Galaxy Tagging: photometric redshift refinement and group richness enhancement

P. R. Kafle,[1⋆] A. S. G. Robotham,[1] S. P. Driver,[1] S. Deeley,[2,3] P. Norberg,[4] M. J. Drinkwater,[2] and L. J. Davies [1]

[1] *ICRAR, The University of Western Australia, 35 Stirling Highway, Crawley, WA 6009, Australia*
[2] *School of Mathematics and Physics, University of Queensland, Brisbane, Queensland 4072, Australia*
[3] *ARC Centre of Excellence for All-Sky Astrophysics (CAASTRO)*
[4] *Institute for Computational Cosmology, Department of Physics, Durham University, South Road, Durham DH1 3LE, UK*

16 June 2018

ABSTRACT

We present a new scheme, *galtag*, for refining the photometric redshift measurements of faint galaxies by probabilistically tagging them to observed galaxy groups constructed from a brighter, magnitude-limited spectroscopy survey. First, this method is tested on the DESI light-cone data constructed on the GALFORM galaxy formation model to tests its validity. We then apply it to the photometric observations of galaxies in the Kilo-Degree Imaging Survey (KiDS) over a 1 deg^2 region centred at 15^h. This region contains Galaxy and Mass Assembly (GAMA) deep spectroscopic observations (i-band<22) and an accompanying group catalogue to r-band<19.8. We demonstrate that even with some trade-off in sample size, an order of magnitude improvement on the accuracy of photometric redshifts is achievable when using *galtag*. This approach provides both refined photometric redshift measurements and group richness enhancement. In combination these products will hugely improve the scientific potential of both photometric and spectroscopic datasets. The *galtag* software will be made publicly available at https://github.com/pkaf/galtag.git.

Key words: galaxies: general – galaxies: haloes – galaxies: groups: general

1 INTRODUCTION

Fundamental to many core aspects of galaxy evolution science is the precise and accurate measurement of the distances to galaxies using redshifts. There are two largely distinct methods for obtaining these redshifts, either using spectroscopically-identified emission and absorption line features (spectroscopic redshift, z_s) or via observed broad-band colours matched to a library of spectral templates targeting the large-scale continuum shape (photometric redshift, z_p). Due to the nature of spectroscopic observations, the former is more precise, but much more observationally costly than the latter. Thus, photometric redshifts can sample orders of magnitude more galaxies for a similar investment of telescope time, but to a lower fidelity. The trade off between sample size and precision when measuring galaxy redshifts, is largely decided based on the specific scientific question being posed (i.e large sample size photometric redshifts for cosmology vs small sample high precision spectroscopic redshifts for group and pair science). However, over the last decade there have been vast improvements in the precision of our photometric redshifts based on improved templates, deep and larger are imaging surveys and improvements to photometry fitting algorithms. This has led to photometric redshifts becoming big business in the field of galaxy formation and evolution, (e.g. Budavári 2009; Carliles et al. 2010; Budavári 2012; Dahlen et al. 2013; Graham et al. 2017, etc), with survey teams pursuing ever more sophisticated approaches to increase the precision of redshift measurements derived from photometry alone.

The different approaches of z_p measurement can be broadly classified into four categories which we discuss below:

(i) spectral energy distributions (SED)/template fitting technique,
(ii) machine learning approach using training and test data,
(iii) moment based clustering, and
(iv) inference from cosmic web constraints.

Thus far, the most commonly used technique in z_p estimation is the template fitting methods. In this method given a library of reference galaxy spectra one fits the

⋆ E-mail: prajwal.kafle@uwa.edu.au; pkafauthor@gmail.com

observed broadband photometry of a galaxy to find the best fit reference spectra to solve for the redshift. The completeness of the template and the imperfect observed fluxes due to biases such as disparate zero-point errors in different photometric bands or underestimated errors limits the use of this method. An advantage of this method is that it provides fully probabilistic treatment to the redshift measurement, allowing to impose priors over the different types, that can further be a function of redshift, of galaxies. Baum (1962); Loh & Spillar (1986); Connolly et al. (1995); Brunner et al. (1997); Furusawa et al. (2000); Benítez (2000); Bolzonella et al. (2000); Fontana et al. (2000); Le Borgne & Rocca-Volmerange (2002); Brammer et al. (2008); Ilbert et al. (2009); Hildebrandt et al. (2012); Laigle et al. (2016) etc are some examples of this category.

In the machine learning approach, first, an empirical model relating galaxy fluxes with redshifts is constructed over the training (trustworthy) data for which the exact redshift is already known. The trained (predictive) model is then run to predict the redshift of the remaining galaxies (target data). With the ever increasing efficiency of computers, as well as due to the surge of the spectroscopic spectra from different observational campaigns boosting the sample size of the training data, the machine learning approach has gained more popularity recently. An advantage of this method is that during the training phase the model learns the complicated relationships within the observables (e.g. fluxes as a function of redshift which is further a function of galaxy types and so on so forth) that is naturally propagated to the final redshift estimation. Firth et al. (2003); Wolf (2009); Budavári (2009); Bonfield et al. (2010); Sadeh et al. (2016); Leistedt & Hogg (2017); Cavuoti et al. (2017); Bilicki et al. (2017) etc are some examples of this category.

In the moment based clustering approach, the position of galaxies in physical space and their proximity to large scale structures of the cosmic web are utilised to constrain the redshifts of galaxies. The applicability of this approach has been limited due to lack of enough overlap between appropriate $z_{\rm s}$ samples and photometric ones, but where there is overlap it is found to yield good constraint on $z_{\rm p}$ (Morrison et al. 2017; Hildebrandt et al. 2017). This approach is not a stand-alone technique to measure $z_{\rm p}$, but more of the ancillary approach to calibrate redshift distribution or to further refine the already measured redshifts. Matthews & Newman (2010); Rahman et al. (2016); Morrison et al. (2017) etc are a few examples of this category.

The last category uses the distribution of the large scale structure of the cosmic web to directly inform the plausible radial positions of galaxies with photometric redshifts (see Kovač et al. 2010; Aragon-Calvo et al. 2015). Of the four techniques discussed here, this family of methods offers the most dramatic refinement possibilities, although it is also the most expensive in terms of data requirements. The method we propose in this paper broadly falls into this category, where we will refine the pre-measured $z_{\rm p}$ using out prior knowledge of the galaxy group distribution rather than the more diffuse cosmic web.

In this paper we describe a complete implementation of photometric redshift refinement method and present the results of applying the technique to realistic mock catalogues as well as observed data as a proof of concept. Throughout the paper, we assume a flat ΛCDM cosmology with $\Omega_m = 0.3$, $\Omega_\Lambda = 0.7$, and Hubble parameter $H_0 = 100\,h\,{\rm km\,s^{-1}, Mpc^{-1}}$, where we have assumed $h = 1$. This paper is organized as follows. In Section 2, we describe the GAMA (Galaxy and Mass Assembly) and KiDS (Kilo Degree Survey) observational data as well as the DESI mock catalogue that are used to test our method. In section 3 we outline the halo based prior that is essentially adopted from the MAGGIE (Models and Algorithms for Galaxy Groups, Interlopers and Environment, Duarte & Mamon 2015), and the redshift refinement method. In section 4 we show the method in-action. Finally, we discuss and summarize our work and provide future prospects in Section 5.

2 DATA

A minimal data set that is required for our redshift refinement scheme is:

(i) a galaxy group catalogue constructed on some apparent magnitude limited galaxy redshift survey and
(ii) a galaxy catalogue, fainter than the group catalogue but covering the same area of sky and with photometric redshift measurement which we wish to refine.

To begin with we construct two independent sets of data obtained from disparate sources, i) a set of observational data includes galaxy catalogue with photometric observations from the KiDS survey ($r > 19.8$ mag) and group catalogue from the GAMA survey ($r < 19.8$ mag) that share identical sky coverage, and ii) a set of theoretical data form the DESI mock catalogue light-cones derived from the GALFORM galaxy formation model. The former forms our test sample to demonstrate the validity of our methods. To match the magnitude limit of the observational data, we also divide the DESI catalogue into two parts separated at an apparent magnitude limit on $r = 19.8$ mag, identical to that of the GAMA survey.

Below, we provide more details about these data, as well as of the derived quantities.

