (Shen et al., 2003; van der Wel et al., 2014; Lange et al., 2015).

The most significant difference between the two plots in Figure 5.10 is the absence of compact massive galaxies in the GSWLC-2 sample in comparison with the VIPERS sample. The canonical explanation for the growth of these galaxies is ongoing minor merger activity and accretion (Naab et al., 2009; Hopkins et al., 2010). The resultant shift between the passive VIPERS clusters and the passive GSWLC-2 clusters is approximately in accordance with the expected redshift evolution of the size-mass relation for early-type, passive galaxies (van Dokkum et al., 2015), though the mass-incompleteness of passive VIPERS galaxies means that this shift is unlikely to have been captured accurately in this chapter. The large overlap of G4 and V4 with their respective passive clusters in Figure 5.10 seems to support the additional "late-track" (late with respect to cosmic time rather than to morphology) of galaxy evolution proposed by Barro et al. (2013) to yield disc-dominated passive galaxies (Ilbert et al. 2010; Carollo et al. 2013; Schawinski et al. 2014). Both sets of SFMS clusters are similarly distributed, capturing the minimal evolution of the sizes of star-forming galaxies between their two redshifts (Lilly et al., 1998; van der Wel et al., 2014).

## 5.5 Summary and conclusions

I present results from the application of the `SEM` clustering algorithm to samples of galaxies at low ($z \sim 0.06$, from GSWLC-2) and intermediate ($z \sim 0.65$, from VIPERS) redshifts. Galaxies are represented using nine UV-through-NIR broadband rest-frame colours, derived from fits of ensembles of synthetic spectra to observed photometry with `CIGALE`. My aims, following Siudek et al. (2018b), were to use unsupervised machine learning to search within these colours for substructures to the established colour bimodality of galaxies, and to understand the evolution of subpopulations of galaxies in terms of these colours over cosmic time. An advantage of `SEM` is its incorporation of dimensionality reduction on the fly, which ensures that it determines clusters using only the most important and discriminative information encoded within the input features. I summarise my results as follows:

1. My cluster evaluation search reveals that both of the samples are best partitioned into seven clusters (Table 5.1). In addition, the best-fitting submodels to each of the samples, identified independently, are closely related, both allowing variation in the shapes of clusters and differing only in their treatment of "noise" among the input features. For both samples, these seven clusters break down into four star-forming clusters and three passive clusters (Figure 5.4).

2. The lack of a cluster in either partition that is confined in its extent to just the green valley (Figure 5.4) rules the green valley out as containing a singular, monolithic population of

galaxies (e.g. Baldry et al. 2004; Taylor et al. 2015; Krywult et al., in prep.). This also confirms that there is more than one evolutionary pathway of galaxies through the green valley (e.g. Faber et al. 2007; Fritz et al. 2014; Schawinski et al. 2014; Smethurst et al. 2015; Moutard et al. 2016b).

3. Overall, `SEM` uses the nine rest-frame colours similarly to determine the partitions (Figure 5.2), reducing the dimensionality of the feature space to 6 in both cases. Altogether, optical colours are most important to the clustering; individually, UV colours are. The availability of photometry with which to constrain the SEDs of galaxies is advantageous to the clustering. UV colours are slightly more important to the clustering in the GSWLC-2 sample, which has more GALEX coverage than the VIPERS sample. Similarly, the lack of any NIR coverage for the GSWLC-2 sample means that NIR colours are less important to its clustering. However, given the broader overall similarity between the clustering structures of the samples (Figure 5.3), it appears that clustering (a statistical method) combined with SED estimation (which can infer rest-frame magnitudes from incomplete photometry) has enabled a partial "filling of the gaps" of missing data in both samples.

4. Star-forming clusters in both samples form clear morphological sequences (Figure 5.7). The correlation between their median Sérsic indices and their median stellar masses captures the growth of the bulges of their galaxies along the SFMS (Figure 5.9). At the highest masses, this growth corresponds with a drop in specific star formation rates. Hence, the quenching of high-mass galaxies is influenced by their inner stellar densities, above a certain threshold, which appears to be linked with nuclear activity (Figure C.4). The retention of discs by the highest-mass galaxies along this morphological sequence indicates that some galaxies quench without fully transforming their morphologies. The lack of a strong trend of these clusters with local environmental overdensity (Figure 5.6) suggests that this evolutionary pathway is dominated by internal processes. This pathway, prominent at the epochs of both samples, appears consistent with "mass quenching', as proposed by Peng et al. (2010). It is expected that the long timescales involved would ultimately lead the VIPERS star-forming clusters to resemble the GSWLC-2 star-forming clusters by the present day.

5. Galaxies in passive clusters in both samples have uniformly high Sérsic indices, indicating a fundamental link between centrally-concentrated morphologies and passiveness (Figure 5.7). Passive clusters in the low-redshift sample are separated from their respective sequence of star-forming clusters, particularly towards lower stellar masses (Figures 5.7 and 5.9). This separation is assumed to originate from the influence of alternative or additional processes to those that dictate the evolution of actively star-forming galaxies. Invoking the offset of these low-redshift passive clusters to high local environmental overdensities (Figure 5.6), I suggest that some of their galaxies are low-mass satellites, and subject to external processes. The homogeneity of their early-type morphologies implies that their

quenching precedes, or is at least simultaneous to, their morphological transformation. In all, this pathway appears consistent with "environment quenching" (Peng et al., 2010). This morphological separation is not as apparent for the passive clusters in the VIPERS sample (Figure 5.7), which is mainly due to incompleteness of low-mass passive galaxies (which would also be expected to trace high-density environments). Hence, I am prohibited from commenting on the prevalence of this evolutionary pathway at intermediate redshifts.

This work confirms the existence of two distinct evolutionary pathways of galaxies through the green valley (Poggianti et al., 1999; Faber et al., 2007; Peng et al., 2010; Barro et al., 2013; Fritz et al., 2014; Schawinski et al., 2014; Moutard et al., 2016b). In addition, these pathways, including their effects on the morphologies of galaxies, appear to be strongly encoded within their SEDs, as estimated from broadband photometry. This invites further investigation of the extent to which a galaxy's assembly history may be discerned purely from its SED.

The use of further ancillary features would be instrumental in substantiating and constraining these pathways. A wealth of such features are available for the GSWLC-2 sample, due to its basis in SDSS. Examples include Galaxy Zoo 2 morphologies (Willett et al., 2013) which include bar and merger classifications, and Yang et al. (2007) group memberships to enable a distinction between central and satellite galaxies. A more detailed analysis of the low-redshift sample in this manner is reserved for a future study. The GAMA survey (Driver et al., 2009) could provide an alternative low-redshift sample, given its panchromatic data release (Driver et al., 2016) and its rich library of value-added catalogues (Baldry et al., 2018). The upcoming Deep Extragalactic VIsible Legacy Survey (DEVILS; Davies et al. 2018), which aims to improve completeness at $0.3 < z < 1.0$, could be the basis for an improved intermediate-redshift sample upon its completion. Furthermore, the Legacy Survey of Space and Time (Ivezić et al., 2019), which will provide galaxy colours and morphologies together, constitutes a particularly promising foundation for a future follow-up study.

The incompleteness of low-mass passive galaxies at intermediate redshifts would be alleviated by moving to deeper surveys such as G10-COSMOS (Andrews et al., 2017) and 3D-HST (Momcheva et al., 2016), both of which also have panchromatic photometric data releases. This would enable an examination of environment quenching at earlier epochs, and of its proposed increase in prevalence at lower redshifts (Fossati et al., 2017; Moutard et al., 2018; Papovich et al., 2018). Surveys like this could also extend the comparison to redshifts as high as $z \sim 2$, thus facilitating the constraint of the changing balance of evolutionary pathways, informed by clustering of rest-frame colours, over a greater extent of cosmic time.

# Chapter 6

# Summary, conclusions, and future prospects

The work presented in this thesis has explored the use of clustering algorithms for study of galaxy evolution. I will begin this chapter by summarising the results and discussion of this thesis thus far.

In Chapter 1, I reviewed the field of galaxy evolution. I covered the advances made by observational campaigns, including the emergence of a rudimentary empirical picture of galaxy evolution on the basis of feature distributions revealed by survey astronomy. I listed theoretical processes that are used to explain the properties of galaxies, and also explained the importance of cosmology as a factor in their development. I visited the progress made by cosmological simulations, and concluded by arguing that multi-dimensional feature spaces must be explored in order to better constrain the interplay of processes that direct galaxy evolution. In Chapter 2, I reviewed machine learning, focusing on clustering and dimensionality reduction and on prior uses of these techniques in the research of galaxy evolution. In addition, I made the case for my use of particular algorithms in Chapters 3-5.

In Chapter 3 and Appendix A, I clustered a pilot sample of galaxies from the GAMA survey using the $k$-means method. Galaxies were represented by five features: stellar masses, $u - r$ colours, Sérsic indices, half-light radii, and specific star formation rates. Clustering was conducted with a unique stability-based cluster evaluation framework, which highlighted outcomes consisting of 2, 3, 5, and 6 clusters as being stable. At $k = 2$ and $k = 3$, the structures of outcomes were dictated by the colours and star formation activity of galaxies. At $k = 5$ and $k = 6$, the sizes and Sérsic indices of galaxies became more important to the clustering. These four outcomes formed a hierarchical structure, dominated at all four levels by a split into two "superclusters" which corresponded with

established notions of bimodalities and/or dichotomies of galaxies in terms of their morphologies and colours. This correspondence was increasingly accurate at higher values of $k$. There was a broad agreement of these four outcomes with Hubble-like morphological classifications, and it was suggested that this agreement might be improved by the addition of further morphological information to the feature set. Outcomes at $k = 5$ and $k = 6$ highlighted the differential role of environment in the evolution of galaxies on their passage through the green valley. Among the four stable outcomes, the $k = 6$ outcome was designated as being optimal as it offered the most detailed and astrophysically meaningful partition while still being highly reproducible.

In Chapter 4 and Appendix B, I developed the approach of Chapter 3 in order to test the cosmological, hydrodynamical EAGLE simulations against the GAMA survey. Clustering was conducted within a five-dimensional feature space shared by both samples. The cluster evaluation framework highlighted a $k = 5$ outcome as being optimal for the simulated galaxies, and a $k = 7$ outcome for the observed galaxies. These outcomes returned an agreement score of $V_a = 0.76$, indicating broad structural similarity, but notable differences in their substructures. These outcomes were then compared on a cluster-by-cluster basis by mapping the clusters determined using EAGLE galaxies onto the GAMA galaxies, which was enabled by the use of a shared feature space. This comparison revealed that the growth of the central bulges of EAGLE galaxies during their evolution along the star-forming main sequence is unrealistic, that EAGLE contains too many low-mass spheroid-dominated galaxies, and that EAGLE produces a subpopulation of high-mass, disc-dominated, star-forming galaxies that do not exist in the GAMA sample. These discrepancies were attributed chiefly to limitations in the resolution of EAGLE, and to insufficiently powerful AGN feedback at high stellar masses.

In Chapter 5 and Appendix C, I compared the clustering structures of samples of galaxies at low ($z \sim 0.06$) and intermediate redshifts ($z \sim 0.67$), from GSWLC-2 and VIPERS respectively. The input features to the clustering were nine rest-frame UV-through-NIR colours. I switched to the Subspace Expectation-Maximisation algorithm for clustering, which incorporates adaptive dimensionality reduction as it iterates. A search over various combinations of submodels and values of $k$ showed that both samples were best fit by seven-cluster partitions, which (in both cases) broke down into four clusters of mainly star-forming galaxies, and three of mainly passive galaxies, and none confined to just the green valley. Star-forming clusters at both epochs exhibited clear morphological trends along the star-forming main sequences of both samples, indicating a link between the central stellar densities of galaxies and their quenching. This trend was also, at low redshifts, linked with an increase in AGN activity. Galaxies in passive clusters in both samples had uniformly high Sérsic indices, suggesting that they follow an alternative evolutionary pathway. Evidence at low redshifts suggests that external processes are involved; mass-incompleteness of the VIPERS sample prohibits the generalisation of this conclusion to intermediate redshifts.

To conclude this thesis, I return to the main questions that I posed at the start of Chapter 1, and comment on how my work addresses them.

**What place does clustering, and by extension unsupervised machine learning in general, have among the arsenal of methods used in future studies of galaxy evolution?**

The results presented in this thesis demonstrate that clustering is a useful method for the analysis of galaxies and their evolution. Use of prototype- and model-based clustering methods leads to clusters that are readily interpretable in terms of their identities. In addition, clustering outcomes from all three of Chapters 3, 4, and 5 can be clearly connected with the evolutionary contexts of galaxies that they contained. The stability-based cluster evaluation approach employed in Chapters 3 and 4 facilitates the robust discovery of reproducible clustering structures in feature spaces of high dimensionalities, and may readily be adapted for use with clustering methods other than the *k*-means method, which may be appropriate for other studies. The utility of this evaluation approach for the multi-dimensional validation of simulations is proven by its ability to highlight specific discrepancies between the EAGLE and GAMA samples of Chapter 4, which can then be attributed directly to particular limitations of the EAGLE simulations. Chapter 5 shows that a combination of clustering and dimensionality reduction are effective in extracting astrophysical information from the SEDs of galaxies, including morphological distinctions between galaxies that would typically be grouped together on the basis of their colours. Hence, unsupervised machine learning techniques are well-poised to assist in the handling of data corresponding to the large numbers of galaxies that will be observed by future large-scale photometric surveys.

**Can clustering in feature spaces of high dimensionalities reveal substructures to the established dichotomies, or bimodalities, of galaxies?**

The broader, global structures of all of the clustering outcomes determined in this thesis correspond well with established notions of dichotomies and bimodalities of galaxies, indicating that these dichotomies and bimodalities are a fundamental characteristic of the overall galaxy popoulation. However, individual clusters single out subpopulations of galaxies that offer a more detailed view of the relationships between the features used to describe galaxies. Clusters from outcomes in Chapter 3 ($Bd_6$) and 5 (G4) include clusters that contained some passive but disc-dominated galaxies. These clusters are more closely associated with the star-forming main sequence than with other clusters of passive galaxies. This suggests that red peak of the colour bimodality does not represent a monolithic subpopulation of galaxies, but that it is partially built up by contributions from different pathways, including galaxies that come from the high-mass end of the star-forming main sequence. In both of Chapters 3 and 4, low-mass star-forming galaxies are distinguished by

the clustering as being either bulge-dominated or disc-dominated, suggesting differing formation mechanisms among galaxies that exhibit similar levels of star formation activity. That the bulge-dominated subpopulation is not distinguished in Chapter 5 likely arises from their similarity in star formation activity with their disc-dominated counterparts. This implies that the most detailed partitions will emerge from the use of a mixture of features which captures various characteristics of the galaxies that they describe.

**If so, can these substructures be used to constrain the balance of theoretical processes that have been proposed as driving galaxy evolution?**

Outcomes from all three of Chapters 3, 4, and 5 segment the star-forming main sequence of galaxies in a similar manner, capturing an increase in the inner stellar densities of galaxies with increasing stellar mass. This increase is, again in all three cases (except for the simulated sample in Chapter 4), accompanied by a decrease in star formation activity – the observed turn-down of the star-forming main sequence. In Chapter 5, this increase is also accompanied by an increase in active galactic nucleus emission. This suggests that the evolution of galaxies along the star-forming main sequence is regulated by active galactic nucleus feedback, for which the inner stellar densities of galaxies are a marker. Clusters of low-mass passive galaxies in Chapters 3 and 4 exhibit a spread in the Sérsic indices and bulge-to-total ratios of their galaxies respectively. Galaxies in passive clusters from the outcomes in Chapter 5 have uniformly high Sérsic indices, but exhibit a spread in their bulge-to-total ratios at low redshift. These results imply a gradual transformation of the morphologies of the galaxies contained by these clusters *following* their quenching. This, in tandem with their likelihood of occupying environments of high densities, suggests that their evolution, generally as satellites, is dictated by external processes, with the action of stripping processes followed by successive minor mergers a possible scenario. Hence, in general, the clusters that comprise the various outcomes determined in this thesis can be linked with different evolutionary pathways and with different processes.

## 6.1 Future prospects

The ends of Chapters 3, 4, and 5 include suggestions for future work that are made on the basis of the work presented in each of those chapters. Here, I make suggestions for future work on the basis of concepts visited throughout this thesis as a whole.

It was suggested at the end of Chapter 4 that machine learning techniques would be useful for linking "hidden" features, measurable only in simulations, with observable features. Here, I build upon this suggestion to explain how this approach might be used to constrain the evolutionary

pathways of galaxies in groups. The influence of external processes upon galaxies in group environments depends upon their situation and motion with respect to the group as a whole and to one another. This information may be described theoretically using a six-dimensional phase space, consisting of three position components and three velocity components. Observationally, only two of these dimensions are available, as part of what is called a "projected phase space" (Gill et al., 2005; Oman et al., 2013; Oman & Hudson, 2016) which is defined by perpendicular separations and line-of-sight velocities of galaxies. Exploration of the six-dimensional phase space is possible through the use of cosmological galaxy simulations. Semi-analytic models are particularly appropriate: dark-matter only simulations may be used to derive the positions and velocities of galaxies, and analytical prescriptions may be used to predict subsequent evolutionary outcomes. In addition, the variation of these analytical prescriptions could then grant insight into balance of processes acting in groups, via their validation against the properties of observed galaxies. Projected phases spaces containing observed galaxies, meanwhile, have previously been used to distinguish infalling galaxies from embedded group members. However, the linking of full six-dimensional phase spaces with two-dimensional projected phase spaces could be further facilitated though the use of machine learning techniques, and especially clustering methods. This would be by clustering galaxies in six dimensions, with a view to distinguishing galaxies at various stages of infall, including "splashback galaxies" (at large radii after having passed through the group centre), and subgroups of galaxies that have been pre-processed together. Then, the distributions of these clusters could be examined in two dimensions to provide a comprehensive map with which to probe the evolution of observed group galaxies (from large-scale surveys like SDSS or the Hyper Suprime-Cam Subaru Strategic Programme; Aihara et al. 2018a,b). Hence, this approach could be important in the validation of the stripping-then-transformation scenario proposed in Chapter 5 by helping to constrain the relative timescales of these two processes.

