

How the spectral energy distribution and galaxy morphology constrain each other, with application to morphological selection using galaxy colours

Emir Uzeirbegovic^{1,2} ,^{1,2}★ Garreth Martin^{3,4} and Sugata Kaviraj^{1,2} 

¹Centre for Astrophysics Research, School of Physics, Engineering & Computer Science, University of Hertfordshire, Hatfield AL10 9AB, UK

²Centre of Data Innovation Research, School of Physics, Engineering & Computer Science, University of Hertfordshire, Hatfield AL10 9AB, UK

³Korea Astronomy and Space Science Institute, 776 Daedeokdae-ro, Yuseong-gu, Daejeon 34055, Republic of Korea

⁴Steward Observatory, University of Arizona, 933 N. Cherry Ave, Tucson, AZ 85719, USA

Accepted 2021 December 13. Received 2021 December 13; in original form 2021 August 24

ABSTRACT

We introduce an empirical methodology to study how the spectral energy distribution (SED) and galaxy morphology constrain each other and implement this on ~ 8000 galaxies from the *HST* CANDELS survey in the GOODS-South field. We show that the SED does constrain morphology and present a method that quantifies the strength of the link between these two quantities. Two galaxies with very similar SEDs are around three times more likely to also be morphologically similar, with SED constraining morphology most strongly for relatively massive red ellipticals. We apply our methodology to explore likely upper bounds on the efficacy of morphological selection using colour. We show that, under reasonable assumptions, colour selection is relatively ineffective at separating homogeneous morphologies. Even with the use of up to six colours for morphological selection, the average purity in the resultant morphological classes is only around 60 per cent. While the results can be improved by using the whole SED, the gains are not significant, with purity values remaining around 70 per cent or below.

Key words: methods: data analysis – Galaxy: formation – galaxies: evolution – galaxies: stellar content – galaxies: structure.

1 INTRODUCTION

The visual appearance of galaxies – commonly referred to as galaxy morphology – correlates with their physical properties, such as stellar mass (e.g. Bundy, Ellis & Conselice 2005), star formation rate (SFR; e.g. Bluck et al. 2014; Kaviraj 2014; Smethurst et al. 2015; Lofthouse et al. 2017), surface brightness (e.g. Martin et al. 2019; Jackson et al. 2021), rest-frame colour (e.g. Strateva et al. 2001; Skibba et al. 2009), and local environment (e.g. Dressler et al. 1997; Postman et al. 2005). It reveals key information about the processes that have shaped the evolution of galaxies over cosmic time (e.g. Martin et al. 2018; Jackson et al. 2020). The literature documents many approaches to measuring galaxy morphology, with historically the most popular being based on visual classification schemes (e.g. Hubble 1926; Lintott et al. 2011; Simmons et al. 2017; Kaviraj, Martin & Silk 2019), light distribution based parametric methods (e.g. de Vaucouleurs 1948; Sérsic 1963; Odewahn et al. 2002; Simard et al. 2002; Lackner & Gunn 2012; Ryan et al. 2012), and non-parametric approaches such as Concentration-Asymmetry-Clumpiness (Abraham et al. 1994; Conselice 2003; Menanteau et al. 2006; Mager et al. 2018) or Gini-M20 (e.g. Lotz, Primack & Madau 2004; Scarlata et al. 2007; Peth et al. 2016). More recently improving computing power has introduced new empirical machine learning methods (e.g. Huertas-Company et al. 2015; Ostrovski et al. 2017; Schawinski et al. 2017; Goulding et al. 2018; Hocking et al. 2018;

Cheng et al. 2020; Martin et al. 2020) often applied to problems of classification or automated clustering in an attempt to make large surveys tractable for morphological analysis.

Although estimating the physical parameters of a galaxy from its spectral energy distribution (SED) is subject to confounders and ambiguity (e.g. Conroy 2013; Magris et al. 2015), the SED does trace the stellar composition of the galaxy and relates to physical quantities like stellar mass, metallicity, ionized-gas properties, and star formation history (e.g. Jones & Lambourne 2004). Several studies exist on the subject of the spectral classification of galaxies, which show that spectral types often have corresponding morphological biases (e.g. Morgan & Mayall 1957; Sodr  & Cuevas 1994; Connolly et al. 1995; Madgwick et al. 2003). In the same vein, derivatives of the SED such as colours are thought to correlate strongly with morphology and have frequently been used by astronomers as a proxy thereof (see Masters et al. 2019 for an extensive list). However, given that the shape of an SED is driven by the galaxy’s star formation history, while morphology depends on dynamical factors, it is not clear to what extent the SED and morphology can be expected to constrain each other.

Results from the spectral classification literature show that morphologies associated with spectral types are impure to varying degrees. For example, Masters et al. (2010) morphologically select 5433 face-on spirals from the Sloan Digital Sky Survey (SDSS) Data Release 6 (Adelman-McCarthy et al. 2007) of which 6 per cent are red spirals, which runs contrary to the expectation that late-types should have bluer colours than early-types. They also show that red spirals primarily manifest themselves in the higher mass range,

* E-mail: e.uzeirbegovic@herts.ac.uk

and motivate the claim that massive galaxies are red regardless of morphology. Skibba et al. (2009) show that colour and morphology appear to be differently related to the local environment and are bound to stray from a one-to-one mapping. Although these results imply that the constraints of SED and morphology on each other are asymmetrical, the question is yet to be explored directly and the constraint of morphology on the SED is yet to be properly considered.

In this work, we introduce an empirical methodology for the analysis of how SEDs and morphologies constrain each other. We utilize the vector space building techniques for galaxy surveys laid out in Uzeirbegovic, Geach & Kaviraj (2020). Galaxy cut-outs are projected to points in a high-dimensional space in which the Euclidean distances between galaxies relate directly to the visual morphological similarity between them. We likewise project galaxy SEDs as proxied by the catalogued rest-frame bands. We use these vector spaces together to demonstrate the implications that the structure of one space has on the other. In contrast to classification or morphological indicator-based methods, vector spaces enable us to compare galaxy similarity on a continuous scale, and – as we will show – conduct analyses without having to commit to specific features such as CAS or a particular typology (e.g. the Hubble sequence; Hubble 1936).

From here, the paper is structured as follows. In Section 2, we describe the source data used, how it is prepared, how the key artefact – the dissimilarity matrices – are derived from it, and how they are interpreted. In Section 3, we analyse the vector spaces to show that: (1) SEDs significantly constrain morphology and the extent to which it happens can be quantified, (2) the constraint is not uniform and some morphologies are especially well-constrained by the SED, (3) galaxy morphology constrains SEDs more than vice versa, (4) there is a visually distinguishable group of galaxies for which the SED and morphology are mutually constraining, and (5) there is an upper bound on coverage and purity when attempting to use rest-frame colours to select homogeneous morphologies, with direct implication on colour-based morphology selection methods. Finally, in Section 4 we conclude by summarizing the main points of our work.

