

## Research Article

# Lyman- $\alpha$ at cosmic noon I: Ly $\alpha$ spectral type selection of $z \sim 2 - 3$ Lyman break galaxies with broadband imaging

Garry Foran<sup>1,2</sup> , Jeff Cooke<sup>1,2</sup>, Naveen Reddy<sup>3</sup>, Charles Steidel<sup>1,4</sup> and Alice Shapley<sup>5</sup>

<sup>1</sup>Centre for Astrophysics and Supercomputing, Swinburne University of Technology, Hawthorn, VIC, Australia, <sup>2</sup>Australian Research Council Centre of Excellence for All-sky Astrophysics in 3 Dimensions (ASTRO-3D), <sup>3</sup>Department of Physics & Astronomy, University of California, Riverside, CA, USA, <sup>4</sup>Cahill Center for Astronomy and Astrophysics, California Institute of Technology, Pasadena, CA, USA and <sup>5</sup>Department of Physics & Astronomy, University of California, Los Angeles, CA, USA

### Abstract

High-redshift Lyman break galaxies (LBGs) are efficiently selected in deep images using as few as three broadband filters, and have been shown to have multiple intrinsic and small- to large-scale environmental properties related to Lyman- $\alpha$ . In this paper we demonstrate a statistical relationship between net Lyman- $\alpha$  equivalent width (net Ly $\alpha$  EW) and the optical broadband photometric properties of LBGs at  $z \sim 2$ . We show that LBGs with the strongest net Ly $\alpha$  EW in absorption (aLBGs) and strongest net Ly $\alpha$  EW in emission (eLBGs) separate into overlapping but discrete distributions in  $(U_n - \mathcal{R})$  colour and  $\mathcal{R}$ -band magnitude space, and use this segregation behaviour to determine photometric selection criteria by which sub-samples with a desired Ly $\alpha$  spectral type can be selected using data from as few as three broadband optical filters. We propose application of our result to current and future large-area and all-sky photometric surveys that will select hundreds of millions of LBGs across many hundreds to thousands of Mpc, and for which spectroscopic follow-up to obtain Ly $\alpha$  spectral information is prohibitive. To this end, we use spectrophotometry of composite spectra derived from a sample of 798 LBGs divided into quartiles on the basis of net Ly $\alpha$  EW to calculate selection criteria for the isolation of Ly $\alpha$ -absorbing and Ly $\alpha$ -emitting populations of  $z \sim 3$  LBGs using *ugri* broadband photometric data from the Vera Rubin Observatory Legacy Survey of Space and Time (LSST).

**Keywords:** Galaxies: fundamental parameters; galaxies: photometry; galaxies: high-redshift

(Received 26 June 2022; revised 28 June 2023; accepted 8 September 2023)

### 1. Introduction

One of the most important and well-studied populations of early star-forming galaxies (SFGs) are the so-called Lyman break galaxies (LBGs) that can be selected based on their rest-frame ultraviolet (UV) colours using as few as three broadband optical filters. The generic Lyman break selection method uses broadband optical photometry sensitive to the discontinuity ('break' or 'drop-out') in the rest-frame UV spectrum of SFGs blueward of the Lyman limit (912 Å), the decrement in flux in the Lyman- $\alpha$  forest blueward of the Lyman- $\alpha$  spectral feature (Ly $\alpha$ , 1216 Å), and the relatively flat rest-frame UV continuum redward of Ly $\alpha$  to efficiently select LBGs in large numbers, on large scales, and across a wide range of redshift pathlengths.

A notable strength of the Lyman break technique is its ability to isolate populations of LBGs at specific redshifts by sampling with different broadband filter combinations. The classic Lyman break technique has been effective at assembling large samples of LBGs in the range  $z \sim 3 - 5$  where the Lyman limit falls at optical wavelengths (e.g., Steidel et al. 2003; Ouchi et al. 2004; Giavalisco et al. 2004; Verma et al. 2007; Iwata et al. 2007; Pentericci et al. 2010; Bielby et al. 2011; Oteo et al. 2013a; Álvarez-Márquez et al. 2016; Malkan et al. 2017), and the use of space-based observatories

has extended the Lyman limit detection window as low as  $z \sim 1$  (e.g., Burgarella et al. 2006; Ly et al. 2009; Basu-Zych et al. 2011; Habertzettl et al. 2012; Oteo et al. 2013b, 2014; Hathi et al. 2013). Modified selection methods exploiting the Lyman- $\alpha$  break that dominates the rest-frame UV at redshifts  $z \gtrsim 5$  have successfully isolated large samples of LBGs at redshifts up to  $z \sim 10$  (e.g., Bouwens et al. 2006, 2010, 2015; McLure et al. 2011; Ellis et al. 2013; Finkelstein 2016; Harikane et al. 2018, 2022b), and the redshift-dependent line blanketing by the Ly $\alpha$  forest, in combination with the relatively flat rest-frame UV continuum, has been used to select LBGs in the range  $1.4 < z < 2.7$  at which redshifts the Lyman limit is not observable from the ground (Adelberger et al. 2004; Steidel et al. 2004).

This feature of the Lyman break selection technique makes it particularly important in terms of the legacy value of the current generation of deep, wide, optical, and near-infrared imaging surveys. Large-area and all-sky optical broadband photometric campaigns such as the Hyper-SuprimeCam Subaru Strategic Program (HSC-SSP: Aihara et al. 2018) and the imminent Vera Rubin Observatory Legacy Survey of Space and Time (LSST: Ivezić et al. 2019) will exploit the Lyman break technique using 3–6 broadband filters across the rest-frame UV to efficiently and inexpensively select hundreds of millions of galaxies in redshift ranges from  $z \sim 2 - 6$  across many hundreds to thousands of Mpc (e.g., Ono et al. 2018; Harikane et al. 2018, 2022b; Wilson & White 2019).

The Lyman break selection method comes with its own set of selection biases in favour of UV-bright, bluer star-forming galaxies

**Corresponding author:** G. Foran; Email: gforan@swin.edu.au

**Cite this article:** Foran G, Cooke J, Reddy N, Steidel C and Shapley A. (2023) Lyman- $\alpha$  at cosmic noon I: Ly $\alpha$  spectral type selection of  $z \sim 2 - 3$  Lyman break galaxies with broadband imaging. *Publications of the Astronomical Society of Australia* 40, e052, 1–13. <https://doi.org/10.1017/pasa.2023.48>

with relatively low dust extinction, resulting in samples that miss a relevant fraction of UV-faint and/or heavily dust-obscured SFGs and passively evolving galaxies, particularly around the peak in cosmic star formation (e.g., Grazian *et al.* 2007; Ly *et al.* 2011; Shapley 2011; Habertzettl *et al.* 2012; Oteo *et al.* 2014, 2015). Nevertheless, LBGs are thought to dominate the UV luminosity density, and possibly the global star formation rate (SFR) density, at  $z \sim 2 - 6$  (e.g., Steidel *et al.* 1999; Giavalisco *et al.* 2004; Bouwens *et al.* 2009; Reddy *et al.* 2008; Reddy & Steidel 2009), and they remain a key target population in recent surveys (e.g., Arrabal Haro *et al.* 2018; Ono *et al.* 2018; Toshikawa *et al.* 2018; Harikane *et al.* 2022a). Moreover, LBGs have been posited as critical populations that meet the demanding requirements of cosmological studies in the era of large-area and all-sky photometric surveys (e.g., Wilson & White 2019; Miyatake *et al.* 2022), especially at higher redshifts where only methods based on Ly $\alpha$  emission or Lyman break detection can be applied in large numbers and over large scales (Finkelstein 2016, and references therein).

Ly $\alpha$  has long been pursued as a potential tool to probe the properties of high-redshift SFGs. This endeavour has been motivated by the fact that Ly $\alpha$  in absorption and/or emission is the dominant feature in the rest-frame UV spectrum of such galaxies, and is typically much stronger than other diagnostic ISM absorption and emission lines. In addition, there are observational advantages that facilitate deep photometric imaging and spectroscopy in the wavelength range corresponding to Ly $\alpha$  at  $z \sim 2 - 3$  (Shapley 2011, and references therein)—a cosmologically critical epoch that spans the peak in SFR density (Madau & Dickinson 2014, and references therein) and during which more than half of the observable stellar mass of the Universe was assembled (e.g., Ilbert *et al.* 2013; Muzzin *et al.* 2013). Moreover, Ly $\alpha$  is the key – and often the only – observable feature in the spectra of Ly $\alpha$  emitters at the highest redshifts ( $z \gtrsim 6$ , Finkelstein 2016; Ouchi *et al.* 2020, and references therein) and, for this reason, has become critical for our understanding of galaxy populations during the epoch of reionisation, and their contribution to the ionising flux budget of the universe (e.g., Dijkstra 2014; Stark *et al.* 2017; Mason *et al.* 2018; Steidel *et al.* 2018).

Due to the resonant character of the Ly $\alpha$  transition, Ly $\alpha$  photons are dispersed in real and frequency space whenever they encounter neutral hydrogen (see Dijkstra 2017, for a comprehensive description). The increased scattering and absorption experienced by Ly $\alpha$  photons under the influence of these radiative transfer processes adversely affect the visibility of Ly $\alpha$  emission, and complicate its spectroscopic interpretation. However, as a direct result of these same processes, the Ly $\alpha$  signal from the central few kpc of high-redshift galaxies encodes information about the structure, kinematics, and ionisation properties of each galaxy and the interstellar, circumgalactic, and intergalactic media through which it propagates (e.g., Shapley *et al.* 2003; Verhamme, Schaerer, & Maselli 2006; Verhamme *et al.* 2008; Dijkstra & Wyithe 2010; Steidel *et al.* 2010; Law *et al.* 2012a; Hayes 2015; Trainor *et al.* 2015, 2019; Gronke & Dijkstra 2016; Byrohl & Gronke 2020; Chen *et al.* 2020).

Relationships between Ly $\alpha$  equivalent width (EW) and the spectral and physical properties of early SFGs have been extensively studied in populations of  $z = 2 - 4$  LBGs (see for example Shapley *et al.* 2003; Reddy *et al.* 2006; Erb *et al.* 2006a; Law *et al.* 2007; Kornei *et al.* 2010; Pentericci *et al.* 2010; Stark *et al.* 2010; Berry *et al.* 2012; Jones, Stark, & Ellis 2012; Law *et al.* 2012b,a; Erb *et al.* 2016; Hathi *et al.* 2016; Trainor *et al.* 2016;

Du *et al.* 2018; Marchi *et al.* 2019), and especially recently in samples of the related Lyman- $\alpha$  emitters (LAEs) at similar redshifts (e.g., Trainor *et al.* 2015, 2019; Oyarzún *et al.* 2017; Guaita *et al.* 2017; Cullen *et al.* 2020; Feltre *et al.* 2020; Santos *et al.* 2020; Matthee *et al.* 2021). In a systematic study at low redshift ( $z \sim 0.1$ ), the Lyman Alpha Reference Sample collaboration investigated all the quantities thought to be involved in the Ly $\alpha$  transport process (Östlin *et al.* 2014; Hayes *et al.* 2014; Pardy *et al.* 2014; Guaita *et al.* 2015; Rivera-Thorsen *et al.* 2015; Duval *et al.* 2016; Herenz *et al.* 2016; Runnholm *et al.* 2020). In both redshift ranges, larger Ly $\alpha$  emission transmission (or Ly $\alpha$  EW) was found to be associated with galaxies with bluer UV colours, lower metallicities, lower stellar masses, lower rest-frame UV luminosities, lower star formation rates, harder ionising field strengths, and more compact morphologies. In addition, observed Ly $\alpha$  emission/absorption strength has been shown to be sensitive to the galactic environment. Not only does Ly $\alpha$  visibility in the early universe reflect the well-established galaxy formation paradigm within which more luminous (massive), older Ly $\alpha$ -absorbing LBGs occupy regions of greater mass overdensity and cluster more strongly than their lower mass, less luminous and younger LAE counterparts (e.g., Ouchi *et al.* 2004, 2010, 2018; Adelberger *et al.* 2005; Jose, Srianand, & Subramanian 2013; Bielby *et al.* 2016; Guaita *et al.* 2017), it is also modulated by the galactic environment on small and large scales (e.g., Cooke *et al.* 2010; Cooke, Omori, & Ryan-Weber 2013; Díaz *et al.* 2014; Muldrew, Hatch, & Cooke 2015; Toshikawa *et al.* 2016; Lemaux *et al.* 2018; Shi *et al.* 2019; Guaita *et al.* 2020).