The morphological features that have been used in thesis, both as inputs to the clustering and for the interpretation of cluster identities, have been monochromatic. It has, however, been shown that quantitative morphological features (and especially Sérsic indices) are dependent upon the effective wavelength of the filter through which they are measured (Kelvin et al., 2012; Häußler et al., 2013; Vika et al., 2013). As a result, Vulcani et al. (2014) and Vika et al. (2015) manufactured a feature, $\mathcal{N}$, which, as the ratio between Sérsic indices in two different filters, enabled a better separation of early-type galaxies over monochromatic Sérsic indices. Machine learning techniques enable the simultaneous combination of Sérsic indices measured through yet more filters, with the potential of offering a finer view of the multi-wavelength morphologies of galaxies. Dimensionality techniques would be particularly appropriate for condensing this multi-wavelength information into a smaller set of summary features. Figure 6.1 shows exploratory two-dimensional projections, determined using uniform manifold approximation and projection (McInnes et al., 2018), of a 10-dimensional subsample (5,606 galaxies) of Chapter 3's pilot GAMA sample. Galaxies are

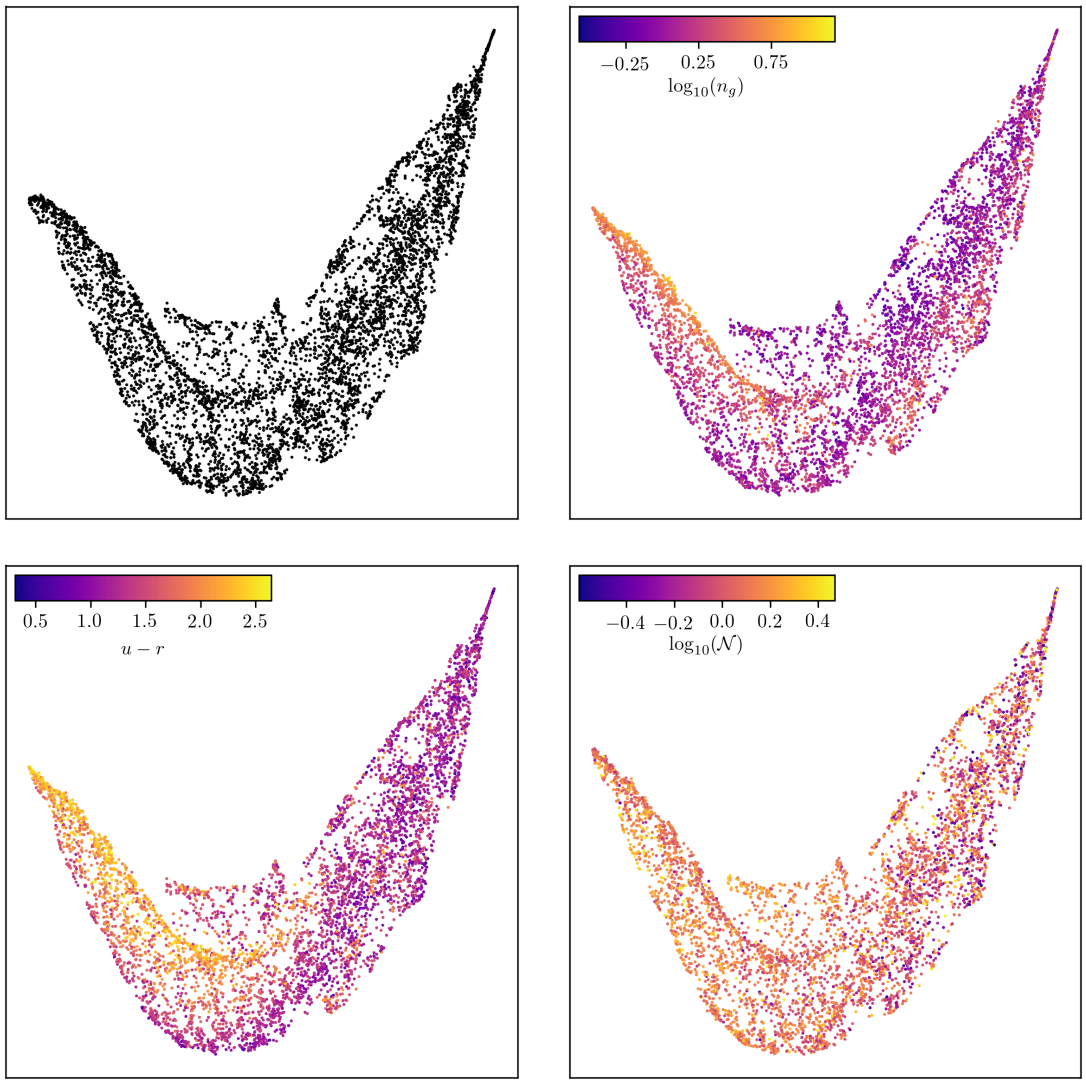


**Figure 6.1:** Two-dimensional projections, via uniform manifold approximation and projection (McInnes et al., 2018), of a 10-dimensional (multi-wavelength Sérsic indices and absolute magnitudes, see main text) sample of galaxies from the GAMA survey. Points are coloured (except in the upper left panel) with respect to certain ancillary features, according to the inset colour bars. Note that the Sérsic indices used to colour points in the upper right panel are measured in the $r$ band; the symbol $n_g$ is used for consistency with notation in the rest of this thesis. In addition, some outlier points have been omitted from the lower right panel in order to ensure representative coverage of $\mathcal{N}$ values by the colour bar.

represented in the 10-dimensional feature space by $u$-, $g$-, $r$-, $i$-, and $z$-band absolute magnitudes and Sérsic indices, via Kelvin et al. (2012). The projections show that the subsample has a non-linear structure in 10 dimensions, but also that it follows a clear sequence. The panels containing coloured points show that there is more structure within this 10-dimensional feature space than can be captured by monochromatic Sérsic indices, $u - r$ colours, and $\mathcal{N}$. The manufacture of further multi-wavelength morphological features from this 10-dimensional feature space through the use of dimensionality reduction could hence promote a better understanding of the growth of the structural components of galaxies.

# Appendix A

# Appendix to Chapter 3

## A.1 Stability Simulation

To demonstrate the use of stability for selecting good values of $k$, I set up a simple simulation (Figure A.1). $5,000$ data points are distributed equally over five two-dimensional Gaussian functions, centred at the vertices of a unit regular pentagon. The standard deviations of the distributions ($\sigma = 0.3$) are set such that they overlap slightly. The value of $k_{true}$ for this simulation is 5. I run `k-means` with $k = 4, 5$, and 6. I initialise 200 times at each $k$ using the Arthur & Vassilvitskii (2007) technique. Cluster names in this section consist of three parts in the format "X$Y_Z$". The first part, either "A" or "B", corresponds to a particular outcome to which the cluster belongs, and is used to identify outcomes in the figures in this section. The second part, a number, corresponds to the individual cluster, also shown in the figures. The third part, another number, indicates the value of $k$ at which the outcome was found.

Figure A.2 shows two examples of the outcomes found at $k = 4 < k_{true}$. In both cases, `k-means`

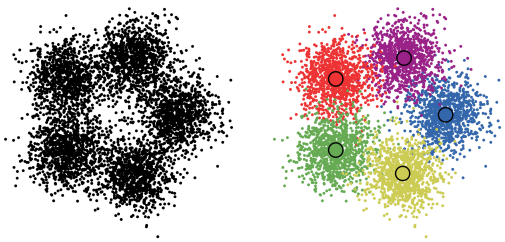

**Figure A.1:** On the left-hand side I display the simple two-dimensional simulation, containing five true clusters. See the main text for information on how it is generated. Points are coloured by their truth labels; all points with the same colour belong to the same true cluster, whose centroid is marked out by a large filled circle, also of the same colour.

**Table A.1:** Contingency table, comparing outcomes A and B generated at $k = 4$ and shown in Figure A.2.

| Cluster | $B1_4$ | $B2_4$ | $B3_4$ | $B4_4$ |
|---|---|---|---|---|
| $A1_4$ | 330 | 0 | 830 | 0 |
| $A2_4$ | 0 | 946 | 0 | 194 |
| $A3_4$ | 808 | 0 | 1 | 839 |
| $A4_4$ | 0 | 245 | 807 | 0 |

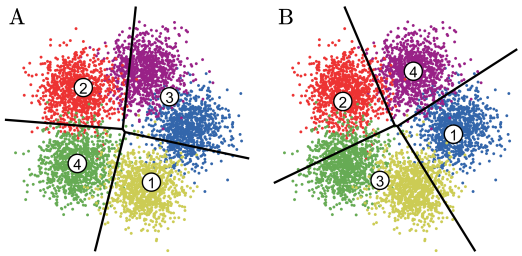


**Figure A.2:** Examples of `k-means` clustering at $k = 4 < k_{true}$. The algorithm has merged the purple and blue true clusters in outcome A on the left, and the yellow and green true clusters in outcome B on the right. The `k-means` centres are marked by filled white circles. The boundaries between `k-means` clusters are marked by straight black lines.

merges two true clusters: purple and blue in outcome A on the left, and yellow and green in outcome B on the right. These mergers have affected the accuracy of the neighbouring `k-means` clusters as well in that they suffer from contamination (in terms of the true cluster structure). Table A.1, a contingency table (a.k.a. a cross-tabulation), shows that the outcomes are only weakly associated with one another. The chi-squared value (Equation 3.4) for these two outcomes (A and B), calculated using the contingency table, is $6,617.95$. From this, using Equation 3.3, I calculate $V_s = 0.66$ (with $N = 5,000$ and $k = 4$).

Figure A.3 shows two examples of the outcomes found when $k = 5 = k_{true}$. While they appear identical, they actually differ by four points (see contingency Table A.1, which shows the near-perfect association between the two outcomes). `k-means` has succeeded in finding the five true clusters in both outcomes. While is not impossible that `k-means` might find an alternative structure in the simulation at $k = 5$ given more initialisations, the rate at which it would do so would be so low (less than at most 0.5 per cent given Figure A.3) that $k = 5$ would still stand out as being particularly stable. For these two $k = 5$ outcomes (A and B), I calculate $\chi^2 = 19,960.03$ and (with $N = 5,000, k = 5$) $V_s = 0.999$.

Figure A.4 shows two examples of the outcomes found at $k = 6 > k_{true}$. The algorithm has split a true cluster in both cases: green in outcome A on the left, and yellow in outcome B on the right. The splits appear to have a lesser effect on neighbouring `k-means` clusters than the mergers at $k = 4$, in that there is less contamination overall. Contingency Table A.3 reveals that the outcomes are more strongly associated with one another than those outcomes found at $k = 4$, as the split in

**Table A.2:** Contingency table, comparing outcomes A and B generated at $k = 5$ and shown in Figure A.3.

| Cluster | $B1_5$ | $B2_5$ | $B3_5$ | $B4_5$ | $B5_5$ |
|---|---|---|---|---|---|
| $A1_5$ | 0 | 985 | 0 | 0 | 0 |
| $A2_5$ | 0 | 3 | 0 | 0 | 1,005 |
| $A3_5$ | 0 | 0 | 0 | 993 | 0 |
| $A4_5$ | 0 | 0 | 1,013 | 1 | 0 |
| $A5_5$ | 1,000 | 0 | 0 | 0 | 0 |

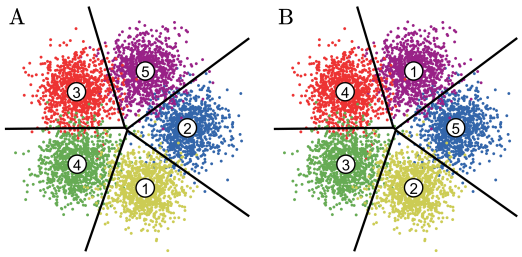


**Figure A.3:** Examples of `k-means` clustering at $k = k_{true} = 5$. The algorithm has correctly found the five true clusters in both outcomes A and B, which differ by only 4 points. The $k$-means centres are marked by filled white circles. The boundaries between `k-means` clusters are marked by straight black lines.

one outcome fits more cleanly into a whole cluster in the other. For these two $k = 6$ outcomes (A and B), I calculate $\chi^2 = 18,017.03$ and (with $N = 5,000, k = 6$) $V_s = 0.85$.

I summarise these results using a stability map (Figure A.5). I reemphasise that the key element of this plot for distinguishing stable and unstable values of $k$ is the gap across all distributions in median $V_s$. The distribution of outcomes at $k = 5$, showing that all 200 initialisations converged to the same stable outcome (within 4 points), is clearly indicative of the true structure of the simulation. The distributions of outcomes at $k = 4$ and $k = 6$, indicating that they are unstable, reflect that there is no objectively correct way to divide the five true clusters into four or six given the symmetry of the simulation. The distribution at $k = 6$ is narrower because splits affect the accuracy of the other `k-means` clusters less than mergers. With the benefit of knowing the true structure of the simulation, The $k = 6$ splits could be remerged and achieve a better approximation to the $k = 5$ outcomes than if the $k = 4$ merges were split. For more complicated samples, involving more features, this effect would be more difficult to discover and exploit.

Given that the same outcome may arise several times over a large number of initialisations, one may opt to select the modal outcome as the most optimal instead of that with the lowest $\phi$. In practice, I find that one criterion implies the other; the most compact clusters tend to emerge most often anyway (thanks to my choice of initialisation technique). I retain $\phi$ as my criterion for optimal clustering at given values of $k$.

**Table A.3:** Contingency table, comparing outcomes A and B generated at $k = 6$ and shown in Figure A.4.

| Cluster | $B1_6$ | $B2_6$ | $B3_6$ | $B4_6$ | $B5_6$ | $B6_6$ |
|---|---|---|---|---|---|---|
| $A1_6$ | 0 | 0 | 658 | 0 | 0 | 8 |
| $A2_6$ | 1 | 988 | 0 | 0 | 0 | 0 |
| $A3_6$ | 258 | 0 | 0 | 0 | 0 | 658 |
| $A4_6$ | 0 | 1 | 0 | 0 | 954 | 0 |
| $A5_6$ | 148 | 3 | 0 | 844 | 0 | 0 |
| $A6_6$ | 34 | 4 | 338 | 0 | 37 | 31 |

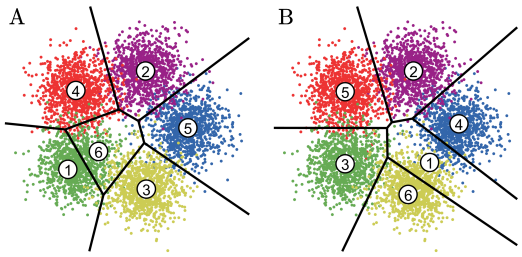


**Figure A.4:** Examples of `k-means` clustering at $k = 6 > k_{true}$. The algorithm has split the green true cluster in the example on the left, and the yellow true cluster in the example on the right. The $k$-means centres are marked by filled white circles. The boundaries between `k-means` clusters are marked by straight black lines.

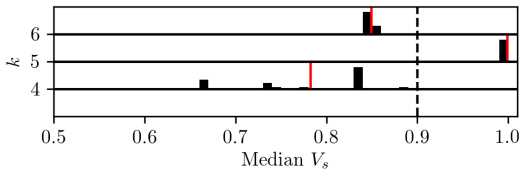


**Figure A.5:** Stability map of `k-means` clustering for the simulated data set at $k = 4, 5$, and 6. I calculate the median $V_s$ of each outcome with respect to all other outcomes at the same $k$. The distributions of all 200 medians at each $k$ are represented using histograms plotted along each of the horizontal black baselines. The heights of the histograms are normalised. Additionally, the means of these distributions are shown as vertical red lines. The outcome at $k = 5$ stands out as being particularly stable, indicative of the true structure of the simulation.

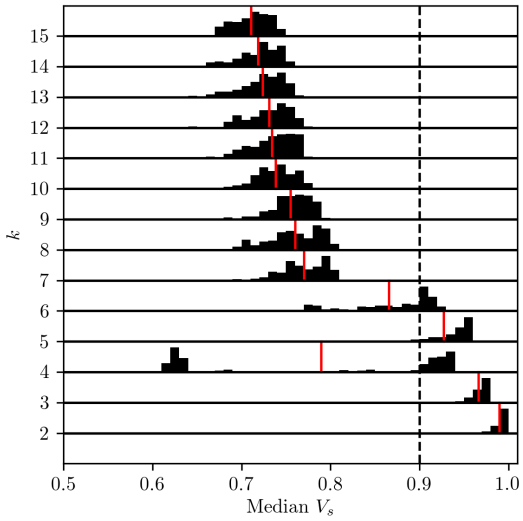


**Figure A.6:** Stability map of `k-means` clustering for the bootstrapped pilot sample at $k = 2$ through $k = 15$. I calculate the median $V_s$ of each outcome with respect to all other outcomes at the same $k$. The distributions of all $7,338$ medians at each $k$ are represented using histograms plotted along each of the horizontal black baselines. The heights of the histograms are normalised. Additionally, the means of these are distributions as vertical red lines. Outcomes at $k = 2, 3, 5$, and 6 remain most stable following application of the bootstrap method to the pilot sample.

## A.2 Bootstrap experiment

In order to estimate the uncertainties on the centroids reported in Table 3.5, I apply the bootstrap method to the pilot sample of galaxies. The method resamples the original sample with replacement, such that the same galaxy may be selected more than once. A total of $7,338$ observations are selected in this manner. I run `k-means` once on this new sample, retaining the centroids, and then partition the original sample according to these centroids. This whole process is itself repeated $7,338$ times. The stability map for the bootstrap experiment is shown in Figure A.6.

The distributions of outcomes at all values of $k$ are shifted to lower stabilities following application of the bootstrap method. This is in comparison with the distributions generated purely from the original sample, shown in Figure 3.4. Outcomes at $k = 2, 3, 5$, and 6 remain the most stable in Figure A.6, though outcomes at $k = 6$ exhibit a more significant reduction in stability than outcomes at $k = 2, 3$, and 5 due to the increased local dependency of `k-means` with a higher numbers of centres. Outcomes at $k = 4$ retain their bimodal structure in stability. The distribution of outcomes at $k = 7$, which exhibited a stable component in Figure 3.4, is now uniformly unstable to the same extent as the distributions at higher values of $k$, justifying the exclusion of outcomes from $k = 7$ from the analyses in Section 3.3.

For the uncertainties in Table 3.5, I calculate the 16th and 84th percentiles of the $7,338$ centroids,

in each of the five features. From these, I subtract the original `k-means` centroids.

## A.3 Postage Stamps

Here I present example postage stamps of galaxies in each of the clusters in each of the clustering outcomes in Chapter 3. The three-colour stamps are made using *r*- and *g*-band imaging from the Kilo-Degree Survey (de Jong et al., 2013), and a mean of the two bands as the central colour channel. The stamps enclose each galaxy to 2.5 times its Kron radius. The examples shown are those that are best represented by the cluster centroids; they are nearest to the centroids in the five-dimensional feature space.