2 DATA AND PREPARATION

2.1 HST-CANDELS

We use the *HST*-CANDELS (Grogin et al. 2011; Koekemoer et al. 2011) survey because it offers a high-resolution probe of galaxy evolution. It is supplied with a corresponding catalogue (Guo et al. 2013; Santini et al. 2015). The survey consists of optical and near-infrared (UVIS/IR) images from the Wide Field Camera 3 (WFC3) and optical images from the Advanced Camera for Surveys in five well-studied extragalactic survey fields. We focus on GOODS-S, one of the deep tier (at least four-orbit effective depth) fields. In order to keep galaxies comparable to each other and minimize the effects of noise, we select only galaxies imaged in the WFC3 *F160W* filter at $z < 3$ with signal-to-noise ratio >20 and $M_* \geq 10^8 M_\odot$. The signal-to-noise ratio is provided per object in the catalogue and is calculated as flux divided by flux error in the *F160W* band. This filtering results in a sample of 7757 galaxies with images in the WFC3 *F160W* filter.

For each object after filtering, we take 31×31 (1.8 arcsec) pixel cut-outs, using the catalogued sky coordinate as a centroid. We also make use of the photometric redshift, mass, SFR, and rest-frame *UBVRIJK* magnitudes provided in the catalogue. The photometric redshifts have normalized median absolute deviation (NMAD) values

better than 0.05 and 0.03 at $z < 1.5$ and $z > 1.5$ respectively and outlier fractions of ~ 5 per cent or better. From herein, we refer to the total number of galaxies in our data set as N , where $N = 7757$, and the pixel width of square cut-outs as n , where $n = 31$.

2.2 The morphological dissimilarity matrix

The representation of (visual) morphological similarity in our method is based on the calculation of a ‘dissimilarity’ matrix, which encodes a measure of the pairwise differences between galaxies. To create a dissimilarity matrix, we begin by projecting all galaxy cut-outs on to a common vector space, in which the squared Euclidean distance between vectors is informative of morphological differences. We call this the *morphology space*. We follow the vector space building methodology for galaxy surveys laid out in Uzeirbegovic et al. (2020) with the following cut-out standardization steps applied sequentially prior to projection:

(i) Masking – Many cut-outs contain background noise that can cause spurious similarity. For each cut-out, we mask away the background by robustly fitting a Gaussian using the median (α) and interquartile range (β) of the flux densities of the pixels in the cut-out, and clipping to zero all pixels below a threshold t given by the solution to $\Phi(X \leq t|\alpha, \beta) = 0.9$, where Φ is a Gaussian CDF. Since a majority of the pixels in most cut-outs are background dominated, this has the effect of zeroing out the background. The process is not overly sensitive to our chosen constant (0.9) and will produce much the same results in the range 0.85–0.95.

(ii) Rotation – The covariance of the coordinates of all the non-zero pixels in *F160W* are used to find the major axis of the data. The cut-out is then rotated so as to bring the galaxy in line with the major axis. This step standardizes the orientation of all cut-outs.

(iii) Flipping – All cut-outs are flipped as necessary horizontally and vertically to make sure that the brightest pixels are in the top left hand corner. This makes it more likely that bright spots in similar galaxies line up. This is achieved by comparing the top/bottom, left/right sides of each cut-out respectively and flipping accordingly to move the brightest quadrant to the top left of each cut-out.

(iv) Normalization – The scale of flux densities in individual images is removed by normalizing each cut-out to the range [0, 1].

We project the N standardized cut-outs by flattening each of them into row vectors and then stacking the vectors into a $N \times n^2$ matrix. We use principal component analysis (PCA) to decompose the matrix into orthogonal basis vectors. We test that PCA is not degenerate by repeatedly leaving out 20 per cent of galaxies at random and making sure the fitting does not change substantially as described in Uzeirbegovic et al. (2020). We decide to retain $k = 40$ basis vectors by trying each value of k in turn and picking the point at which additional dimensions stop making a difference to how the dissimilarities are distributed.

A dissimilarity of zero implies that two galaxies are the same, and otherwise the closer to zero the more similar galaxies are. That is enough for our analysis as we are only interested in the rank of a galaxy relative to another (detailed in Section 3), and not the absolute distances. However, for the astronomer in search of a more intuitive interpretation of dissimilarities it should be noted that $Cor(X, Y) \propto 1 - d^2(\bar{X}, \bar{Y})$ where X, Y are row vectors, \bar{X}, \bar{Y} are z -scaled versions, and d^2 is the squared Euclidean distance function. That is, a dissimilarity may be interpreted as the correlation between the pixels in two images.

2.3 The SED dissimilarity matrix

We use rest-frame magnitudes in the *UBVRIJK* bands as a proxy for the SED of any given galaxy. We shall refer to this proxy as the SED for brevity. One objection may be that a 7D proxy of the SED underestimates the variance available in a fuller sampling of the spectrum. To investigate, we conduct a principal component analysis of the *UBVRIJK* bands. The first component accounts for over 96 per cent of explained variance, increasing to 99.9 per cent when a second component is added. The extreme redundancy suggests there would be limited benefit to a finer sampling of the spectrum between bands *U* and *K*. However, if the spectrum beyond *U* – *K* is discontinuous with *U* – *K*, then it could affect the results.

The SEDs are treated as 7D vectors and the SED dissimilarity matrix is calculated as the pairwise squared Euclidean distances between all SEDs. We call this the *SED space*. Note further that since colour is just the magnitude in two bands subtracted from each other (e.g. *U* – *R*), nearby SEDs imply similar bands which in turn imply similar colours on average. It may thus help readers to think about SED distance as an indicator of how closely matched two galaxies are across all the colours implied by the rest-frame bands in use.

3 ANALYSIS

In this section, we show how SED and morphological dissimilarity matrices can be used together to produce insights regarding the relationship between the SED and morphology. We note that all the graphs featuring density plots are produced by resampling the original series 1000 times, calculating 1000 histograms and reporting back the 16th–84th percentiles as an area to give the reader a clearer sense of sample uncertainty. The densities per histogram are just the underlying frequencies, normalized such that the integral over all the bins equals 1.

3.1 SED constraints on morphology

To determine whether a galaxy’s SED constrains its morphology, it helps to consider the extreme cases wherein (1) the SED completely determines the morphology and (2) the SED is completely independent of morphology. Suppose we draw a galaxy at random and then find its nearest neighbour in SED space (that is, another galaxy which has an SED most similar to its own). In case 1, we may expect the two galaxies to look very similar (that is, to have a small morphological dissimilarity distance), while in case 2 we may expect that the morphological dissimilarity distance could be anything at all.