2.1 Galaxy and Mass Assembly (GAMA) survey

The GAMA survey is a spectroscopic and multi-wavelength survey of $\sim$ 300,000 galaxies down to Petrosian r-band magnitude $m_r = 19.8$ over $\sim 286\,{\rm deg}^2$ with high spatial completeness carried out on the Anglo-Australian Telescope (Driver et al. 2011; Liske et al. 2015). Details of the GAMA survey characteristics are given in Driver et al. (2011), with the survey input catalogue described in Baldry et al. (2010), the spectroscopic processing outlined in Hopkins et al. (2013), and the spectroscopic tiling algorithm explained in Robotham et al. (2010), while the group catalogue is provided in Robotham et al. (2011). The group catalogue is constructed using an adaptive Friends-of-Friends (FoF) algorithm, linking galaxies in projected and line-of-sight separations. For the full details about the algorithm, diagnostic tests, construction and

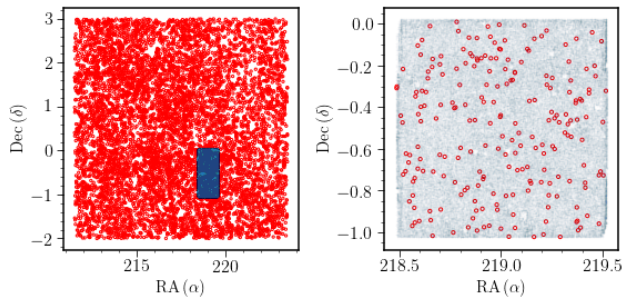

Figure 1. Galaxy and group samples. Left: position of KiDS galaxies in G15SQRDEG region (blue region) and overlapping galaxy groups (represented by the positions of the central galaxies) from the GAMA group catalogue (red dots) in the entire G15 region shown in the equatorial coordinates. Right: zoomed-in version of the left panel at G15SQRDEG region. RA and Dec are equatorial angles in degrees.

caveats of the group catalogue we refer the reader to Robotham et al. (2011). As such we only use the galaxy group data from the northern equatorial region of the GAMA survey field centred at $15^{\rm h}$, i.e., $218.5^\circ < \rm R.A. < 219.5^\circ$ and $-1.09^\circ < \rm Dec < 0.0^\circ$ and refer to it as the G15SQRDEG region. In the G15SQRDEG region we have 1712 galaxies with $r < 19.8$ mag of which $\sim 55\%$ galaxies are present in 236 galaxy groups with richness ≥ 2 whereas remaining galaxies are singleton i.e. with no observed satellites within the magnitude depth of the survey. We describe the relevant properties of the group galaxies in Section 2.4.

The 1 $\rm deg^2$ field centred at G15 region aka G15SQRDEG is selected mainly because in this region we have galaxies spectra out to a deeper magnitude limit in i-band $m_i = 22$ mag than the formal limit of the GAMA survey, providing us with spectroscopic redshifts to compare against our refined photometric redshift and to establish the robustness of our method. For simplicity, we refer this set of data as a G15SQRDEG-DEEP spectroscopic data.

Spectroscopic observations of the G15SQRDEG-DEEP region were undertaken using the AAT AA_OMEGA+2DF system in July-Sept 2014. Targets were selected to $i < 22$ ($r < 24$) mag and assigned to fibres using a nightly feedback method, where initially sources were tiled as described in Robotham et al. (2010). Pointings were observed for 40 minute intervals. Following each pointing spectra were automatically reduced using 2DFDR and assigned redshifts and confidences using AUTOZ (Baldry et al. 2014). Sources with secure redshifts were removed from the target list and those without redshifts were re-observed. Once multiple observations of the same source were acquired, they were S/N weighted stacked prior to redshifting. This process was repeated to allow variable integration times depending on the ability to obtain a redshift for a particular source. Once completed, all sources were visually inspected and redshifts adjusted accordingly. The catalogue contains 3,241 targeted sources of which 2,289 have a secure redshift (VIS_CLASS==‘Y’).

2.2 Kilo-Degree Survey (KiDS)

In the G15SQRDEG-DEEP region we constructed a photometric catalogue of fainter galaxies with $r > 19.8$ mag obtained from the Kilo-Degree Survey (KiDS, Kuijken et al. 2015). KiDS is an optical wide-field imaging survey carried out with the VLT Survey Telescope and the OmegaCAM camera. To obtain the photometric measurements of G15SQRDEG-DEEP galaxies we undertook following steps. First, in the image cut-out centred at the G15SQRDEG-DEEP region we fixed the apertures manually and then measured the photometry using the Lambda Adaptive Multi Band Deblending Algorithm in R (LAMBDAR) software (Wright et al. 2016). LAMBDAR requires at least the image from which one wants photometry measurements and also a corresponding catalogue of aperture parameters. Then it places the apertures over the image and measures the flux within them. Also, it performs deblending for those apertures which intersect with each other and provides the sky background noise to subtract from the galaxies. It then estimates noise correlation, calculate flux accounting for local backgrounds. Finally, we get fluxes and flux uncertainties over the 4 optical u, g, r and i bands observation from the KiDS and 5 near-infrared $Z, J, H, K_{\rm s}$ and Y bands from VISTA Kilo-Degree Infrared Galaxy Survey (VIKING, Edge et al. 2013).

The complete G15SQRDEG-DEEP photometric catalogue consists of 164,581 galaxies; removing those with incomplete photometric measurements and with $i > 22$ mag (to match the magnitude limit of the spectroscopic sample) results in a final sample of 59,134 galaxies. The left panel of Fig. 1 shows the entire G15 region of the GAMA survey, where the red dots represent the group central and singleton galaxies whereas the blue mask depict the G15SQRDEG region. The right panel is the zoomed in version of the left panel centred at G15SQRDEG region, where blue dots show galaxies in G15SQRDEG-DEEP photometric catalogue. Next, we use the derived photometry measurements of this sample to estimate their photometric-redshift.

2.2.1 Photometric redshift measurement

In this work we mainly rely on the machine learning approach of ANNz2 (Sadeh et al. 2016) to derive photometric redshifts. ANNz2 is a new implementation of the code of Collister & Lahav (2004), which utilizes methods such as artificial neural networks and boosted decision/regression tree, and is freely available software package. To recap, the algorithm uses machine learning methods to learn the relation between photometry and redshift from an appropriate training set of galaxies for which the redshift is already known. The trained model is then used to predict the photometric redshift of the galaxies for which spectroscopic measurements are lacking.

The data we use here to train the ANNz2 networks and generate a catalogue of photometric redshifts consists of galaxies in G15SQRDEG-DEEP region, a subset for which spectroscopic redshifts have been determined (described in Section 2.1). This catalogue consists of 3241 galaxies with $i < 22$ mag, out of which 2289 galaxies have a high quality spectroscopic redshift measurement. Matching these galaxies up to their corresponding entries in the G15SQRDEG-DEEP photometric catalogue provide us with photometric

measurements in the u, g, r, i, Z, Y, J, H and K_s bands for most galaxies. Removing those with missing or incomplete photometric measurements leaves us with 2,188 galaxies, this being the final sample used in the training and validation runs of ANNz2. Half of these galaxies are randomly selected for training with the other half used for validation. Finally, we apply the trained ANNz2 networks to the G15SQRDEG-DEEP photometric catalogue to determine their photometric redshifts.

Methods

ANNz2 employs two different approaches which can be selected by the user, namely, Artificial Neural Networks (ANN) or Boosted Decision Trees (BDT). Both approaches consist of a training phase where the networks are trained on data with known spectroscopic redshifts, a validation phase and an evaluation phase where the resulting trained networks are applied to a new photometric dataset where the redshifts are unknown. In this section we apply both methods and determine which provides the most consistent results for our dataset. In both cases we used 50 iterations in the training phase, as additional iterations resulted in limited improvements and increased the risk of biases introduced from over-training, given our limited training sample.

The ANN approach uses at least three layers of nodes, the input layer (consisting of the same number of nodes as the number of input variables), at least 1 hidden layer and a final node which outputs the calculated photometric redshift. In each instance of the training run, the number of hidden layers and the number of nodes in each hidden layer are randomly set, along with weightings in the various connections between nodes in neighbouring layers. The probability distribution function (pdf) of the galaxy's redshift is determined from the distribution of the weighted photometric redshift estimates from the ensemble of trained networks.

In contrast, the BDT approach takes the input through an initial root node and passes it through branching linkages of internal nodes before arriving at a final output node, or 'leaf'. Similarly to the ANN approach, each BDT training run initializes a new tree with different weightings of the input data. This results in a 'forest' of decision trees, from which the weighted distribution of redshift estimates can be used to determine the pdf for the galaxy's redshift.

Training Sample

First, we look at at the results of the ANNz2 algorithm for the sub-sample with spectroscopic redshifts, and compare the derived photometric redshifts with the spectroscopically determined values. Here we used 50 training runs for both the ANN and BDT methods.

Fig. 2 compares the distributions of the BDT and ANN photometric redshifts z_p with the spectroscopic redshift z_s distribution, highlighting that the overall distribution of redshifts is reproduced well. This figure highlights the scattering of galaxies with low z_s values towards higher z_p values, resulting in an under representation of galaxies at low redshift in the photometric distribution. Both methods produce z_p which are closely correlated with the spectroscopic value, with the BDT results featuring slightly less scattering. However, the distributions for both z_p sets are slightly skewed

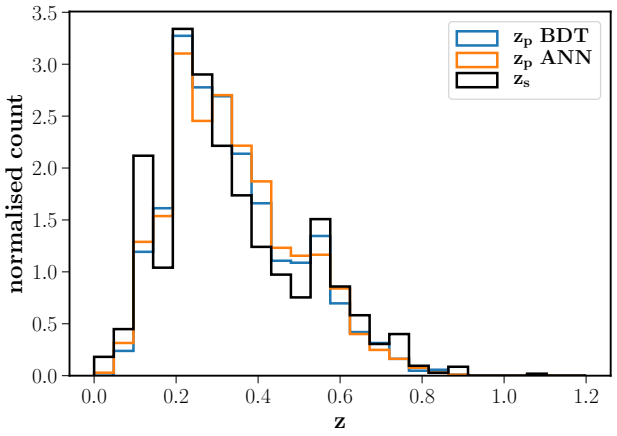

Figure 2. The redshift distributions of the two (ANN and BDT) photometric redshift estimates compared to the spectroscopic redshifts. Both ANNz2 methods produce a similar distribution which follows the spectroscopic distribution.