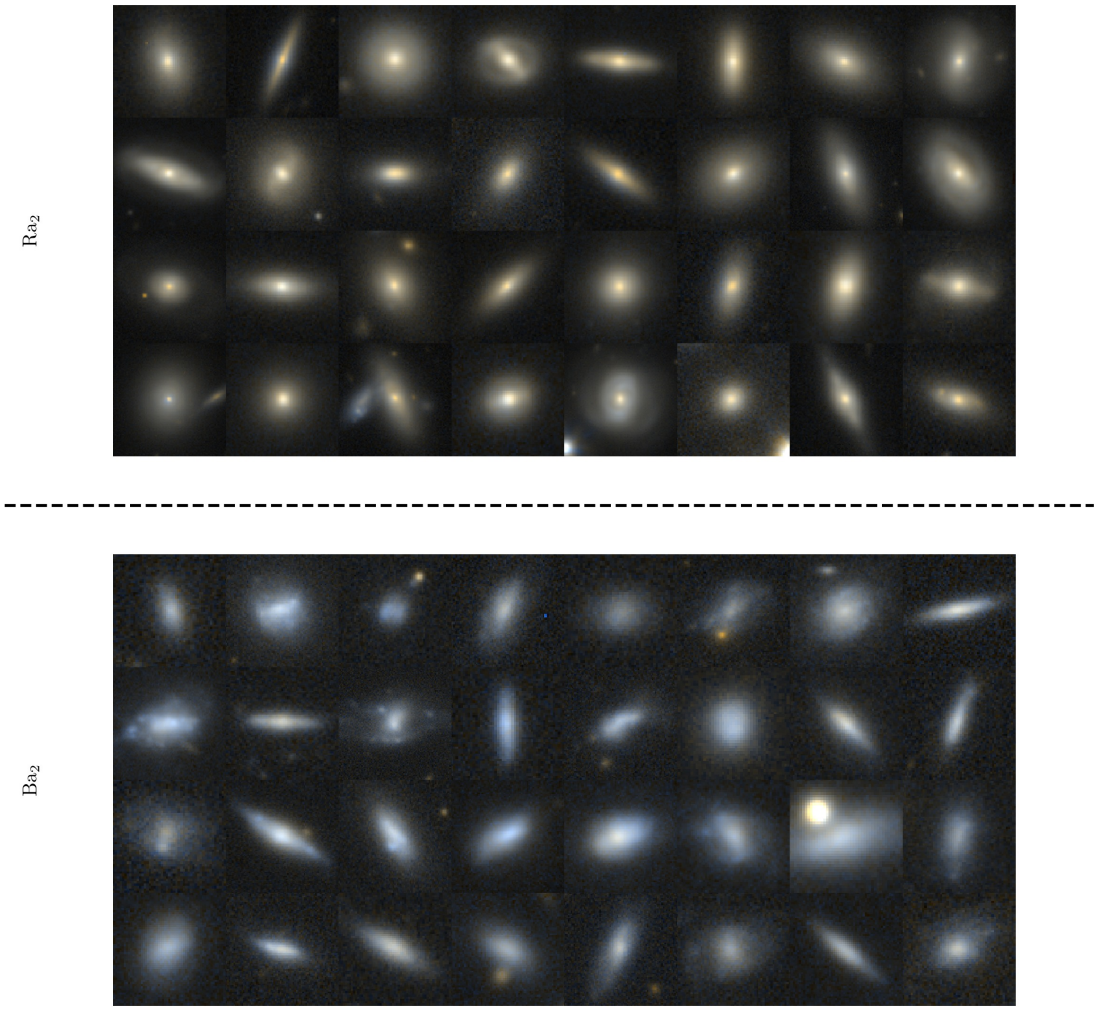


**Figure A.7:** Example postage stamps of galaxies in each of the clusters in $k = 2$. The dashed black line separates the two superclusters that `k-means` finds. See Section 3.3.1 for discussion.

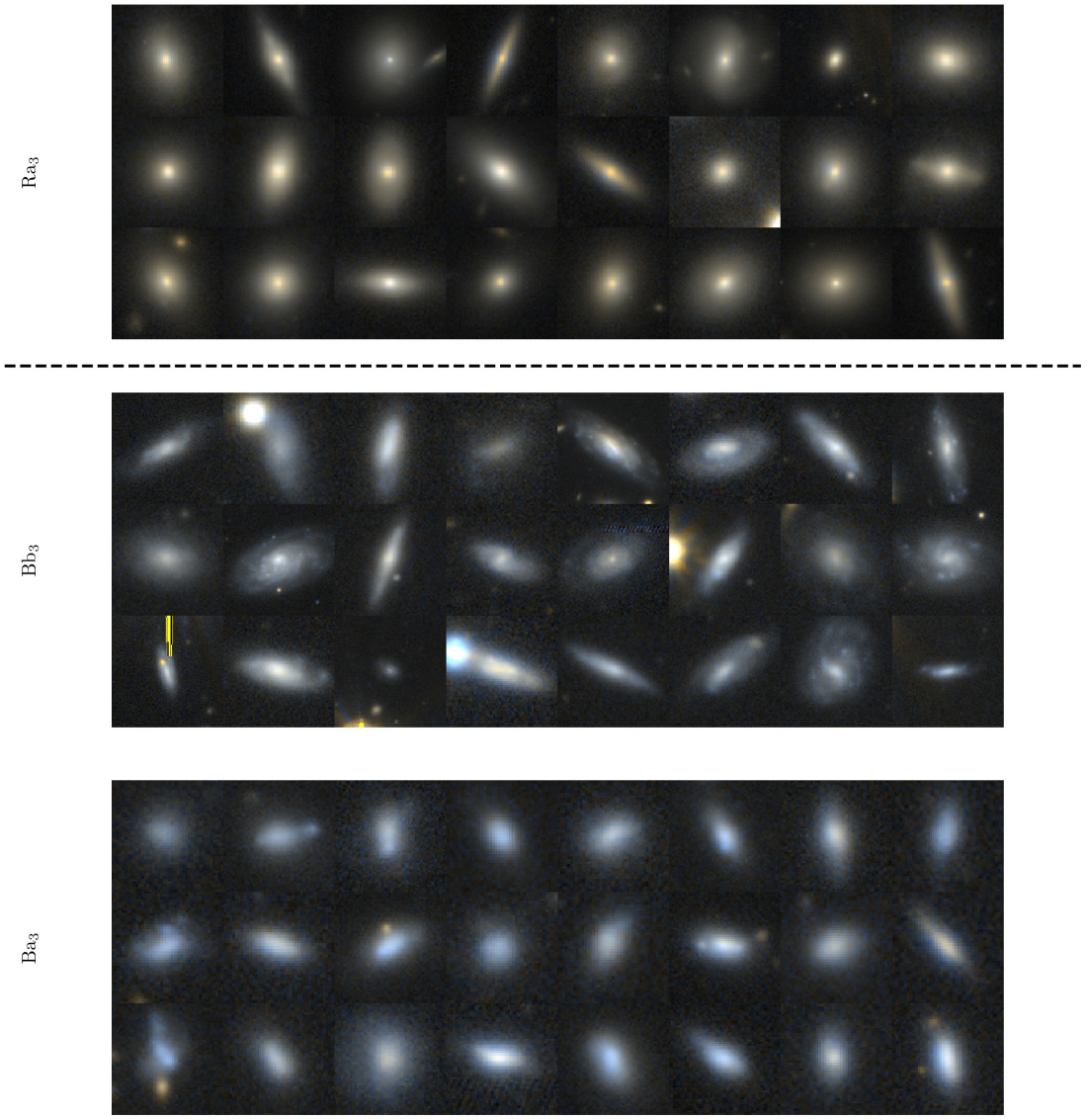


**Figure A.8:** Example postage stamps of galaxies in each of the clusters in $k = 3$. The dashed black line separates the two superclusters that `k-means` finds. See Section 3.3.2 for discussion.

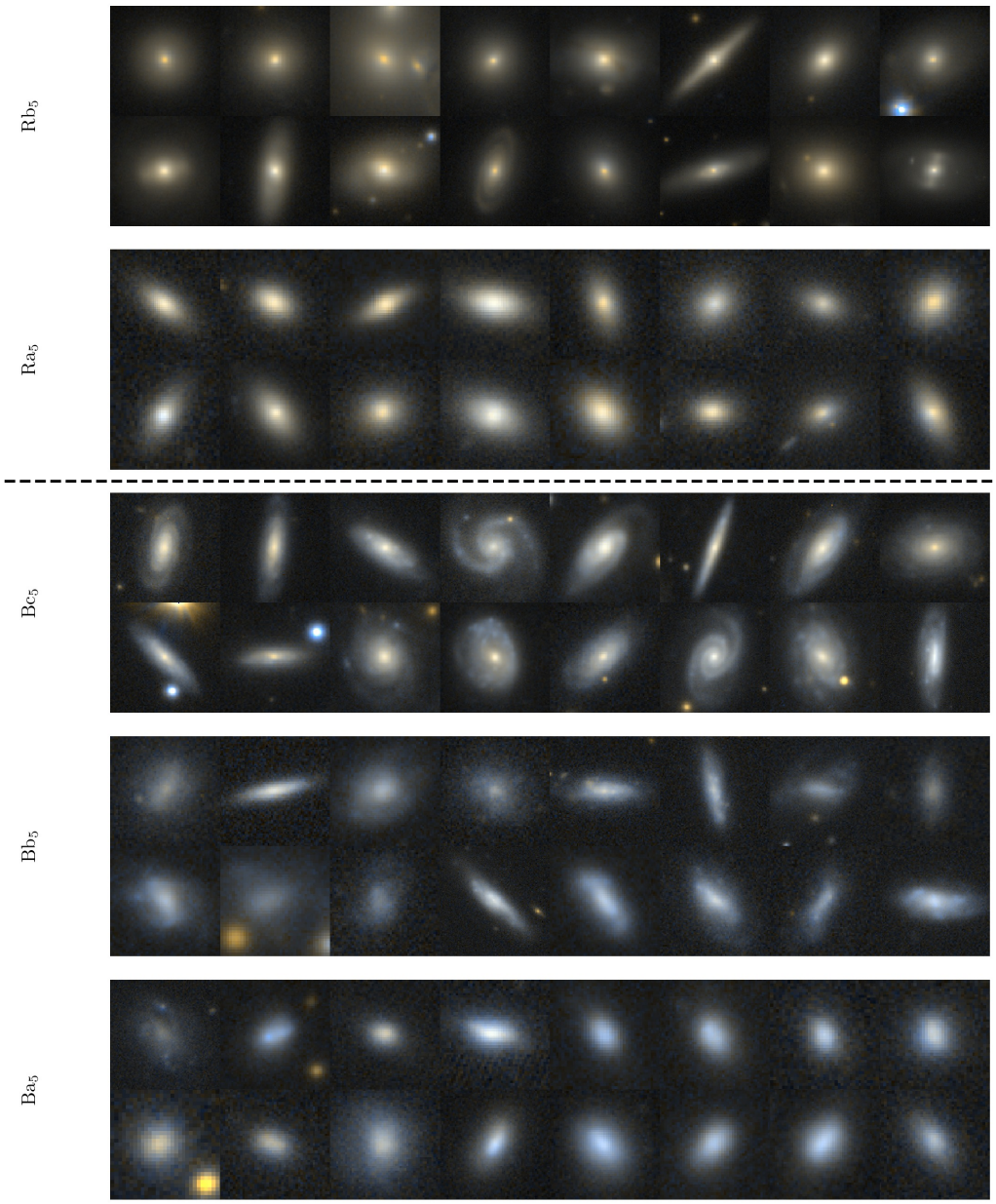


**Figure A.9:** Example postage stamps of galaxies in each of the clusters in $k = 5$. The dashed black line separates the two superclusters that `k-means` finds. See Section 3.3.3 for discussion.

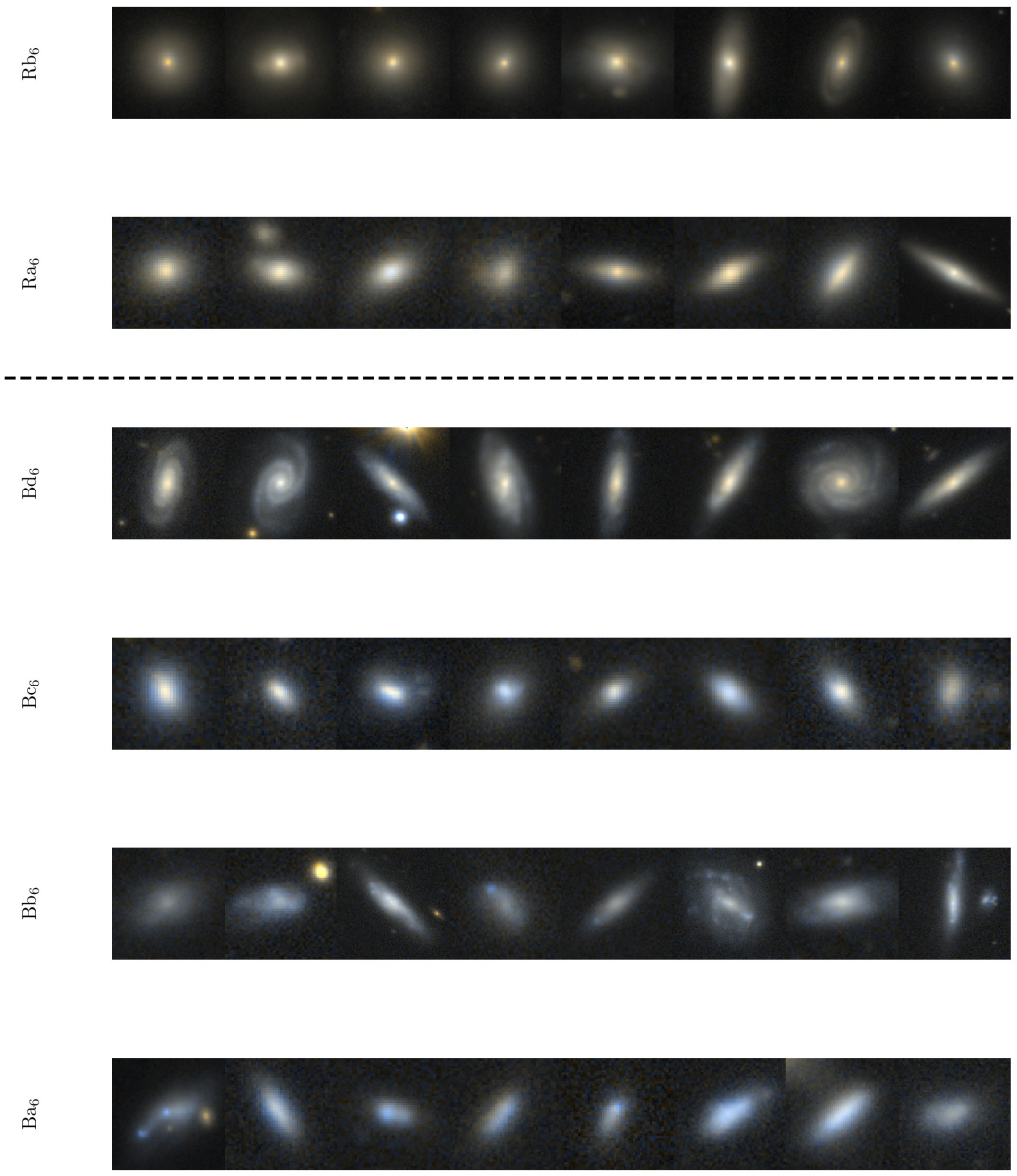


**Figure A.10:** Example postage stamps of galaxies in each of the clusters in $k = 6$. The dashed black line separates the two superclusters that `k-means` finds. See Section 3.3.4 for discussion.

# Appendix B

# Appendix to Chapter 4

## B.1 Measuring agreement between clustering outcomes using $V_a$

I use Cramer's $V$ index in two separate contexts in Chapter 4. The first context, denoted by the symbol $V_s$, is the identification of values of $k$ at which `k-means` clustering outcomes are stable for both of our samples (see Section 3.1). Values of $k$ are judged to be stable if their distributions in median $V_s$ peak strongly at a value close to 1 (see Figs. 3.4 and 4.2). A guide to the interpretation of $V_s$ is provided in Section A.1. The second context, denoted by the symbol $V_a$, is the measurement of the agreement between different clustering outcomes, potentially consisting of different numbers of clusters. Specifically, I measure the agreement between outcomes determined within the GAMA and EAGLE samples in terms of how their centroids partition the GAMA sample. For this comparison, I measure $V_a = 0.75$. In this appendix, I provide a guide to the interpretation of $V_a$.

For a sample of $N$ observations, partitioned into $k_X$ clusters by outcome X and into $k_Y$ clusters by outcome Y, $V_a$ is calculated thusly:

$$V_a = \sqrt{\frac{\chi^2}{N \cdot \min(\mathrm{k_X} - 1, \mathrm{k_Y} - 1)}}. \tag{B.1}$$

Here, $\chi^2$ is the chi-squared value for outcomes X and Y, calculated from a $k_X \times k_Y$ contingency table and given by:

$$\chi^2_{X,Y} = \sum_{x,y} \frac{(o_{x,y} - e_{x,y})^2}{e_{x,y}}. \tag{B.2}$$

Here, $e_{x,y}$ represents the expected number or fraction of observations shared by clusters $x$ and $y$ (equal for all combinations of $x$ and $y$, given a null hypothesis of independence of X and Y) and $o_{x,y}$ represents the actual observed number or fraction. In general, the value of $V_a$ rises with the extent to which clusters from one outcome correspond to clusters from another, such as if two clusters from outcome Y map exactly onto one cluster from outcome X.

In the remainder of this appendix I analyse pairs of partitions, determined within a simple two-dimensional data set, that yield a range of values of $V_a$. The data set is a projection of Chapter 4's five-dimensional GAMA sample of $3,724$ galaxies onto its $M_*$ and $sSFR$ axes, and it is shown in panel (a) of Figure B.1. The data set is rescaled ahead of partitioning such that each of its two axes spans the range 0 to 1. The basis of the analysis in this section is a $k = 5$ reference partition of this data set. Its borders are shown in blue in panel (b) of Figure B.1. This reference partition is generated via a run of `k-means` with $k = 5$ that is initialised using the Arthur & Vassilvitskii (2007) technique.

In panels (c)-(h), I plot a series of $k = 7$ partitions (red) for visual comparison with the $k = 5$ reference partition (blue). The $V_a$ value that each $k = 7$ partition yields with respect to the $k = 5$ partition is shown above its respective panel. The methods used to generate these $k = 7$ partitions are listed in Table B.1. These methods are contrived purely to generate a range of $V_a$ values, with a view to building an intuition for the level of agreement that each value captures. In addition, clustering cannot naturally produce outcomes that would yield the highest and lowest values of $V_a$ shown in this appendix (see below). Hence, some of the $k = 7$ partitions are artificially generated. This appendix is intended mostly to facilitate the interpretation of results that we present in Section 4.2, but its conclusions are also readily generalisable to pairs of clustering outcomes with any two $k$ values. I do not attempt to astrophysically interpret the partitions in this appendix in detail. Instead, I consider only their visual appearances in the $sSFR$ versus $M_*$ plane, and their $V_a$ values with respect to the $k =$ reference partition (shown above each panel). Hence, the panels in Figure B.1 do not includes axis labels and ticks.