We generalize this intuition as follows. For every galaxy *g*, we find its nearest neighbour in SED space, \bar{g} . We then order all the galaxies in morphology space by distance to *g* and calculate the fraction *x* that are closer to *g* than \bar{g} is. For example, if for some galaxy *g*, the nearest neighbour in SED space \bar{g} has *x* = 0.1, it means that 10 per cent of other galaxies are more visually similar to *g* than \bar{g} is. We call *x* the *morphology space fraction*. It should be noted that all comparisons and nearest neighbour searches are restricted to be within 0.15 of the photo-*z* of *g* to keep comparisons within similar visual conditions. In case 1 we should expect *x* = 0. In case 2 – since \bar{g} might end up ranked anywhere in the ordered set – we should expect *x* = 0.5 on average. We calculate *x* as above, for every galaxy. Fig. 1 shows a density plot of the resultant distribution at various mass cuts. All distributions are heavily skewed towards smaller fractions, peaking at *x* < 0.1. The skew can be seen to increase at the highest mass

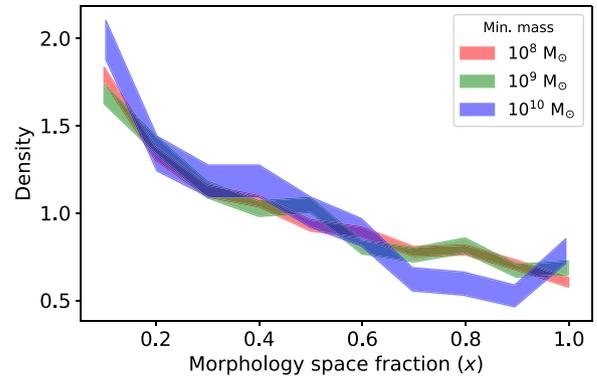


Figure 1. A density plot of the distribution of the morphology space fraction for various stellar mass cuts. The areas represent the density between the 16th and 84th percentiles. All distributions are heavily skewed towards smaller fractions, peaking at *x* < 0.1. The skew can be seen to increase at the highest mass cut, which may be because larger galaxies are better resolved and therefore less ambiguous in morphology space. The overall median *x* value is 0.37, and the smallest sample size is 1222. The results are significant at a *p*-value approaching zero.

cut, which may be because larger galaxies are better resolved and therefore less ambiguous in morphology space. The overall median *x* value is 0.37 and the smallest sample size is 1222. For any given mass cut, it is straightforward to show using the binomial distribution with $p = \frac{1}{m}$, where *m* is sample size, that the probability of these results occurring by chance is approximately zero. Thus, SED similarity between galaxies implies a significantly greater likelihood of similar morphology.

Another useful way to view the result is in terms of the probability that galaxies which are most similar to each other via the SED are also most similar to each other via visual morphology. Let ϵ be a morphology space fraction threshold below which a galaxy is considered suitably similar. For example, $\epsilon \leq 0.01$ implies that there are 1 per cent or fewer galaxies more similar to *g* than \bar{g} is. Under the null hypothesis that the SED and morphology are independent, the probability that \bar{g} (the most SED similar galaxy to *g*) is visually similar enough, is ϵ . In our data, at $\epsilon = 0.005$, the empirical probability is approximately 3ϵ : an improvement by a factor of 3.

In the next section, we consider the group of galaxies at *x* < 0.1 wherein the SED is particularly constraining.

3.2 The *x* < 0.1 group in morphology space

In this section, we consider in more detail the galaxies for which morphology appears particularly well constrained by the SED. Fig. 2 shows group comparisons of redshift, mass, specific star formation rate (SSFR), and *U* – *R* colour. The *x* < 0.1 group is overrepresented in $0.25 < z < 1$ and at higher masses but present throughout the redshift and mass range of the population, which indicates that this is not an artificial result brought about by the fact that more nearby objects are better resolved and therefore easier to differentiate. This group has a lower SSFR and a significantly different rest-frame *U* – *R* colour distribution that is bimodal, compared with the unimodal *x* ≥ 0.1 distribution. There are 1390 galaxies in the *x* < 0.1 group, which comprises around 18 per cent of the data set. Thus, the galaxies for which morphology has the greatest link with the SED are typically redder and more massive compared with more visually dissimilar objects.

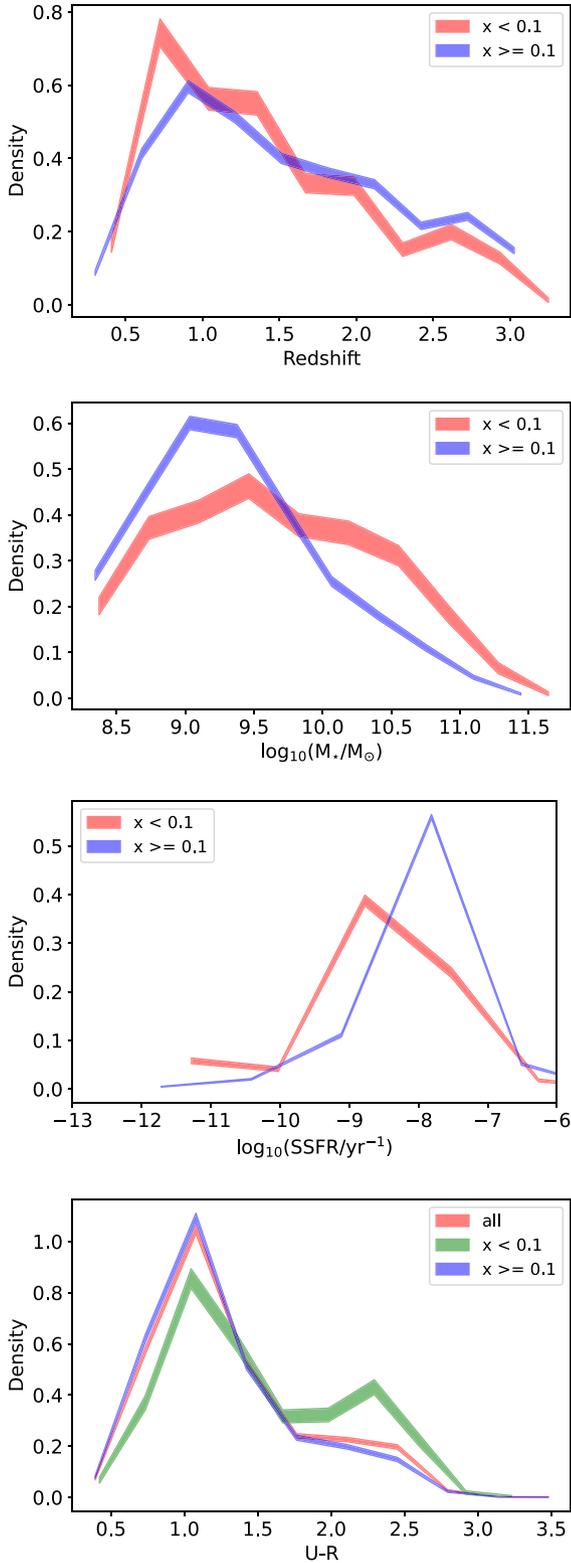


Figure 2. Comparisons of redshift, stellar mass, SSFR, and rest-frame $U - R$ colour for different groups. The areas represent the density between the 16th and 84th percentile. The $x < 0.1$ group is overrepresented in the redshift range $0.25 < z < 1$ and at higher stellar masses but present throughout the redshift and mass range of the population. It has a lower SSFR and a significantly different $U - R$ colour distribution wherein bimodality with peaks at 1 and 2.1 is clearly present for the $x < 0.1$ group, in contrast to the unimodal distribution of the $x \geq 0.1$ group.