Such relationships suggest the tantalising prospect of using Ly $\alpha$  as a multi-purpose tool to elucidate the physical, environmental, and large-scale clustering properties of primordial galaxies. To properly explore these relationships however, large samples that reflect the spectral characteristics of the selected population are necessary and, in most cases, spectroscopic measurement of Ly $\alpha$  is required in order to extract the physical properties of interest. In an approach that addressed this problem, Cooke (2009, hereafter C09) reported a method of Ly $\alpha$  spectral type classification for a population of  $z \sim 3$  LBGs by which pure LBG samples displaying either dominant Ly $\alpha$  in absorption (aLBGs) or dominant Ly $\alpha$  in emission (eLBGs) could be isolated using only broadband information. One example of the power of this approach was demonstrated on large scales by Cooke *et al.* (2013, hereafter C13) who performed an auto- and cross-correlation function analysis of pure aLBG and eLBG samples photometrically selected from  $\sim 55000$   $z \sim 3$  LBGs. C13 found that aLBGs preferentially reside in group and cluster environments, eLBGs reside on the outskirts of groups and in the field, the two spectral types avoid each other on small, single halo scales, and that without accounting for the anti-correlation between aLBGs and eLBGs, masses for LBG populations were underestimated.

One motivation for this paper is to extend the method developed by C09 to other redshifts, especially to  $z \sim 2$ , where the availability of a statistical sample of LBGs with consistent multi-band rest-frame UV broadband photometry, uniformly measured net Ly $\alpha$  EWs, and kinematic classifications quantitatively determined from IFU-based spectroscopy, prompted the investigation of the relationship between Ly $\alpha$  spectral type and galaxy kinematics described in Paper II in this series (Foran *et al.* 2023b, submitted).

More broadly, we aim to develop a method that can be applied to large samples of  $z \sim 2 - 6$  LBGs selected from current and

future large-area and all-sky photometric campaigns. The current generation of deep, large-area photometric surveys such as HSC-SSP and LSST will probe cosmic volumes and transverse areas on scales that transcend the cosmic variance of even the largest legacy fields (see e.g., Arcila-Osejo & Sawicki 2013). The challenge for spectroscopy and targeted deep multiband/multiwavelength imaging is the huge datasets that will derive from surveys conducted on such scales; in the LSST 10-yr data for example, there will be 20 billion sources that will need to be processed in order to identify hundreds of millions of high-redshift galaxies. Optimising the discovery potential of these investments requires new techniques to statistically characterise the huge datasets they will deliver, and to efficiently select from these the most promising samples for expensive follow-up observations. Here we explore how inexpensive broadband photometric information that is sensitive to the Ly $\alpha$  properties of LBGs might be used to address these challenges, and suggest a means by which the Ly $\alpha$ -related physical and spectroscopic properties and environments of  $z \sim 2 - 3$  LBGs might be explored on the basis of broadband photometric information alone.

In this paper we demonstrate a statistical relationship between net Ly $\alpha$  EW and the optical broadband photometric properties of  $z \sim 2$  LBGs. We characterise the segregation of spectroscopic Ly $\alpha$ -absorbing, and Ly $\alpha$ -emitting spectral types in colour-magnitude space, and define photometric criteria by which pure sub-samples of LBGs with Ly $\alpha$  dominant in absorption (p-aLBGs), and Ly $\alpha$  dominant in emission (p-eLBGs), can be selected using only broadband imaging data. As a first step toward the application of our approach to large-area and all-sky surveys, we also present here a set of ugrizy photometric selection criteria by which pure samples of p-aLBGs and p-eLBGs might be isolated from datasets derived from the LSST.

This paper is structured as follows: in Section 2, we present the photometric and spectroscopic data used in the subsequent sections. Section 3 describes the segregation versus net Ly $\alpha$  EW of  $z \sim 2$  LBGs in colour-magnitude space, and the application of this result to determine criteria for the selection of photometric Ly $\alpha$  spectral type sub-samples. We summarise the important conclusions and potential applications of this work in Section 4. We assume a  $\Lambda$ CDM cosmology with  $\Omega_M = 0.3$ ,  $\Omega_\Lambda = 0.7$  and  $H_0 = 70$  km s $^{-1}$  Mpc $^{-1}$ . All magnitudes are quoted in the AB system of Oke & Gunn (1983).

## 2. Data

Broadband optical photometry and rest-frame net Ly $\alpha$  equivalent width (hereafter ‘net Ly $\alpha$  EW’) data for a sample of 557 rest-frame UV colour-selected  $z \sim 2$  galaxies in the redshift range  $1.7 < z < 2.5$  were extracted from the spectroscopic catalog of Steidel et al. (2004), Reddy et al. (2008). Similarly, we make use of the rest-frame net Ly $\alpha$  EW measurements of Shapley et al. (2003) for a sample of 775 LBGs in the redshift range  $2.5 < z < 3.5$  drawn from the catalog of Steidel et al. (2003). Values for net Ly $\alpha$  EW – which incorporate information about Ly $\alpha$  in both emission and absorption – were measured uniformly across both redshift ranges in their respective source studies using the method described by Kornei et al. (2010). Typical uncertainties in absolute Ly $\alpha$  EW are  $\sim 25$ – $50\%$  for galaxies with absorption profiles, and  $\sim 25\%$  for galaxies with Ly $\alpha$  dominant in emission (Shapley et al. 2003).

The parent catalogs of the  $z \sim 2$  and  $z \sim 3$  samples derive from an observational campaign that targeted 14 uncorrelated fields with a total survey area of 1 900 arcmin $^2$ , resulting in samples that are minimally affected by systematic biases due to cosmic variance or clustering. The survey used the  $U_nGR$  photometric system (Steidel et al. 2003), and the rest-frame UV colour selection criteria of Steidel et al. (2003) ( $z \sim 3$  LBGs) and Steidel et al. (2004) ( $z \sim 2$  BX galaxies). These criteria were designed to recover objects with intrinsic properties, particularly UV luminosity and reddening by dust, that were similar across both redshift ranges. Accordingly, and although the  $z \sim 2$  BX selection method does not probe the Lyman break, we henceforth refer to both samples as ‘LBGs’. These selection criteria result in a net Ly $\alpha$  EW distribution for the  $\mathcal{R} < 25.5$  samples that is representative of the intrinsic distribution for the parent population of galaxies (Reddy et al. 2008). The mean redshift of our extracted  $z \sim 2$  sample is  $z = 2.16 \pm 0.20$ , corresponding to a mean absolute magnitude sensitivity 0.58 mag fainter in the observed  $\mathcal{R}$ -band imaging than at  $z = 2.96$ , the mean redshift of the  $z \sim 3$  LBG sample.

The bulk of galaxies in the  $z \sim 2$  LBG sample have stellar masses in the range  $9 \lesssim \log(M_*/M_\odot) \lesssim 11$  (Shapley et al. 2005; Erb et al. 2006b; Reddy et al. 2006; Reddy & Steidel 2009) and star formation rates inferred from rest-frame UV luminosities (uncorrected for extinction) in the range  $3 \lesssim M_\odot \text{ yr}^{-1} \lesssim 60$  (Steidel et al. 2004). Accordingly, our  $z \sim 2$  sample is typical of LBGs/SFGs at these redshifts (Álvarez-Márquez et al. 2016, and references therein) and lies with a range of properties (see Reddy et al. 2006) on the main sequence of stellar mass and star formation rate for  $z \sim 2$  SFGs (Daddi et al. 2007).

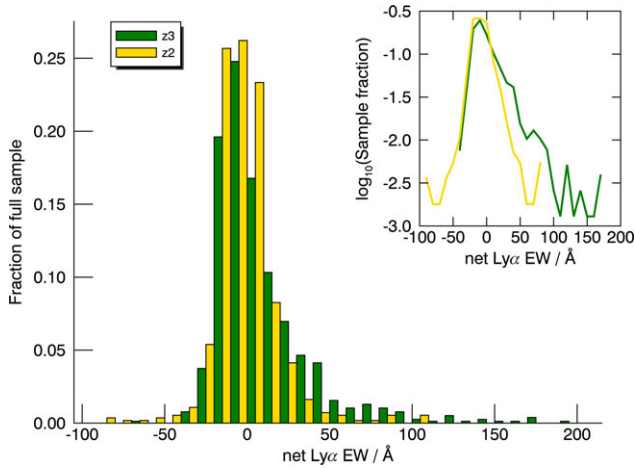
The  $z \sim 2$  parent sample has  $\mathcal{R}$ -band apparent magnitudes in the range  $22.0 < \mathcal{R} < 25.5$ , corresponding to rest-frame UV luminosities (absolute magnitudes) of  $-22.6 < M_{UV} < -19.1$ . The faint end magnitude cut of  $\mathcal{R} \leq 25.5$  was determined by signal-to-noise requirements of the spectroscopic measurements. Given our need for accurate ( $U_n - \mathcal{R}$ ) colours, and the different  $U_n$ -band depths for the fields targeted by Steidel et al. (2003, 2004), we applied a further (conservative)  $U_n < 26.5$  cut to the  $z \sim 2$ ,  $U_n$ -band data to ensure that our sample included only the most reliable photometry. Appendix 1 describes the derivation of indicative photometric uncertainties for our  $z \sim 2$  and  $z \sim 3$  samples.

## 3. Analysis and results

### 3.1. Net Ly $\alpha$ EW distribution and spectral type classification

The profile of Ly $\alpha$  in the spectrum of high-redshift LBGs manifests in absorption, emission, or a combination of both. In the  $z \sim 3$  LBG sample of Shapley et al. (2003, hereafter S03), for example, the distribution of net Ly $\alpha$  EWs is centred near zero and varies from  $\lesssim -50$  Å to  $\gtrsim +200$  Å. Net Ly $\alpha$  EW values for our 557  $z \sim 2$  LBGs span a similar range ( $-85.0$  Å to  $+108.7$  Å) and, like the S03 sample, are asymmetrically dispersed toward higher net Ly $\alpha$  EWs around a median near zero ( $-4.42$  Å at  $z \sim 2$  and  $+0.56$  Å at  $z \sim 3$ ). These similarities, however, belie a change in the shape of the distribution that is evidenced by a shift in the mean net Ly $\alpha$  EW for the respective full samples from  $+10.3$  Å at  $z \sim 3$  to  $-2.2$  Å at  $z \sim 2$ .

The changing shape of the net Ly $\alpha$  EW distribution with redshift is readily apparent in Fig. 1 that shows normalised histogram



**Figure 1.** Normalised histograms showing the distribution of net Ly $\alpha$  EWs for  $z \sim 3$  (green) and  $z \sim 2$  (gold) LBG samples. Inset: The same distributions plotted with a logarithmic ordinate axis to accentuate the ‘tails’ of the Ly $\alpha$  EW distributions. Net Ly $\alpha$  EWs less than zero are essentially identical between the two populations, while the  $z \sim 3$  sample has a significantly larger fraction of net Ly $\alpha$ -emitters (see Table 1).

**Table 1.** Statistics for sub-samples of  $z \sim 2$  and  $z \sim 3$  LBGs divided on the basis of net Ly $\alpha$  EW.

Redshift range								
$z \sim 3^a$			$z \sim 2^b$			$z \sim 2^c$		
Nd	$N/N_{tot}^e$	Ly $\alpha$ EW <sup>f</sup> (Å)	N	N	$N/N_{tot}$ (Å)	Ly $\alpha$ EW	$N/N_{tot}$	Ly $\alpha$ EW (Å)
775	1.0	+10.3	557	1.0	-2.2	557	1.0	-2.2
194	0.25	-16.7	140	0.25	-21.5	188	0.34	-18.9
193	0.25	-4.8	139	0.25	-8.6	146	0.26	-5.3
194	0.25	+8.3	139	0.25	+0.7	176	0.32	+6.9
194	0.25	+54.3	139	0.25	+20.8	47	0.08	+40.6

<sup>a</sup> $z \sim 3$  LBGs from Shapley et al. (2003).

<sup>b</sup>Our  $z \sim 2$  sample divided into numerical quartiles.

<sup>c</sup>Our  $z \sim 2$  sample divided into aLBG, G<sub>a</sub>, G<sub>e</sub> and eLBG Ly $\alpha$  spectral types as per the definitions given in Section 3.1.

<sup>d</sup>Number of galaxies.

<sup>e</sup>Fraction of full sample ( $N_{tot}$ ) in sub-sample ( $N$ ).

<sup>f</sup>Mean net Ly $\alpha$  EW for each (sub)sample.

plots for the  $z \sim 2$  and  $z \sim 3$  samples. Consistent with the result of Reddy et al. (2008), we find that the two distributions are very similar at net Ly $\alpha$  EWs  $\lesssim 0$  Å, but there is a sharp drop off at  $z \sim 2$  toward higher values of net Ly $\alpha$  EW that is largely responsible for the difference in overall mean net Ly $\alpha$  EW between the two samples.