Lower values of $V_a$ are given by the divergence of the structures of the $k = 7$ partitions from that of the $k = 5$ reference partition. The $k = 7$ partitions in panels (c)-(f) all generally capture the bimodal structure of the underlying data like the $k = 5$ partition does, but those in panels (g) and (h) do not. Hence, values of $V_a$ above $\sim 0.75$ indicate agreement between partitions in terms of broad, global structures. However, the $k = 7$ partitions in panels (d)-(f) exhibit some differences from the $k = 5$ partition at the substructure level, with the extent of these differences increases with decreasing $V_a$. For panels (d) and (e), this corresponds with a change in initialisation technique; while the $k = 7$ partition in panel (d) is generated following the initialisation of `k-means` with the Arthur & Vassilvitskii (2007) technique, the one in panel (e) is generated instead following initialisation using a uniform random selection of points from the data set. Uniformly random

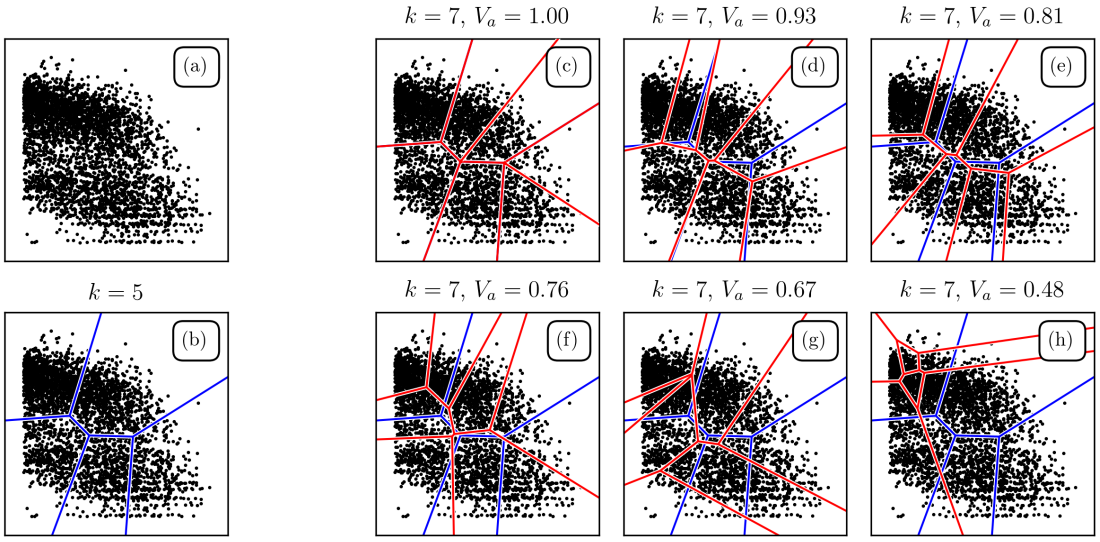


**Figure B.1:** Comparisons of partitions that yield a range of values of $V_a$. In panel (a), I show a projection of the GAMA sample (see Section 4.1.1) onto its $M_*$ and $sSFR$ axes. In panel (b), I show a reference $k = 5$ partition (blue) determined within this projection. Panels (c)-(h) show $k = 7$ partitions (red), also determined within this projection, that yield a range of $V_a$ values with respect to the $k = 5$ reference partition (blue). Their $V_a$ values are shown above their respective panels. I focus on the visual appearances of the partitions in this figure, rather on any astrophysical interpretation; hence, axis ticks and labels are omitted from the panels. Note that the $k = 5$ partition is not visible in panel (c) because its borders are covered perfectly by those of the $k = 7$ partition.

**Table B.1:** A summary of the methods used to generate each of the $k = 7$ partitions shown in panels (c)-(h) of Figure B.1. Included are the values of $V_a$ that they yield with respect to the $k = 5$ reference partition. The $k = 5$ reference partition is generated via a run of `k-means` with $k = 5$ and Arthur & Vassilvitskii (2007) initialisation.

| Figure B.1 Panel | $V_a$ | Method |
|---|---|---|
| (c) | 1.00 | Splitting of two $k = 5$ reference clusters |
| (d) | 0.93 | `k-means` with $k = 7$ and Arthur & Vassilvitskii (2007) initialisation |
| (e) | 0.81 | `k-means` with $k = 7$ and uniform random initialisation |
| (f) | 0.76 | `k-means` with $k = 7$ and Arthur & Vassilvitskii (2007) initialisation within the EAGLE $M_*$ versus $sSFR$ plane, mapped onto the GAMA $M_*$ versus $sSFR$ plane |
| (g) | 0.67 | Voronoi tessellation about a uniform random selection of points in the GAMA $M_*$ versus $sSFR$ plane |
| (h) | 0.48 | Voronoi tessellation about a uniform random selection of points in the GAMA $M_*$ versus $sSFR$ plane |

initialisation is less likely to produce globally optimal outcomes. The $k = 7$ partition in panel (f) is generated via a run of `k-means` within the projection of our EAGLE sample onto its $M_*$ and $sSFR$ axes, with the centroids mapped onto the GAMA projection.

The $k = 7$ partition in panel (c) produces $V_a = 1$ with respect to the $k = 5$ partition, indicating perfect agreement (i.e. that each of its clusters corresponds to exactly one $k = 5$ cluster). `k-means` cannot naturally determine two outcomes with different $k$ that yield $V_a = 1.0$ in continuous data because its minimisation of within-cluster variances means that the cluster centres would tend to be spread evenly throughout the data set. As a result, the "extra" clusters determined as part of a $k = 7$ outcome tend to occupy more space than a single cluster from a $k = 5$ outcome. Hence, the $k = 7$ partition in panel (c) was generated artificially, by splitting two of the $k = 5$ clusters. The $k = 7$ partitions shown in panels (g) and (h) are also unrealistic in that their clusters have highly disparate shapes and sizes, which `k-means` is unlikely to produce. Hence, in order to produce these two partitions which particularly low $V_a$ values, "centroids" were selected from the input data set uniformly at random, and assigned points to their nearest centroid selected in this manner (i.e. Voronoi tessellation).

# Appendix C

# Appendix to Chapter 5

## C.1 Iterations of SEM

In Figure C.1, I show ICL scores reported at each of up to 25 iterations by various combinations of submodel and $k$ for the GSWLC-2 sample. These 'iteration profiles' are mostly quite flat; hence, 25 iterations are more than sufficient for allowing SEM to stabilise to an outcome. In addition, the bulk of the clustering structure appears to be determined during the k-means initialisation step, which spreads the cluster centres out ahead of the first iteration. The ICL criterion rewards separated clusters, so k-means initialisations are particularly well suited to yielding useful clustering outcomes. Trials of the use of uniform random initialisations resulted in more combinations of submodels and $k$ failing to converge.

Variations in the ICL values reported by individual combinations of submodel and $k$ over successive iterations arise due to the Subspace step of SEM, in which the subspace within which the clusters are to be modelled is found. Hence, the updating of the model parameters during the Maximisation step is *indirectly* related to the probabilities calculated in the Expectation step. For traditional EM algorithms, these steps are directly related and thereby guarantee convergence. The large changes between successive iterations exhibited by some combinations (e.g. $\Sigma$, $\delta_k$, $k = 9$) are most often due to the emptying of clusters; a reduction in the number of clusters used by SEM leads, in these cases, to a sudden increase in ICL.

## C.2 Smoothing of feature data for the GSWLC-2 sample

Preliminary tests revealed that a truncated, bimodal substructure among passive galaxies within the nine-dimensional colour space representing the GSWLC-2 sample (see the left-hand plot of

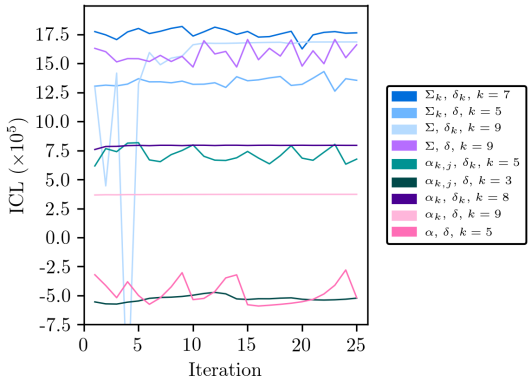


**Figure C.1:** ICL scores reported at iterations 1 through 25 by various combinations of submodel and $k$ for the GSWLC-2 sample. For each submodel, I show the value of $k$ which yields the highest ICL score. These iteration profiles are generally quite flat, indicating that `SEM` quickly converges to a stable outcome. The large changes exhibited by $\Sigma$, $\delta_k$, $k = 9$ are due to the emptying of clusters as it iterates.

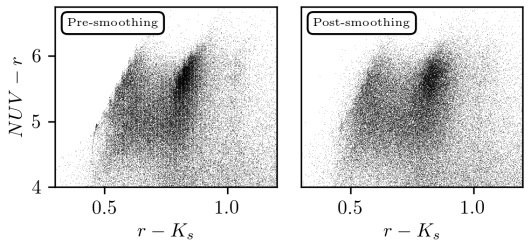


**Figure C.2:** The effect of my smoothing on the distribution of GSWLC-2 galaxies in the passive region of the $NUV - r - K_s$ colour-colour plane. Substructures in the distribution of galaxies within this region are preserved post-smoothing.

Figure C.2; also visible in Figure 5.3) led to an inability of `SEM` to converge for the majority of submodels and values of $k$. This truncated bimodal substructure is due to the lack of input NIR photometry to the `CIGALE` SED estimation of GSWLC-2 galaxies, such that their NIR SEDs must be inferred from UV and optical photometry. This, in turn, leads to poorly constrained, discretised metallicities: galaxies at $r - K_s \lesssim 0.67$ peak strongly at $\log_{10}(Z) \sim -2.4$, and those at $r - K_s \gtrsim 0.67$ at $\log_{10}(Z) \sim -2.1$. The NIR SEDs of VIPERS galaxies, on the other hand, are constrained by $K_s$-band photometry and hence have slightly more freedom to vary. This smooths their colour and metallicity distributions.

I hence opt to apply a small level of Gaussian smoothing to the GSWLC-2 distributions of the rest-frame absolute magnitudes reported by CIGALE. The smoothing scale for the rest-frame absolute magnitude of a given galaxy is given by its Bayesian error. These errors are winsorised at the mean value of the logarithmic distribution of errors (i.e. errors larger than the mean value are set to the mean value). This winsorisation ensures that the smoothing scale is kept small enough to avoid the potential loss of astrophysically meaningful substructures, while still enabling `SEM`

to converge more readily. The absolute rest-frame magnitude most affected by this smoothing is $FUV$, whose errors are winsorised at a maximum value of 0.25 (all other magnitudes have a maximum error < 0.1 after winsorisation). The right-hand plot of Figure C.2 demonstrates the effect of my smoothing, showing that the bimodality in the colours of passive galaxies is retained post-smoothing. While this bimodality is likely to be an artefact, trends in the astrophysical features of galaxies between its peaks are still likely to be genuine (see also Section 5.3.4).

## C.3 Behaviour of the various submodels of `SEM` for the samples

My model selection approach considers ICL scores for 72 different combinations of submodel and $k$ for *each* of the samples. The comparison of these 72 combinations is simplified greatly by the realisation that several submodels exhibit consistent patterns of behaviour across all values of $k$.

`SEM` is unable to converge to an outcome for several combinations of submodel and $k$. The most common diagnosis made by `SEM` in the case of non-convergence is that a cluster has become empty (i.e., that it no longer contains galaxies). Table 5.1 shows that several submodels are unable to converge beyond a maximum value of $k$, suggesting a limit to their ability to properly partition the samples. Alternatively, submodels that converge at $k$, but fail to converge at $k-1$ and $k+1$ appear to be striking a 'sweet spot' in terms of this ability. Different combinations are generally very consistent with respect to convergence, converging for either all or none of the 100 initialisations.

Given their flexibility and their high levels of parametrisation, the $\Sigma_k$, $\delta_k$ and $\Sigma_k$, $\delta$ submodels offer the greatest promise among all of the `SEM` submodels for yielding detailed and astrophysically meaningful partitions of the samples. The outcomes they produce are similar; they exhibit near-identical trends in their ICL scores for $k = 2$ through $k = 5$ for the GSWLC-2 sample in Table 5.1. They differ only in their treatment of the noise terms, which appears to be a minor detail in comparison with their shared use of full, unique covariance matrices. Outcomes at higher values of $k$ generally consist of splits of clusters present in outcomes at lower values of $k$.

Submodels featuring non-unique covariance matrices for the Gaussian density functions representing the clusters (i.e. submodels with $\Sigma$ and $\alpha$, such that they all have the same shape) consistently produce clusters with highly disparate sizes. Some clusters are large, containing 30 to 60 per cent of the galaxies in the samples each (and each often spanning both blue and red galaxies); others are empty or nearly empty, containing $\lesssim 1$ per cent of the galaxies in the samples each. Nearly-empty clusters appear to capture small, undesirable artefacts in the structure of the samples within their input feature spaces. While it is unclear why `SEM` registers a valid ICL score for these outcomes when they include empty clusters (often cited as a cause for the failure of `SEM`; see above), it is clear that these submodels are too crude to return more than a very broad partition of the samples,

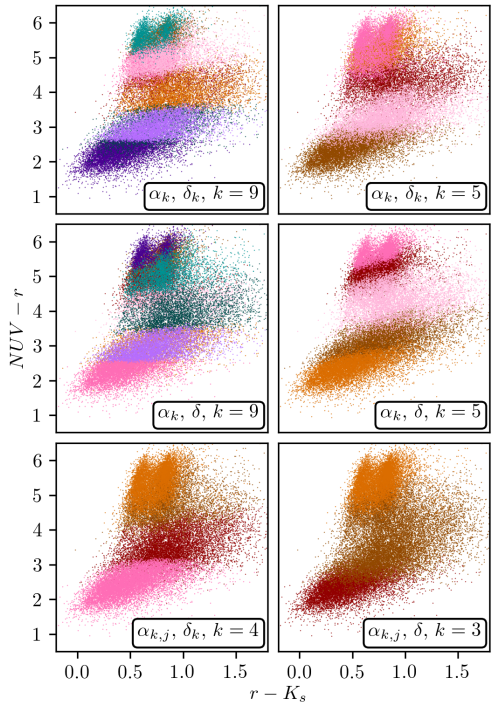


**Figure C.3:** Examples of the clustering structures determined by $\alpha_{k,j}$ and $\alpha_k$ submodels for the GSWLC-2 sample, shown in the $NUV - r - K_s$ colour-colour plane. The combination of submodel and $k$ for each outcome is shown to the lower-right of each plot. Individual galaxies are coloured in accordance with the cluster to which they belong. The choice of colours in this figure is not intended to imply any trends within or between plots. The horizontal striping pattern exhibited by these examples in these plots, which is a general property of $\alpha_{k,j}$- and $\alpha_k$-based outcomes, indicates segmentation mainly along a single axis.

and that their outcomes are limited in their capacity for astrophysical interpretation. All of this is also true for the $\Sigma$, $\delta$ clustering outcome at $k = 9$ for the GSWLC-2 sample, which achieved the highest ICL score in my model selection search despite including empty and nearly-empty clusters. For these reasons, I reject this outcome for analysis.

A general property of clustering outcomes reported by submodels which assume diagonal covariance matrices ($\alpha_{k,j}$, $\alpha_k$) for the Gaussian density functions within the discriminative latent subspace is that they segment the samples principally along a single dimension. Representative examples of their clustering structures are shown in Figure C.3, revealing that this single dimension is most strongly associated with the UV colours among the 9 input features, with little-to-no distinction made between galaxies based on their NIR colours. I note that these submodels scored highest when I tested clustering of the samples using *i*-band magnitudes of galaxies as a reference point for defining colours (as in Siudek et al. 2018b; see also Section 5.2.4), producing the same striping pattern within the $NUV - r - K_s$ plane. While this simple segmentation does correspond broadly with incremental changes in the star formation activity of galaxies within the samples, other submodels (with $\Sigma_k$) return more detailed partitions and achieve higher ICL scores anyway.

The large spread in the ICL scores reported in Table 5.1 arises directly from a large spread in the log-likelihood values of the fits. This large spread in the log-likelihood values arises, in turn, primarily from a $1/\delta_k$ coefficient in the log-likelihood function of DLM model (which may be seen in full in appendix 2 of Bouveyron & Brunet 2012). Submodels which yield very large but negative log-likelihood (and hence, ICL) values tend to have very small $\delta_k$ values for most (if not all) of their clusters; usually 0.001, which is the floor that `SEM` imposes upon the value of $\delta_k$. Very small values of $\delta_k$ produce very large, positive values of $1/\delta_k$, and (via a $-1/2$ coefficient of the log-likelihood function) very large, negative values of the log-likelihood and, thus, of the ICL criterion. The addition of this especially low-variance "noise" to subspace Gaussians leads to highly peaked full space Gaussians which are unlikely to reflect the more continuous distributions of both samples (see Figure 5.3).

## C.4 Active galactic nuclei

In this section, I examine the emission-line properties of star-forming galaxies in the GSWLC-2 sample, with a view to establishing the influence of active galactic nuclei upon their evolution. Figure C.4 shows the distributions of clusters G1-4 within the emission-line classification diagram of Lamareille (2010) – the region labels and boundaries are explained in the caption of Figure C.4. This diagram is chosen with a view to its applicability to galaxies at higher redshifts as well. The equivalent widths of the relevant emission lines were determined by Brinchmann et al. (2004), and are available for 94 per cent of the galaxies in G1-4. Spectroscopy of these emission lines exists for VIPERS galaxies as well (Garilli et al., 2014), but only for 34 per cent of them, such that I would not be confident in the significance of any trend of the VIPERS clusters within the Lamareille (2010) diagram. I note, however, that the few VIPERS galaxies for which this spectroscopy does exist tend to lie within the 'SF' region of the plot, above the 'Comp.' region (i.e. as in figure 10 of Siudek et al. 2018b). Hence, I tentatively suggest a minimal influence of active galactic nuclei upon their current evolution, but reiterate that more data is needed to confirm this.

Clusters G1-4 are all centred upon the 'Comp.' region of Figure C.4, indicating that galactic nuclei are prevalent throughout them. G4 in particular extends well into the 'LINERs' region of the diagram. Given the enhancement in the Sérsic indices of G4 galaxies over those belonging to G1-3 (Table 5.3 and Figure 5.7), this is consistent with previous studies which find that low-ionisation nuclear emission-line regions are more common in galaxies with earlier-type morphologies (Heckman, 1980). In addition, this increase in nuclear activity for G4 galaxies coincides with their decrease in $sSFR$ in comparison with G1-3 galaxies (Table 5.2), validating the suggestion that the supermassive black hole is involved in the quenching of these galaxies (Croton et al., 2006; Vergani et al., 2018; Moutard et al., 2020).