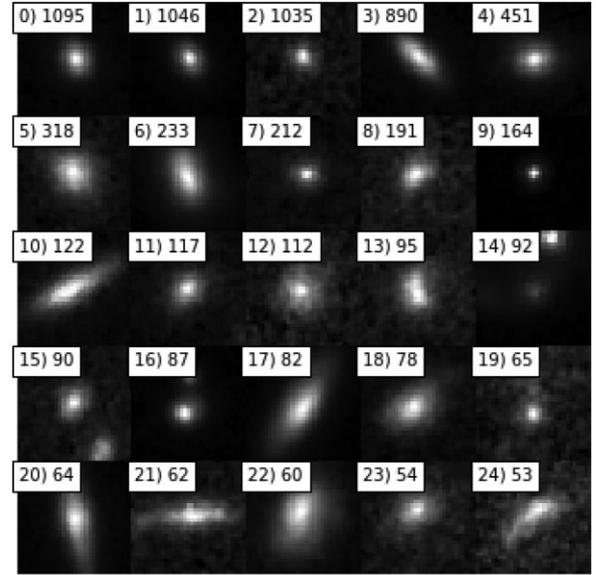


Figure 3. The exemplars (prototypical galaxies) of the top 25 partitions ordered by size from left to right, top to bottom. The labels indicate the partition index and the number of galaxies in the partition. All partitions contain more than 50 members and together make up more than 90 per cent of the data set.

We next consider how this group differs from the rest of the galaxies in terms of morphology. In order to investigate possible patterns in morphology amongst the $x < 0.1$ group, we empirically generate a set of morphological classes and then cross tabulate these against $x < 0.1$ group membership to see whether in-group galaxies are over/underrepresented anywhere.

We first explain how the morphological classes are generated and then how they are used in our analysis. We use the morphology space to naturally partition galaxies around a set of exemplars (i.e. prototypical galaxies) such that galaxies in the same partition are all close to the same exemplar. Let $\{d_{i,j}\}$ be our $N \times N$ dissimilarity matrix in morphology space, the objective then is to choose a set of exemplars such that the sum of the distances between each galaxy and its closest exemplar is minimized. Formally, for some set of galaxy vectors $s \in S$:

$$\min_{\{q_i\}_{i=1}^m} \left(\sum_{s \in S} \min_i d_{s,q_i} + \sum d_{q_i,q_i} \right), \quad (1)$$

where $q_1, \dots, q_m \in S$ are $m = |S|$ exemplars. The diagonal d_{q_i,q_i} is set to the median distance. The optimization above is usually intractable but the affinity propagation (AP; Frey & Dueck 2007, see Appendix A5 for a short overview) algorithm offers a good approximation. A crucial advantage with AP is that the number of partitions are automatically determined. We run AP on our data set that returns 68 partitions, of which 25 have at least 50 galaxies and subsume more than 90 per cent of the data set. Fig. 3 shows the top 25 partitions along with their indices and the number of galaxies in them. Full panels containing all the galaxies in each partition are available as part of the supplementary material.

We next produce a cross tabulation of all the galaxies, counting each into a partition and either in or out of group $x < 0.1$. We then conduct a contingency table analysis, producing Pearson’s residuals for each cell. A Pearson residual can be interpreted as the difference (measured in standard deviations) between the observed and expected counts in each cell, on the assumption (for the expected counts) that

Table 1. The table shows the top 25 partitions by size along with the Pearson’s residuals for the $x \geq 0.1$ and $x < 0.1$ groups where x is the morphology space fraction. Additionally, for the $x < 0.1$ group, the table shows the sample size, mean redshift, mean stellar mass, and the number of galaxies with a $U - R > 2$. Its noteworthy that the $x < 0.1$ group is significantly over/underrepresented in several partitions, and that the redder peak in the colour distribution given in Fig. 2 is mostly accounted for by the overrepresentation in Partition 1.

Partition	$x \geq 0.1$	$x < 0.1$	#	z	Mass	$(U - R) > 2$
0	0.86	-1.83	171	1.02	1.64e+09	40
1	-1.3	2.78	226	1.16	7.58e+09	103
2	0.72	-1.53	165	1.41	1.04e+09	35
3	0.73	-1.57	140	0.93	1.70e+09	9
4	0.52	-1.11	71	1.03	6.64e+09	21
5	-0.06	0.12	58	0.86	5.07e+09	3
6	-0.37	0.8	47	0.93	4.68e+09	5
7	0.69	-1.47	29	1.47	1.33e+09	7
8	0.74	-1.59	25	1.57	9.74e+08	1
9	-2.55	5.45	59	0.5	1.28e+09	31
10	-2.11	4.51	43	0.87	6.82e+09	4
11	0.21	-0.44	19	1.41	1.22e+09	1
12	-0.2	0.42	22	1.09	4.44e+09	2
13	0.35	-0.74	14	1.51	1.33e+09	0
14	-0.29	0.61	19	2.03	3.89e+08	2
15	0.02	-0.04	16	0.95	4.17e+08	0
16	0.9	-1.93	8	1.97	4.46e+09	1
17	-0.52	1.11	19	0.74	6.64e+09	2
18	0	-0	14	1.03	2.67e+09	1
19	0.91	-1.95	5	1.92	3.13e+09	0
20	-1.31	2.8	21	0.63	6.96e+09	4
21	0.16	-0.34	10	1.52	2.12e+09	0
22	-0.32	0.68	13	0.68	3.7e+09	0
23	-0.05	0.1	10	1.2	1.40e+10	0
24	0.99	-2.11	3	1.02	1.09e+09	1

there is no dependence between the partitions and the $x < 0.1$, $x \geq 0.1$ groups. For example, a residual of 2 or more could be interpreted as a likely overrepresentation of galaxies in that cell, and a residual of -2 or less could be interpreted as an under representation.