The utility of dividing a population of rest-frame UV-colour selected galaxies into sub-samples based on observed net Ly $\alpha$  EW was first demonstrated by S03, and it continues to be a useful approach in the study of relationships between Ly $\alpha$  and the physical and spectral properties of LBGs (e.g., Du et al. 2018; Pahl et al. 2020). In their discovery of the broadband photometric segregation versus Ly $\alpha$  EW in the S03 sample, C09 exploited the same approach to derive the method of photometric Ly $\alpha$  spectral-type classification (see Section 3.2).

Table 1 shows a comparison of the statistics for our  $z \sim 2$  and  $z \sim 3$  LBG samples divided into numerical quartiles on the basis of

net Ly $\alpha$  EW. There is a shift toward more negative mean net Ly $\alpha$  EW ( $\Delta \sim -5$  Å) of the most absorbing quartile at  $z \sim 2$  compared to the same quartile at  $z \sim 3$ . The two  $z \sim 2$  quartiles that span the more Ly $\alpha$ -emitting end of the distribution show a larger (and increasing) shift to lower average net Ly $\alpha$  EW compared to the analogous quartiles of S03 ( $\Delta -9.0$  Å and  $\Delta -32.5$  Å for q3 and q4, respectively).

Motivated by these observations, and our results showing a relationship between net Ly $\alpha$  EW and nebular emission-line kinematics (Foran et al. 2023b, submitted), we applied to our  $z \sim 2$  sample the same net Ly $\alpha$  EW cuts used by S03 to generate numerical quartiles at  $z \sim 3$ . We define the most absorbing fraction of galaxies with net Ly $\alpha$  EW  $\leq -10.0$  Å as ‘aLBGs’, and the most strongly emitting fraction with net Ly $\alpha$  EW  $\geq +20.0$  Å as ‘eLBGs’. We further divide the remaining LBGs into G<sub>a</sub> and G<sub>e</sub> spectral types with net Ly $\alpha$  EWs  $-10.0$  Å  $<$  net Ly $\alpha$  EW  $<$  0.0 Å and 0.0 Å  $<$  net Ly $\alpha$  EW  $<$  +20.0 Å, respectively. Table 1 summarises the population statistics of the  $z \sim 2$  sample and our Ly $\alpha$  spectral types compared to the  $z \sim 3$  LBGs of S03.

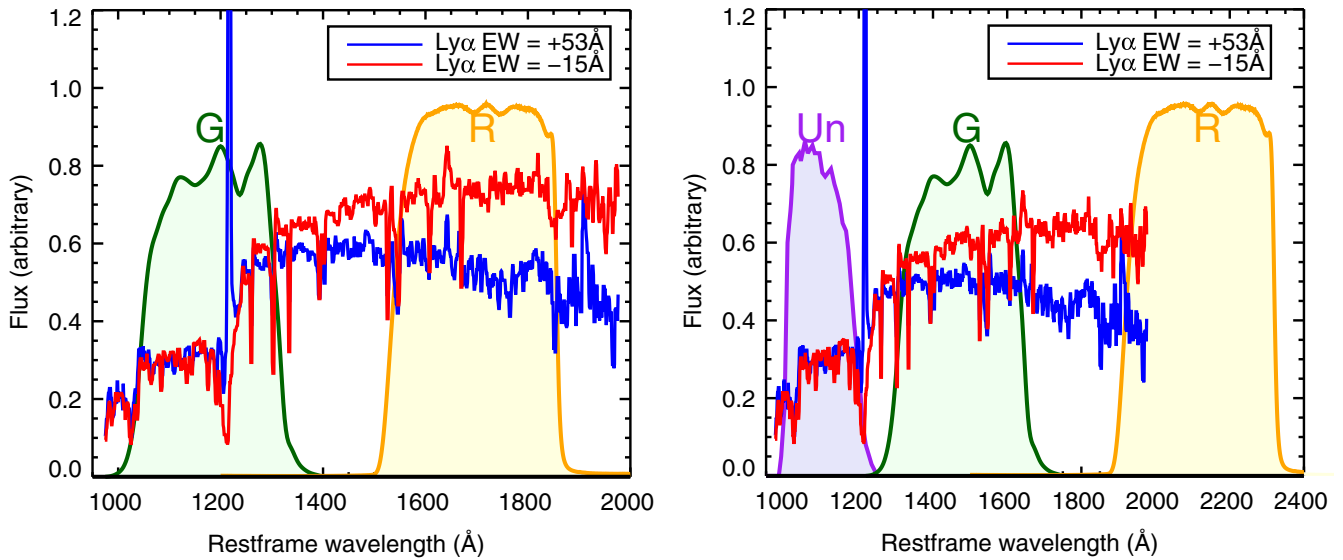
Given that we have defined our spectral types using the same net Ly $\alpha$  EW cuts as S03, it is not surprising that the central (G<sub>a</sub> and G<sub>e</sub>) spectral types have mean net Ly $\alpha$  EWs similar ( $\Delta \sim -1$  Å) to the equivalent quartiles (q2 and q3) in the  $z \sim 3$  sample. It is noteworthy, however, that despite the overall shift in mean net Ly $\alpha$  EW of  $\sim -12.5$  Å between the  $z \sim 3$  and  $z \sim 2$  samples, the mean net Ly $\alpha$  EW of the  $z \sim 2$  aLBGs is similarly only  $\sim -2$  Å more negative than the equivalent S03 quartile, indicative of a compression of the LBG population toward the Ly $\alpha$ -absorbing end of the distribution. This behaviour is likely due to the fact that the measured net Ly $\alpha$  EW becomes insensitive to the total absorption once the Ly $\alpha$  absorption feature is saturated. That is, beyond that point, any further decrease in measured EW would reflect only the contribution of the damping wings, and depend weakly on increasing HI column density (see Section 3.2 for manifestation of this effect in the broadband imaging data).

Conversely, the mean net Ly $\alpha$  EW of the eLBG spectral type sub-sample at  $z \sim 2$  is  $\sim -14$  Å more negative than the analogous (most strongly Ly $\alpha$ -emitting) S03 quartile, and the relative fraction of eLBGs at  $z \sim 2$  is 0.08 compared to 0.25 at  $z \sim 3$  – a clear reflection of the lower relative abundance of net Ly $\alpha$  emitting LBGs in the universe and/or within the LBG selection function at  $z \sim 2$  compared to  $z \sim 3$ .

### 3.2. Segregation of $z \sim 2$ LBGs in colour-magnitude space

C09 discovered that over the redshift path  $z \sim 3.0 \pm 0.3$ , the relationship between rest-frame UV continuum slope and net Ly $\alpha$  EW leads to a photometric dispersion of LBGs, and an ability to separate LBG spectral types on a broadband colour-magnitude plane based on their net Ly $\alpha$  EW. At  $z \sim 3$ , aLBGs are (on average) brighter in  $\mathcal{R}$  magnitude and redder than eLBGs. They separate in  $(G - \mathcal{R})$  colour as a result of the redder UV continuum slopes of aLBGs, combined with an additional small red enhancement as a result of the Ly $\alpha$  absorption in the G-band, as compared to eLBGs with bluer UV continuum slopes, combined with an additional blue enhancement from the Ly $\alpha$  emission in the G-band. Together, these behaviours enable a statistical segregation of the two populations on a  $(G - \mathcal{R})$  vs  $\mathcal{R}$  colour-magnitude diagram (CMD), with subsets containing pure samples of each spectral type. Fig. 2 illustrates the origin of the  $z \sim 3$  broadband imaging segregation of Ly $\alpha$ -absorbing and Ly $\alpha$ -emitting





**Figure 2.** Illustration of the origin of the colour separation of Lyman break galaxy (LBG) Ly $\alpha$  spectral types. Left: Plotted are the G and  $\mathcal{R}$  filter transmission curves (Steidel et al. 2003) in green and orange, respectively, shifted to the  $z \sim 3$  rest-frame. Overlaid are the (smoothed) quartile 1 (red, representative of aLBGs) and quartile 4 (blue, representative of eLBGs) composite spectra of Shapley et al. (2003). The composite spectra consist of  $\sim 200$   $z \sim 3$  LBG spectra with similar Ly $\alpha$  EW, with the mean values indicated in the legend. The spectra are shown normalised over the G filter to help illustrate the (G –  $\mathcal{R}$ ) colour difference in the two spectral types for a given G magnitude. The origin of the Ly $\alpha$  spectral type photometric segregation on the (G –  $\mathcal{R}$ ) vs  $\mathcal{R}$  CMD results from their colour differences based on the UV continuum slope relationship with spectral type and a small (and inverse) contribution from the Ly $\alpha$  emission/absorption feature and the magnitude differences in spectral type, in that  $z \sim 3$  aLBGs are brighter on average than eLBGs. Right: Similar to the left plot, but for LBGs at  $z \sim 2$ . The composite spectra are shown normalised over the  $U_n$  filter (violet, see text). Note: the composite spectra and the normalisation are shown for illustrative purposes and extend to 2000 $\text{\AA}$ , rest-frame. However, the UV continuum slopes of quartiles 1 (red) and 4 (blue) maintain a significant difference in  $\mathcal{R}$  that is sufficient to separate aLBG and eLBG spectral types in ( $U_n$  –  $\mathcal{R}$ ) colour and  $\mathcal{R}$  magnitude on the CMD. Depending on the redshift of  $z \sim 2$  LBGs, the Ly $\alpha$  feature may fall in or out of the  $U_n$  filter (see Section 3.4).

LBGs that enables the determination of photometric Ly $\alpha$  spectral types.

To test whether a similar relationship between broadband photometry and net Ly $\alpha$  EW might exist at  $z \sim 2$ , we use ( $U_n$  –  $\mathcal{R}$ ) colours and  $\mathcal{R}$ -band magnitudes to construct a CMD for our sample of 557  $z \sim 2$  LBGs. We use ( $U_n$  –  $\mathcal{R}$ ) rather than ( $U_n$  – G), to sample the rest-frame UV continuum farther redward of the Ly $\alpha$  feature so as to increase the segregation between the redder-sloped aLBGs and the bluer-sloped eLBGs, and to avoid any possibility of contamination of our redward filter by Ly $\alpha$  emission. The separation in wavelengths probed by the  $U_n$  and G filters at  $z \sim 2$  is smaller than the separation of the G and  $\mathcal{R}$  filters at  $z \sim 3$  (see Fig. 2).

The left panel of Fig. 3 shows the spectroscopic  $z \sim 2$  LBG sample dispersed in colour ( $U_n$  –  $\mathcal{R}$ ) and magnitude  $\mathcal{R}$ , with symbols colour-coded on a red-to-blue gradient according to their measured net Ly $\alpha$  EW. For visualisation purposes the colour table is scaled to map the range  $-35.0 \text{ \AA} < \text{net Ly}\alpha \text{ EW} < +40.0 \text{ \AA}$  which encompasses  $\gtrsim 95\%$  of the galaxies in our sample. Plotting the data in this way (i.e., downplaying the colour effect of the few extreme net Ly $\alpha$  EW cases), the bulk of the LBG sample manifests on the CMD as a visible colour gradient from red to blue moving diagonally from roughly the top left to bottom right. The galaxies in our sample with the most negative net Ly $\alpha$  EW (dark red symbols) do not lie at the extreme end of the colour gradient direction as might be expected for a simple monotonic relationship. While they are certainly well within the ‘absorbing’ half of the CMD, they lie toward the centre of the distribution, and approximately along a line orthogonal to the underlying trend. This apparently anomalous behaviour of the most absorbing galaxies in our sample notwithstanding, the overall trend is confirmed by the points

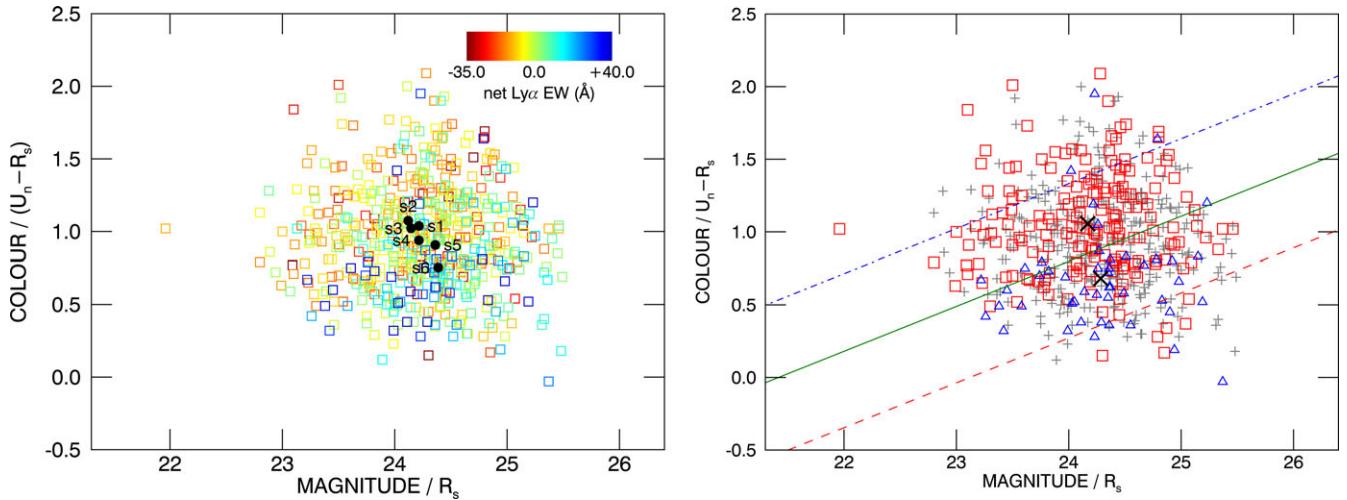
**Table 2.** Statistics for the dispersion of  $z \sim 2$  LBGs in colour ( $U_n$  –  $\mathcal{R}$ )–magnitude ( $\mathcal{R}$ ) space divided into numerical sextiles based on net Ly $\alpha$  EW.