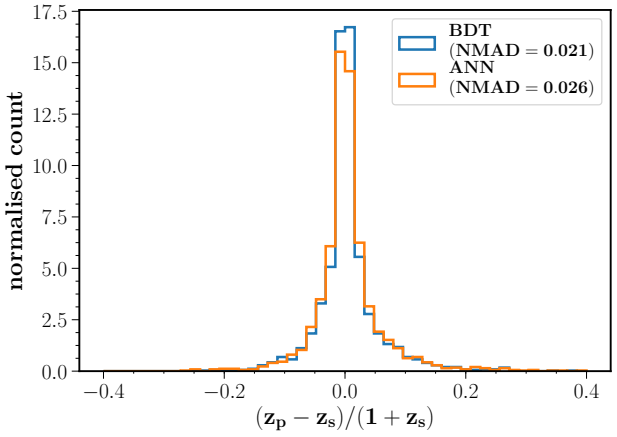

Figure 3. Distributions of scaled biases for the ANN and BDT methods. The BDT distribution, judged from its NMAD value, is marginally better compared to the ANN, indicating more accurate redshift estimates.

towards higher values at low redshifts and lower values at high redshifts. The scatter is greater at the high end of the z_s due to the small number of training galaxies in this region, and the lower quality photometric measurements for these generally dimmer galaxies.

The quality of the photometric redshift estimates can be quantified using the normalised median absolute deviation ($\sigma_{\rm NMAD}$ or for simplicity, just NMAD), defined as

$$\sigma_{\rm NMAD} = 1.48 \times \text{median}\left(\left|\frac{\Delta z - \text{median}\,(\Delta z)}{1 + z_s}\right|\right), \qquad (1)$$

where $\Delta z = z_p - z_s$ and lower NMAD values indicate more accurate redshift estimates. The NMAD values for the two ANNz2 methods ANN and BDT are 0.026 and 0.021 respectively. These calculations were done using galaxies in the validation set i.e. those not used for training the algorithms. Fig. 3 illustrates the distributions of scaled bias $\Delta z/(1 + z_s)$ for the ANN and BDT methods. The distribution for the BDT-derived photometric redshifts is more

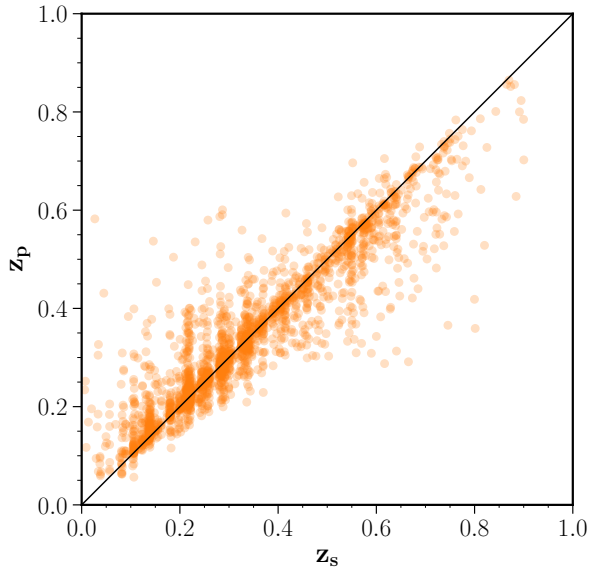

Figure 4. Spectroscopic versus photometric redshifts of the G15SQRDEG-DEEP galaxies produced by the BDT method.

sharply peaked at $\Delta z/(1 + z_s) = 0$ relative to the other two, furfther indicating that the BDT approach is producing the more accurate redshift estimates. For our sample we find that the BDT method gives more accurate redshift estimates than the ANN, with its NMAD statistics comparing favourably to other photometric implementations (see for example Dahlen et al. 2013).

We also run the template fitting scheme, EAZY (Brammer et al. 2008) with empirical templates of Brown et al. (2014) and obtain inferior NMAD value of 0.041 compared to the machine-learning approach. We also find that the EAZY photometric redshifts are slightly asymmetric around $\Delta z/(1 + z_s) = 0$ while both ANN and BDT produce a symmetric distribution. Given the z_p distributions with higher NMAD values produced by EAZY and ANN schemes, from this point on we do not consider the results produced by them and only make use of the BDT outputs. The relationship between the spectroscopic and photometric redshift for the G15SQRDEG-DEEP spectroscopic catalogue for the BDT approach is shown in Fig. 4.

Probability Distribution Functions

In addition to the photometric redshift point-estimate, ANNz2 also produces a probability distribution functions (pdfs) of possible redshifts for each galaxy. Fig. 5 shows a representative sample of the pdfs for galaxy redshifts as determined by the BDT method, with the galaxies taken from the training sample. These pdfs feature strong peaks in most cases, however many pdfs are evidently not very smooth. For a minority of galaxies, the pdfs found by the BDT method feature a double peak, though only on rare occasions are the two peaks near equal in amplitude. These features may nevertheless have an impact on the next stage of this project,

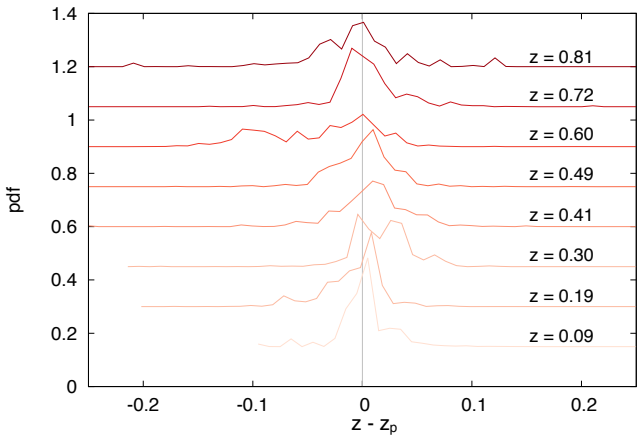

Figure 5. The probability distribution functions found by the BDT method for randomly selected galaxies across the redshift range, centred on the photometric redshift.

particularly if there is a galaxy group located around the secondary peak.

To get the greatest improvement from the new refinement method, we require pdfs which are unbiased and representative of the actual distribution of true redshifts around the photometric redshifts. The pdfs generated by the BDT approach were tested for uniformity by looking at the $C(z_s)$ statistic, which gives the total predicted probability that z_s was located somewhere between zero and the z_s value which was actually observed:

$$C(z_s) = \int_0^{z_s} p(z)dz. \quad (2)$$

If the generated pdfs were unbiased and correctly representative of the distribution of z_s about z_p, one would expect to find that 10 percent of galaxies would have measured z_s values located in the first 10 percent of their pdf (corresponding to a $C(z_s)$ value ≤ 0.1), another 10 percent would have $C(z_s)$ values between 0.1 and 0.2, and so on. Therefore, finding $C(z_s)$ for all galaxies in the sample and then finding an empirical cumulative distribution function (ECDF) of all the $C(z_s)$ values should result in a straight line.

When applying this test to the BDT pdfs, we found that the ECDF deviated from a straight line, indicating that they are indeed biased. Since the deviation was found to be systematic, we corrected for this bias by converting the pdf of each individual galaxy into a cumulative distribution function and, at each point along the distribution, correcting it's value to the corresponding value of the global ECDF. The full details of this correction process are given in a separate paper, Deeley et al. (in prep).

2.3 Theoretical data

The DESI light-cone mock catalogues are based on the GALFORM galaxy formation model of Gonzalez-Perez et al. (2014). The outputs of the model are placed in a light-cone using the technique described in Merson et al. (2013). The light-cone has a circular field of view of radius 4.0 degrees, and only galaxies with apparent

magnitudes brighter than $r \leq 23.8$ mag, i.e., 4 magnitudes fainter than the G15SQRDEG data, and cosmological redshifts less than 2.5 are included. Similar to the case with the G15 data, here also we construct two separate sub-catalogues, which include

- sets of fainter ($19.8 < r/\mathrm{mag} < 19.8+i$) with $i \in 1 \rightarrow 4$ galaxy catalogues with synthetic photometric redshift and
- a common corresponding halo catalogue with $r \leq 19.8$ mag.