Table 1 shows the top 25 partitions by size along with the Pearson’s residuals for the $x \geq 0.1$ and $x < 0.1$ groups. Additionally, for the $x < 0.1$ group, the table shows the sample size, mean redshift, mean stellar mass, and the number of galaxies with rest-frame $U - R > 2$. Its noteworthy that the $x < 0.1$ group is significantly over/underrepresented in several partitions, and that the redder peak in the colour distribution from Fig. 2 is mostly accounted for by the overrepresentation in Partition 1. Thus, SED constrains morphology most strongly for galaxies that are typically more massive, redder, and morphologically elliptical.

3.3 Morphology constraints on SED

We now consider how SEDs may constrain morphology. Just as before, but in the opposite direction, for every galaxy g , we find its nearest neighbour in morphology space \bar{g} . We then order all the galaxies in SED space by distance to g and calculate the fraction x that are closer to g than \bar{g} is. We call this the *SED space fraction*. It should be noted that all comparisons and nearest neighbour searches are restricted to be within 0.15 of the photo- z for g to keep comparisons within similar visual conditions. Fig. 4 shows the distribution of the SED fraction at various mass cuts. Unlike Fig. 1, as the minimum mass increases, the left skew decreases, likely because more massive galaxies have less diverse SEDs therefore increasing ambiguity.

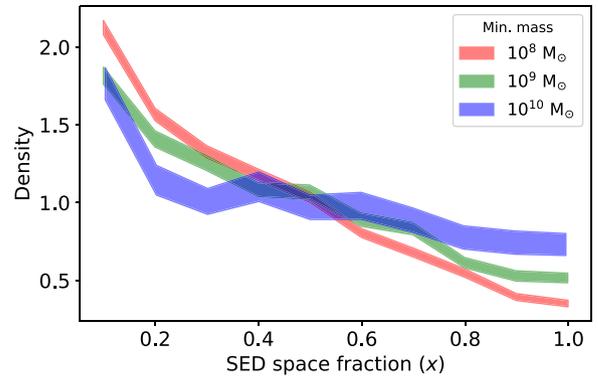


Figure 4. The SED space fraction density distribution at a selection of mass cuts. The areas represent the density between the 16th and 84th percentile. Unlike Fig. 1, as the mass minimum increases, the left skew decreases, likely because more massive galaxies have less diverse SEDs therefore increasing ambiguity. Overall the median x value is 0.30, and the smallest sample size is 1222.

It should be noted that, as with the morphology space fraction distribution, the SED space fraction distribution is most dense at $x < 0.1$. Overall the median x value is 0.30, and the smallest sample size is 1222.

As before, another useful way to view the result is in terms of the probability that galaxies which are most similar to each other via morphology are also most similar to each other via SED. Let ϵ be an SED space fraction threshold below which a galaxy is considered suitably similar. Under the null hypothesis that the SED and morphology are independent, the probability that \bar{g} (the most morphologically similar galaxy to g) is SED similar enough, is ϵ . In our data, at $\epsilon = 0.005$, the empirical probability is approximately 2.75ϵ : an improvement by a factor of 2.75.

Fig. 5 shows the overall morphology fraction distribution and SED fraction distribution. Both density series show a definitive skew towards the smaller fractions. However, it is noteworthy that morphology constrains the SED significantly more than vice versa on average (0.30 versus 0.37 median fractions). The bottom plot shows the multiple of the gain in probability compared to the null hypothesis [i.e. $(\alpha\epsilon)$ of \bar{g} being within the ϵ threshold]. It shows that even though the SED constraint results in a higher peak probability, morphological constraints are more limiting over all but the very beginning of the ϵ range.

3.4 A mutually constrained group

In this section, we investigate whether the galaxies that fall into the $x < 0.1$ group in both the morphology and SED spaces are different from the rest. Fig. 6 shows the redshift, stellar mass, and colour distributions for the $x < 0.1$ group in morphology space, SED space, and for galaxies that fall into both groups. Its noteworthy that the ‘Both’ subset is more massive and bluer, than the other two. This suggests that the ‘Both’ group is not just a random sample from the other two groups. Interestingly, the overlap between the two $x < 0.1$ groups (that is, the ‘Both’ group) is only about 24 per cent, underlining that a morphology heavily constrained by the SED does not imply an SED heavily constrained by morphology, and vice versa.

We produce a contingency table for the mutually constrained group (MCG) but it does not reveal significant over indexing and suggests instead a heterogeneous mix of galaxies. We instead search for common visual features by training a classifier to distinguish

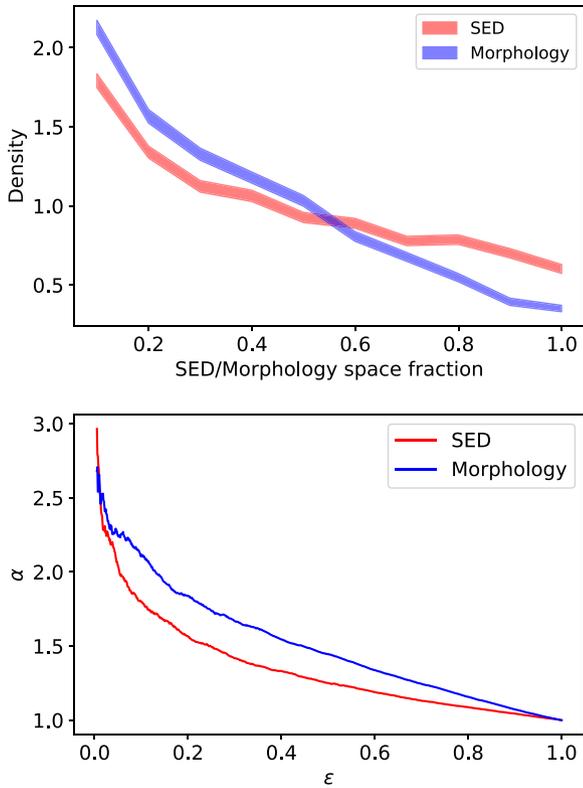


Figure 5. Density plot (top) of morphology fraction distribution and SED fraction distribution. The areas represent the density between the 16th and 84th percentile. Both series show a definitive skew towards smaller fractions. However, it is noteworthy that on average morphology constrains the SED significantly more than vice versa: the median fractions are 0.30 for morphology and 0.37 for SED. The bottom plot shows the relative gain in probability ($\alpha\epsilon$) of \bar{g} being within the ϵ threshold. It shows that even though the SED constraint results in a higher peak probability, morphology constraints are more limiting over all but the very beginning of the threshold range.

between members of the MCG and the rest of the population. A novelty here is that we use the distances to the partition exemplars as features thus allowing feature importances to be interpreted in terms of similarity of MCG members to particular exemplars. We use a ‘ridge’ classifier (Hoerl & Kennard 1970, see Appendix A1 for a short overview) with 100-fold cross-validation (see Appendix A4 for a short overview) to confirm an average classification accuracy¹ of about 64 per cent, which suggests that some visual features are more abundant in the MCG than the general population.