Sextiles <sup>a</sup>	Mean $\mathcal{R}$ mag.	Mean ( $U_n$ – $\mathcal{R}$ ) colour	Mean Ly $\alpha$ EW ( $\text{\AA}$ )
s1	24.22	1.04	–25.08
s2	24.12	1.07	–12.85
s3	24.15	1.02	–7.21
s4	24.21	0.94	–0.86
s5	24.36	0.91	5.66
s6	24.39	0.75	27.44

<sup>a</sup> Full  $z \sim 2$  spectroscopic LBG sample (557 galaxies) divided into sextiles of  $\sim 93$  galaxies each.

labelled s1 to s6 on the colour gradient plot, that indicate the positions of the magnitude and colour distribution means for the numerical sextiles (of  $\sim 93$  galaxies each) grouped according to their net Ly $\alpha$  EW (see Table 2 for a summary of the sextile statistics). The more positive net Ly $\alpha$  EW LBGs (weaker absorption and more emission) show an overall trend toward fainter  $\mathcal{R}$ -band magnitudes and bluer ( $U_n$  –  $\mathcal{R}$ ) colours. Indeed, only the most absorbing sextile (s1 in Fig. 3) does not follow this monotonic trend. That being said, the colours and magnitudes of the  $z \sim 2$  sextiles converge with increasing Ly $\alpha$  absorption strength (i.e., from s6 to s1), unlike at  $z \sim 3$ , where the mean colours (magnitudes) continue to redden (brighten) monotonically (cf. C09).

We can speculate that the ‘off trend’ positions of the strongest Ly $\alpha$  absorbers on the CMD is a manifestation of the environmental effect proposed by C13 over their intrinsic net Ly $\alpha$  EWs. In



**Figure 3.** Rest-frame UV colour ( $U_n - \mathcal{R}$ )-magnitude ( $\mathcal{R}$ ) diagrams (CMDs) for Lyman break galaxies (LBGs) in the redshift range  $1.7 < z < 2.5$ , and with magnitude cuts of  $\mathcal{R} < 25.5$  and  $U_n < 26.5$ . Left:  $z \sim 2$  LBGs dispersed in colour-magnitude space with symbols colour-coded on a red-blue gradient according to their measured net Ly $\alpha$  EW. The colour table maps the range  $-35.0 \text{ \AA} < \text{net Ly}\alpha \text{ EW} < +40.0 \text{ \AA}$ , which encompasses  $\geq 95\%$  of the sample. Points labelled s1 to s6 indicate the colour and magnitude distribution means of the numerical sextiles of the LBG sample divided on the basis of net Ly $\alpha$  EW. Right: Grey plus (+) marks denote the 557 galaxies in the  $z \sim 2$  spectroscopic sample. Galaxies with net Ly $\alpha$  EW  $\leq -10.0 \text{ \AA}$  (aLBGs) are overlaid with red squares, and those with net Ly $\alpha$  EW  $\geq +20.0 \text{ \AA}$  (eLBGs) are overlaid with blue triangles. The mean value for each distribution is marked with a black cross (X), with aLBG mean indicated by the upper cross and eLBG mean by the lower. The dotted-dashed blue and dashed red lines indicate a  $1.5\sigma$  dispersion in colour from the primary cut (green line) that divides the aLBG and eLBG distributions, respectively (see text).

this scenario, the most negative net Ly $\alpha$  EWs observed, with otherwise typical aLBG colour and magnitude, may be a result of their environment near the cores of groups and proto-clusters and the presence of larger column densities of intra-group/cluster neutral gas at or near the systemic velocity of the LBGs along the line of sight (e.g., Muldrew et al. 2015; Toshikawa et al. 2016; Lemaux et al. 2018). More prosaically, it is also plausible that this behaviour, and the grouping of the three most absorbing sextiles (s1–s3) on the CMD, is a reflection of the compression of the  $z \sim 2$  sample towards the more negative end of the net Ly $\alpha$  EW distribution (as described in Section 3.1), combined with the inherently greater uncertainties (25–50%) associated with the net Ly $\alpha$  EW measurements for the most absorbing systems (S03) and the inherent photometric scatter (see Appendix 1).

The association between net Ly $\alpha$  EW and the photometric properties of  $z \sim 2$  LBGs suggested by the trend shown in the left panel of Fig. 3, prompts a statistical examination using the Ly $\alpha$  spectral type classification scheme described in Section 3.1 and the method demonstrated by C09. In the right panel of Fig. 3 we plot on the CMD the LBG spectral types as described in Section 3.1, i.e., aLBGs with net Ly $\alpha$  EW  $\leq -10 \text{ \AA}$  and eLBGs with net Ly $\alpha$  EW  $\geq +20 \text{ \AA}$ , and show that they segregate into two cohesive, albeit overlapping, distributions.

We define a primary cut (solid green line) that passes through the midpoint between the mean colour and magnitude values of the aLBG and eLBG distributions (black crosses), and has slope that maximises the difference in mean net Ly $\alpha$  EW and spectral-type purity between the sub-samples that lie above and below the broken blue and red lines, respectively. These dashed (red) and dotted-dashed (blue) lines indicate an offset of  $1.5\sigma$  in colour dispersion from the primary cut for the aLBG and eLBG distributions, respectively (see Section 3.3.1). Statistics for the segregation of the aLBG and eLBG spectral types shown in the right panel of Fig. 3 are summarised in Table 3 together with (for comparison) the segregation statistics for the  $z \sim 3$  sample of C09.

**Table 3.** Statistics for the photometric segregation of Ly $\alpha$ -absorbing and Ly $\alpha$ -emitting spectral types in  $z \sim 2$  and  $z \sim 3$  LBGs.

	Redshift			
	$z \sim 2$		$z \sim 3$	
Colour	$(U_n - \mathcal{R})$		$(G - \mathcal{R})$	
Magnitude	$\mathcal{R}$		$\mathcal{R}$	
Total LBGs	557		775	
	aLBGs	eLBGs	aLBGs	eLBGs
Number	188	47	144	150
Magnitude mean	24.17	24.28	24.43	24.94
Colour mean	1.06	0.68	0.78	0.40
Colour $\sigma$	0.35	0.36	0.24	0.31
Slope	0.31		0.40	
Intercept	-6.62		-9.38	

Although the slope and intercept for the  $z \sim 2$  segregation quoted in Table 3 are the values that give the maximum difference in mean net Ly $\alpha$  EW between the photometrically selected sub-samples, the maximum is shallow, asymmetric, and relatively insensitive to the choice of slope. For example, in the optimal case where  $c_\sigma = 1.25$  (see Section 3.3.1), the maximum difference in mean net Ly $\alpha$  EW is  $17.1 \text{ \AA}$  at a slope of 0.31. We note, however, that the difference in mean net Ly $\alpha$  EW is greater than  $16.0 \text{ \AA}$  for all slopes between 0.17 and 0.40. Thus we might quote an uncertainty (or ‘range of confidence’) of slope =  $0.31^{+0.09}_{-0.13}$ , within which any choice of slope would result in photometrically selected sub-samples with a difference in mean net Ly $\alpha$  EW that is within  $\sim 5\%$  of the maximum. Constraining the primary cut to pass through

the mid-point of the aLBG and eLBG distribution means similarly gives intercept values in the range  $-6.62_{-2.30}^{+3.43}$ .

### 3.3. Photometric Ly $\alpha$ spectral type selection and sub-sample purity

#### 3.3.1. $z \sim 2$ LBGs

Following the method used by C09 at  $z \sim 3$ , we use the parameters of the segregated aLBG and eLBG distributions shown in Fig. 3 to isolate sub-samples with Ly $\alpha$  dominant in absorption ('photometric' aLBGs, or p-aLBGs) and with Ly $\alpha$  dominant in emission ('photometric' eLBGs, or p-eLBGs) from the parent  $z \sim 2$  LBG sample. Invoking the primary cut slope and intercept values from Table 3, and the supplied broadband photometry, the following relationships can be used to extract sub-samples of the desired photometric Ly $\alpha$  spectral type.

For p-aLBGs,

$$(U_n - \mathcal{R}) \geq 0.3091 \cdot \mathcal{R} - 6.6208 + c_\sigma \cdot \sigma_e \quad (1)$$

and for p-eLBGs,

$$(U_n - \mathcal{R}) \leq 0.3091 \cdot \mathcal{R} - 6.6208 - c_\sigma \cdot \sigma_a \quad (2)$$

where  $c_\sigma$  is the coefficient of colour standard deviation by which boundaries used to isolate the photometric spectral type sub-samples are offset from the primary cut on the CMD, and  $\sigma_a$  (0.3509) and  $\sigma_e$  (0.3558) are the  $1\sigma$  standard deviations of the  $(U_n - \mathcal{R})$  colour distributions for the aLBG and eLBG subsets, respectively.

We use  $\sigma_a$  and  $\sigma_e$  to estimate the density of aLBGs and eLBGs on the CMD. This approach implies that any  $\sigma$  should extend around a distribution mean in some circular (or similar) contour. The primary cut we make between the aLBG and eLBG distribution means (and its use as the basis for estimating photometric spectral type purity) is a line for which our assumptions only formally apply at the point of closest approach (tangent to a circular contour) of our lines to the respective distribution means. Thus, the multiples of  $c_\sigma$  (1.5 in Fig. 3) applied to  $\sigma_a$  and  $\sigma_e$ , plus the fraction of  $\sigma$  by which the primary cut is removed from the respective distribution mean positions ( $\sim 0.5\sigma_a$  and  $\sim 0.5\sigma_e$ ), represent a minimum coefficient of  $\sigma$  that can be used to estimate the extent and purity of different Ly $\alpha$  spectral types on the CMD. For example, cuts on the CMD for which  $c_\sigma = 1.5$  along the same slope as the primary cut approximate (for Gaussian distributions) criteria for selecting Ly $\alpha$  spectral type sub-samples  $\gtrsim 2\sigma$  from the mean value of the opposite distribution.

In theory, the above criteria can be made stricter (or relaxed) by varying the value of  $c_\sigma$ , thereby trading sub-sample size for sub-sample purity according to the properties of the parent sample, and the requirements of the intended application. In practice, the range of  $c_\sigma$  values that can be meaningfully employed is limited by the degree to which the aLBG and eLBG distributions deviate from Gaussian behaviour, and by small-number statistics at higher values of  $c_\sigma$  – especially for eLBGs which are  $\sim 4$  times less abundant than aLBGs in our  $z \sim 2$  sample. Table 4 summarises the statistics for p-aLBG and p-eLBG Ly $\alpha$  spectral type sub-samples selected from the parent  $z \sim 2$  LBGs using the selection criteria given in Equations (1) and (2), respectively, and a range of  $c_\sigma$  values.

We estimate the purity of each photometric spectral type sub-sample by calculating the degree to which they exclude galaxies with opposite spectral type as determined by their measured net

**Table 4.** Statistics for photometric sub-samples with Ly $\alpha$  dominant in absorption (p-aLBGs) and Ly $\alpha$  dominant in emission (p-eLBGs) selected from the parent  $z \sim 2$  LBG sample using Equations (1) & (2) and different values of  $c_\sigma$ .

Photometric spectral type						
$c_\sigma^a$	$N^b$	p-aLBGs		$N$	p-eLBGs	
		Purity <sup>c</sup> (%)	Ly $\alpha$ EW <sup>d</sup> (Å)		Purity (%)	Ly $\alpha$ EW (Å)
0.0	311	97.1(70.7)	-6.25	246	76.0(53.6)	3.03
0.5	223	98.2(73.1)	-7.59	156	80.1(62.2)	5.06
0.75	178	97.8(74.7)	-8.00	104	84.6(66.3)	7.03
1.0	138	97.8(79.0)	-8.23	69	85.5(65.2)	7.73
1.25	101	97.0(79.2)	-7.42	50	84.0(64.0)	9.59
1.5	72	95.8(77.8)	-5.80	36	83.3(61.1)	2.94
2.0	38	97.4(76.3)	-7.19	11	54.5(45.4)	-4.35

<sup>a</sup>Coefficient of colour standard deviations ( $\sigma_a$  &  $\sigma_e$ ) by which boundaries used to isolate the photometric spectral type sub-samples are offset from the CMD primary cut.