To estimate the synthetic photometric redshifts for the DESI galaxies, first we take an approach similar to the one for the G15SQRDEG-DEEP data, i.e., given colours and magnitudes employ photometric redshift determination software. However, we note that irrespective of machine learning and template fitting based photo-z software the yielded photometric redshifts have large variance and significant systematics. There could be many reasons for this such as imperfect stellar population synthesis models, or simply because there is not enough non-degenerate information present in broadband photometry. To minimise the unknown systematics and have a controlled sample, we generate a pseudo photometric redshift by applying a random error to each mock galaxy redshift randomly generated from a normal distribution. This is repeated for normal distributions with two choices of NMAD (or simply, standard deviation as we do not simulate outliers), 0.02 and 0.04, to investigate how the precision of photometric redshift affects out method.

2.4 Intricacies of the data: deriving galaxy and group properties

There are a number of key inputs required for our refinement method, including the properties of (i) the group central galaxy (stellar mass and position) (ii) the group (velocity dispersion, virial mass and virial radius), and (iii) the fainter galaxies for which we wish to refine z_p (projected distance from the group centre). Below the derivation of each of these properties is described in detail.

2.4.1 Group centric distance and velocity

We consider the brightest galaxy in a group (BGG) as its central galaxy. The projected separation (R) of a galaxy from the centre of the group are calculated using the cosmological formulae for distance estimation:

$$R = \theta \; d_\mathrm{ang} \tag{3}$$

where the cosmological angular distance

$$d_\mathrm{ang}(z_\mathrm{G}) = \frac{c}{1+z} \int \frac{\mathrm{d}z'}{H(z')}, \tag{4}$$

z_G is the central group galaxy redshift and c is the speed of light. The angle θ is the angular separation between the galaxy (α_g, δ_g) and central group galaxy (α_G, δ_G), where α and δ are the equatorial coordinates representing the Right Ascension and Declination angles respectively. Note the projected distance R has to be calculated for all combination of galaxies and central-galaxies. Fortunately, R does not depend on the galaxy redshift and therefore, the distance matrix can be calculated once for each data set and later looked-up when needed. Similarly, velocity of any galaxy relative to the group centre is given by

$$v/c = \frac{z - z_\mathrm{G}}{1 + z_\mathrm{G}}. \tag{5}$$

2.4.2 Halo properties

Finally, we determine the mass of each group using the theoretical relation between central galaxy stellar mass and halo mass, that is one derived from the abundance matching. For this we rank order match central galaxy stellar masses against the expected number density of halos in the comoving volume. For halo number density we use the halo mass predictions of Sheth et al. 2001, as taken from HMFcalc (Murray et al. 2013). A singleton galaxy that is not assigned to any group in a group catalogue could be a potential central galaxy of the group containing unobserved fainter satellites. Hence, we also treat a singleton galaxy as a potential group. From the derived halo mass (M_{200}) we estimate the group virial radius (r_{200}) using

$$r_{200} = \sqrt[3]{\frac{2GM_{200}}{\Delta H^2(z)}}. \tag{6}$$

$$H(z) = H_0 \sqrt{\Omega_m (1+z)^3 + 1 - \Omega_m}, \tag{7}$$

where Ω_m is the cosmological density parameter at $z = 0$ and the value for the virial over-density parameter $\Delta = 200$. Similarly, virial velocity is calculated using the relation

$$v_{200} = 10 H(z) r_{200}, \tag{8}$$

whereas concentration parameter c_{200} is derived from the concentration-virial-mass relation obtained from Duffy et al. (2008) given by

$$c_{200}(M_{200}, z_\mathrm{G}) = 6.71 \; (0.5\,h\,M_{200}/10^{12})^{-0.091} (1+z_\mathrm{G})^{-0.44}. \tag{9}$$

3 METHOD: GALAXY-TO-GROUP ASSIGNMENT

We now present the description of the different steps involved in *galtag*. First, we outline the prescription for the phase space distributions of the halo member galaxies and interlopers, where interlopers mean the galaxies that lie outside the virial sphere of the group, but within the cone circumscribing the virial sphere. Second, we show how galaxies are probabilistically tagged to the potential group. Finally, we illustrate the photometric redshift refinement process.

We obtain the ansatz for the halo and interloper models from Duarte & Mamon (2015, 2016), who developed it as a part of the Models and Algorithms for Galaxy Groups, Interlopers and Environment (MAGGIE). MAGGIE is a prior- and halo-based abundance matching group finding algorithm, showing a promising alternative to the conventional crispy group-finding scheme such as the friends-of-friends. For the purpose of our paper we only need and make use of the halo and interloper models given in MAGGIE and not of its group finding aptitude. While we refer to the above papers for the full derivation, tests and justification of parameters assumed, below we outline minimal complete information that is relevant to our work.

3.1 Halo surface density

Following Mamon et al. (2013), the density of halo member galaxies $g_{\rm h}(R,v)$ in projected phase-space limited to the virial sphere can be written as

$$g_{\rm h}(R,v) = 2\int_R^{r_{200}} \rho(r)\; h(v|R,r)\; \frac{r}{\sqrt{r^2-R^2}}\;{\rm d}r, \tag{10}$$

where $\rho(r)$ is a galaxy number density profile. Assuming that a galaxy group is a self-consistent system, i.e. the galaxy number distribution follows the mass distribution we can consider that $\rho(r)$ follows a NFW profile (Navarro et al. 1996), given by

$$\rho(r) = \left(\frac{N_{200}}{4\pi r_{200}^3}\right)\frac{f(c_{200})}{x(x+1/c)^2}, \tag{11}$$

Here N_{200} stands for the number of galaxies within the virial sphere, which as we will see later cancel out and hence, can be assumed to be an arbitrary number at this stage. Also, $x = r/r_{200}$ and the function

$$f(c_{200}) = \frac{1}{\ln(1+c_{200}) - c_{200}(1+c_{200})}.$$

In equation 10, $h(v|R,r)$ is the probability of observing a line-of-sight velocity at the position (r,R), which is assumed to have a Gaussian distribution written as

$$h(v|R,r) = \frac{1}{\sqrt{2\pi\sigma_z^2(R,r)}}\exp\left(-\frac{v^2}{2\sigma_z^2(R,r)}\right), \tag{12}$$

with the squared velocity dispersion run given by

$$\sigma_z^2(R,r) = \left(1-\beta(r)\frac{R^2}{r^2}\sigma_r^2(r)\right). \tag{13}$$

Here,

$$\beta = 1 - \frac{\sigma_\theta^2}{\sigma_r^2}$$

is the velocity anisotropy parameter with σ_r and σ_θ being the second moments of radial and tangential components of the velocity vector in spherical coordinates relative to the centre of the halo at a rest frame. It is clear that to calculate $\sigma_z(R,r)$ we must know the $\beta(r)$ and $\sigma_r(r)$ runs of each halo. Unfortunately, due to the lack of the peculiar velocity information of galaxies, β and σ_r are not directly observable quantities. For this we resort to the theoretical data of the ΛCDM cosmological simulations, and choose the following form for the β profile

$$\beta(r) = \frac{r}{2(r+r_{200}/c_{200})}, \tag{14}$$

which is taken from Mamon & Łokas (2005) and has been shown to agree with a list of different cosmological ΛCDM simulations. Moreover, it is also the recommended $\beta(r)$ profile in MAGGIE. When $\beta(r)$ is assumed, we can substitute in the spherical Jeans equation and determine the other known unknown, $\sigma_r(r)$. Thankfully, Duarte & Mamon (2015) already provide the solution for us in the set of equations (A1-A5) from the appendix section, which is terms of halo virial properties can be summarised as

$$\begin{aligned}\sigma_r^2(r) = \left(\frac{GM_{200}}{r_{200}}\right)\frac{c_{200}\; f(c_{200})}{6y(y+1)}\times\\ \left[6y^2(1+y)^2{\rm Li}_2(-y) + 6y^4\coth^{-1}(1+2y)\right.\\ -3y^2(1+2y)\ln y + y^2(1+y)^2\{\pi^2+3\ln^2(1+y)\}\\ \left.-3(-1+2y^2)\ln(1+y) - 3y(1+y)(1+3y)\right],\end{aligned} \tag{15}$$

where $y = c_{200}\; r/r_{200}$. Here, ${\rm Li}_2$ is a dialogarithm function defined as

$${\rm Li}_2(-x) = \begin{cases}\sum_{i=1}^{10}(-1)^i\frac{x^i}{i^2} & x < 0.35\\ -\frac{\pi^2}{12} + \sum_{i=1}^{10}\left(\frac{\ln 2}{i} - \frac{a_i}{b_i}\right)(1-x)^i & 0.35 \le x < 1.95\\ -\frac{\pi^2}{6} - \frac{1}{2}\ln^2(x) - \sum_{i=1}^{10}(-1)^i\frac{x^{-i}}{i^2} & x \ge 1.95.\end{cases} \tag{16}$$

[1]. The values for the coefficients a_i and b_i are given in Table A1 of Duarte & Mamon (2015).