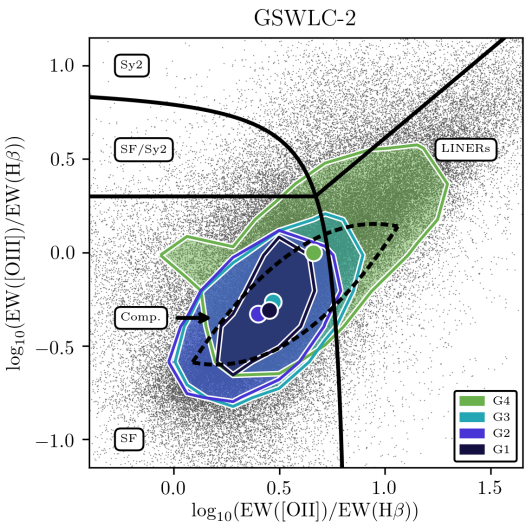


**Figure C.4:** A diagram for the classification of emission-line galaxies (Lamareille, 2010) in the GSWLC-2 sample. Different regions, labelled and demarcated by black lines, correspond to different types of galaxy: 'Sy2' to type II Seyfert galaxies, 'SF' to purely star-forming galaxies, 'SF/Sy2' to a mixture of type II Seyfert and star-forming galaxies, 'LINERs' to galaxies containing low-ionisation nuclear emission-line regions, and 'Comp.' to a mixture of LINERs and star-forming galaxies. The distributions of clusters are shown using coloured, filled contours (drawn at a relative density of 0.4), and the coloured, circular markers show their medians.

# Bibliography


Abell G. O., 1958, The Astrophysical Journal Supplement Series, 3, 211

Abell G. O., Corwin Jr. H. G., Olowin R. P., 1989, The Astrophysical Journal Supplement Series, 70, 1

Abraham R. G., van den Bergh S., Nair P., 2003, The Astrophysical Journal, 588, 218

Aceves H., Velázquez H., Cruz F., 2006, Monthly Notices of the Royal Astronomical Society, 373, 632

Aggarwal C. C., 2014, Data Classification: Algorithms and Applications. Chapman & Hall/CRC

Aguirre A., Hernquist L., Schaye J., Weinberg D. H., Katz N., Gardner J., 2001, The Astrophysical Journal, 560, 599

Ahn C. P., et al., 2014, The Astrophysical Journal Supplement Series, 211, 17

Aihara H., et al., 2018a, Publications of the Astronomical Society of Japan, 70, S4

Aihara H., et al., 2018b, Publications of the Astronomical Society of Japan, 70, S8

Almaini O., et al., 2017, Monthly Notices of the Royal Astronomical Society, 472, 1401

Alpher R. A., Bethe H., Gamow G., 1948, Physical Review, 73, 803

Andrews S. K., Driver S. P., Davies L. J. M., Kafle P. R., Robotham A. S. G., Wright A. H., 2017, Monthly Notices of the Royal Astronomical Society, 464, 1569

Antonucci R., 1993, Annual Review of Astronomy and Astrophysics, 31, 473

Aristotle, 350BC, Meteorology

Arnaud M., Pratt G. W., Piffaretti R., Böhringer H., Croston J. H., Pointecouteau E., 2010, Astronomy & Astrophysics, 517, A92

Arnouts S., et al., 2013, Astronomy & Astrophysics, 558, A67

Arp H., 1966, The Astrophysical Journal Supplement Series, 14, 1

Arthur D., Vassilvitskii S., 2007, in Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms. p. 1027

Baes M., Verstappen J., De Looze I., Fritz J., Saftly W., Vidal Pérez E., Stalevski M., Valcke S., 2011, The Astrophysical Journal Supplement Series, 196, 22

Bahé Y. M., McCarthy I. G., 2015, Monthly Notices of the Royal Astronomical Society, 447, 969

Bahé Y. M., et al., 2017, Monthly Notices of the Royal Astronomical Society, 470, 4186

Baldry I. K., 2008, Astronomy and Geophysics, 49, 5.25

Baldry I. K., Glazebrook K., Brinkmann J., Ivezić Ž., Lupton R. H., Nichol R. C., Szalay A. S., 2004, The Astrophysical Journal, 600, 681

Baldry I. K., Balogh M. L., Bower R. G., Glazebrook K., Nichol R. C., Bamford S. P., Budavari T., 2006, Monthly Notices of the Royal Astronomical Society, 373, 469

Baldry I. K., Glazebrook K., Driver S. P., 2008, Monthly Notices of the Royal Astronomical Society, 388, 945

Baldry I. K., et al., 2010, Monthly Notices of the Royal Astronomical Society, 404, 86

Baldry I. K., et al., 2012, Monthly Notices of the Royal Astronomical Society, 421, 621

Baldry I. K., et al., 2018, Monthly Notices of the Royal Astronomical Society, 474, 3875

Baldwin J. A., Phillips M. M., Terlevich R., 1981, Publications of the Astronomical Society of the Pacific, 93, 5

Ball N. M., Brunner R. J., 2010, International Journal of Modern Physics D, 19, 1049

Ball N. M., Loveday J., Fukugita M., Nakamura O., Okamura S., Brinkmann J., Brunner R. J., 2004, Monthly Notices of the Royal Astronomical Society, 348, 1038

Balogh M. L., McGee S. L., 2010, Monthly Notices of the Royal Astronomical Society, 402, L59

Balogh M. L., Morris S. L., Yee H. K. C., Carlberg R. G., Ellingson E., 1999, The Astrophysical Journal, 527, 54

Balogh M. L., Navarro J. F., Morris S. L., 2000, The Astrophysical Journal, 540, 113

Balogh M. L., Pearce F. R., Bower R. G., Kay S. T., 2001, Monthly Notices of the Royal Astronomical Society, 326, 1228

Balogh M. L., Baldry I. K., Nichol R., Miller C., Bower R., Glazebrook K., 2004, The Astrophysical Journal Letters, 615, L101

Bamford S. P., et al., 2009, Monthly Notices of the Royal Astronomical Society, 393, 1324

Banerji M., et al., 2010, Monthly Notices of the Royal Astronomical Society, 406, 342

Barchi P. H., da Costa F. G., Sautter R., Moura T. C., Stalder D. H., Rosa R. R., de Carvalho R. R., 2016, Journal of Computational Interdisciplinary Sciences, 7, 114

Barnes J. E., 1988, The Astrophysical Journal, 331, 699

Barnes J. E., 1992, The Astrophysical Journal, 393, 484

Barnes J. E., 2002, Monthly Notices of the Royal Astronomical Society, 333, 481

Barnes J. E., Hernquist L. E., 1991, The Astrophysical Journal Letters, 370, L65

Barnes J. E., Hernquist L., 1996, The Astrophysical Journal, 471, 115

Barnes D. J., et al., 2017, Monthly Notices of the Royal Astronomical Society, 471, 1088

Baron D., 2019, Machine Learning in Astronomy: a practical overview

Barro G., et al., 2013, The Astrophysical Journal, 765, 104

Bartelmann M., 2010, Reviews of Modern Physics, 82, 331

Baugh C. M., 2006, Reports on Progress in Physics, 69, 3101

Bayes T., 1763, Philosophical Transactions of the Royal Society, 53, 370

Bekki K., Couch W. J., Shioya Y., 2002, The Astrophysical Journal, 577, 651

Bell E. F., 2008, The Astrophysical Journal, 682, 355

Bell E. F., de Jong R. S., 2001, The Astrophysical Journal, 550, 212

Bellman R., Corporation R., Collection K. M. R., 1957, Dynamic Programming. Princeton University Press

Bertin E., Arnouts S., 1996, Astronomy & Astrophysics Supplement Series, 117, 393

Bertone G., Hooper D., 2018, Reviews of Modern Physics, 90, 045002

Biernacki C., Celeux G., Govaert G., 2000, IEEE Transactions on Pattern Analysis and Machine Intelligence, 22, 719

Bignone L. A., Pedrosa S. E., Trayford J. W., Tissera P. B., Pellizza L. J., 2020, Monthly Notices of the Royal Astronomical Society, 491, 3624

Binggeli B., Sandage A., Tammann G. A., 1988, Annual Review of Astronomy and Astrophysics, 26, 509

Binney J., 1977, The Astrophysical Journal, 215, 483

Binney J., 1982, Annual Review of Astronomy and Astrophysics, 20, 399

Birnboim Y., Dekel A., 2003, Monthly Notices of the Royal Astronomical Society, 345, 349

Birnboim Y., Dekel A., 2011, Monthly Notices of the Royal Astronomical Society, 415, 2566

Blanton M. R., Berlind A. A., 2007, The Astrophysical Journal, 664, 791

Bluck A. F. L., Mendel J. T., Ellison S. L., Moreno J., Simard L., Patton D. R., Starkenburg E., 2014, Monthly Notices of the Royal Astronomical Society, 441, 599

Blumenthal G. R., Faber S. M., Primack J. R., Rees M. J., 1984, Nature, 311, 517

Boquien M., Burgarella D., Roehlly Y., Buat V., Ciesla L., Corre D., Inoue A. K., Salas H., 2019, Astronomy & Astrophysics, 622, A103

Boselli A., Gavazzi G., 2006, Publications of the Astronomical Society of the Pacific, 118, 517

Boselli A., Boissier S., Cortese L., Gavazzi G., 2009, Astronomische Nachrichten, 330, 904

Bourlard H., Kamp Y., 1988, Biological Cybernetics, 59, 291

Bournaud F., Combes F., 2002, Astronomy & Astrophysics, 392, 83

Bournaud F., Elmegreen B. G., 2009, The Astrophysical Journal Letters, 694, L158

Bournaud F., Combes F., Jog C. J., 2004, Astronomy & Astrophysics, 418, L27

Bournaud F., Jog C. J., Combes F., 2005, Astronomy & Astrophysics, 437, 69

Bournaud F., Jog C. J., Combes F., 2007a, Astronomy & Astrophysics, 476, 1179

Bournaud F., Elmegreen B. G., Elmegreen D. M., 2007b, The Astrophysical Journal, 670, 237

Bournaud F., Dekel A., Teyssier R., Cacciato M., Daddi E., Juneau S., Shankar F., 2011, The Astrophysical Journal Letters, 741, L33

Bournaud F., et al., 2012, The Astrophysical Journal, 757, 81

Bouveyron C., Brunet C., 2012, Statistics and Computing, 22, 301

Bower R. G., Lucey J. R., Ellis R. S., 1992a, Monthly Notices of the Royal Astronomical Society, 254, 589

Bower R. G., Lucey J. R., Ellis R. S., 1992b, Monthly Notices of the Royal Astronomical Society, 254, 601

Bower R. G., Benson A. J., Malbon R., Helly J. C., Frenk C. S., Baugh C. M., Cole S., Lacey C. G., 2006, Monthly Notices of the Royal Astronomical Society, 370, 645

Boylan-Kolchin M., Springel V., White S. D. M., Jenkins A., Lemson G., 2009, Monthly Notices of the Royal Astronomical Society, 398, 1150

Brammer G. B., et al., 2009, The Astrophysical Journal Letters, 706, L173

Bremer M. N., et al., 2018, Monthly Notices of the Royal Astronomical Society, 476, 12

Brinchmann J., Charlot S., White S. D. M., Tremonti C., Kauffmann G., Heckman T., Brinkmann J., 2004, Monthly Notices of the Royal Astronomical Society, 351, 1151

Brough S., et al., 2013, Monthly Notices of the Royal Astronomical Society, 435, 2903

Brown T., et al., 2017, Monthly Notices of the Royal Astronomical Society, 466, 1275

Bruce V. A., et al., 2012, Monthly Notices of the Royal Astronomical Society, 427, 1666

Bruzual G., Charlot S., 2003, Monthly Notices of the Royal Astronomical Society, 344, 1000

Bubeck S., Meila M., von Luxburg U., 2012, ESAIM: Probability and Statistics, 16, 436

Bundy K., et al., 2006, The Astrophysical Journal, 651, 120

Bundy K., et al., 2010, The Astrophysical Journal, 719, 1969

Byrd G., Valtonen M., 1990, The Astrophysical Journal, 350, 89

Cacciato M., Dekel A., Genel S., 2012, Monthly Notices of the Royal Astronomical Society, 421, 818

Calzetti D., Armus L., Bohlin R. C., Kinney A. L., Koornneef J., Storchi-Bergmann T., 2000, The Astrophysical Journal, 533, 682

Campello R., Moulavi D., Sander J., 2013. p. 160

Camps P., Baes M., 2015, Astronomy and Computing, 9, 20

Camps P., Trayford J. W., Baes M., Theuns T., Schaller M., Schaye J., 2016, Monthly Notices of the Royal Astronomical Society, 462, 1057

Camps P., et al., 2018, The Astrophysical Journal Supplement Series, 234, 20

Cano-Díaz M., Ávila-Reese V., Sánchez S. F., Hernández-Toledo H. M., Rodríguez-Puebla A., Boquien M., Ibarra-Medel H., 2019, Monthly Notices of the Royal Astronomical Society, 488, 3929

Cappellari M., 2016, Annual Review of Astronomy and Astrophysics, 54, 597

Cappellari M., et al., 2011, Monthly Notices of the Royal Astronomical Society, 416, 1680

Carollo C. M., et al., 2013, The Astrophysical Journal, 773, 112

Cattaneo A., Dekel A., Devriendt J., Guiderdoni B., Blaizot J., 2006, Monthly Notices of the Royal Astronomical Society, 370, 1651

Cattaneo A., Dekel A., Faber S. M., Guiderdoni B., 2008, Monthly Notices of the Royal Astronomical Society, 389, 567

Cayatte V., van Gorkom J. H., Balkowski C., Kotanyi C., 1990, The Astronomical Journal, 100, 604

Chabrier G., 2003, Publications of the Astronomical Society of the Pacific, 115, 763

Chang J., Macciò A. V., Kang X., 2013, Monthly Notices of the Royal Astronomical Society, 431, 3533

Charlot S., Fall S. M., 2000, The Astrophysical Journal, 539, 718

Charlot S., Longhetti M., 2001, Monthly Notices of the Royal Astronomical Society, 323, 887

Chary R., Elbaz D., 2001, The Astrophysical Journal, 556, 562

Cheng T.-Y., Huertas-Company M., Conselice C. J., Aragón-Salamanca A., Robertson B. E., Ramachandra N., 2020, arXiv e-prints, p. arXiv:2009.11932

Chester C., Roberts M. S., 1964, The Astronomical Journal, 69, 635

Cheung E., et al., 2012, The Astrophysical Journal, 760, 131

Cheung E., et al., 2013, The Astrophysical Journal, 779, 162

Cheung E., et al., 2015, Monthly Notices of the Royal Astronomical Society, 447, 506

Churazov E., Sazonov S., Sunyaev R., Forman W., Jones C., Böhringer H., 2005, Monthly Notices of the Royal Astronomical Society, 363, L91

Cid Fernandes R., Stasińska G., Schlickmann M. S., Mateus A., Vale Asari N., Schoenell W., Sodré L., 2010, Monthly Notices of the Royal Astronomical Society, 403, 1036

Cid Fernandes R., Stasińska G., Mateus A., Vale Asari N., 2011, Monthly Notices of the Royal Astronomical Society, 413, 1687

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

Clauwens B., Schaye J., Franx M., Bower R. G., 2018, Monthly Notices of the Royal Astronomical Society, 478, 3994

Cochrane R. K., Best P. N., 2018, Monthly Notices of the Royal Astronomical Society, 480, 864

Cole S., Lacey C. G., Baugh C. M., Frenk C. S., 2000, Monthly Notices of the Royal Astronomical Society, 319, 168

Colless M., et al., 2001, Monthly Notices of the Royal Astronomical Society, 328, 1039

Combes F., Dupraz C., Casoli F., Pagani L., 1988, Astronomy & Astrophysics, 203, L9

Connolly A. J., Szalay A. S., Bershady M. A., Kinney A. L., Calzetti D., 1995, The Astronomical Journal, 110, 1071

Conroy C., 2013, Annual Review of Astronomy and Astrophysics, 51, 393

Conseil S., Vibert D., Amouts S., Milliard B., Zamojski M., Liebaria A., Guillaume M., 2011, EMphot — Photometric Software with Bayesian Priors: Application to GALEX. p. 107

Conselice C. J., 2003, The Astrophysical Journal Supplement Series, 147, 1

Conselice C. J., et al., 2004, The Astrophysical Journal Letters, 600, L139

Conselice C. J., Wilkinson A., Duncan K., Mortlock A., 2016, The Astrophysical Journal, 830, 83

Correa C. A., Schaye J., Clauwens B., Bower R. G., Crain R. A., Schaller M., Theuns T., Thob A. C. R., 2017, Monthly Notices of the Royal Astronomical Society, 472, L45

Cowie L. L., Songaila A., 1977, Nature, 266, 501

Cowie L. L., Lilly S. J., Gardner J., McLean I. S., 1988, The Astrophysical Journal Letters, 332, L29

Cowie L. L., Songaila A., Hu E. M., Cohen J. G., 1996, The Astronomical Journal, 112, 839

Cox T. J., Jonsson P., Somerville R. S., Primack J. R., Dekel A., 2008, Monthly Notices of the Royal Astronomical Society, 384, 386

Crain R. A., et al., 2015, Monthly Notices of the Royal Astronomical Society, 450, 1937

Cramér H., 1946, Mathematical Methods of Statistics. Princeton Mathematical Series, Princeton University Press

Croton D. J., et al., 2006, Monthly Notices of the Royal Astronomical Society, 365, 11

Cucciati O., et al., 2017, Astronomy & Astrophysics, 602, A15

Da Cunha E., Charlot S., Elbaz D., 2008, Monthly Notices of the Royal Astronomical Society, 388, 1595

Daddi E., et al., 2010, The Astrophysical Journal, 713, 686

Dalla Vecchia C., Schaye J., 2012, Monthly Notices of the Royal Astronomical Society, 426, 140

Darg D. W., et al., 2010, Monthly Notices of the Royal Astronomical Society, 401, 1552

Davé R., Thompson R., Hopkins P. F., 2016, Monthly Notices of the Royal Astronomical Society, 462, 3265

Davé R., Anglés-Alcázar D., Narayanan D., Li Q., Rafieferantsoa M. H., Appleby S., 2019, Monthly Notices of the Royal Astronomical Society, 486, 2827

Davidzon I., et al., 2013, Astronomy & Astrophysics, 558, A23

Davidzon I., et al., 2016, Astronomy & Astrophysics, 586, A23

Davidzon I., et al., 2019, Monthly Notices of the Royal Astronomical Society, 489, 4817

Davies L. J. M., et al., 2018, Monthly Notices of the Royal Astronomical Society, 480, 768

Davies J. J., Crain R. A., McCarthy I. G., Oppenheimer B. D., Schaye J., Schaller M., McAlpine S., 2019, Monthly Notices of the Royal Astronomical Society, 485, 3783

Davies J. J., Crain R. A., Pontzen A., 2020a, arXiv e-prints, p. arXiv:2006.13221

Davies J. J., Crain R. A., Oppenheimer B. D., Schaye J., 2020b, Monthly Notices of the Royal Astronomical Society, 491, 4462

Davis M., Geller M. J., 1976, The Astrophysical Journal, 208, 13

Davis M., Efstathiou G., Frenk C. S., White S. D. M., 1985, The Astrophysical Journal, 292, 371

Debattista V. P., Carollo C. M., Mayer L., Moore B., 2004, The Astrophysical Journal Letters, 604, L93

Debattista V. P., Mayer L., Carollo C. M., Moore B., Wadsley J., Quinn T., 2006, The Astrophysical Journal, 645, 209

Dekel A., Birnboim Y., 2006, Monthly Notices of the Royal Astronomical Society, 368, 2

Dekel A., Birnboim Y., 2008, Monthly Notices of the Royal Astronomical Society, 383, 119

Dekel A., Sari R., Ceverino D., 2009, The Astrophysical Journal, 703, 785

Dempster A. P., Laird N. M., Rubin D. B., 1977, Journal of the Royal Statistical Society. Series B (Methodological), 39, 1

Deng J., Dong W., Socher R., Li L.-J., Li K., Fei-Fei L., 2009, in CVPR09.