<sup>b</sup>Number of galaxies in the photometric sub-samples.

<sup>c</sup>For p-aLBGs: Percent purity with respect to eLBG and (eLBG + G<sub>e</sub>) spectral types. For p-eLBGs: Percent purity with respect to aLBG and (aLBG + G<sub>a</sub>) spectral types.

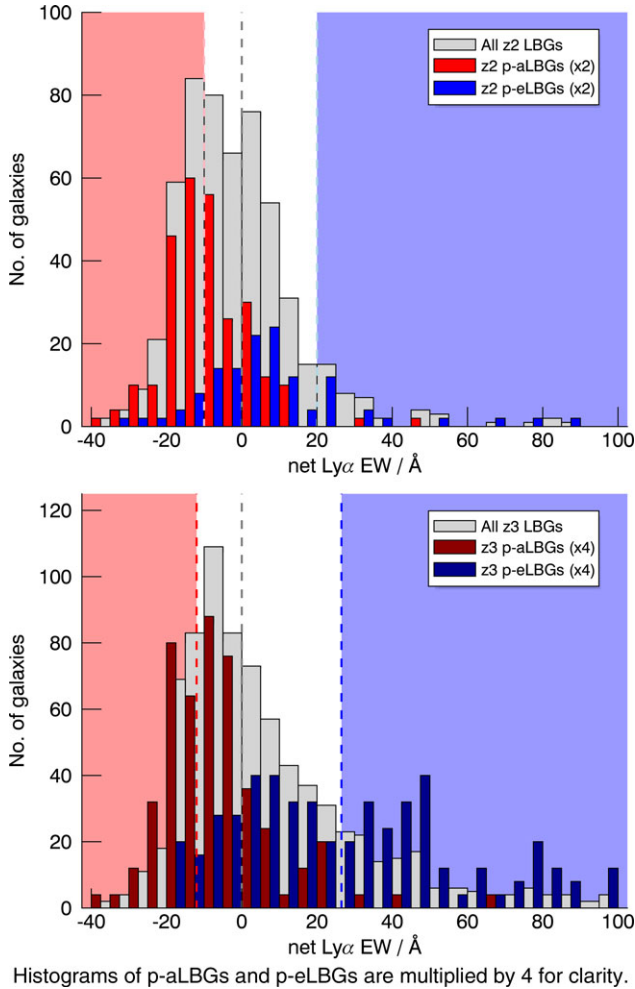
<sup>d</sup>Mean net Ly $\alpha$  EW for each sub-sample.

Ly $\alpha$  EW and our classification scheme described in Section 3.1. That is, for example, for each p-aLBG sub-sample selected using a different value of  $c_\sigma$ , we calculate the contamination fraction of spectroscopic eLBGs and eLBG + G<sub>e</sub> spectral types. The purity of the p-aLBG sub-sample thus determined is quoted as a percentage with respect to eLBGs and with respect to eLBG + G<sub>e</sub> spectral types (parenthesised) in Table 4. The mean net Ly $\alpha$  EW of the p-aLBG and p-eLBG sub-samples (also listed in Table 4) is a further measure of the quality of the broadband photometric segregation, and the average properties of the respective sub-samples.

Across a wide range of  $c_\sigma$  values, we select high-purity p-aLBG sub-samples, particularly with respect to contamination by spectroscopic eLBGs. Indeed, even using the primary cut between the aLBG and eLBG spectral types (i.e.,  $c_\sigma = 0.0$ ), results in a large sub-sample of 311 p-aLBGs ( $\gtrsim 55\%$  of total LBGs) that is  $\gtrsim 97\%$  free of eLBGs and  $\gtrsim 70\%$  pure with respect to galaxies with any detectable net Ly $\alpha$  emission. The practical upper limit of  $c_\sigma$  for the selection of p-aLBGs appears to be restricted only by the diminishing return of smaller sub-sample sizes. As a result, large broadband photometric samples can greatly benefit from stricter cuts. The optimal coefficient for the dataset here of  $\sigma_e$  ( $c_\sigma \approx 1.0 - 1.25$ ) selects  $\sim 100-140$  photometric aLBGs that are  $\gtrsim 97\%$  and  $\gtrsim 79\%$  pure with respect to spectroscopic eLBGs, and eLBG + G<sub>e</sub> spectral types respectively, and for which the mean net Ly $\alpha$  EW is  $\sim -8$  Å.

With  $c_\sigma \approx 1.0 - 1.25$  we select a sample of  $\sim 50-70$  LBGs with Ly $\alpha$  dominant in emission (p-eLBGs) that are  $\gtrsim 85\%$  and  $\gtrsim 65\%$  pure with respect to spectroscopic aLBGs and aLBG + G<sub>a</sub> spectral types, respectively, with a mean net Ly $\alpha$  EW of  $\sim +8$  Å. These purities represent a significant enhancement over the native (full sample) abundances of eLBGs ( $\sim 8\%$ ) and the sum of eLBG and G<sub>e</sub> spectral types ( $\sim 40\%$ ).

The segregation versus net Ly $\alpha$  EW of photometrically selected  $z \sim 2$  p-aLBG and p-eLBG sub-samples with  $c_\sigma = 1.0$  is plotted in the top panel of Fig. 4 compared to the distribution versus net Ly $\alpha$  EW of the parent  $z \sim 2$  LBG sample. In the optimal case, we select sub-samples with a desired Ly $\alpha$  spectral type at  $z \sim 2$  that are for p-aLBGs, comparable to, and for p-eLBGs  $\sim 10\%$  less pure



**Figure 4.** Histograms of p-aLBGs and p-eLBGs are multiplied by 4 for clarity. Histograms versus net Ly $\alpha$  EW of  $z \sim 2$  and  $z \sim 3$  ‘photometric’ aLBG (p-aLBG) and ‘photometric’ eLBG (p-eLBG) spectral type sub-samples overlaid on the distribution versus net Ly $\alpha$  EW of their respective parent samples shown in grey. Vertical dashed lines indicate the net Ly $\alpha$  EW thresholds used here to divide the spectroscopic sample into aLBG, G $_a$ , G $_e$  and eLBG Ly $\alpha$  spectral types. Red and blue shaded regions indicate aLBGs and eLBGs, respectively. Top:  $z \sim 2$  p-aLBGs and p-eLBGs selected from the parent sample of 557  $z \sim 2$  LBGs using the selection criteria given in Equations (1) & (2) with  $c_\sigma = 1.0$ . Histograms of p-aLBGs and p-eLBGs are multiplied by 2 for clarity. Bottom:  $z \sim 3$  p-aLBGs and p-eLBGs selected from the parent sample of 775  $z \sim 3$  LBGs using the selection criteria given in Equations (3) & (4) with  $c_\sigma = 1.5$ .

than, the optimised  $z \sim 3$  result of C09 (see Section 3.3.2). The lower optimised purity of p-eLBGs at  $z \sim 2$  is attributable to the intrinsic overlap of the aLBG and eLBG distributions, and the relatively lower fraction of Ly $\alpha$ -emitting LBGs selected at this redshift. That is, the ratio of aLBGs to eLBGs has increased from around 1:1 at  $z \sim 3$  to more than 4:1 at  $z \sim 2$  when comparing samples to the same absolute magnitude. This is not specific to our sample. On the contrary, a reduced fraction of Ly $\alpha$ -emitting galaxies with decreasing redshift is expected from the findings of Stark et al. (2010), Stark, Ellis, & Ouchi (2011), Mallery et al. (2012), Cassata et al. (2015) who consistently report an evolutionary decrease in the fraction of LBGs with Ly $\alpha$  in emission from  $z \sim 6$  to  $z \sim 2$  at fixed luminosity.

**Table 5.** Statistics for photometric sub-samples with Ly $\alpha$  dominant in absorption (p-aLBGs) and Ly $\alpha$  dominant in emission (p-eLBGs) selected from the parent sample of 775  $z \sim 3$  LBGs using the spectral type criteria of C09<sup>a</sup> and different values of  $c_\sigma$ .

$c_\sigma$ <sup>b</sup>	$N$ <sup>c</sup>	Photometric spectral type				
		p-aLBGs		p-eLBGs		
		Purity <sup>d</sup> (%)	Ly $\alpha$ EW <sup>e</sup> ( $\text{\AA}$ )	$N$	Purity (%)	Ly $\alpha$ EW ( $\text{\AA}$ )
0.0	419	92.6 (64.4)	0.0	356	91.0 (69.1)	+22.4
1.0	200	94.5 (70.5)	-2.9	192	94.3 (74.5)	+27.3
1.5	119	96.6 (76.5)	-5.3	131	95.4 (82.4)	+33.7
2.0	56	98.2 (73.2)	-6.4	91	94.5 (82.4)	+37.7
2.5	20	95.0 (65.0)	-7.3	55	98.2 (92.7)	+51.4

<sup>a</sup>For the purposes of determining photometric segregation criteria, C09 defined aLBGs and eLBGs as having net Ly $\alpha$  EW  $\leq -12.0$  and  $\geq +26.5$   $\text{\AA}$  respectively.

<sup>b</sup>Coefficient of colour standard deviations ( $\sigma_a$  &  $\sigma_e$ ) by which boundaries used to isolate the photometric spectral type sub-samples are offset from the CMD primary cut.

<sup>c</sup>Number of galaxies in the photometric sub-samples.

<sup>d</sup>For p-aLBGs: Percent purity with respect to eLBG and (eLBG + G $_e$ ) spectral types. For p-eLBGs: Percent purity with respect to aLBG and (aLBG + G $_a$ ) spectral types. Mean net Ly $\alpha$  EW for each sub-sample

### 3.3.2. $z \sim 3$ LBGs

For the purposes of reference and direct comparison, we present here the Ly $\alpha$  spectral type photometric selection results for  $z \sim 3$  LBGs analysed and presented in the same format as the  $z \sim 2$  result above.

Parameters for the photometric segregation of  $z \sim 3$  LBG Ly $\alpha$ -absorbing and Ly $\alpha$ -emitting spectral types in  $(G - \mathcal{R})$  vs  $\mathcal{R}$  colour-magnitude space as determined by C09 are listed in Table 3, and we re-produce in Equations (3) & (4) criteria for the photometric selection of p-aLBG and p-eLBG spectral type sub-samples at  $z \sim 3$ .

For p-aLBGs,

$$(G - \mathcal{R}) \geq 0.4047 \cdot \mathcal{R} - 9.3760 + c_\sigma \cdot \sigma_e \quad (3)$$

and for p-eLBGs,

$$(G - \mathcal{R}) \leq 0.4047 \cdot \mathcal{R} - 9.3760 - c_\sigma \cdot \sigma_a \quad (4)$$

where for the parent sample of  $z \sim 3$  LBGs,  $\sigma_a = 0.2392$  and  $\sigma_e = 0.3095$ .

Table 5 summarises the statistics for p-aLBG and p-eLBG Ly $\alpha$  spectral type sub-samples selected from a parent sample of 775  $z \sim 3$  LBGs using the above relationships and a range of  $c_\sigma$  values. The bottom panel of Fig. 4 shows the segregation versus net Ly $\alpha$  EW of  $z \sim 3$  p-aLBG and p-eLBG sub-samples selected with  $c_\sigma = 1.5$  compared to the distribution versus net Ly $\alpha$  EW of the parent  $z \sim 3$  LBG sample.

The optimal coefficient of  $\sigma_e$  ( $c_\sigma \approx 1.5$ ) selects  $\sim 120$  photometric aLBGs that are  $\gtrsim 96\%$  and  $\gtrsim 76\%$  pure with respect to spectroscopic eLBGs, and eLBG + G $_e$  spectral types respectively, and for which the mean net Ly $\alpha$  EW is  $\sim -5$   $\text{\AA}$ .

With any coefficient of  $\sigma_a \gtrsim 1.0$ , we select large samples of photometric eLBGs that are  $\sim 94$ – $98\%$  pure with respect to spectroscopic aLBGs. Over the range  $c_\sigma = 1.0 - 2.5$ , the purity of the p-eLBG sample with respect to all net Ly $\alpha$ -absorbers (aLBG + G $_a$



**Table 6.** Statistics for the segregation of  $z \sim 2$  LBGs in colour ( $U_n - \mathcal{R}$ )–magnitude ( $\mathcal{R}$ ) space over different redshift ranges.