To further investigate which exemplars most affect the model, we randomly shuffle each feature in turn and measure how much the intervention affects coverage (the fraction of true positives correctly identified). This procedure is known as permutation feature importance (Breiman 1996, see Appendix A2 for a short overview). Bigger changes in coverage indicate that the feature is more important for picking out MCG members. Fig. 7 shows the partition exemplars ordered by effect on coverage. Based on these results, there are several features that are particularly important in the MCG. Objects with the greatest effect typically fall into partitions containing fewer

¹Since the MCG is small compared to the population, balanced classification accuracy is used, which takes into account the class frequencies to produce an accuracy figure centred on 50 per cent.

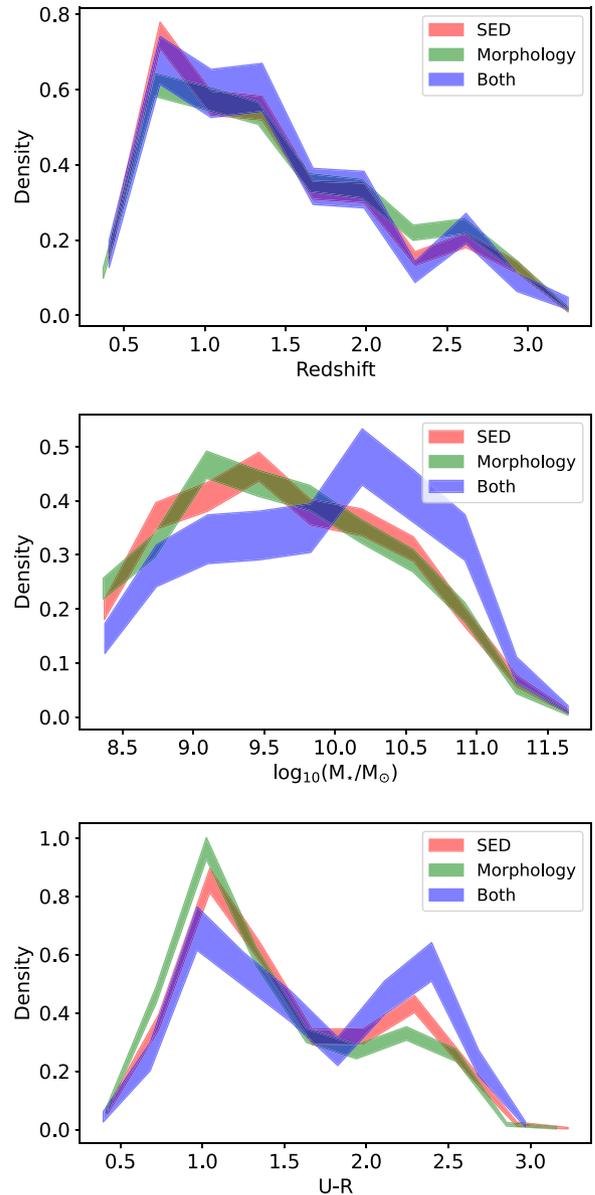


Figure 6. The redshift, mass, and rest-frame $U - R$ colour of galaxies in the $x < 0.1$ group by SED and morphology, and in both. The areas represent the density between the 16th and 84th percentile. It is noteworthy that the ‘Both’ subset contains galaxies that are more massive and bluer than the other two.

than 50 galaxies and appear more morphologically disturbed or clumpy compared with later exemplars. A plausible explanation is that the MCG objects are affected by processes which have a simultaneous effect on both the SED and morphological properties of the galaxy, likely on short time-scales, for example mergers, interactions or clumpy star formation. This is particularly clear to see when comparing the exemplars found in Fig. 2 to those found in Fig. 7. We leave further analysis, classifier improvement and better description of the MCG for future work.

3.5 Implications for morphology selection by colour

The classification of morphology using colour works by establishing constraints of the form $a < y < b$ for one or more colours that best

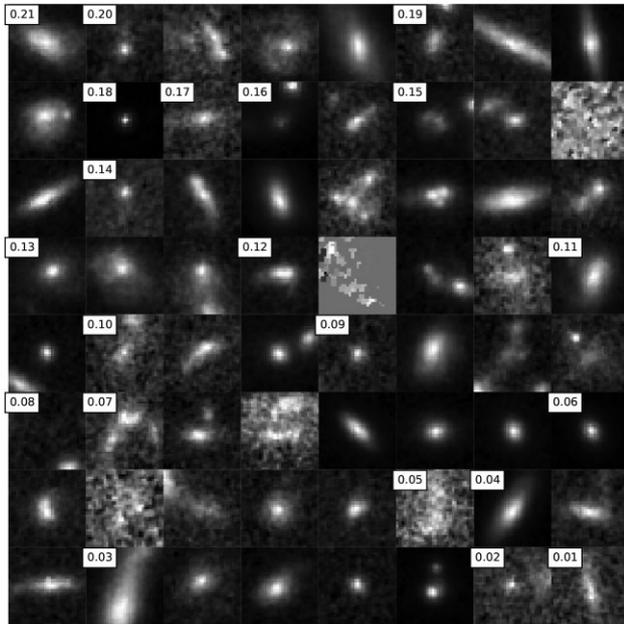


Figure 7. The partition exemplars ordered by effect size on a ridge classifier. The thumbnails are laid out left to right, top to bottom. The numbers indicate the difference in coverage when the feature associated to the exemplar is randomly permuted. Labels are omitted where they would be the same as the last label in the sequence. Objects with the greatest effect size typically fall into partitions containing fewer than 50 galaxies and appear more morphologically disturbed or clumpy compared with later exemplars.

separate the target population from the rest. For example, Strateva et al. (2001) use rest-frame $U - R \geq 2.22$ to separate early-type galaxies from the rest with a claim of 80 per cent coverage and 62 per cent purity (the overall fraction correctly classified). The process of classifying by colour can be iterative, with each subsequent threshold attempting to strike a balance between purity (the rate of true positives) and coverage. This bears direct analogy to the operation of a decision tree classifier (Breiman et al. 2017, see Appendix A3 for a short overview) which on each iteration – in the binary case – picks the best of its features according to some coverage/purity criterion to make a threshold cut, repeating the process at every partition, until there are no more gains to be made or some limiting condition – such as maximum depth – has been reached. We can use the parallel between selection by colour as practiced by astronomers and the operations of a decision tree classifier, together with the empirical partitions and SED space derived above, to investigate whether we can establish an upper bound to the efficacy of methods that use colour selection to classify morphology.