3.2 Interloper surface density

In their study Mamon et al. (2010) analyse the distribution of dark matter particles from a cosmological hydrodynamical simulation and predict that the universal distribution of halo interlopers in projected phase-space can be represented by a Gaussian line-of-sight distribution velocity plus a constant term as follows

$$g_{\rm i}(R,v) = \frac{N_{200}}{r_{200}^2 v_{200}}\left(A(x)\exp\left[-\frac{1}{2}\frac{(v/v_{200})^2}{\sigma_i^2(x)}\right] + B\right). \tag{17}$$

Calibrating with the galaxies of the semi-analytic model of Guo et al. (2011) at redshift zero, Duarte & Mamon (2015, 2016) determine that the terms A, σ_i and B obey the following forms

$$\log(A(x)) = -1.092 - 0.01922x^3 + 0.1829x^6, \tag{18}$$

$$\sigma_i(x) = 0.6695 - 0.1004x^2, \text{ and} \tag{19}$$

$$B = 0.0067, \tag{20}$$

where $x = R/r_{200}$.

Finally, utilising the halo member (galaxies within the virial sphere) and interloper (galaxies within the virial cone, but residing outside the periphery of the virial sphere) density distributions, the probability that a galaxy at projected radius R and a relative distance z from the group centre to belong to a given group (to the virial sphere of the real-space group) can be written as

$$p_{\rm G}(\theta,v|\Theta) = \begin{cases}\frac{g_{\rm h}(R,v)}{g_{\rm h}(R,v)+g_{\rm i}(R,v)} & R \le r_{200}\\ 0 & R > r_{200}\end{cases}. \tag{21}$$

The total assignment probability is non-zero only within the virial cone ($R > r_{200}$), therefore, for a practical purpose the galaxy by central-galaxy dimensional distance matrix (Equation 3) has to be only calculated for cases where $R \le r_{200}$ making it a highly sparse matrix with roughly 95% sparsity. Here, the distribution is conditioned over Θ, consisting of a set of group properties such as the position of the group centre (RA, Dec, $z_{\rm G}$) and group virial properties (primarily, M_{200}). It is to be noted here that the normalization N_{200} appears both in the $g_{\rm h}$ and $g_{\rm i}$ distributions, hence, cancels out when we write the probability term $p_{\rm G}(R,v|\Theta)$.

[1] note, Equation A5 in Duarte & Mamon (2015) has a factor of 1/2 missing, and also, the sign shown in the dialogarithm function for $x \ge 1.95$ case should be negative

3.3 Photometric redshift refinement

From the photometric redshift measurement method we obtain a normalised probability distribution of the galaxy redshift that can be denoted as $p_{\rm g}(z|z_{\rm t})$, where $z_{\rm t}$, a latent variable, is the error free true redshift that can not be observed. With $p_{\rm g}$ and a model for the galaxy group distribution $p_{\rm G}(\theta, z|\Theta)$ (equation 21), we can express a joint galaxy-group distribution as

$$p(\theta, z|\Theta) = \int p_{\rm G}(\theta, v(z_{\rm t})|\Theta)\ p_{\rm g}(z|z_{\rm t}) dz_{\rm t}. \tag{22}$$

This allows us to calculate the likelihood for a galaxy to belong to a given group as

$$p_{\rm tot} = \int p(\theta, z|\Theta)\ {\rm d}z, \tag{23}$$

which gives a measure of correlation of the galaxy and group redshift distributions. Finally, the resultant refined probability distribution of galaxy redshift will be given by

$$p_{\rm ref}(\theta, z|\Theta) = p(\theta, z|\Theta)/p_{\rm tot}. \tag{24}$$

Probabilistically, every galaxy will have some finite probability to belong to all the groups. But, in the end we aim to find the best-matching galaxy-group pair, that is, to apply a hard assignment. Hard assignment in our case is a two-step process. First, we apply a relative criteria, in which we only consider a group for which a galaxy has the highest assignment probability as the best match. Second an absolute measure, where out of the best-matching galaxy-group pairs we only consider pairs for which the assignment probability is greater than some threshold value. All the remaining galaxies, with an assignment probability less than a threshold value, are considered un-grouped or a singleton. The optimal value for the threshold assignment probability is determined from the tests done in the synthetic data as we discuss in the later section. In cases where we only aim to refine the photometric-redshift, we can skip the second step and for all the existing best-matching pair we can directly calculate the expected value of the redshift for the galaxy given a group using the following formula,

$$z_{\rm r} = \langle z \rangle = \int_0^{z_{\rm max}} z\ p_{\rm ref}(\theta, z|\Theta)\ {\rm d}z, \tag{25}$$

where $z_{\rm max}$ can be some arbitrarily large redshift, which should at least accommodate the full range of the $p_{\rm g}(z)$ distribution. For our fainter galaxies limited to $r < 23.8$ mag, $z_{\rm max} = 2$ is a large enough value.

3.4 Group assignment purity

At this point it is interesting to explore how accurately *galtag* can assign fainter galaxies. Strictly speaking *galtag* does not assign galaxies to groups, but galaxies have some probability $p_{\rm tot}$ (given by equation 23) to belong to the virial sphere of a given group. Nevertheless to gauge the accuracy, we hard assign galaxies to the highest probable group and compare the purity of the predicted classification with the true group membership. For this we first construct a confusion matrix, which is a square matrix of order two providing the true positive (TP), false positive (FP), true negative (TN) and false negative (FN) counts. For the DESI mock ..... are typical values for the TP, FP, TN and FN rate respectively. Finally, the group purity fraction or the fraction of correct assignments [2] is given by $\frac{\rm TP+TN}{\rm TP+TN+FP+FN}$, the value for which ranges from 0 to 1, where 1 means all galaxies are correctly assigned to its true group. This purity fraction can only be calculated for the DESI mock data where we know the true partition for all galaxies.

4 *galtag* IN ACTION

The software *galtag* is written in in Python 2.7 and includes both the halo and interloper models from MAGGIE as well as the refinement step discussed above. The software will be made available at `https://github.com/pkaf/galtag.git` [3]. Below, we highlight key diagnostic results demonstrating the application of *galtag* in the DESI mock and G15SQRDEGDEEP data.

4.1 With DESI synthetic data

In Fig. 6 we show the quantitative analysis of refined redshift and group purity for two representative sets of DESI data. The results in the left panels are for the data with intrinsic NMAD of 0.04 whereas one on the right panels are for the case with NMAD=0.02. The panels in the top two rows respectively show the biases in redshifts $|z_{\rm p} - z_{\rm s}| - |z_{\rm r} - z_{\rm s}|$ and the post-refinement NMAD trends as a function of assignment probabilities ($p_{\rm tot} = p$) in the logarithmic scale. The solid lines of different shades in panels (b) and (f) represent cases with different limiting magnitude ranging between $r < 20.8$ mag to $r < 23.8$ mag. However, to avoid crowding in the panels (a) and (e) we only show two cases: $r < 23.8$ mag case in darker shade and $r < 20.8$ mag case with fainter shades. The solid and dashed lines in these panels represent the respective running medians of $|z_{\rm p} - z_{\rm s}| - |z_{\rm r} - z_{\rm s}|$ as a function of log p. In panels (a) and (e) we observe that only at $\log p \gtrsim -7$ the median biases in redshift measurements are close to zero and at lower probabilities the bias is significantly high. Moreover, the darker points have much longer low probability tail compare to the fainter points. Similarly, in panels (b) and (f) we observe that only at $\log p \gtrsim -7$ are the NMAD values of refined redshifts found to improve compared to the intrinsic photometric redshift. Here, we see that at lower probabilities the NMAD gets much worse than the intrinsic NMAD of 0.04 (left panel) or 0.02 (right panel) and it further worsens with increasing depth of the limiting magnitudes. These discrepancy are due to the physical effect that most faint galaxies tend to be at larger redshifts, and we force them to match groups at low redshift, leading to the underestimation of $z_{\rm r}$ compare to $z_{\rm s}$ or $z_{\rm p}$.

In panels (c) and (g) we show the fractional cumulative count of the galaxies which have been refined above a given value of $\log p$ whereas in panels (d) and (h) we show the group purity fraction all as a function of log p. In both the cases again solid lines with different shades represent cases with different limiting magnitude. In the figure we see that at $\log p = -7$ we have approximately $50 - 70$ per cent of the

[2] Also known as a rand index

[3] under GNU general public license (GPL), which guarantees end users the freedom to run, study, share and modify the software.