	Redshift range					
	1.7 < z < 2.5		1.7 < z < 2.17		2.17 < z < 2.5	
Total LBGs	557		266		291	
z mean	2.158		1.983		2.318	
$\mathcal{R}$ mean	24.24		24.18		24.29	
$(U_n - \mathcal{R})$ mean	0.96		0.83		1.07	
Net Ly $\alpha$ EW mean						
( $\text{\AA}$ )	-2.15		-3.85		-0.60	
	aLBGs	eLBGs	aLBGs	eLBGs	aLBGs	eLBGs
Number	188	47	96	20	92	27
$\mathcal{R}$ mean	24.17	24.28	24.13	24.21	24.21	24.34
$(U_n - \mathcal{R})$ mean	1.06	0.68	0.91	0.67	1.21	0.69
Colour $\sigma$	0.35	0.36	0.31	0.41	0.32	0.32
Net Ly $\alpha$ EW mean						
( $\text{\AA}$ )	-18.87	+40.57	-19.84	+38.13	-17.86	+42.38
Slope	0.31		0.11		0.16	
Intercept	-6.62		-1.75		-2.89	

spectral types) increases monotonically from  $\sim 74\%$  to  $\sim 93\%$ , with a commensurate increase in mean net Ly $\alpha$  EW from  $\sim +27$  to  $\sim +51 \text{\AA}$ .

### 3.4. The contribution of Ly $\alpha$

The segregation of  $z \sim 3$  aLBGs and eLBGs on the CMD is enhanced by the contribution of the Ly $\alpha$  feature itself when it falls within the bandpass of the G filter (see Section 3.2 and Fig. 2). The contribution of Ly $\alpha$  to the observed luminosities was estimated by C13 to be  $\sim -0.1$  mags for aLBGs,  $\sim +0.1$  mags for eLBGs, and negligible for LBGs with net Ly $\alpha$  EW near zero. A similar effect might be anticipated at  $z \sim 2$  when Ly $\alpha$  falls within the bandpass of the relevant ( $U_n$ ) filter.

Unlike the  $z \sim 3$  case where the Ly $\alpha$  spectral feature lies within the G-band filter across the full redshift range of the sample ( $2.5 < z < 3.5$ ), about half ( $\sim 52\%$ ) of the  $z \sim 2$  LBG sample is in the redshift range  $2.17 \leq z \leq 2.50$ , where the Ly $\alpha$  feature lies outside the half-power bandpass limits of the  $U_n$  filter. Reddy et al. (2008) showed that the ratio of strong emitters to absorbers for LBGs at redshifts  $2.17 \leq z \leq 2.48$  is approximately the same as for those selected by the same set of colour criteria at  $z < 2.17$  (see the respective population statistics in Table 6). Thus, there is no underlying selection bias of aLBGs versus eLBGs that could affect the segregation properties in the different redshift ranges. This result does not, however, preclude the possibility that the segregation statistics across the full  $z$ -range of the sample may be variably affected by the contribution of Ly $\alpha$  to the measured  $U_n$ -band photometry. For this reason – and because the measured segregation at  $z \sim 2$  is less well resolved than at  $z \sim 3$  – we look to quantify the effect of Ly $\alpha$  on the observed broadband segregation for galaxies in the  $z \sim 2$  sample in different redshift ranges and with different Ly $\alpha$  spectral type.

To this end, we divide the  $z \sim 2$  LBG sample into two subsets: one containing only galaxies in the range  $1.7 < z < 2.17$  where Ly $\alpha$  falls within the bandpass of the  $U_n$  filter (3250–3850  $\text{\AA}$ ), and another comprising galaxies in the  $2.17 < z < 2.5$  range for which Ly $\alpha$  lies beyond the red half-power bandpass limit of the same filter (see Fig. 2). We then optimise the primary cut slope in each redshift bin in the same manner as for the sample as a whole (see Section 3.2), and compare the segregation statistics for the two subsets with each other, and with those for the full  $z$ -range sample (see Table 6).

We find a significantly stronger segregation in the  $2.17 < z < 2.5$  sample, most apparent in the greater average colour segregation between aLBGs and eLBGs in this redshift range (0.52) compared to that in the lower redshift bin (0.24). This effect is likely due to the larger contribution of the Ly $\alpha$  forest to the  $U_n$ -band of the  $2.17 < z < 2.5$  sample, leading to redder ( $U_n - \mathcal{R}$ ) colours on average. There is also a larger colour dispersion of eLBGs ( $\sigma = 0.41$ ) in the lower redshift range that blurs the photometric segregation.

This difference in the degree of segregation between aLBGs and eLBGs translates into the purity of p-aLBG and p-eLBG sub-samples that can be selected from the two redshift ranges. Using the same methodology as was applied to the full  $z \sim 2$  and  $z \sim 3$  samples in Section 3.3, we determined the purity of p-aLBG and p-eLBG sub-samples selected from each redshift bin using the segregation parameters listed in Table 6, and a range of  $c_\sigma$  values (see Table 7). The optimised purity of the p-aLBGs and p-eLBGs selected from the  $2.17 < z < 2.5$  sample ( $\sim 98\%$  and  $\sim 94\%$ , respectively) is significantly better than can be achieved from the lower redshift bin ( $\sim 94\%$  for p-aLBGs and  $\sim 84\%$  for p-eLBGs). In fact, with  $c_\sigma$  values of 1.0–1.5, the p-aLBG and p-eLBG sub-samples in the  $2.17 < z < 2.5$  range have purities that are essentially indistinguishable from those achievable in the  $z \sim 3$  sample, and comparably high  $\Delta_{Ly\alpha EW}$  between them, indicating that any direct contribution of the Ly $\alpha$  feature to the segregation properties of aLBGs and eLBGs is dominated by other redshift-dependent spectrophotometric effects such as the one described above.

### 3.5. LSST photometric selection criteria for $z \sim 3$ LBG Ly $\alpha$ spectral types

A key objective of this work is to develop a method that can be applied to large samples of  $z \sim 2 - 6$  LBGs identified from current and future large-area and all-sky photometric campaigns. As a first step toward this goal, we adapt the spectrophotometric method of C13 to model photometric selection criteria by which populations of  $z \sim 3$  LBGs with Ly $\alpha$  dominant in absorption and Ly $\alpha$  dominant in emission might be selected from the broadband *ugri* photometric data of the Vera Ruben Observatory Legacy Survey of Space and Time (VRO/LSST).

In order to model the segregation statistics for different Ly $\alpha$  spectral types in the LSST *ugri* photometric system, it is first necessary to calculate *ugri* magnitudes for each galaxy in our  $z \sim 3$  sample. Our photometric segregation method relies on the fact that LBGs with different net Ly $\alpha$  EW have different spectral properties – in particular rest-frame UV continuum slope – that give rise to different rest-frame UV colours depending on their Ly $\alpha$  absorbing/emitting properties (see Fig. 2). Thus, in order to convert from  $U_nGR$  to *ugri* magnitudes via spectrophotometry,

**Table 7.** Statistics for p-aLBG and p-eLBG sub-samples photometrically selected from the parent  $z \sim 2$  LBG sample using segregation parameters optimised in different redshift ranges.

Redshift range : $1.70 < z < 2.17$							
p-aLBGs				p-eLBGs			
$c_\sigma^a$	$N^b$	Purity <sup>c</sup> (%)	Ly $\alpha$ EW <sup>d</sup> (Å)	$N$	Purity	Ly $\alpha$ EW	$\Delta_{Ly\alpha EW}^e$
					(%)	(Å)	
0.00	142	96.5 (71.8)	-7.5	126	71.4 (46.8)	0.1	7.6
0.25	113	96.5 (71.7)	-6.8	97	79.4 (54.6)	1.5	8.3
0.50	82	95.1 (69.5)	-6.7	75	82.7 (56.0)	2.8	9.4
0.75	50	94.0 (62.0)	-6.4	57	84.2 (57.9)	3.8	10.2
1.00	32	93.8 (75.0)	-9.7	38	81.6 (44.7)	0.4	10.1
1.25	21	90.5 (71.4)	-4.6	23	78.3 (52.2)	1.0	5.6
1.50	13	84.6 (61.5)	0.0	16	68.8 (43.8)	-1.1	1.1
2.00	5	100.0 (80.0)	-14.5	6	66.7 (66.7)	-3.5	11.1

Redshift range : $2.17 < z < 2.50$							
p-aLBGs				p-eLBGs			
$c_\sigma$	$N$	Purity (%)	Ly $\alpha$ EW (Å)	$N$	Purity	Ly $\alpha$ EW	$\Delta_{Ly\alpha EW}$
					(%)	(Å)	
0.00	174	98.9 (70.7)	-6.0	115	80.9 (63.5)	7.7	13.7
0.25	157	98.7 (72.6)	-6.5	98	84.7 (67.3)	9.7	16.2
0.50	131	99.2 (77.1)	-8.2	76	84.2 (73.7)	11.1	19.3
0.75	107	99.1 (75.7)	-7.9	52	90.4 (80.8)	13.8	21.8
1.00	84	98.8 (78.6)	-7.8	31	93.5 (87.1)	19.9	27.7
1.25	65	98.5 (81.5)	-8.6	19	94.7 (94.7)	19.4	28.0
1.50	49	98.0 (79.6)	-8.2	12	100.0 (100.0)	25.3	33.5
2.00	25	96.0 (76.0)	-6.0	2	100.0 (100.0)	10.5	16.4

<sup>a</sup>Coefficient of colour standard deviations by which boundaries used to isolate the photometric spectral type sub-samples are offset from the primary cut in each redshift range.

<sup>b</sup>Number of galaxies in the photometric sub-samples.

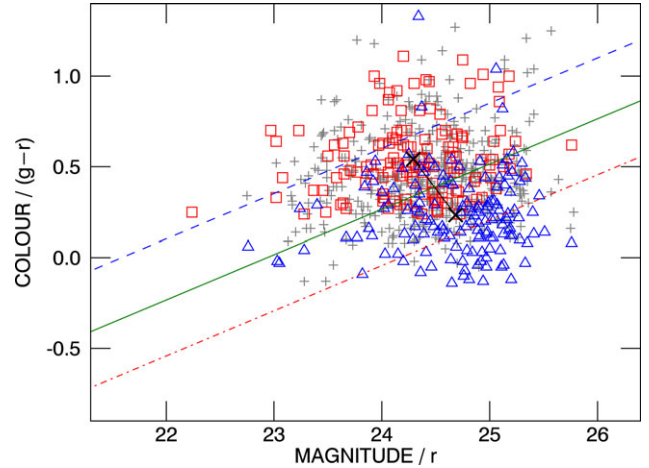
<sup>c</sup>For p-aLBGs: Percent purity with respect to eLBG and (eLBG +  $G_e$ ) spectral types. For p-eLBGs: Percent purity with respect to aLBG and (aLBG +  $G_a$ ) spectral types.

<sup>d</sup>Mean net Ly $\alpha$  EW for each sub-sample.

<sup>e</sup>Difference in mean net Ly $\alpha$  EW between the p-aLBG and p-eLBG sub-samples. In the case of Full<sub>bins</sub>,  $\Delta_{Ly\alpha EW}$  is the weighted average of  $\Delta_{Ly\alpha EW}$  values for each magnitude bin.

we must be able to assign to each galaxy in our sample, an appropriate spectrum corresponding to its Ly $\alpha$  spectral type. For this purpose, we make use of the four composite spectra of Shapley et al. (2003), derived from the  $z \sim 3$   $U_nGR$  LBGs described in Sections 2 & 3.1 divided into quartiles on the basis of net Ly $\alpha$  EW. C13 showed that spectrophotometry of these composite spectra accurately reproduces the magnitude and colour means and dispersions of each of the four net Ly $\alpha$  EW quartile samples, as well as the full distribution of  $z \sim 3$  LBGs on the  $(G - \mathcal{R})/G$  CMD when combined. Thus, although the colours and magnitudes of individual galaxies vary within each quartile, the composite spectra can be used to compute net Ly $\alpha$  EW means and dispersions on the CMD for our  $z \sim 3$  LBG sample when viewed through the VRO/LSST *ugri* filters<sup>a</sup>. Fig. 5 shows our  $z \sim 3$   $U_nGR$  LBG sample dispersed in

<sup>a</sup>LSST filter bandpasses and throughputs (31 May 2021 updates) downloaded from: <https://github.com/lst/throughputs/tree/main/baseline>.