For each empirical partition, we train a decision tree classifier with a maximum depth of 6 using all the possible colours in the $UBVRIJK$ bands as features: this equates to looking for an optimal set of thresholds on up to six colours. We repeat the fitting for every redshift window with a width of 0.5, in which the partition has at least 50 members. That is, for every partition, we try to find a redshift slice in which some set of up to six colour constraints best separate galaxies belonging to that partition from the rest. We use 10-fold cross-validation to establish the key metrics. Table 2 shows optimal redshift, sample size, coverage, and purity for all partitions big enough to test. With the exception of Partition 9, the purity of classifications is around the 60–70 per cent level. Several partitions have coverage of at or above 75 per cent but it is 60 per cent or below for most partitions. The 11 partitions tested span more than 75 per

Table 2. The best redshift range, sample size, coverage, and purity achieved by a decision tree classifier based on colour features, for each partition. Purity is notably below 64 per cent for most partitions implying an upper bound limit to colour-based methods.

Partition	Best redshift range	#	Coverage	Purity
0	0.94–1.44	278	0.6	0.61
1	0.53–1.03	336	0.4	0.64
2	1.91–2.41	190	0.47	0.62
3	0.78–1.28	239	0.67	0.59
4	0.85–1.35	133	0.55	0.62
5	0.05–0.55	73	0.62	0.61
6	0.80–1.30	55	0.71	0.67
7	2.24–2.74	50	0.34	0.56
8	0.59–1.09	50	0.76	0.62
9	0.23–0.73	96	0.91	0.92
10	0.27–0.77	54	0.57	0.6

Table 3. The best redshift range, sample size, coverage, and purity achieved by a ridge classifier on rest-frame band features, for each partition. Both coverage and purity are improved for almost all partitions when compared to the decision tree classifier on colour features.

Partition	Best redshift range	#	Coverage	Purity
0	0.00–0.50	60	0.62	0.61
1	0.58–1.08	327	0.54	0.64
2	0.16–0.66	73	0.82	0.73
3	0.05–0.55	109	0.7	0.62
4	0.87–1.37	134	0.68	0.68
5	0.79–1.29	92	0.74	0.71
6	0.71–1.21	63	0.76	0.73
7	1.89–2.39	50	0.46	0.53
8	0.57–1.07	50	0.8	0.68
9	0.03–0.53	78	0.98	0.9
10	0.33–0.83	50	0.66	0.65

cent of the data so it appears that morphological classification using colour results in generally poor efficacy as measured by coverage and purity.

Efficacy can be improved by abandoning colours in favour of any linear combination of rest-frame bands. We fit a Ridge classifier using the SEDs as features and repeat the fitting for every redshift window with a width of 0.5 in which the partition has at least 50 members. That is, for every partition, we try to find a redshift slice in which some combination of bands best separates galaxies belonging to that partition from the rest. As before, we use 10-fold cross-validation to establish the key metrics. Table 3 shows the optimal redshift range, sample size, coverage, and purity for all partitions tested. Its noteworthy that almost all coverage and purity figures are improved by using all the bands, tipping purity for many partitions past the 70 per cent level.

Since the empirical partitions are homogeneous morphologies which together span the whole data set, one would expect that coarser groupings (e.g. early versus late types) could be produced by pooling partitions together into fewer morphological classes. However, since the efficacy – as measured by coverage and purity – is relatively low for both methods, it is unlikely that pooling would result in better morphological classification, and could make it worse by decreasing in-class homogeneity. It should be further noted that the numbers for coverage and purity presented here are effectively *upper bounds*, since we report only the results for the optimal redshift

slice. Hence, significantly worse efficacy could be expected in other redshift ranges. In general, therefore, pure morphological classes cannot be selected very effectively using colours.

4 SUMMARY

In this work, we have introduced an empirical methodology for the analysis of how galaxy SEDs and morphologies constrain each other, using the vector space building techniques for galaxy surveys laid out in Uzeirbegovic et al. (2020). Our main results can be summarized as follows:

(i) Two galaxies with very similar SEDs are around three times more likely to also be most morphologically similar, compared to the null hypothesis that SED and morphology are independent. Massive red ellipticals are especially likely to be well-identified their SEDs.

(ii) Two morphologically similar galaxies are slightly under three times more likely to be most SED similar, compared to the null hypothesis that SED and morphology are independent. However, morphology constrains the SED more strongly than vice versa on average (with around a 7 per cent improvement in the space fraction).

(iii) Disturbed or interacting systems – systems that are experiencing processes like mergers that affect both the SED and the morphology simultaneously – are prominent amongst galaxies for which the SED and morphology are mutually constraining (i.e. in which the SED implies a relatively similar morphology and vice versa).

(iv) No combination of colour cuts is able to strongly constrain galaxy morphology. On average, purity is around 64 per cent, if up to three colours are used to try and select homogeneous morphological classes. While the results can be improved by considering linear combinations of the whole SED, the improvements are not significant and purity levels for most morphological classes remain at the 70 per cent or lower.

ACKNOWLEDGEMENTS

SK acknowledges support from the STFC (ST/S00615X/1) and a Senior Research Fellowship from Worcester College Oxford. This work is based on observations taken by the CANDELS Multi-Cycle Treasury Program with the NASA/ESA *HST*, which is operated by the Association of Universities for Research in Astronomy, Inc., under NASA contract NAS5-26555. This research made use of ASTROPY,² a community-developed core Python package for astronomy (Astropy Collaboration 2013; Price-Whelan et al. 2018).

DATA AVAILABILITY

Code for all results and figures can be found at <https://emiruz.com/similarity>.

REFERENCES

Abraham R. G., Valdes F., Yee H. K. C., van den Bergh S., 1994, *ApJ*, 432, 75
 Adelman-McCarthy J. K. et al., 2007, *ApJS*, 172, 634
 Astropy Collaboration, 2013, *A&A*, 558, A33
 Bluck A. F. L., Mendel J. T., Ellison S. L., Moreno J., Simard L., Patton D. R., Starkeburg E., 2014, *MNRAS*, 441, 599
 Breiman L., 1996, *Mach. Learn.*, 24, 123