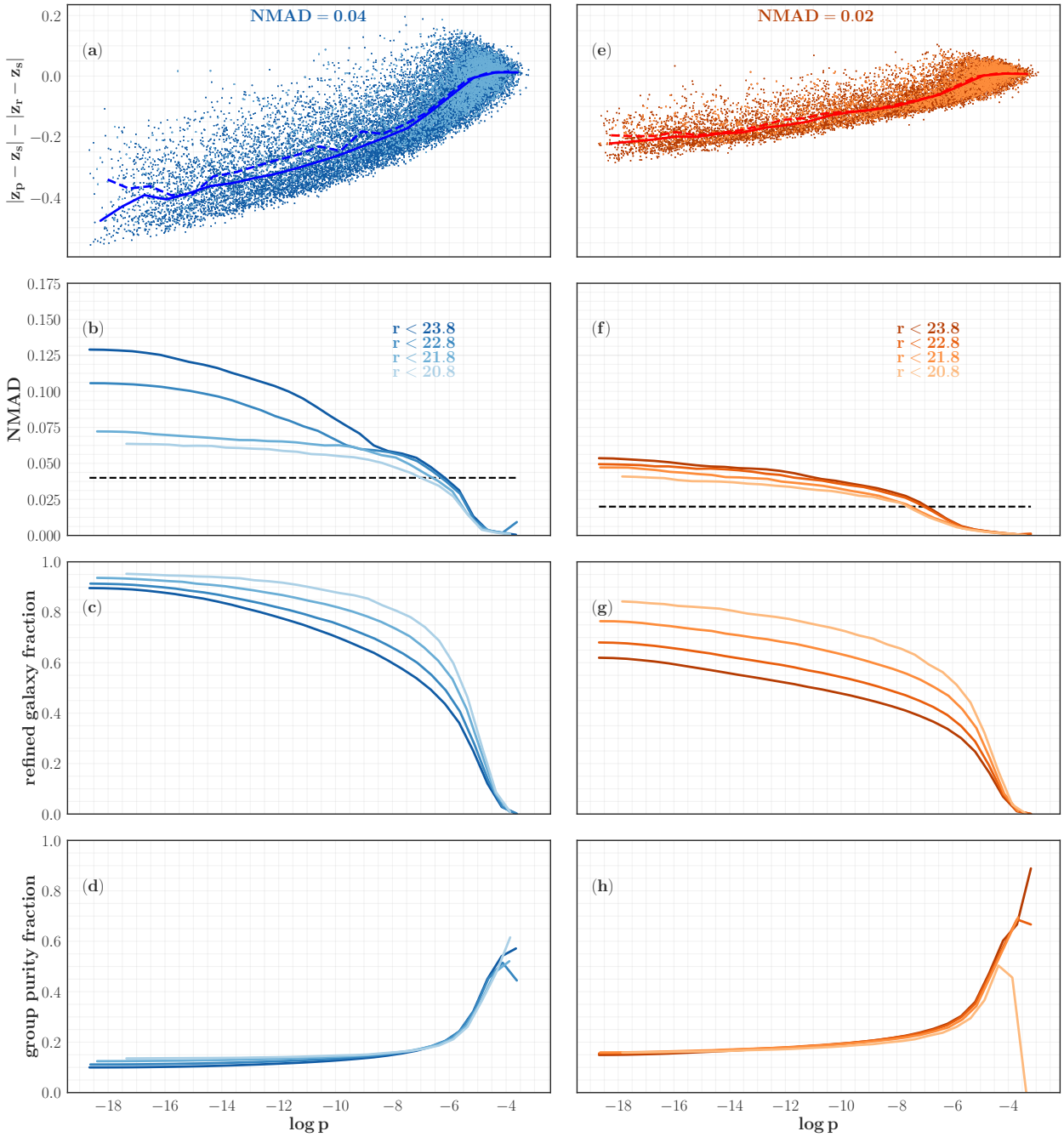

Figure 6. Post-refinement assessment of the DESI data, with intrinsic NMAD=0.04 (left panels) and 0.02 (right panels) as a function of assignment probabilities.

total galaxies that are matched to groups while the group purity fraction is 10 per cent. However, the trends suggest that for stricter probability cut, with some sample size trade-off, higher group purity is achievable.

The choice of minimum $\log p$ in *galtag* is left up to the users so that they can choose based on their science case. In cases where the user only desires the refined redshift, they can set a generous limit. However, for projects demanding higher assignment purity one can set a higher threshold probability. As a guide in Fig. 7 we once again present post-refinement NMAD (along y-axis) as a function of $\log\ p$ (along x-axis), where we also show an additional intermediate case (NMAD=0.03). Different shades of solid lines represent different limiting magnitude whereas different colours display different intrinsic NMAD. The dots from left to right in each case can be used to infer the threshold $\log\ p$ for which NMAD can be improved by 1, 1/2, 1/3, 1/5 and 1/10 times the intrinsic NMAD. We see that, for example, in a case with intrinsic NMAD=0.02 and the limiting magnitude of 23.8 mag setting $\log p \simeq -4.5$ will give an order of magnitude improvement in the NMAD value. For this case the group purity fraction is approximately 60 per cent and is compara-

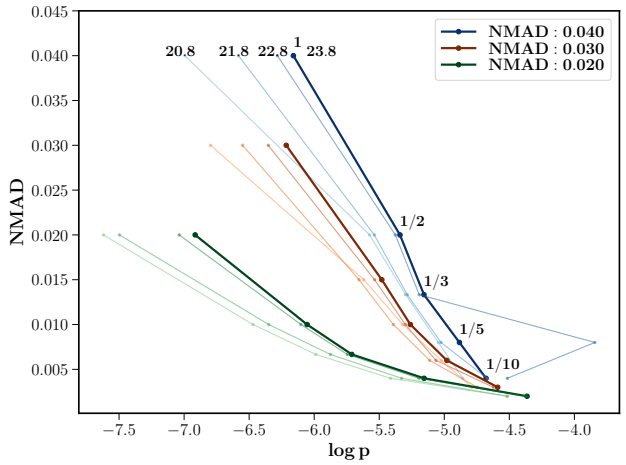

Figure 7. NMAD as a function of minimum probability in the case of DESI data. Different shades of same colour represent data limited to the labelled magnitude limit, where the darkest shades are for cases with $r < 23.8$ mag. The labels 1/2, 1/3 etc, representing the fraction of input NMAD, are guides to determine threshold $\log p$ that one should set to obtain corresponding gain in photometric redshift accuracy.

ble to the ~ 80 per cent halo assignment accuracy from the input group catalogue (Robotham et al. 2011). The average value of precision (TP/TP+FP) and recall (TP/TP+FN) in this case are 0.45 and 0.54. The fraction of refined sample compare to the total number of galaxies within the limiting magnitude in this case is only 10 per cent, which on its face value seem small. However, this forms the 85 per cent of the total sample of galaxies within the group redshift range and these are the only galaxies for which we expect any improvement.

In summary, to highlight the improvement in the redshifts from our refinement, Fig. 8 we show the quintessential redshift correlations between $z_{\rm p} - z_{\rm s}$ (left panels) and $z_{\rm r} - z_{\rm s}$ (mid panels). The top row shows the case of input NMAD=0.04 whereas the bottom row shows the case of input NMAD=0.02. Here, we have only shown the galaxies with assignment probability $\log\ p > -4.9$ (-5.2), the probability at which the NMAD value post-refinement is $1/5^{\rm th}$ compared to the intrinsic value. The darker points in the mid-panel, along the 1:1 correlation line that represent higher density of points, is enhanced compare to the left panel. This qualitatively shows the improvement in redshift values due to refinement. The right most panels show the distributions of the scaled bias where solid lines are for biases in photometric redshifts whereas dashed lines show the biases in the refined redshift, both compared to the spectroscopic measurements. As expected, we see that the distributions for scaled refined redshifts are much narrower and peaky in comparison with the scaled photometric redshift distributions. Note, the wings of the distributions of the scaled biases in the refined cases are slightly asymmetrical due to the underestimation of $z_{\rm r}$, as discussed earlier in Fig. 6 (a) and (b), which can be eliminated by imposing stricter $\log\ p$ cut.

We re-run the analysis for the DESI mock data with a different definition of the halo virial masses to understand its effect on our final result. The results from this additional exercise is shown in Fig. 9. The figure shows the relation between derived NMAD from the refined redshift sample as a function of $\log\ p$ for two different definitions of halo virial properties. Here again the blue and orange lines represent the sets of DESI data with photometric NMAD values of 0.04 and 0.02 respectively. The darker lines are when we consider the intrinsic halo virial masses and radii provided by the halo catalogues whereas the fainter lines are when we use the values of halo virial properties derived from the line-of-sight velocity dispersion of group members. We observe that at any assignment probability NMAD values for the intrinsic case is always slightly smaller than for the derived case suggesting that the results obtained from the former case is marginally better. Importantly, the improvement is marginal, which allows us to confidently apply our method to the real data where virial properties are largely inferred from the group velocity dispersions.