**Figure 5.** Rest-frame UV  $(g - r)$  vs  $r$  colour-magnitude diagram (CMD) for Steidel et al. (2003)  $z \sim 3$  LBGs, constructed with VRO/LSST *ugri* photometry derived from the net Ly $\alpha$  EW quartile composite spectra of Shapley et al. (2003). The CMD shows the segregation with net Ly $\alpha$  EW of galaxies with Ly $\alpha$  dominant in absorption (aLBGs, red squares) and Ly $\alpha$  dominant in emission (eLBGs, blue triangles). Grey plus (+) symbols denote galaxies with intermediate values of net Ly $\alpha$  EW. Black crosses mark the mean positions of the aLBG and eLBG distributions. The dashed red and dotted-dashed blue lines indicate a  $1.5\sigma$  dispersion in colour from the primary cut (green line) for the aLBG and eLBG distributions, respectively.

**Table 8.** Statistics for the segregation of LSST  $z \sim 3$  LBG Ly $\alpha$  spectral types in  $(g - r)/r$  colour-magnitude space.

	Ly $\alpha$ spectral type <sup>a</sup>	
	aLBGs	eLBGs
Mag. mean	24.29	24.68
Mag. $1\sigma$	0.54	0.54
Col. mean	0.54	0.23
Col. $1\sigma$	0.21	0.22
Slope	0.25	
Intercept	-5.72	

<sup>a</sup>Consistent with C09, we define aLBGs and eLBGs as having net Ly $\alpha$  EW  $\leq -12.0$  and  $\geq +26.5$  Å respectively, for the purposes of constructing the CMD and determining photometric segregation criteria.

VRO/LSST  $(g - r)$  vs  $r$  colour-magnitude space with photometry thus derived from spectrophotometry of the net Ly $\alpha$  EW quartile composite spectra.

Following the method described in Section 3.3, we use the parameters of the segregated aLBG and eLBG distributions shown in Fig. 5 and tabulated in Table 8 to determine photometric selection criteria by which pure populations of  $z \sim 3$  LBGs with Ly $\alpha$  dominant in absorption (p-aLBGs; Equation (5)) and Ly $\alpha$  dominant in emission (p-eLBGs; Equation (6)), might be selected from VRO/LSST LBG data dispersed in  $(g - r)$  vs  $r$  space.

Specifically, for p-aLBGs:

$$(g - r) \geq 0.25 \cdot r - 5.72 + c_\sigma \cdot \sigma_e \quad (5)$$

and for p-eLBGs:

$$(g - r) \leq 0.25 \cdot r - 5.72 - c_\sigma \cdot \sigma_a \quad (6)$$

where  $c_\sigma$ ,  $\sigma_a$ , and  $\sigma_e$  are the coefficient and respective standard deviations of colour dispersion as described in Section 3.3.1.

These selection criteria provide a useful starting point for the isolation of p-aLBG and p-eLBG populations from LSST photometry. They will be confirmed and/or refined, and selection criteria in other redshift ranges added – especially at  $z \sim 2$  – once LSST data of sufficient depth has been measured in fields within which Ly $\alpha$  spectroscopic data are available.

#### 4. Summary and conclusions

The Ly $\alpha$  observables from a given galaxy are known to be sensitive to a wide range of galactic physical, spectral, and environmental properties. Net Ly $\alpha$  EW in particular has been shown to correlate with, for example, galaxy morphology, rest-frame UV colour, ISM line strengths, gas kinematics, and the large-scale spatial distribution of populations of  $z \gtrsim 2$  LAEs and LBGs. Accordingly, the ability to select pure statistical sub-samples of a desired Ly $\alpha$  spectral type from large photometric datasets facilitates the study of a variety of intrinsic and small- to large-scale environmental galactic properties that are related to Ly $\alpha$ , in large numbers and over distance scales for which ancillary multi-wavelength/multi-band photometry and/or spectroscopic information is not usually available (Cooke 2009; Cooke et al. 2013).

In this paper we characterise the broadband imaging segregation of a spectroscopic sample of 557  $z \sim 2$  LBGs using sub-samples with Ly $\alpha$  dominant in absorption (aLBGs), and Ly $\alpha$  dominant in emission (eLBGs), and determine photometric criteria by which relatively pure sub-samples with desired Ly $\alpha$  spectral properties can be selected using imaging data in as few as three optical broadband filters.

We draw the following specific conclusions from our study:

- $z \sim 2$  LBGs segregate according to their net Ly $\alpha$  EW properties in rest-frame UV colour ( $U_n - \mathcal{R}$ ) and magnitude ( $\mathcal{R}$ ) space in a manner similar to their  $z \sim 3$  counterparts in the  $(G - \mathcal{R})/\mathcal{R}$  plane (see Section 3.2 and cf. Cooke 2009).
- Using the segregation statistics for our sample of 557 LBGs in the range  $1.7 < z < 2.5$ , we determine photometric criteria for the selection of sub-samples of LBGs with Ly $\alpha$  dominant in absorption (p-aLBGs) and Ly $\alpha$  dominant in emission (p-eLBGs). These criteria select sub-samples of p-aLBGs and p-eLBGs that are respectively  $\gtrsim 97\%$  and  $\sim 85\%$  pure with respect to contamination by galaxies with the opposite spectral type. The mean net Ly $\alpha$  EW of the optimised p-aLBG and p-eLBG sub-samples selected from the  $z \sim 2$   $U_n GR$  LBGs is  $\sim -8 \text{ \AA}$  and  $\sim +8 \text{ \AA}$ , respectively (Section 3.3.1).
- Sub-dividing the  $z \sim 2$  sample into two redshift bins, we find that the degree of photometric segregation in the range  $2.17 < z < 2.5$  (Ly $\alpha$  outside the  $U_n$  filter) is significantly greater than in the range  $1.70 < z < 2.17$  (Ly $\alpha$  within the  $U_n$  filter). We attribute this difference to a larger contribution of the Ly $\alpha$  forest leading to greater dispersion in  $(U_n - \mathcal{R})$  colour at higher redshifts. In the range  $2.17 < z < 2.5$ , we select sub-samples of p-aLBGs and p-eLBGs that are  $\gtrsim 95\%$  pure with respect to galaxies of the opposite spectral type, and which segregate in mean net Ly $\alpha$  EW ( $\Delta_{Ly\alpha EW} \approx 30 \text{ \AA}$ ) on the same order as the  $z \sim 3$  LBGs (Section 3.4).
- Using the result of C09 and spectrophotometry of the composite spectra of  $z \sim 3$  LBGs with different Ly $\alpha$  spectral type, we

calculate photometric criteria by which populations of p-aLBGs and p-eLBGs can be selected from the *ugri* broadband imaging of the LSST (Section 3.5).

One motivation for this work is to provide the statistical foundation for application of the result described in Paper II in this series (Foran et al. 2023b, submitted) wherein we report a relationship between net Ly $\alpha$  EW and galaxy kinematics, and demonstrate how the photometric segregation described here can be used to predict the kinematic type (and other properties) of large numbers of  $z \sim 2 - 3$  LBGs without the need for spectroscopic information.

More broadly, we propose that this method has strong potential to expand the legacy value of the current generation of deep, wide, optical and near-infrared, large-area and all-sky photometric campaigns such as the Hyper-SuprimeCam Subaru Strategic Program (HSC-SSP: Aihara et al. 2018) and the upcoming Vera Rubin Observatory Legacy Survey of Space and Time (LSST: Ivezić et al. 2019) that will exploit the Lyman break technique using 3–5 broadband filters across the rest-frame UV to select hundreds of millions of galaxies in redshift ranges from  $z \sim 2 - 6$  across many hundreds to thousands of Mpc. Optimising the discovery potential of such programs requires new techniques to statistically characterise such huge datasets, and to efficiently select from these the most promising samples for expensive follow-up observations. The techniques and insights presented here and in Paper II, explore how inexpensive broadband photometric information that is sensitive to the Ly $\alpha$  properties of LBGs might address this challenge. This approach also provides a statistical framework within which  $z \sim 2 - 3$  LBGs will serve as low-redshift reference samples for the study of galaxy populations at higher redshifts where only selection methods based on Ly $\alpha$  emission or Lyman break detection can be applied in large numbers and over large scales (Finkelstein 2016).

Specific applications of this approach might include:

- study of the environments of Ly $\alpha$  absorbers and emitters on small and large scales out to hundreds and thousands of Mpc;
- generation of the large samples of Ly $\alpha$  absorbers and emitters required for three-point correlation function analysis, whereby the geometry, spatial shape, and distribution of the different spectral types might be mapped relative to the filaments and nodes of the cosmic web;
- investigation of the origins and character of the morphology–density relation at  $z \sim 2$  and beyond;
- furnishing of the kinematic properties of large numbers of early galaxies of known Ly $\alpha$  spectral-type to aid halo-matching between observations and cosmological simulations; and
- cosmological studies in which tailored samples of  $z \sim 2 - 5$  LBGs with varying Ly $\alpha$  EW are used in combination with cosmic microwave background lensing cross-correlation analysis, to infer the time evolution of matter-density fluctuations, and to carry out compelling tests of horizon-scale general relativity, neutrino masses, and inflation (e.g., Wilson & White 2019).

As a first step toward these goals, we present here photometric criteria by which populations of  $z \sim 3$  LBGs with Ly $\alpha$  dominant in absorption, and Ly $\alpha$  dominant in emission, might be selected from *ugri* photometric data from the LSST (Section 3.5).