Di Matteo T., Springel V., Hernquist L., 2005, Nature, 433, 604

Dicke R. H., Peebles P. J. E., Roll P. G., Wilkinson D. T., 1965, The Astrophysical Journal, 142, 414

Dickinson H., et al., 2018, The Astrophysical Journal, 853, 194

Dieleman S., Willett K. W., Dambre J., 2015, Monthly Notices of the Royal Astronomical Society, 450, 1441

Digges T., 1571, A Geometrical Practise Named Pantometria

Digges L., Digges T., 1576, A Prognostication Everlasting of Right Good Effect

Donnari M., et al., 2019, Monthly Notices of the Royal Astronomical Society, 485, 4817

Dressler A., 1980a, The Astrophysical Journal Supplement Series, 42, 565

Dressler A., 1980b, The Astrophysical Journal, 236, 351

Dreyer J. L. E., 1888, Memoirs of the Royal Astronomical Society, 49, 1

Driver S. P., et al., 2006, Monthly Notices of the Royal Astronomical Society, 368, 414

Driver S. P., et al., 2009, Astronomy and Geophysics, 50, 5.12

Driver S. P., et al., 2011, Monthly Notices of the Royal Astronomical Society, 413, 971

Driver S. P., et al., 2016, Monthly Notices of the Royal Astronomical Society, 455, 3911

Driver S. P., et al., 2018, Monthly Notices of the Royal Astronomical Society, 475, 2891

Dubois Y., et al., 2014, Monthly Notices of the Royal Astronomical Society, 444, 1453

Duda R. O., Hart P. E., Stork D. G., 2000, Pattern Classification. Wiley-Interscience

Eales S., de Vis P., Smith M. W. L., Appah K., Ciesla L., Duffield C., Schofield S., 2017, Monthly Notices of the Royal Astronomical Society, 465, 3125

Eales S. A., et al., 2018, Monthly Notices of the Royal Astronomical Society, 481, 1183

Ebeling H., Stephenson L. N., Edge A. C., 2014, The Astrophysical Journal Letters, 781, L40

Einstein A., 1917, Sitzungsberichte der Königlich Preußischen Akademie der Wissenschaften, p. 142

Ellis S. C., Driver S. P., Allen P. D., Liske J., Bland-Hawthorn J., De Propris R., 2005, Monthly Notices of the Royal Astronomical Society, 363, 1257

Ellison S. L., Patton D. R., Simard L., McConnachie A. W., 2008, The Astronomical Journal, 135, 1877

Ellison S. L., Patton D. R., Simard L., McConnachie A. W., Baldry I. K., Mendel J. T., 2010, Monthly Notices of the Royal Astronomical Society, 407, 1514

Elmegreen D. M., Elmegreen B. G., Ravindranath S., Coe D. A., 2007, The Astrophysical Journal, 658, 763

Elmegreen B. G., Bournaud F., Elmegreen D. M., 2008a, The Astrophysical Journal, 684, 829

Elmegreen B. G., Bournaud F., Elmegreen D. M., 2008b, The Astrophysical Journal, 688, 67

Emsellem E., et al., 2004, Monthly Notices of the Royal Astronomical Society, 352, 721

Emsellem E., et al., 2011, Monthly Notices of the Royal Astronomical Society, 414, 888

Eskridge P. B., et al., 2000, The Astronomical Journal, 119, 536

Ester M., Kriegel H.-P., Sander J., Xu X., 1996. AAAI Press, p. 226

Faber S. M., 1972, Astronomy & Astrophysics, 20, 361

Faber S. M., Gallagher J. S., 1979, Annual Review of Astronomy and Astrophysics, 17, 135

Faber S. M., et al., 2007, The Astrophysical Journal, 665, 265

Fabian A. C., 2012, Annual Review of Astronomy and Astrophysics, 50, 455

Fall S. M., Efstathiou G., 1980, Monthly Notices of the Royal Astronomical Society, 193, 189

Fang J. J., Faber S. M., Koo D. C., Dekel A., 2013, The Astrophysical Journal, 776, 63

Fisher D. B., Drory N., 2008, The Astronomical Journal, 136, 773

Forman W., Kellogg E., Gursky H., Tananbaum H., Giacconi R., 1972, The Astrophysical Journal, 178, 309

Fossati M., et al., 2017, The Astrophysical Journal, 835, 153

Freeman K. C., 1970, The Astrophysical Journal, 160, 811

Freeman P. E., Izbicki R., Lee A. B., Newman J. A., Conselice C. J., Koekemoer A. M., Lotz J. M., Mozena M., 2013, Monthly Notices of the Royal Astronomical Society, 434, 282

Frey B. J., Dueck D., 2007, Science, 315, 972

Frieman J. A., Turner M. S., Huterer D., 2008, Annual Review of Astronomy and Astrophysics, 46, 385

Fritz A., et al., 2014, Astronomy & Astrophysics, 563, A92

Fujita Y., 1998, The Astrophysical Journal, 509, 587

Fujita Y., 2004, Publications of the Astronomical Society of Japan, 56, 29

Fukugita M., et al., 2007, The Astronomical Journal, 134, 579

Fukushima K., 1980, Biological Cybernetics, 36, 193

Furlong M., et al., 2015, Monthly Notices of the Royal Astronomical Society, 450, 4486

Furlong M., et al., 2017, Monthly Notices of the Royal Astronomical Society, 465, 722

Gabor J. M., Bournaud F., 2013, Monthly Notices of the Royal Astronomical Society, 434, 606

Gabor J. M., Davé R., 2015, Monthly Notices of the Royal Astronomical Society, 447, 374

Gabor J. M., Davé R., Oppenheimer B. D., Finlator K., 2011, Monthly Notices of the Royal Astronomical Society, 417, 2676

Galilei G., 1610, Siderius Nuncius

Gallagher J. S., Bushouse H., Hunter D. A., 1989, The Astronomical Journal, 97, 700

Gallazzi A., Charlot S., Brinchmann J., White S. D. M., 2006, Monthly Notices of the Royal Astronomical Society, 370, 1106

Galloway M. A., et al., 2015, Monthly Notices of the Royal Astronomical Society, 448, 3442

Garilli B., et al., 2014, Astronomy & Astrophysics, 562, A23

Gavazzi G., Contursi A., Carrasco L., Boselli A., Kennicutt R., Scodeggio M., Jaffe W., 1995, Astronomy & Astrophysics, 304, 325

Geach J. E., 2012, Monthly Notices of the Royal Astronomical Society, 419, 2633

Geach J. E., et al., 2014, Nature, 516, 68

Geller M. J., Huchra J. P., 1983, The Astrophysical Journal Supplement Series, 52, 61

Geller M. J., Huchra J. P., 1989, Science, 246, 897

Genel S., et al., 2018, Monthly Notices of the Royal Astronomical Society, 474, 3976

Genzel R., et al., 2011, The Astrophysical Journal, 733, 101

Gilbank D. G., Baldry I. K., Balogh M. L., Glazebrook K., Bower R. G., 2010, Monthly Notices of the Royal Astronomical Society, 405, 2594

Gilbank D. G., Baldry I. K., Balogh M. L., Glazebrook K., Bower R. G., 2011a, Monthly Notices of the Royal Astronomical Society, 412, 2111

Gilbank D. G., et al., 2011b, Monthly Notices of the Royal Astronomical Society, 414, 304

Gill S. P. D., Knebe A., Gibson B. K., 2005, Monthly Notices of the Royal Astronomical Society, 356, 1327

Giovanelli R., Haynes M. P., 1983, The Astronomical Journal, 88, 881

Gonzalez-Perez V., Lacey C. G., Baugh C. M., Lagos C. D. P., Helly J., Campbell D. J. R., Mitchell P. D., 2014, Monthly Notices of the Royal Astronomical Society, 439, 264

Goto T., Yamauchi C., Fujita Y., Okamura S., Sekiguchi M., Smail I., Bernardi M., Gomez P. L., 2003, Monthly Notices of the Royal Astronomical Society, 346, 601

Gott III J. R., Jurić M., Schlegel D., Hoyle F., Vogeley M., Tegmark M., Bahcall N., Brinkmann J., 2005, The Astrophysical Journal, 624, 463

Graham A. W., Driver S. P., 2005, Publications of the Astronomical Society of Australia, 22, 118

Graham A. W., Janz J., Penny S. J., Chilingarian I. V., Ciambur B. C., Forbes D. A., Davies R. L., 2017, The Astrophysical Journal, 840, 68

Gribbin J., 2003, Science: A History. Gardners' Books

Grogin N. A., et al., 2011, The Astrophysical Journal Supplement Series, 197, 35

Gu Y., Fang G., Yuan Q., Lu S., Li F., Cai Z.-Y., Kong X., Wang T., 2019, The Astrophysical Journal, 884, 172

Gunn J. E., Gott III J. R., 1972, The Astrophysical Journal, 176, 1

Guth A. H., 1981, Physical Review D (Particles and Fields), 23, 347

Guzzo L., et al., 2014, Astronomy & Astrophysics, 566, A108

Haines C. P., et al., 2017, Astronomy & Astrophysics, 605, A4

Häring N., Rix H.-W., 2004, The Astrophysical Journal Letters, 604, L89

Harris C. R., et al., 2020, Nature, 585, 357

Hart R. E., et al., 2017, Monthly Notices of the Royal Astronomical Society, 472, 2263

Häußler B., et al., 2013, Monthly Notices of the Royal Astronomical Society, 430, 330

Hawarden T. G., Mountain C. M., Leggett S. K., Puxley P. J., 1986, Monthly Notices of the Royal Astronomical Society, 221, 41P

Hayward C. C., Hopkins P. F., 2017, Monthly Notices of the Royal Astronomical Society, 465, 1682

Heckman T. M., 1980, Astronomy & Astrophysics, 500, 187

Heckman T. M., Best P. N., 2014, Annual Review of Astronomy and Astrophysics, 52, 589

Hemmati S., et al., 2019, The Astrophysical Journal Letters, 881, L14

Henriques B. M. B., White S. D. M., Thomas P. A., Angulo R., Guo Q., Lemson G., Springel V., Overzier R., 2015, Monthly Notices of the Royal Astronomical Society, 451, 2663

Henriques B. M. B., Yates R. M., Fu J., Guo Q., Kauffmann G., Srisawat C., Thomas P. A., White S. D. M., 2020, Monthly Notices of the Royal Astronomical Society, 491, 5795

Herschel J. F. W., 1864, Philosophical Transactions of the Royal Society of London Series I, 154, 1

Hester J. A., 2006, The Astrophysical Journal, 647, 910

Hickox R. C., et al., 2009, The Astrophysical Journal, 696, 891

Hill D. T., et al., 2011, Monthly Notices of the Royal Astronomical Society, 412, 765

Hlavacek-Larrondo J., Fabian A. C., Edge A. C., Ebeling H., Sanders J. S., Hogan M. T., Taylor G. B., 2012, Monthly Notices of the Royal Astronomical Society, 421, 1360

Hocking A., Geach J. E., Davey N., Sun Y., 2017, in 2017 International Joint Conference on Neural Networks. p. 4179

Hocking A., Geach J. E., Sun Y., Davey N., 2018, Monthly Notices of the Royal Astronomical Society, 473, 1108

Holmberg E., 1958, Meddelanden fran Lunds Astronomiska Observatorium Serie II, 136, 1

Hopkins P. F., Hernquist L., Cox T. J., Di Matteo T., Martini P., Robertson B., Springel V., 2005, The Astrophysical Journal, 630, 705

Hopkins P. F., Hernquist L., Cox T. J., Di Matteo T., Robertson B., Springel V., 2006, The Astrophysical Journal Supplement Series, 163, 1

Hopkins P. F., et al., 2009a, Monthly Notices of the Royal Astronomical Society, 397, 802

Hopkins P. F., Cox T. J., Younger J. D., Hernquist L., 2009b, The Astrophysical Journal, 691, 1168

Hopkins P. F., et al., 2010, The Astrophysical Journal, 715, 202

Hopkins P. F., Quataert E., Murray N., 2011, Monthly Notices of the Royal Astronomical Society, 417, 950

Hopkins P. F., Quataert E., Murray N., 2012, Monthly Notices of the Royal Astronomical Society, 421, 3522

Hopkins P. F., Kereš D., Oñorbe J., Faucher-Giguère C.-A., Quataert E., Murray N., Bullock J. S., 2014, Monthly Notices of the Royal Astronomical Society, 445, 581

Hopkins P. F., et al., 2018, Monthly Notices of the Royal Astronomical Society, 480, 800

Hotelling H., 1936, Biometrika, 28, 321

Hubble E. P., 1922, The Astrophysical Journal, 56, 162

Hubble E. P., 1925, Popular Astronomy, 33

Hubble E. P., 1926a, The Astrophysical Journal, 63, 236

Hubble E. P., 1926b, The Astrophysical Journal, 64, 321

Hubble E. P., 1927, The Observatory, 50, 276

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

Hubble E. P., 1930, The Astrophysical Journal, 71, 231

Hubble E., 1935, The Astrophysical Journal, 81, 334

Hubble E. P., 1936, Realm of the Nebulae

Hubble E., Humason M. L., 1931, The Astrophysical Journal, 74, 43

Huertas-Company M., Rouan D., Tasca L., Soucail G., Le Fèvre O., 2008, Astronomy & Astrophysics, 478, 971

Huertas-Company M., Aguerri J. A. L., Bernardi M., Mei S., Sánchez Almeida J., 2011, Astronomy & Astrophysics, 525, A157

Huertas-Company M., et al., 2015, The Astrophysical Journal Supplement Series, 221, 8

Huertas-Company M., et al., 2019, Monthly Notices of the Royal Astronomical Society, 489, 1859

Huggins W., Miller W. A., 1864, Philosophical Transactions of the Royal Society of London Series I, 154, 437

Hunter J. D., 2007, Computing In Science & Engineering, 9, 90

Ichimaru S., 1977, The Astrophysical Journal, 214, 840

Ilbert O., et al., 2006, Astronomy & Astrophysics, 457, 841

Ilbert O., et al., 2010, The Astrophysical Journal, 709, 644

Ilbert O., et al., 2013, Astronomy & Astrophysics, 556, A55

Ilbert O., et al., 2015, Astronomy & Astrophysics, 579, A2

Immeli A., Samland M., Gerhard O., Westera P., 2004, Astronomy & Astrophysics, 413, 547

Ivezić Ž., et al., 2019, The Astrophysical Journal, 873, 111

Jaccard P., 1901, Bulletin de la Societe Vaudoise des Sciences Naturelles, 37, 547

Jarvis M. J., et al., 2013, Monthly Notices of the Royal Astronomical Society, 428, 1281

Jeans J. H., 1928, Astronomy and Cosmogony

Jones E., Oliphant T., Peterson P., et al., 2001, SciPy: Open source scientific tools for Python. "http://www.scipy.org/"

Kant I., 1755, Allgemeine Naturgeschichte und Theorie des Himmels

Katz N., White S. D. M., 1993, The Astrophysical Journal, 412, 455

Katz N., Keres D., Dave R., Weinberg D. H., 2003, How Do Galaxies Get Their Gas?. Astrophysics and Space Science Library Vol. 281

Kauffmann G., White S. D. M., 1993, Monthly Notices of the Royal Astronomical Society, 261, 921

Kauffmann G., White S. D. M., Guiderdoni B., 1993, Monthly Notices of the Royal Astronomical Society, 264, 201

Kauffmann G., et al., 2003a, Monthly Notices of the Royal Astronomical Society, 341, 54

Kauffmann G., et al., 2003b, Monthly Notices of the Royal Astronomical Society, 346, 1055

Kauffmann G., White S. D. M., Heckman T. M., Ménard B., Brinchmann J., Charlot S., Tremonti C., Brinkmann J., 2004, Monthly Notices of the Royal Astronomical Society, 353, 713

Kaufmann L., Rousseeuw P., 1987, Data Analysis based on the L1-Norm and Related Methods, p. 405

Kaviraj S., 2014, Monthly Notices of the Royal Astronomical Society, 437, L41

Kaviraj S., et al., 2017, Monthly Notices of the Royal Astronomical Society, 467, 4739

Keller B. W., Wadsley J., Couchman H. M. P., 2016, Monthly Notices of the Royal Astronomical Society, 463, 1431

Kelvin L. S., et al., 2012, Monthly Notices of the Royal Astronomical Society, 421, 1007

Kelvin L. S., et al., 2014a, Monthly Notices of the Royal Astronomical Society, 439, 1245

Kelvin L. S., et al., 2014b, Monthly Notices of the Royal Astronomical Society, 444, 1647

Kelvin L. S., et al., 2018, Monthly Notices of the Royal Astronomical Society, 477, 4116

Kenney J. D. P., Rubin V. C., Planesas P., Young J. S., 1995, The Astrophysical Journal, 438, 135

Kennicutt Robert C. J., 1998, Annual Review of Astronomy and Astrophysics, 36, 189

Kennicutt Robert C. J., Tamblyn P., Congdon C. E., 1994, The Astrophysical Journal, 435, 22

Kereš D., Katz N., Weinberg D. H., Davé R., 2005, Monthly Notices of the Royal Astronomical Society, 363, 2

Kereš D., Katz N., Fardal M., Davé R., Weinberg D. H., 2009, Monthly Notices of the Royal Astronomical Society, 395, 160

Kewley L. J., Nicholls D. C., Sutherland R. S., 2019, Annual Review of Astronomy and Astrophysics, 57, 511

Klypin A. A., Trujillo-Gomez S., Primack J., 2011, The Astrophysical Journal, 740, 102

Kodama T., et al., 2005, Publications of the Astronomical Society of Japan, 57, 309

Koekemoer A. M., et al., 2011, The Astrophysical Journal Supplement Series, 197, 36