4.2 With G15SQRDEG-DEEP data

Similar to the DESI mock data, we also process the G15SQRDEG-DEEP data with *galtag*, and present the results in Fig. 10. We observe trends consistent to one observed in the DESI mock data. Such as the median bias $|z_{\rm p} - z_{\rm s}| - |z_{\rm r} - z_{\rm s}|$ shown with black solid line in panel (a) ceases to zero at larger values of $\log\ p$. Also, as shown in panel (b) the NMAD value for the refined data (shown in black solid line) improves at larger values of $\log\ p$. For sufficiently large cut-off values for $\log\ p$, we can see that even an order of magnitude gain in NMAD values is achievable. Additionally, the panel (c) shows the fraction of refined galaxy again as a function of $\log\ p$. Furthermore, to give the sense of improvement in the redshift measurements, in Fig. 11 we show the redshift correlation between the z_p, z_s and z_r. Here we have only considered galaxies that have group matching probability of $\log p > -7$, resulting reduction of NMAD by $1/5^{\rm th}$. The improvement in redshift measurements post-refinement can also be gauged from the enhanced number density at 1:1 correlation line seen in the mid-panel compare to the left-most one. Similarly, the leaner and peaky distribution of scaled bias $(z_{\rm r} - z_{\rm s})/(1 + z_{\rm s})$ compare to the distribution of the $(z_{\rm p} - z_{\rm s})/(1 + z_{\rm s})$, shown in the right-most panel, also demonstrate the improvement achieved post-refinement. We note that the cross-over point, that is point where NMAD value for refined sample is same as the value for photometric sample, happens at $\log p \simeq -8$, which is achieved sooner than in the case of the DESI data. This is akin to observational uncertainties in various derived quantities that the observed data possess, which get propagated to the final measurements of $\log p$ values.

5 DISCUSSION, SUMMARY AND SCIENCE EXPLOITATION

Before we summarize, we would like to point out the main limitations of our work. The input photometric redshift and group catalogues both have their own caveats that *galtag* will naturally inherit. For example, Robotham et al. (2011), in their studies of GAMA mock catalogues conclude that the halo assignment accuracy with spectroscopic redshifts is

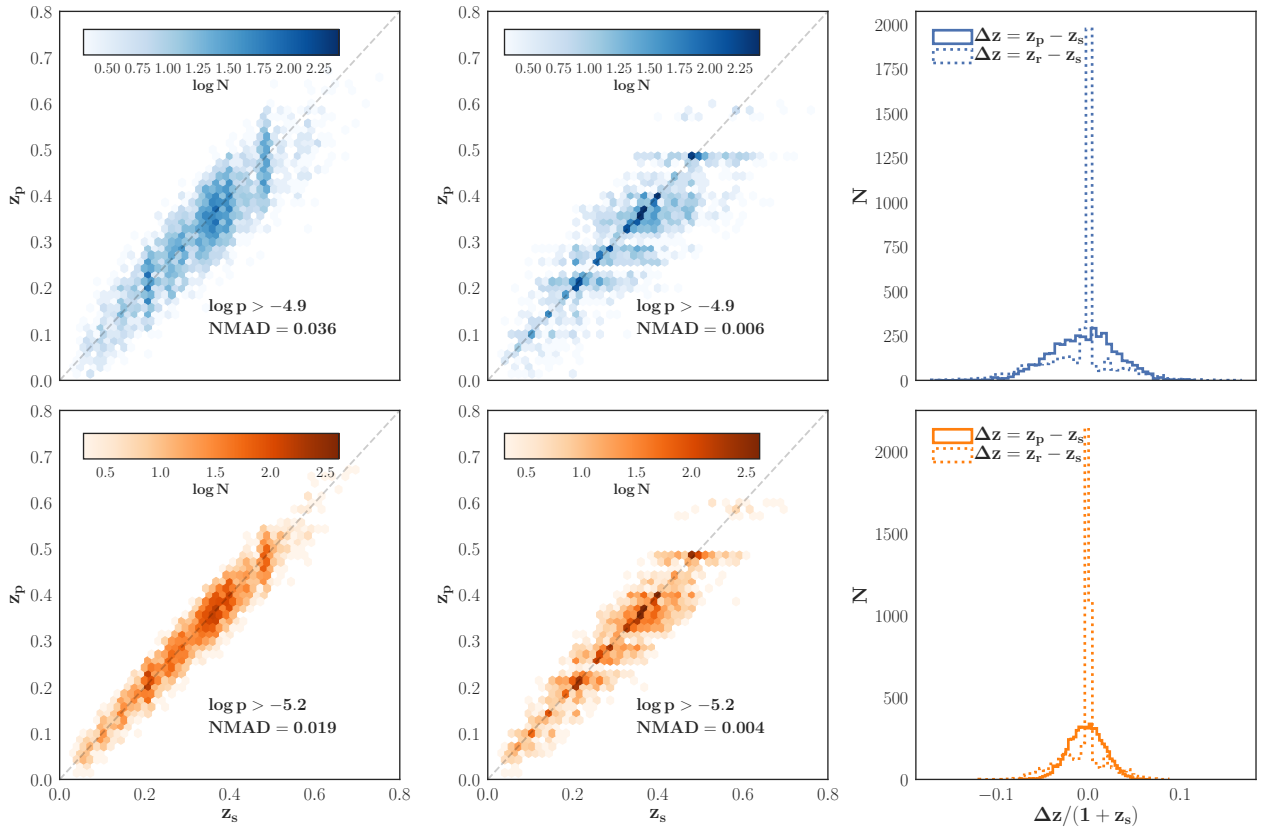

Figure 8. Photometric (left panels) and refined (central panels) redshifts correlations with spectroscopic redshifts, and distributions of scaled-biases before and after refinement (right panels) for the DESI data with intrinsic NMAD=0.04 (top row) and 0.02 (bottom row).

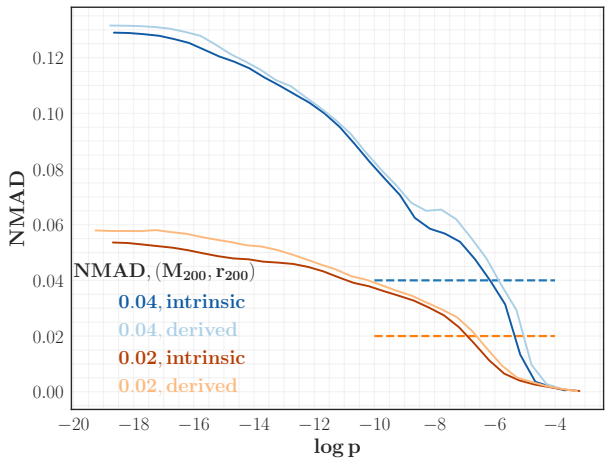

Figure 9. NMAD of the refined redshift of the DESI data generated with intrinsic NMAD=0.04 (blue lines) and 0.02 (orange lines) as a function of assignment probability for the two cases of intrinsic (solid lines) and derived halo virial properties.

only $\sim$ 80 per cent. Furthermore, the accuracy of low occupancy groups ($<$ 5 members) worsens to $\lesssim$ 50 per cent, and moreover, they form the significant $\sim$ 90 per cent of the total group population. Fidelity of input group catalogue is just the first tier of the issue which is further complicated by the need to define and estimate contentious quantities such as the centre and global group properties. There is no unique way to define the group centre, as such any of either centre of mass/light, geometric centre, or brightest group galaxy can be considered as a group centre. 'Correct' selection of a group centre is crucial particularly for very low-occupancy groups (say with $\lesssim$ 3 members) where perhaps the only other general way to estimate the halo mass is to map their central galaxy stellar-mass to halo-mass assuming the theoretical stellar-mass–halo-mass relation. Approximately 60% of galaxies in GAMA within a magnitude limit of $r < 19.8$ mag are singletons and are potential group centre for fainter satellite galaxies. Again, we have to use the theoretical stellar-mass–halo-mass relation to predict halo masses for singletons. In order to predict the concentration parameter from the group virial mass, yet another theoretical relation we have to assume is the concentration-virial mass relation. The above discussed theoretical predictions are for pure dark matter simulations, and are prone to serious systematics as they do not include baryonic processes such as cooling, star formation, and feedback. For example, the collapse of gas due to cooling leads to adiabatic contraction of the dark matter halo, which increases its concentration. Feedback, on the other hand, can have the reverse effect. Also, it has been observed that even in the cases of the Milky Way and M31, galaxies that can be studied in greatest details, the derived concentration-virial mass relations do not agree with the theoretical prediction (Kafle et al. 2014, 2018). Similarly, the relationship between dark matter halos

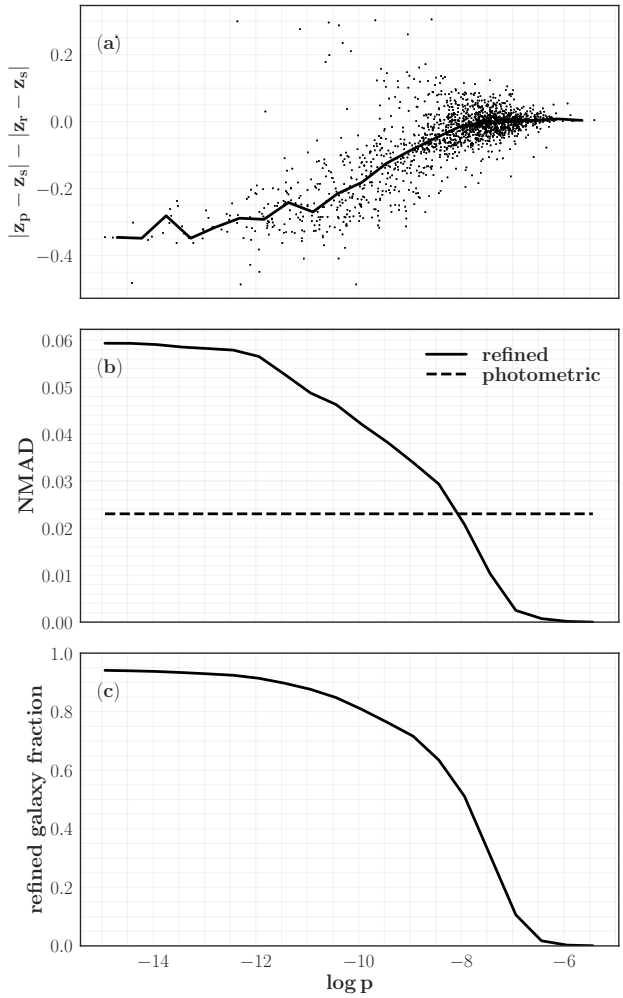

Figure 10. Post-refinement assessment of G15SQRDEG-DEEP data as a function of assignment probabilities. Labels are similar to Fig. 6.