## References

- Adelberger, K. L., et al. 2005, *ApJ*, **619**, 697  
 Adelberger, K. L., et al. 2004, *ApJ*, **607**, 226  
 Aihara, H., et al. 2018, *PASJ*, **70**, S4  
 Álvarez-Márquez, J., et al. 2016, *A&A*, **587**, A122  
 Arcila-Osejo, L., & Sawicki, M. 2013, *MNRAS*, **435**, 845  
 Arrabal Haro, P., et al. 2018, *MNRAS*, **478**, 3740  
 Basu-Zych, A. R., Hornschemeier, A. E., Hoversten, E. A., Lehmer, B., & Gronwall, C. 2011, *ApJ*, **739**, 98  
 Berry, M., et al. 2012, *ApJ*, **749**, 4  
 Bielby, R. M., et al. 2011, *MNRAS*, **414**, 2  
 Bielby, R. M., et al. 2016, *MNRAS*, **456**, 4061  
 Bouwens, R. J., Illingworth, G. D., Blakeslee, J. P., & Franx, M. 2006, *ApJ*, **653**, 53  
 Bouwens, R. J., et al. 2009, *ApJ*, **705**, 936  
 Bouwens, R. J., et al. 2010, *ApJ*, **709**, L133  
 Bouwens, R. J., et al. 2015, *ApJ*, **803**, 34  
 Burgarella, D., et al. 2006, *A&A*, **450**, 69  
 Byrohl, C., & Gronke, M. 2020, arXiv e-prints, arXiv:2006.10041  
 Cassata, P., et al. 2015, *A&A*, **573**, A24  
 Chen, Y., et al. 2020, arXiv e-prints, arXiv:2006.13236  
 Cooke, J. 2009, *ApJ*, **704**, L62  
 Cooke, J., Berrier, J. C., Barton, E. J., Bullock, J. S., & Wolfe, A. M. 2010, *MNRAS*, **403**, 1020  
 Cooke, J., Omori, Y., & Ryan-Weber, E. V. 2013, *MNRAS*, **433**, 2122  
 Cullen, F., et al. 2020, *MNRAS*, **495**, 1501  
 Daddi, E., et al. 2007, *ApJ*, **670**, 156  
 Daz, C. G., et al. 2014, *MNRAS*, **442**, 946  
 Dijkstra, M. 2014, *PASA*, **31**, e040  
 Dijkstra, M. 2017, arXiv e-prints, arXiv:1704.03416  
 Dijkstra, M., & Wyithe, J. S. B. 2010, *MNRAS*, **408**, 352  
 Du, X., et al. 2018, *ApJ*, **860**, 75  
 Duval, F., et al. 2016, *A&A*, **587**, A77  
 Ellis, R. S., et al. 2013, *ApJ*, **763**, L7  
 Erb, D. K., et al. 2016, *ApJ*, **830**, 52  
 Erb, D. K., et al. 2006 a, *ApJ*, **644**, 813  
 Erb, D. K., et al. 2006 b, *ApJ*, **646**, 107  
 Feltre, A., et al. 2020, *A&A*, **641**, A118  
 Finkelstein, S. L. 2016, *PASA*, **33**, e037  
 Giavalisco, M., et al. 2004, *ApJ*, **600**, L103  
 Grazian, A., et al. 2007, *A&A*, **465**, 393  
 Gronke, M., & Dijkstra, M. 2016, *ApJ*, **826**, 14  
 Guaita, L., et al. 2015, *A&A*, **576**, A51  
 Guaita, L., et al. 2017, *A&A*, **606**, A19  
 Guaita, L., et al. 2020, *A&A*, **640**, A107  
 Habertzell, L., Williger, G., Lehnert, M. D., Nesvadba, N., & Davies, L. 2012, *ApJ*, **745**, 96  
 Harikane, Y., et al. 2018, *PASJ*, **70**, S11  
 Harikane, Y., et al. 2022 a, *ApJ*, **929**, 1  
 Harikane, Y., et al. 2022 b, *ApJS*, **259**, 20  
 Hathi, N. P., et al. 2013, *ApJ*, **765**, 88  
 Hathi, N. P., et al. 2016, *A&A*, **588**, L26  
 Hayes, M. 2015, *PASA*, **32**, e027  
 Hayes, M., et al. 2014, *ApJ*, **782**, 6  
 Herenz, E. C., et al. 2016, *A&A*, **587**, A78  
 Ilbert, O., et al. 2013, *A&A*, **556**, A55  
 Ivezić, Ž., et al. 2019, *ApJ*, **873**, 111  
 Iwata, I., et al. 2007, *MNRAS*, **376**, 1557  
 Jones, T., Stark, D. P., & Ellis, R. S. 2012, *ApJ*, **751**, 51  
 Jose, C., Srianand, R., & Subramanian, K. 2013, *MNRAS*, **435**, 368  
 Kornei, K. A., et al. 2010, *ApJ*, **711**, 693  
 Law, D. R., et al. 2007, *ApJ*, **656**, 1  
 Law, D. R., et al. 2012 a, *ApJ*, **759**, 29  
 Law, D. R., et al. 2012 b, *ApJ*, **745**, 85  
 Lemaux, B. C., et al. 2018, *A&A*, **615**, A77  
 Ly, C., et al. 2011, *ApJ*, **735**, 91  
 Ly, C., et al. 2009, *ApJ*, **697**, 1410  
 Madau, P., & Dickinson, M. 2014, *ARA A*, **52**, 415  
 Malkan, M. A., et al. 2017, *ApJ*, **850**, 5  
 Mallery, R. P., et al. 2012, *ApJ*, **760**, 128  
 Marchi, F., et al. 2019, *A&A*, **631**, A19  
 Mason, C. A., et al. 2018, *ApJ*, **857**, L11  
 Matthee, J., et al. 2021, arXiv e-prints, arXiv:2102.07779  
 McLure, R. J., et al. 2011, *MNRAS*, **418**, 2074  
 Miyatake, H., et al. 2022, *PhRvL*, **129**, 061301  
 Muldrew, S. I., Hatch, N. A., & Cooke, E. A. 2015, *MNRAS*, **452**, 2528  
 Muzzin, A., et al. 2013, *ApJ*, **777**, 18  
 Oke, J. B., & Gunn, J. E. 1983, *ApJ*, **266**, 713  
 Ono, Y., et al. 2018, *PASJ*, **70**, S10  
 Östlin, G., et al. 2014, *ApJ*, **797**, 11  
 Oteo, I., et al. 2015, *MNRAS*, **452**, 2018  
 Oteo, I., et al. 2013 a, *A&A*, **554**, L3  
 Oteo, I., et al. 2013 b, *MNRAS*, **433**, 2706  
 Oteo, I., et al. 2014, *MNRAS*, **439**, 1337  
 Ouchi, M., Ono, Y., & Shibuya, T. 2020, *ARA&A*, **58**, 617  
 Ouchi, M., et al. 2004, *ApJ*, **611**, 660  
 Ouchi, M., et al. 2010, *ApJ*, **723**, 869  
 Ouchi, M., et al. 2018, *PASJ*, **70**, S13  
 Oyarzún, G. A., Blanc, G. A., González, V., Mateo, M., & Bailey, JOHN I., I. 2017, *ApJ*, **843**, 133  
 Pahl, A. J., et al. 2020, *MNRAS*, **493**, 3194  
 Pardy, S. A., et al. 2014, *ApJ*, **794**, 101  
 Pentericci, L., et al. 2010, *A&A*, **514**, A64  
 Reddy, N. A., & Steidel, C. C. 2009, *ApJ*, **692**, 778  
 Reddy, N. A., Steidel, C. C., Erb, D. K., Shapley, A. E., & Pettini, M. 2006, *ApJ*, **653**, 1004  
 Reddy, N. A., et al. 2008, *ApJS*, **175**, 48  
 Rivera-Thorsen, T. E., et al. 2015, *ApJ*, **805**, 14  
 Runnholm, A., et al. 2020, *ApJ*, **892**, 48  
 Santos, S., et al. 2020, *MNRAS*, **493**, 141  
 Shapley, A. E. 2011, *ARA&A*, **49**, 525  
 Shapley, A. E., et al. 2005, *ApJ*, **626**, 698  
 Shapley, A. E., Steidel, C. C., Pettini, M., & Adelberger, K. L. 2003, *ApJ*, **588**, 65  
 Shi, K., et al. 2019, *ApJ*, **879**, 9  
 Stark, D. P., Ellis, R. S., Chiu, K., Ouchi, M., & Bunker, A. 2010, *MNRAS*, **408**, 1628  
 Stark, D. P., Ellis, R. S., & Ouchi, M. 2011, *ApJ*, **728**, L2  
 Stark, D. P., et al. 2017, *MNRAS*, **464**, 469  
 Steidel, C. C., Adelberger, K. L., Giavalisco, M., Dickinson, M., & Pettini, M. 1999, *ApJ*, **519**, 1  
 Steidel, C. C., et al. 2003, *ApJ*, **592**, 728  
 Steidel, C. C., et al. 2018, *ApJ*, **869**, 123  
 Steidel, C. C., et al. 2010, *ApJ*, **717**, 289  
 Steidel, C. C., et al. 2004, *ApJ*, **604**, 534  
 Toshikawa, J., et al. 2016, *ApJ*, **826**, 114  
 Toshikawa, J., et al. 2018, *PASJ*, **70**, S12  
 Trainor, R. F., Steidel, C. C., Strom, A. L., & Rudie, G. C. 2015, *ApJ*, **809**, 89  
 Trainor, R. F., Strom, A. L., Steidel, C. C., & Rudie, G. C. 2016, *ApJ*, **832**, 171  
 Trainor, R. F., et al. 2019, *ApJ*, **887**, 85  
 Verhamme, A., Schaerer, D., Atek, H., & Tapken, C. 2008, *A&A*, **491**, 89  
 Verhamme, A., Schaerer, D., & Maselli, A. 2006, *A&A*, **460**, 397  
 Verma, A., Lehnert, M. D., Förster Schreiber, N. M., Bremer, M. N., & Douglas, L. 2007, *MNRAS*, **377**, 1024  
 Wilson, M. J., & White, M. 2019, *JCAP*, **2019**, 015

## Appendix 1. Photometric Uncertainties

As part of corrections for photometric incompleteness in their study of the rest-frame UV luminosity function at  $z \sim 1.9 - 3.4$ , Reddy et al. (2008) applied a Monte Carlo (MC) galaxy population simulation method to joint photometric and spectroscopic samples of  $z \sim 2$   $U_nGR$  LBGs to assess the systematic effects of



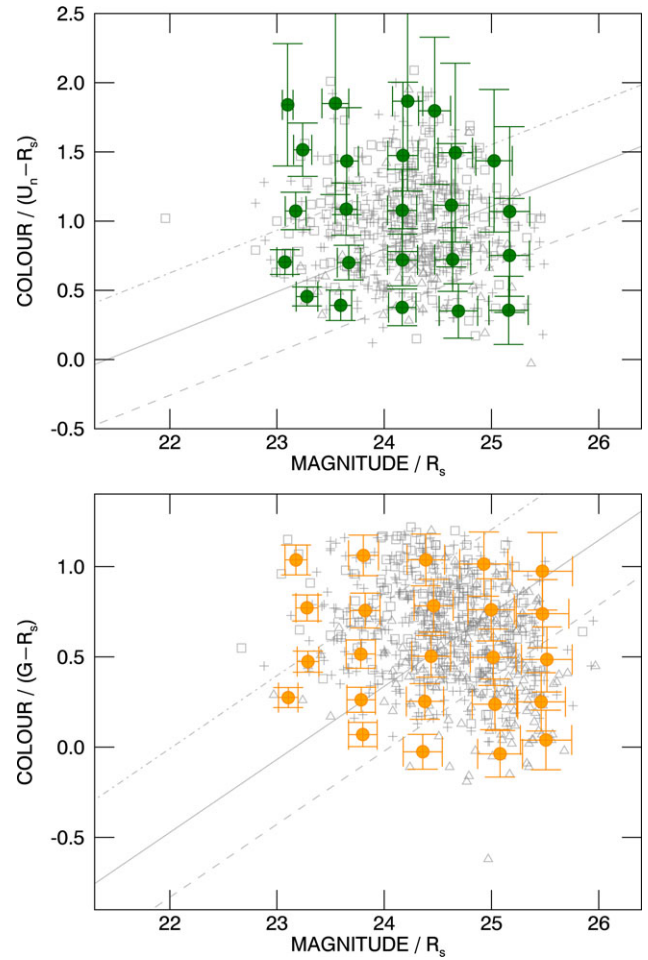
photometric scatter and the intrinsic variation in colours due to Ly $\alpha$  line emission and absorption. These simulations yielded statistical estimates of the photometric uncertainties for the imaging data used to select the  $z \sim 2$  and  $z \sim 3$   $U_nGR$  LBG samples used in this work. Tables of these uncertainties were supplied (N. Reddy, private communication) in 0.5 mag bins of  $G$  and  $\mathcal{R}$  magnitude and 0.2 mag bins of  $(U_n - G)$  and  $(G - \mathcal{R})$  colour for all observed fields in the  $z \sim 2$   $U_nGR$  survey. In the absence of source-by-source photometric errors, we calculated from the MC simulation data indicative photometric uncertainties for the  $z \sim 2$  and  $z \sim 3$  LBGs dispersed in  $(U_n - \mathcal{R})/\mathcal{R}$  and  $(G - \mathcal{R})/\mathcal{R}$  colour/magnitude space, respectively.

For the  $z \sim 2$  LBGs,  $\mathcal{R}$ -band and  $(G - \mathcal{R})$  uncertainties for each galaxy were extracted from the MC simulation tables for the relevant field according to their observed  $\mathcal{R}$ -band luminosity and  $(G - \mathcal{R})$  colour, and added in quadrature to give calculated estimates of  $G$ -band uncertainty. These  $G$ -band uncertainties were in turn added in quadrature with the tabulated  $(U_n - G)$  uncertainties to give an estimate of the  $U_n$ -band uncertainty for each galaxy. Finally,  $(U_n - \mathcal{R})$  uncertainties were estimated by subtracting in quadrature the  $\mathcal{R}$ -band uncertainty from that of the  $U_n$ -band.

For  $z \sim 3$  LBGs in the three fields where the  $z \sim 2$  and  $z \sim 3$  catalogs overlap (i.e., HDF/GOODS-N, Q0933 and Q1422)  $\mathcal{R}$ -band and  $(G - \mathcal{R})$  uncertainties for each galaxy were extracted similarly to the  $z \sim 2$  sample. For all other  $z \sim 3$  LBGs,  $\mathcal{R}$ -band and  $(G - \mathcal{R})$  uncertainties were estimated by averaging values for the HDF/GOODS-N, Q0933 and Q1422 fields at the relevant luminosity and colour.

For the purpose of illustrating the representative photometric uncertainties thus calculated, the  $z \sim 2$  and  $z \sim 3$  LBG samples were divided into a  $5 \times 5$  grid on their respective CMDs. Fig. A.1 shows the mean colour, magnitude, and associated uncertainties, for the galaxies in each grid element overlaid on the full sample for both redshift ranges.

The estimated typical  $\mathcal{R}$ -band uncertainty of  $\lesssim 0.2$  up to the  $\mathcal{R} = 25.5$  limit, gives confidence for the use of the  $z \sim 2$   $U_nGR$  LBG sample in our analysis. On the other hand, the estimated uncertainties in  $(U_n - \mathcal{R})$  suggest that  $U_n$ -band magnitudes fainter than  $\sim 26.0$ – $26.5$  introduce photometric errors  $\gtrsim 0.5$  that are potentially problematic for the colour-magnitude segregation approach investigated here. Given our need for reliable  $(U_n - G)$  and/or  $(U_n - \mathcal{R})$  colours, and in the absence of source-by-source



**Figure A.1.** Indicative photometric uncertainties for  $z \sim 2$  (top) and  $z \sim 3$  (bottom)  $U_nGR$  LBGs dispersed in colour-magnitude space and divided into a  $5 \times 5$  grid on the CMD. The green and orange symbols indicate the mean colour, magnitude, and associated uncertainties for the galaxies in each grid element. In each case, the representative symbols are overlaid on their respective full sample (grey symbols).

photometric uncertainties, the estimates derived from the MC simulations motivated the decision to limit the  $z \sim 2$  sample to galaxies with  $U_n$ -band magnitudes  $\leq 26.5$ .