Kohonen T., 1982, Biological Cybernetics, 43, 59

Kormendy J., 1993, in Dejonghe H., Habing H. J., eds, IAU Symposium Vol. 153, Galactic Bulges. p. 209

Kormendy J., Bender R., 1996, The Astrophysical Journal Letters, 464, L119

Kormendy J., Bender R., 2012, The Astrophysical Journal Supplement Series, 198, 2

Kormendy J., Ho L. C., 2013, Annual Review of Astronomy and Astrophysics, 51, 511

Kormendy J., Kennicutt Robert C. J., 2004, Annual Review of Astronomy and Astrophysics, 42, 603

Kovač K., et al., 2010, The Astrophysical Journal, 708, 505

Kraft R. P., et al., 2017, The Astrophysical Journal, 848, 27

Kragh H., 2013, Astronomy and Geophysics, 54, 2.28

Kramer M. A., 1991, AIChE Journal, 37, 233

Kruk S. J., Erwin P., Debattista V. P., Lintott C., 2019, Monthly Notices of the Royal Astronomical Society, 490, 4721

Krywult J., et al., 2017, Astronomy & Astrophysics, 598, A120

Kukstas E., McCarthy I. G., Baldry I. K., Font A. S., 2020, Monthly Notices of the Royal Astronomical Society, 496, 2241

Kullback S., Leibler R. A., 1951, Ann. Math. Statist., 22, 79

Lacey C., Cole S., 1993, Monthly Notices of the Royal Astronomical Society, 262, 627

Lahav O., et al., 1995, Science, 267, 859

Lamareille F., 2010, Astronomy & Astrophysics, 509, A53

Lambas D. G., Alonso S., Mesa V., O'Mill A. L., 2012, Astronomy & Astrophysics, 539, A45

Lang D., Hogg D. W., Schlegel D. J., 2016, The Astronomical Journal, 151, 36

Lange R., et al., 2015, Monthly Notices of the Royal Astronomical Society, 447, 2603

Larson R. B., Tinsley B. M., Caldwell C. N., 1980, The Astrophysical Journal, 237, 692

Laureijs R., et al., 2011, Euclid Definition Study Report

Le Fèvre O., et al., 2003, in Iye M., Moorwood A. F. M., eds, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series Vol. 4841, Instrument Design and Performance for Optical/Infrared Ground-based Telescopes. p. 1670

LeCun Y., 1987, Ph.D. thesis: Modeles connexionnistes de l'apprentissage. Universite P. et M. Curie (Paris 6)

LeCun Y., Bottou L., Bengio Y., Haffner P., 1998, Proceedings of the IEEE, 86, 2278

Leavitt H. S., Pickering E. C., 1912, Harvard College Observatory Circular, 173, 1

Legendre A., 1805, Nouvelles méthodes pour la détermination des orbites des comètes. F. Didot

Lemaître G., 1927, Annales de la Société; Scientifique de Bruxelles, 47, 49

Lemaître G., 1931a, Monthly Notices of the Royal Astronomical Society, 91, 483

Lemaître G., 1931b, Nature, 127, 706

Lemson G., Springel V., 2006, in Gabriel C., Arviset C., Ponz D., Enrique S., eds, Astronomical Society of the Pacific Conference Series Vol. 351, Astronomical Data Analysis Software and Systems XV. p. 212

Lilly S., et al., 1998, The Astrophysical Journal, 500, 75

Lintott C. J., et al., 2008, Monthly Notices of the Royal Astronomical Society, 389, 1179

Lintott C., et al., 2011, Monthly Notices of the Royal Astronomical Society, 410, 166

Lisboa P., Ellis I., Green A., Ambrogi F., Dias M., 2008, Pattern Recognition Letters, 29, 1814

Lisboa P. J., Etchells T. A., Jarman I. H., Chambers S. J., 2013, BMC Bioinformatics, 14, S8

Liske J., Lemon D. J., Driver S. P., Cross N. J. G., Couch W. J., 2003, Monthly Notices of the Royal Astronomical Society, 344, 307

Liske J., et al., 2015, Monthly Notices of the Royal Astronomical Society, 452, 2087

Liu Y., Li Z., Xiong H., Gao X., Wu J., 2010, in Proceedings of the 2010 IEEE International Conference on Data Mining. ICDM '10. IEEE Computer Society, p. 911

Lloyd S., 1982, IEEE Transactions on Information Theory, 28, 129

Lotz J. M., Primack J., Madau P., 2004, The Astronomical Journal, 128, 163

Lotz J. M., Jonsson P., Cox T. J., Primack J. R., 2010a, Monthly Notices of the Royal Astronomical Society, 404, 575

Lotz J. M., Jonsson P., Cox T. J., Primack J. R., 2010b, Monthly Notices of the Royal Astronomical Society, 404, 590

Lotz J. M., Jonsson P., Cox T. J., Croton D., Primack J. R., Somerville R. S., Stewart K., 2011, The Astrophysical Journal, 742, 103

Ludlow A. D., Schaye J., Schaller M., Bower R., 2020, Monthly Notices of the Royal Astronomical Society, 493, 2926

Luo Y., et al., 2020, Monthly Notices of the Royal Astronomical Society, 493, 1686

Lynden-Bell D., 1969, Nature, 223, 690

MacQueen J., 1967, in Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics. p. 281

Madau P., Dickinson M., 2014, Annual Review of Astronomy and Astrophysics, 52, 415

Madau P., Ferguson H. C., Dickinson M. E., Giavalisco M., Steidel C. C., Fruchter A., 1996, Monthly Notices of the Royal Astronomical Society, 283, 1388

Madau P., Pozzetti L., Dickinson M., 1998, The Astrophysical Journal, 498, 106

Madgwick D. S., et al., 2002, Monthly Notices of the Royal Astronomical Society, 333, 133

Madgwick D. S., Somerville R., Lahav O., Ellis R., 2003, Monthly Notices of the Royal Astronomical Society, 343, 871

Maiolino R., et al., 2012, Monthly Notices of the Royal Astronomical Society, 425, L66

Malmquist K. G., 1922, Meddelanden fran Lunds Astronomiska Observatorium Serie I, 100, 1

Maraston C., 2005, Monthly Notices of the Royal Astronomical Society, 362, 799

Marchetti A., et al., 2013, Monthly Notices of the Royal Astronomical Society, 428, 1424

Marcolini A., Brighenti F., D'Ercole A., 2003, Monthly Notices of the Royal Astronomical Society, 345, 1329

Marinoni C., et al., 2008, Astronomy & Astrophysics, 487, 7

Martig M., Bournaud F., Teyssier R., Dekel A., 2009, The Astrophysical Journal, 707, 250

Martig M., Bournaud F., Croton D. J., Dekel A., Teyssier R., 2012, The Astrophysical Journal, 756, 26

Martig M., et al., 2013, Monthly Notices of the Royal Astronomical Society, 432, 1914

Martin C. L., 1999, The Astrophysical Journal, 513, 156

Martin C. L., 2005, The Astrophysical Journal, 621, 227

Martin D. C., et al., 2005, The Astrophysical Journal Letters, 619, L1

Martin D. C., et al., 2007, The Astrophysical Journal Supplement Series, 173, 342

Martin G., Kaviraj S., Hocking A., Read S. C., Geach J. E., 2020, Monthly Notices of the Royal Astronomical Society, 491, 1408

Martins L. P., González Delgado R. M., Leitherer C., Cerviño M., Hauschildt P., 2005, Monthly Notices of the Royal Astronomical Society, 358, 49

Masters K. L., et al., 2010, Monthly Notices of the Royal Astronomical Society, 405, 783

Masters K. L., et al., 2011, Monthly Notices of the Royal Astronomical Society, 411, 2026

Masters K. L., et al., 2012, Monthly Notices of the Royal Astronomical Society, 424, 2180

Masters D., et al., 2015, The Astrophysical Journal, 813, 53

Matzner C. D., 2002, The Astrophysical Journal, 566, 302

Mayer L., Mastropietro C., Wadsley J., Stadel J., Moore B., 2006, Monthly Notices of the Royal Astronomical Society, 369, 1021

McAlpine S., et al., 2016, Astronomy and Computing, 15, 72

McCarthy I. G., Frenk C. S., Font A. S., Lacey C. G., Bower R. G., Mitchell N. L., Balogh M. L., Theuns T., 2008, Monthly Notices of the Royal Astronomical Society, 383, 593

McCarthy I. G., et al., 2010, Monthly Notices of the Royal Astronomical Society, 406, 822

McConnell N. J., Ma C.-P., 2013, The Astrophysical Journal, 764, 184

McCulloch W. S., Pitts W., 1943, The Bulletin of Mathematical Biophysics, 5, 115

McGee S. L., Balogh M. L., Bower R. G., Font A. S., McCarthy I. G., 2009, Monthly Notices of the Royal Astronomical Society, 400, 937

McInnes L., Healy J., Melville J., 2018, arXiv e-prints, p. arXiv:1802.03426

McLachlan G. J., Basford K. E., 1988, Mixture models. Inference and applications to clustering

McNamara B. R., et al., 2000, The Astrophysical Journal Letters, 534, L135

McPartland C., Sanders D. B., Kewley L. J., Leslie S. K., 2019, Monthly Notices of the Royal Astronomical Society, 482, L129

Merritt D., 1984, The Astrophysical Journal, 276, 26

Mignoli M., et al., 2009, Astronomy & Astrophysics, 493, 39

Mihos J. C., 2004, in Mulchaey J. S., Dressler A., Oemler A., eds, Clusters of Galaxies: Probes of Cosmological Structure and Galaxy Evolution. p. 277

Mihos J. C., Hernquist L., 1994a, The Astrophysical Journal Letters, 425, L13

Mihos J. C., Hernquist L., 1994b, The Astrophysical Journal Letters, 431, L9

Mihos J. C., Hernquist L., 1996, The Astrophysical Journal, 464, 641

Mishra P. K., Wadadekar Y., Barway S., 2017, Monthly Notices of the Royal Astronomical Society, 467, 2384

Mishra P. K., Wadadekar Y., Barway S., 2018, Monthly Notices of the Royal Astronomical Society, 478, 351

Mishra P. K., Wadadekar Y., Barway S., 2019, Monthly Notices of the Royal Astronomical Society, 487, 5572

Mo H. J., Mao S., White S. D. M., 1998, Monthly Notices of the Royal Astronomical Society, 295, 319

Moffett A. J., et al., 2016, Monthly Notices of the Royal Astronomical Society, 457, 1308

Moffett A. J., et al., 2019, Monthly Notices of the Royal Astronomical Society, 489, 2830

Molinari E., Smareglia R., 1998, Astronomy & Astrophysics, 330, 447

Momcheva I. G., et al., 2016, The Astrophysical Journal Supplement Series, 225, 27

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

Moore B., Lake G., Katz N., 1998, The Astrophysical Journal, 495, 139

Moore B., Lake G., Quinn T., Stadel J., 1999, Monthly Notices of the Royal Astronomical Society, 304, 465

Morrissey P., et al., 2007, The Astrophysical Journal Supplement Series, 173, 682

Moutard T., et al., 2016a, Astronomy & Astrophysics, 590, A102

Moutard T., et al., 2016b, Astronomy & Astrophysics, 590, A103

Moutard T., Sawicki M., Arnouts S., Golob A., Malavasi N., Adami C., Coupon J., Ilbert O., 2018, Monthly Notices of the Royal Astronomical Society, 479, 2147

Moutard T., Malavasi N., Sawicki M., Arnouts S., Tripathi S., 2020, Monthly Notices of the Royal Astronomical Society, 495, 4237

Murray N., Quataert E., Thompson T. A., 2010, The Astrophysical Journal, 709, 191

Murray N., Ménard B., Thompson T. A., 2011, The Astrophysical Journal, 735, 66

Muzzin A., et al., 2013, The Astrophysical Journal, 777, 18

Muzzin A., et al., 2014, The Astrophysical Journal, 796, 65

Naab T., Trujillo I., 2006, Monthly Notices of the Royal Astronomical Society, 369, 625

Naab T., Johansson P. H., Ostriker J. P., 2009, The Astrophysical Journal Letters, 699, L178

Naim A., et al., 1995, Monthly Notices of the Royal Astronomical Society, 274, 1107

Nair P. B., Abraham R. G., 2010, The Astrophysical Journal Letters, 714, L260

Nandra K., et al., 2007, The Astrophysical Journal Letters, 660, L11

Navarro J. F., Steinmetz M., 2000, The Astrophysical Journal, 538, 477

Nelson D., et al., 2018, Monthly Notices of the Royal Astronomical Society, 475, 624

Netzer H., 2015, Annual Review of Astronomy and Astrophysics, 53, 365

Newnham L., Hess K. M., Masters K. L., Kruk S., Penny S. J., Lingard T., Smethurst R. J., 2020, Monthly Notices of the Royal Astronomical Society, 492, 4697

Nipoti C., Binney J., 2007, Monthly Notices of the Royal Astronomical Society, 382, 1481

Noeske K. G., et al., 2007, The Astrophysical Journal Letters, 660, L43

Noll S., Burgarella D., Giovannoli E., Buat V., Marcillac D., Muñoz-Mateos J. C., 2009, Astronomy & Astrophysics, 507, 1793

Nolte A., Wang L., Biehl M., 2018, ESANN, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, 26, 339

Nulsen P. E. J., 1982, Monthly Notices of the Royal Astronomical Society, 198, 1007

Oemler Augustus J., 1974, The Astrophysical Journal, 194, 1

Oepik E., 1922, The Astrophysical Journal, 55, 406

Oh K., Choi H., Kim H.-G., Moon J.-S., Yi S. K., 2013, The Astronomical Journal, 146, 151

Okada T., Tomita S., 1985, Pattern Recognition, 18, 139

Oliphant T., 2006, A guide to NumPy. Trelgol Publishing

Oman K. A., Hudson M. J., 2016, Monthly Notices of the Royal Astronomical Society, 463, 3083

Oman K. A., Hudson M. J., Behroozi P. S., 2013, Monthly Notices of the Royal Astronomical Society, 431, 2307

Oppenheimer B. D., et al., 2020, Monthly Notices of the Royal Astronomical Society, 491, 2939

Ostriker J. P., Peebles P. J. E., 1973, The Astrophysical Journal, 186, 467

Pacifici C., et al., 2016, The Astrophysical Journal, 832, 79

Paolillo M., Fabbiano G., Peres G., Kim D. W., 2002, The Astrophysical Journal, 565, 883

Papovich C., Dickinson M., Ferguson H. C., 2001, The Astrophysical Journal, 559, 620

Papovich C., et al., 2018, The Astrophysical Journal, 854, 30

Parsons W., 1850, Philosophical Transactions of the Royal Society of London Series I, 140, 499

Pearson K., 1901, The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 2, 559

Pedregosa F., et al., 2011, Journal of Machine Learning Research, 12, 2825

Peebles P. J. E., 1968, The Astrophysical Journal, 153, 1

Peebles P. J. E., 1982, The Astrophysical Journal Letters, 263, L1

Peek J. E. G., Schiminovich D., 2013, The Astrophysical Journal, 771, 68

Peng C. Y., Ho L. C., Impey C. D., Rix H.-W., 2002, The Astronomical Journal, 124, 266

Peng Y.-j., et al., 2010, The Astrophysical Journal, 721, 193

Peng Y.-j., Lilly S. J., Renzini A., Carollo M., 2012, The Astrophysical Journal, 757, 4

Peng Y., Maiolino R., Cochrane R., 2015, Nature, 521, 192

Penzias A. A., Wilson R. W., 1965, The Astrophysical Journal, 142, 419

Perlmutter S., et al., 1999, The Astrophysical Journal, 517, 565

Petrosian V., 1976, The Astrophysical Journal Letters, 210, L53

Pickles A. J., 1998, Publications of the Astronomical Society of the Pacific, 110, 863

Pillepich A., et al., 2018, Monthly Notices of the Royal Astronomical Society, 473, 4077

Planck Collaboration et al., 2014, Astronomy & Astrophysics, 571, A1

Planck Collaboration et al., 2018, arXiv e-prints, p. arXiv:1807.06209

Poggianti B. M., Barbaro G., 1997, Astronomy & Astrophysics, 325, 1025

Poggianti B. M., Smail I., Dressler A., Couch W. J., Barger A. J., Butcher H., Ellis R. S., Oemler Augustus J., 1999, The Astrophysical Journal, 518, 576

Poggianti B. M., et al., 2016, The Astronomical Journal, 151, 78

Poggianti B. M., et al., 2017, The Astrophysical Journal, 844, 48

Popesso P., et al., 2019a, Monthly Notices of the Royal Astronomical Society, 483, 3213

Popesso P., et al., 2019b, Monthly Notices of the Royal Astronomical Society, 490, 5285

Pozzetti L., et al., 2010, Astronomy & Astrophysics, 523, A13

Puget P., et al., 2004, WIRCam: the infrared wide-field camera for the Canada-France-Hawaii Telescope. p. 978

Qu Y., et al., 2017, Monthly Notices of the Royal Astronomical Society, 464, 1659

R Core Team 2019, R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org

Read J. I., Wilkinson M. I., Evans N. W., Gilmore G., Kleyna J. T., 2006, Monthly Notices of the Royal Astronomical Society, 366, 429

Refregier A., 2003, Annual Review of Astronomy and Astrophysics, 41, 645

Renaud F., Bournaud F., Kraljic K., Duc P. A., 2014, Monthly Notices of the Royal Astronomical Society, 442, L33

Renzini A., 1999, in Carollo C. M., Ferguson H. C., Wyse R. F. G., eds, The Formation of Galactic Bulges. p. 9

Reynolds J. H., 1920, Monthly Notices of the Royal Astronomical Society, 80, 746

Riess A. G., et al., 1998, The Astronomical Journal, 116, 1009

Ritchey G. W., 1917, Publications of the Astronomical Society of the Pacific, 29, 210

Roberts M. S., 1963, Annual Review of Astronomy and Astrophysics, 1, 149

Robotham A. S. G., et al., 2011, Monthly Notices of the Royal Astronomical Society, 416, 2640

Robotham A. S. G., Taranu D. S., Tobar R., Moffett A., Driver S. P., 2017, Monthly Notices of the Royal Astronomical Society, 466, 1513

Robotham A. S. G., Davies L. J. M., Driver S. P., Koushan S., Taranu D. S., Casura S., Liske J., 2018, Monthly Notices of the Royal Astronomical Society, 476, 3137

Rodriguez-Gomez V., et al., 2019, Monthly Notices of the Royal Astronomical Society, 483, 4140

Roediger E., Hensler G., 2005, Astronomy & Astrophysics, 433, 875

Rosenblatt F., 1957, The Perceptron, a perceiving and recognizing automaton, Cornell Aeronautical Laboratory report

Rosito M. S., Pedrosa S. E., Tissera P. B., Avila-Reese V., Lacerna I., Bignone L. A., Ibarra-Medel H. J., Varela S., 2018, Astronomy & Astrophysics, 614, A85