²<http://www.astropy.org>

Breiman L., Friedman J. H., Olshen R. A., Stone C. J., 2017, *Classification and Regression Trees*. Chapman and Hall/CRC, United Kingdom
 Bundy K., Ellis R. S., Conselice C. J., 2005, *ApJ*, 625, 621
 Cheng T.-Y. et al., 2020, *MNRAS*, 493, 4209
 Connolly A. J., Szalay A. S., Bershady M. A., Kinney A. L., Calzetti D., 1995, *AJ*, 110, 1071
 Conroy C., 2013, *Annu. Rev. Astron. Astrophys.*, 51, 393
 Conselice C. J., 2003, *ApJS*, 147, 1
 de Vaucouleurs G., 1948, *Ann. Astrophys.*, 11, 247
 Dressler A. et al., 1997, *ApJ*, 490, 577
 Frey B. J., Dueck D., 2007, *Science*, 315, 972
 Goulding A. D. et al., 2018, *Publ. Astron. Soc. Japan*, 70, S37
 Grogan N. A. et al., 2011, *ApJS*, 197, 35
 Guo Y. et al., 2013, *ApJS*, 207, 24
 Hocking A., Geach J. E., Sun Y., Davey N., 2018, *MNRAS*, 473, 1108
 Hoerl A. E., Kennard R. W., 1970, *Technometrics*, 12, 55
 Hubble E. P., 1926, *ApJ*, 64, 321
 Hubble E. P., 1936, *Realm of the Nebulae*. Yale University Press
 Huertas-Company M. et al., 2015, *ApJS*, 221, 8
 Jackson R. A., Martin G., Kaviraj S., Laigle C., Devriendt J. E. G., Dubois Y., Pichon C., 2020, *MNRAS*, 494, 5568
 Jackson R. A. et al., 2021, *MNRAS*, 502, 4262
 Jones M. H., Lambourne R. J. A., 2004, *An Introduction to Galaxies and Cosmology*. Open University, Cambridge
 Kaviraj S., 2014, *MNRAS*, 440, 2944
 Kaviraj S., Martin G., Silk J., 2019, *MNRAS*, 489, L12
 Koekemoer A. M. et al., 2011, *ApJS*, 197, 36
 Lackner C. N., Gunn J. E., 2012, *MNRAS*, 421, 2277
 Lintott C. et al., 2011, *MNRAS*, 410, 166
 Lofthouse E. K., Kaviraj S., Conselice C. J., Mortlock A., Hartley W., 2017, *MNRAS*, 465, 2895
 Lotz J. M., Primack J., Madau P., 2004, *AJ*, 128, 163
 Madgwick D. S. et al., 2003, *ApJ*, 599, 997
 Mager V. A., Conselice C. J., Seibert M., Gusbar C., Katona A. P., Villari J. M., Madore B. F., Windhorst R. A., 2018, *ApJ*, 864, 123
 Magris G., Mateu P. J., Mateu C., Bruzual A. G., Cabrera-Ziri I., Mejía-Narváez A., 2015, *PASP*, 127, 16
 Martin G., Kaviraj S., Devriendt J. E. G., Dubois Y., Pichon C., 2018, *MNRAS*, 480, 2266
 Martin G. et al., 2019, *MNRAS*, 485, 796
 Martin G., Kaviraj S., Hocking A., Read S. C., Geach J. E., 2020, *MNRAS*, 491, 1408
 Masters K. L. et al., 2010, *MNRAS*, 405, 783
 Masters K. L. et al., 2019, *MNRAS*, 487, 1808
 Menanteau F., Ford H. C., Motta V., Benítez N., Martel A. R., Blakeslee J. P., Infante L., 2006, *AJ*, 131, 208
 Morgan W. W., Mayall N. U., 1957, *PASP*, 69, 291
 Odewahn S. C., Cohen S. H., Windhorst R. A., Philip N. S., 2002, *ApJ*, 568, 539
 Ostrowski F. et al., 2017, *MNRAS*, 465, 4325
 Pearl J., 1982, in *Proceedings of the Second National Conference on Artificial Intelligence*. AAAI Press, Menlo Park, California, p. 133
 Peth M. A. et al., 2016, *MNRAS*, 458, 963
 Postman M. et al., 2005, *ApJ*, 623, 721
 Price-Whelan A. M. et al., 2018, *AJ*, 156, 123
 Ryan R. E. J. et al., 2012, *ApJ*, 749, 53
 Santini P. et al., 2015, *ApJ*, 801, 97
 Scarlata C. et al., 2007, *ApJS*, 172, 406
 Schawinski K., Zhang C., Zhang H., Fowler L., Santhanam G. K., 2017, *MNRAS*, 467, L110
 Sérsic J. L., 1963, *Boletín de la Asociación Argentina de Astronomía La Plata Argentina*, 6, 41
 Simard L. et al., 2002, *ApJS*, 142, 1
 Simmons B. D. et al., 2017, *MNRAS*, 464, 4420
 Skibba R. A. et al., 2009, *MNRAS*, 399, 966
 Smethurst R. J. et al., 2015, *MNRAS*, 450, 435
 Sodr e L. J., Cuevas H., 1994, *Vistas Astron.*, 38, 287
 Strateva I. et al., 2001, *AJ*, 122, 1861

Uzeirbegovic E., Geach J. E., Kaviraj S., 2020, *MNRAS*, 498, 4021

SUPPORTING INFORMATION

Supplementary data are available at *MNRAS* online.

supplementary.zip

Please note: Oxford University Press is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

APPENDIX A

A1 Ridge classifier

A ridge regression (Hoerl & Kennard 1970) is an alternative to ordinary least-squares regression for fitting multiple regression models which is particularly suitable when the independent variables are highly correlated. The regression is made into a binary classifier by coding the dependent variable as either 1 (true) or -1 (false) and then conducting a regression. Ridge regression is well suited to the task at hand because (1) it out-performs more complex algorithms that we have also tried, and (2) we are dealing with a large number of correlated independent variables.

A2 Permutation feature importance

The permutation feature importance (Breiman 1996) is the decrease in a metric used to score a model (coverage in our case) when a single independent variable is randomly permuted. The permutation breaks the relationship between the independent and the dependent variables, allowing the drop in the scoring metric to be used as an indicator of the extent to which the model depends on the feature. Since permutations are random, the processes must be repeated many times per independent variable to calculate an average effect.

A3 Decision trees

Classification and regression trees (Breiman et al. 2017) are an umbrella term for various ways to construct decision trees for the purposes of classification and regression. We utilize a relatively basic procedure for our binary classification case, wherein independent variables are evaluated one at a time, and the one which maximizes a metric such as information gain (difference in information entropy) is selected as a splitting criteria. For each split the procedure is recursively repeated until some stopping criteria – such as maximum tree depth or too small a gain – is reached.

A4 K -fold cross validation

K -fold cross validation refers to the splitting of the data into K pieces wherein each piece is used for testing in turn, while the $K - 1$ other pieces are used for training. The individual evaluations may be combined together to produce an overall result.

A5 Affinity propagation

Affinity propagation (Frey & Dueck 2007) is a clustering algorithm that partitions data around a set of ‘exemplars’ by solving the optimization problem presented in equation (1). The problem is known to be intractable, but affinity propagation uses a way of reformulating the problem known as ‘message passing’, such that a solution can be approximated by iteration. More particularly, affinity propagation makes extensive use of the sum-product rule (Pearl 1982) to minimize the number of computations required on each iteration in order to make the algorithm tractable. Key benefits of affinity propagation for our problem are that (1) it is exemplar based, and (2) that it discovers the number of exemplars.

This paper has been typeset from a $\text{\TeX}/\text{\LaTeX}$ file prepared by the author.