Rosito M. S., Tissera P. B., Pedrosa S. E., Lagos C. D. P., 2019, Astronomy & Astrophysics, 629, L3

Rousseeuw P. J., 1987, Journal of Computational and Applied Mathematics, 20, 53

Rubin V. C., Ford W. Kent J., 1970, The Astrophysical Journal, 159, 379

Rubin V. C., Ford W. K. J., Thonnard N., 1980, The Astrophysical Journal, 238, 471

Sachs R. K., Wolfe A. M., 1967, The Astrophysical Journal, 147, 73

Salim S., 2014, Serbian Astronomical Journal, 189, 1

Salim S., Rich R. M., 2010, The Astrophysical Journal Letters, 714, L290

Salim S., et al., 2007, The Astrophysical Journal Supplement Series, 173, 267

Salim S., et al., 2016, The Astrophysical Journal Supplement Series, 227, 2

Salim S., Boquien M., Lee J. C., 2018, The Astrophysical Journal, 859, 11

Salimbeni S., et al., 2008, Astronomy & Astrophysics, 477, 763

Salpeter E. E., 1955, The Astrophysical Journal, 121, 161

Sammon J. W., 1969, IEEE Transactions on Computers, C-18, 401

Sánchez Almeida J., Aguerri J. A. L., Muñoz-Tuñón C., de Vicente A., 2010, The Astrophysical Journal, 714, 487

Sandage A., 1961, The Hubble Atlas of Galaxies

Sandage A., 2005, Annual Review of Astronomy and Astrophysics, 43, 581

Sandage A., Binggeli B., 1984, The Astronomical Journal, 89, 919

Scannapieco C., White S. D. M., Springel V., Tissera P. B., 2009, Monthly Notices of the Royal Astronomical Society, 396, 696

Scannapieco C., Gadotti D. A., Jonsson P., White S. D. M., 2010, Monthly Notices of the Royal Astronomical Society, 407, L41

Scarlata C., et al., 2007, The Astrophysical Journal Supplement Series, 172, 406

Schaller M., Dalla Vecchia C., Schaye J., Bower R. G., Theuns T., Crain R. A., Furlong M., McCarthy I. G., 2015, Monthly Notices of the Royal Astronomical Society, 454, 2277

Schawinski K., et al., 2006, Nature, 442, 888

Schawinski K., Thomas D., Sarzi M., Maraston C., Kaviraj S., Joo S.-J., Yi S. K., Silk J., 2007, Monthly Notices of the Royal Astronomical Society, 382, 1415

Schawinski K., et al., 2009a, Monthly Notices of the Royal Astronomical Society, 396, 818

Schawinski K., et al., 2009b, The Astrophysical Journal, 690, 1672

Schawinski K., et al., 2010, The Astrophysical Journal, 711, 284

Schawinski K., et al., 2014, Monthly Notices of the Royal Astronomical Society, 440, 889

Schaye J., et al., 2015, Monthly Notices of the Royal Astronomical Society, 446, 521

Schiminovich D., et al., 2007, The Astrophysical Journal Supplement Series, 173, 315

Schubert E., Sander J., Ester M., Kriegel H. P., Xu X., 2017, ACM Trans. Database Syst., 42

Schwarz G., 1978, Ann. Statist., 6, 461

Scodeggio M., et al., 2018, Astronomy & Astrophysics, 609, A84

Scott D., Thompson J., 1983, Computer Science and Statistics: Proceedings of the Fifteenth Symposium on the Interface

Selim S. Z., Ismail M. A., 1984, IEEE Transactions on Pattern Analysis and Machine Intelligence, 6, 81

Sellwood J. A., Wilkinson A., 1993, Reports on Progress in Physics, 56, 173

Serra P., et al., 2012, Monthly Notices of the Royal Astronomical Society, 422, 1835

Sérsic J. L., 1963, Boletin de la Asociacion Argentina de Astronomia La Plata Argentina, 6, 41

Sérsic J. L., 1968, Atlas de Galaxias Australes

Shakura N. I., Sunyaev R. A., 1973, Astronomy & Astrophysics, 500, 33

Shapley H., 1943, Galaxies. p. 50

Shapley H., Ames A., 1926, Harvard College Observatory Circular, 294, 1

Shapley H., Curtis H. D., 1921, Bulletin of the National Research Council, 2, 171

Shapley H., Paraskevopoulos J. S., 1940, Proceedings of the National Academy of Science, 26, 31

Sharp R., et al., 2006, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series. p. 62690G

Shen S., Mo H. J., White S. D. M., Blanton M. R., Kauffmann G., Voges W., Brinkmann J., Csabai I., 2003, Monthly Notices of the Royal Astronomical Society, 343, 978

Sheth R. K., et al., 2003, The Astrophysical Journal, 594, 225

Sheth K., Vogel S. N., Regan M. W., Thornley M. D., Teuben P. J., 2005, The Astrophysical Journal, 632, 217

Sibson R., 1973, The Computer Journal, 16, 30

Silk J., 1967, Nature, 215, 1155

Silk J., 1968, The Astrophysical Journal, 151, 459

Silk J., 1977, The Astrophysical Journal, 211, 638

Silk J., Rees M. J., 1998, Astronomy & Astrophysics, 331, L1

Silva A., et al., 2018, The Astrophysical Journal, 868, 46

Simard L., Mendel J. T., Patton D. R., Ellison S. L., McConnachie A. W., 2011, The Astrophysical Journal Supplement Series, 196, 11

Simmons B. D., et al., 2017, Monthly Notices of the Royal Astronomical Society, 464, 4420

Siudek M., et al., 2017, Astronomy & Astrophysics, 597, A107

Siudek M., et al., 2018a, arXiv e-prints, p. arXiv:1805.09905

Siudek M., et al., 2018b, Astronomy & Astrophysics, 617, A70

Skelton R. E., et al., 2014, The Astrophysical Journal Supplement Series, 214, 24

Skibba R. A., et al., 2009, Monthly Notices of the Royal Astronomical Society, 399, 966

Skrutskie M. F., et al., 2006, The Astronomical Journal, 131, 1163

Slipher V. M., 1914, Lowell Observatory Bulletin, 2, 66

Slipher V. M., 1915, Popular Astronomy, 23, 21

Smethurst R. J., et al., 2015, Monthly Notices of the Royal Astronomical Society, 450, 435

Smethurst R. J., Lintott C. J., Bamford S. P., Hart R. E., Kruk S. J., Masters K. L., Nichol R. C., Simmons B. D., 2017, Monthly Notices of the Royal Astronomical Society, 469, 3670

Smith R., et al., 2015, Monthly Notices of the Royal Astronomical Society, 454, 2502

Smoot G. F., et al., 1992, The Astrophysical Journal Letters, 396, L1

Sodré L. J., Cuevas H., 1994, Vistas in Astronomy, 38, 287

Sofue Y., Rubin V., 2001, Annual Review of Astronomy and Astrophysics, 39, 137

Somerville R. S., Davé R., 2015, Annual Review of Astronomy and Astrophysics, 53, 51

Somerville R. S., Kolatt T. S., 1999, Monthly Notices of the Royal Astronomical Society, 305, 1

Somerville R. S., Primack J. R., 1999, Monthly Notices of the Royal Astronomical Society, 310, 1087

Somerville R. S., Hopkins P. F., Cox T. J., Robertson B. E., Hernquist L., 2008, Monthly Notices of the Royal Astronomical Society, 391, 481

Songaila A., 2005, The Astronomical Journal, 130, 1996

Songaila A., 2006, The Astronomical Journal, 131, 24

Speagle J. S., Steinhardt C. L., Capak P. L., Silverman J. D., 2014, The Astrophysical Journal Supplement Series, 214, 15

Spindler A., Geach J. E., Smith M. J., 2020, arXiv e-prints, p. arXiv:2009.08470

Spinrad H., 1962, The Astrophysical Journal, 135, 715

Spinrad H., Taylor B. J., 1971, The Astrophysical Journal Supplement Series, 22, 445

Springel V., 2005, Monthly Notices of the Royal Astronomical Society, 364, 1105

Springel V., Hernquist L., 2003, Monthly Notices of the Royal Astronomical Society, 339, 289

Springel V., Di Matteo T., Hernquist L., 2005a, Monthly Notices of the Royal Astronomical Society, 361, 776

Springel V., et al., 2005b, Nature, 435, 629

Springel V., Di Matteo T., Hernquist L., 2005c, The Astrophysical Journal Letters, 620, L79

Springel V., Frenk C. S., White S. D. M., 2006, Nature, 440, 1137

Steinhardt C. L., Weaver J. R., Maxfield J., Davidzon I., Faisst A. L., Masters D., Schemel M., Toft S., 2020, The Astrophysical Journal, 891, 136

Stoughton C., et al., 2002, The Astronomical Journal, 123, 485

Strateva I., et al., 2001, The Astronomical Journal, 122, 1861

Sunyaev R. A., Zeldovich Y. B., 1972, Comments on Astrophysics and Space Physics, 4, 173

Sutherland W., et al., 2015, Astronomy & Astrophysics, 575, A25

Tacconi L. J., et al., 2010, Nature, 463, 781

Taylor M. B., 2005, in Shopbell P., Britton M., Ebert R., eds, Astronomical Society of the Pacific Conference Series Vol. 347, Astronomical Data Analysis Software and Systems XIV. p. 29

Taylor E. N., et al., 2011, Monthly Notices of the Royal Astronomical Society, 418, 1587

Taylor E. N., et al., 2015, Monthly Notices of the Royal Astronomical Society, 446, 2144

Thob A. C. R., et al., 2019, Monthly Notices of the Royal Astronomical Society, 485, 972

Tinsley B. M., 1968, The Astrophysical Journal, 151, 547

Tonini C., Mutch S. J., Croton D. J., Wyithe J. S. B., 2016, Monthly Notices of the Royal Astronomical Society, 459, 4109

Tonnesen S., Bryan G. L., 2012, Monthly Notices of the Royal Astronomical Society, 422, 1609

Toomre A., 1977, in Tinsley B. M., Larson Richard B. Gehret D. C., eds, Evolution of Galaxies and Stellar Populations. p. 401

Toomre A., Toomre J., 1972, The Astrophysical Journal, 178, 623

Trayford J. W., et al., 2015, Monthly Notices of the Royal Astronomical Society, 452, 2879

Trayford J. W., Theuns T., Bower R. G., Crain R. A., Lagos C. d. P., Schaller M., Schaye J., 2016, Monthly Notices of the Royal Astronomical Society, 460, 3925

Trayford J. W., et al., 2017, Monthly Notices of the Royal Astronomical Society, 470, 771

Tremonti C. A., Moustakas J., Diamond-Stanic A. a. M., 2007, The Astrophysical Journal Letters, 663, L77

Turner S., et al., 2019, Monthly Notices of the Royal Astronomical Society, 482, 126

Valiante E., et al., 2016, Monthly Notices of the Royal Astronomical Society, 462, 3146

Vergani D., et al., 2018, Astronomy & Astrophysics, 620, A193

Vika M., Bamford S. P., Häußler B., Rojas A. L., Borch A., Nichol R. C., 2013, Monthly Notices of the Royal Astronomical Society, 435, 623

Vika M., Vulcani B., Bamford S. P., Häußler B., Rojas A. L., 2015, Astronomy & Astrophysics, 577, A97

Vogelsberger M., et al., 2014, Monthly Notices of the Royal Astronomical Society, 444, 1518

Vogelsberger M., Marinacci F., Torrey P., Puchwein E., 2020, Nature Reviews Physics, 2, 42

Voit G. M., 2005, Reviews of Modern Physics, 77, 207

Vulcani B., et al., 2014, Monthly Notices of the Royal Astronomical Society, 441, 1340

Vulcani B., et al., 2018, The Astrophysical Journal Letters, 866, L25

Wada K., Norman C. A., 2001, The Astrophysical Journal, 547, 172

Wake D. A., van Dokkum P. G., Franx M., 2012, The Astrophysical Journal Letters, 751, L44

Walcher J., Groves B., Budavári T., Dale D., 2011, Astrophysics and Space Science, 331, 1

Walker I. R., Mihos J. C., Hernquist L., 1996, The Astrophysical Journal, 460, 121

Walmsley M., et al., 2020, Monthly Notices of the Royal Astronomical Society, 491, 1554

Ward Jr. J. H., 1963, Journal of the American Statistical Association, 58, 236

Weigel A. K., et al., 2017, The Astrophysical Journal, 845, 145

Weinberger R., et al., 2017, Monthly Notices of the Royal Astronomical Society, 465, 3291

Weiner B. J., et al., 2009, The Astrophysical Journal, 692, 187

Werbos P., 1975, Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences. Harvard University

Wetzel A. R., Tinker J. L., Conroy C., 2012, Monthly Notices of the Royal Astronomical Society, 424, 232

Wetzel A. R., Tinker J. L., Conroy C., van den Bosch F. C., 2013, Monthly Notices of the Royal Astronomical Society, 432, 336

Wetzel A. R., Deason A. J., Garrison-Kimmel S., 2015, The Astrophysical Journal, 807, 49

Whitaker K. E., et al., 2015, The Astrophysical Journal Letters, 811, L12

White S. D. M., Rees M. J., 1978, Monthly Notices of the Royal Astronomical Society, 183, 341

Wijesinghe D. B., Hopkins A. M., Kelly B. C., Welikala N., Connolly A. J., 2010, Monthly Notices of the Royal Astronomical Society, 404, 2077

Wild V., Hewett P. C., 2005, Monthly Notices of the Royal Astronomical Society, 358, 1083

Wild V., Kauffmann G., Heckman T., Charlot S., Lemson G., Brinchmann J., Reichard T., Pasquali A., 2007, Monthly Notices of the Royal Astronomical Society, 381, 543

Wild V., Walcher C. J., Johansson P. H., Tresse L., Charlot S., Pollo A., Le Fèvre O., de Ravel L., 2009, Monthly Notices of the Royal Astronomical Society, 395, 144

Wild V., et al., 2014, Monthly Notices of the Royal Astronomical Society, 440, 1880

Willett K. W., et al., 2013, Monthly Notices of the Royal Astronomical Society, 435, 2835

Willett K. W., et al., 2017, Monthly Notices of the Royal Astronomical Society, 464, 4176

Williams R. J., Quadri R. F., Franx M., van Dokkum P., Labbé I., 2009, The Astrophysical Journal, 691, 1879

Wolf M., 1908, Publikationen des Astrophysikalischen Instituts Koenigstuhl-Heidelberg, 3, 109

Woo J., et al., 2013, Monthly Notices of the Royal Astronomical Society, 428, 3306

Woo J., Carollo C. M., Faber S. M., Dekel A., Tacchella S., 2017, Monthly Notices of the Royal Astronomical Society, 464, 1077

Worthey G., 1994, The Astrophysical Journal Supplement Series, 95, 107

Wright T., 1750, An Original Theory or New Hypothesis of the Universe

Wright E. L., et al., 1992, The Astrophysical Journal Letters, 396, L13

Wright E. L., et al., 2010, The Astronomical Journal, 140, 1868

Wright A. H., et al., 2016, Monthly Notices of the Royal Astronomical Society, 460, 765

Wuyts S., et al., 2007, The Astrophysical Journal, 655, 51

Wyder T. K., et al., 2007, The Astrophysical Journal Supplement Series, 173, 293

Xia X. Y., Xue S. J., Mao S., Boller T., Deng Z. G., Wu H., 2002, The Astrophysical Journal, 564, 196

Yang X., Mo H. J., van den Bosch F. C., Jing Y. P., 2005, Monthly Notices of the Royal Astronomical Society, 356, 1293

Yang X., Mo H. J., van den Bosch F. C., Pasquali A., Li C., Barden M., 2007, The Astrophysical Journal, 671, 153

Yang Y., Zabludoff A. I., Zaritsky D., Mihos J. C., 2008, The Astrophysical Journal, 688, 945

Yip C. W., et al., 2004, The Astronomical Journal, 128, 585

York D. G., et al., 2000, The Astronomical Journal, 120, 1579

Yuan H. B., Liu X. W., Xiang M. S., 2013, Monthly Notices of the Royal Astronomical Society, 430, 2188

Zabludoff A. I., Zaritsky D., Lin H., Tucker D., Hashimoto Y., Shectman S. A., Oemler A., Kirshner R. P., 1996, The Astrophysical Journal, 466, 104

Zhu X., 2005, Semi-supervised learning literature survey

Zinger E., Dekel A., Kravtsov A. V., Nagai D., 2018, Monthly Notices of the Royal Astronomical Society, 475, 3654

Zwicky F., 1933, Helvetica Physica Acta, 6, 110

Zwicky F., 1952, Publications of the Astronomical Society of the Pacific, 64, 247

de Jong J. T. A., Verdoes Kleijn G. A., Kuijken K. H., Valentijn E. A., 2013, Experimental Astronomy, 35, 25

de Jong J. T. A., et al., 2015, Astronomy & Astrophysics, 582, A62

de Jong J. T. A., et al., 2017, Astronomy & Astrophysics, 604, A134

de Souza R. S., et al., 2017, Monthly Notices of the Royal Astronomical Society, 472, 2808

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

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

de Vaucouleurs G., de Vaucouleurs A., Corwin Herold G. J., Buta R. J., Paturel G., Fouque P., 1991, Third Reference Catalogue of Bright Galaxies

van Dokkum P. G., et al., 2014, The Astrophysical Journal, 791, 45

van Dokkum P. G., et al., 2015, The Astrophysical Journal, 813, 23

van Maanen A., 1916, The Astrophysical Journal, 44, 210

van Maanen A., 1935, The Astrophysical Journal, 81, 336

van de Sande J., et al., 2019, Monthly Notices of the Royal Astronomical Society, 484, 869

van den Bergh S., 1976, The Astrophysical Journal, 206, 883

van den Bergh S., Abraham R. G., Ellis R. S., Tanvir N. R., Santiago B. X., Glazebrook K. G., 1996, The Astronomical Journal, 112, 359

van der Maaten L., Hinton G., 2008, Journal of Machine Learning Research, 9, 2579

van der Wel A., 2008, The Astrophysical Journal Letters, 675, L13

van der Wel A., Bell E. F., van den Bosch F. C., Gallazzi A., Rix H.-W., 2009, The Astrophysical Journal, 698, 1232

van der Wel A., et al., 2014, The Astrophysical Journal, 788, 28

von Luxburg U., 2010, Foundations and Trends in Machine Learning, 2, 235

von der Linden A., Wild V., Kauffmann G., White S. D. M., Weinmann S., 2010, Monthly Notices of the Royal Astronomical Society, 404, 1231

*"So this marks the end of my doctorate,*
*And I hope that my claims are all accurate.*
*If they're not I will sob*
*Before getting a job*
*At a limerick-writing conglomerate."*

- Alfred Lawrence Athelstan Champion