and galaxy stellar masses from the halo abundance matching technique rely on the accuracies of observed stellar mass function, the theoretical halo mass function and techniques of abundance matching. However, for groups with high number of occupants, the line-of-sight group velocity dispersion can provide unbiased and robust handle on the dynamical mass of the groups even in the case of weak perturbations in group membership (Beers et al. 1990). That being said, Robotham et al. (2011) find that in 80 per cent of all mock groups the recovered velocity dispersion is only within ~ 50 per cent of the true value and are as likely to have underestimate as overestimate of the velocity dispersion. Furthermore, our scheme is not provisioned to discover any new groups whose even central galaxies were unobserved in the input magnitude-limited group catalogue. Both the central and satellite members of such groups are either matched to observed groups or left unassigned, determined by probability cuts. The z_p of galaxy members of such groups can still get improved, provided they are correctly matched to group that is nearby in physical space. However, inability to identify such fainter group as a stand-alone individual group will hamper the completeness of group catalogue resulting from the matching process. However, this should not have significant impact on group-by-group studies, provided the fainter photometric galaxies are constructed out of surveys with high spatial completeness. Therefore, we can summarise with the remark that even at best the accuracy of processed redshifts and groups matching are limited by the pitfalls of input catalogues, like in the case of any other scientific exploitation of a group and/or photometric redshift catalogues.

In addition to the observational limitations discussed above, our approach is also likely to be impacted from the modelling approximations that we have to make. As we have highlighted earlier, the probabilistic halo model that we adopt for group matching are of Duarte & Mamon (2015), who conduct extensive tests with the cosmological simulations to establish the model. However, the veracity of a few fundamental assumptions made in the model can still be questioned. For example, the three-dimensional velocity distribution of the galaxies in ΛCDM haloes are not strictly Gaussian (Wojtak et al. 2005). Similarly, the model assumes the galaxy number density distribution within the haloes to have the NFW (Navarro et al. 1996) form, while Navarro et al. (2004) suggest the Einasto model to be a better fit. Moreover, the velocity anisotropy profile is assumed to represent the particles in cluster mass ΛCDM haloes (Mamon & Łokas 2005). Firstly, the galaxy groups are considered to be a less evolved object and whether they follow the similar dynamics to that of the rich clusters and whether our knowledge about the clusters can be scaled and extended to groups or not is still an open question. Secondly, in any case, the velocity anisotropy is not directly observable for the galaxy groups as we can only measure the line-of-sight component of the velocity vector of the galaxies in groups. Therefore, the correctness of the assumed profile is unknown. An incorrect assumption about the velocity anisotropy could lead to the notorious mass-anisotropy degeneracy, meaning underestimation of the anisotropy results the overestimation of groups mass profile and vice-versa.

This is a proof of concept paper that present a new scheme *galtag* to refine galaxy photometric redshift and enhance group membership based on our prior knowledge of the galaxy group distribution. Here, we forgo an explicit conclusion as we attempt to summarise the paper in the abstract. However, we like to briefly discuss potential scientific objectives of the project that we will pursue in future works.

In a forthcoming paper (Kafle et al. 2018, in prep) we aim to extend the method to two independent sets of observed data namely, the ~ 300 deg^2 of KiDS data overlapping the entire GAMA fields and the ~ 6 deg^2 of The Deep Extragalactic Visible Legacy Survey (DEVILS[4]) data overlapping the COSMOS fields (Capak et al. 2007). The key science we will carry out with this new extended group catalogue is a robust measurement of the galaxy occupation frequency for a large dynamic range of stellar mass and halo masses, including Local Group mass systems down to sub 1/10 times Magellanic Cloud mass galaxies. Having an $r < 19.8$ mag group catalogue over ~ 300 deg^2 and populated additional satellites by adding photo-z galaxies with the *galtag* method described above, we will use this com-

[4] https://devilsurvey.org/wp/

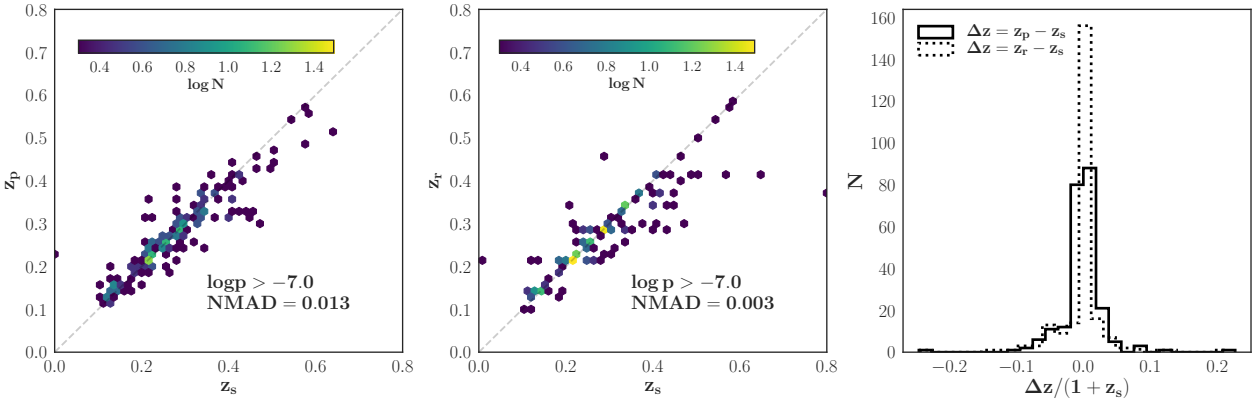

Figure 11. G15SQRDEG-DEEP spectroscopic versus photometric (left panel) and refined (centre panel) redshifts, and distributions of scaled-biases before and after refinement (right panel).

bined data to probe significantly further down the luminosity distribution for a large range of halo masses. As well as measuring the luminosity distribution for fainter satellites, the data will also allow for a much more accurate measurement of the luminosity distribution throughout the full range of halo masses that host galaxies. This will place the Milky-Way halo in context, and provide new data for modern galaxy formation models. Furthermore, by being able to measure the luminosity distribution of individual halos rather than a statistical stack (which is the approach used in clustering based halo occupation distribution work) we will be able to identify whether any individual groups share the luminosity distribution characteristics of the Milky-Way halo. This is important since we might well find that the distribution of the faintest satellites is more or less likely given the presence of the bright Magellanic satellites. As well as pushing the direct measurement of the halo luminosity distribution into a new regime, this dataset will also open up numerous fresh avenues of scientific exploration. Future work could explore the stars, dust, gas, shape, colour, structure and spatial distribution of low mass satellites. All of this information is available to GAMA and already exists for Local Group dwarf galaxies, opening up multiple future avenues of comparative exploration. In short, we can utilise the data products for projects that do not require exemplary redshift such as to address the missing satellite problem (Klypin et al. 1999), the too-big-to-fail problem (Boylan-Kolchin et al. 2011, 2012) and also to test the lopsidedness of satellite galaxy systems (Libeskind et al. 2016; Pawlowski et al. 2017) as well as to look for the Local Group analogues to put our Milky Way and neighbouring galaxies in a cosmological context (Robotham et al. 2012; Geha et al. 2017).

Beyond comparisons to the Local Group, this new assortment of halo luminosity distributions will serve as a key reference point for future simulation and theory work. By combining the data in the manner described we can do much more than present a simple 'average' luminosity distribution, in- stead we will also measure the allowed distribution space that individual halo luminosity distributions are allowed to occupy. This will allow us to characterise sub-populations for different halo masses, information that is entirely lost with current statistical stacking techniques and in broad-brush halo occupation distribution techniques.

In this work we have focussed on optimising the behaviour of *galtag* for global outcomes for different photometric sample limits. As mentioned, it is possible to modify *galtag* parameters such that you obtain the most overall improvement in refined redshift, or more accurate assignment of satellites to known halos. We do not investigate how to optimise *galtag* for a particular range of halo mass or group richness (e.g. clusters of low mass groups), but this is certainly possible in future applications.

Regarding future improvements, there are a number of plausible additional priors that could be utilised in the galaxy tagging framework presented here. Recent work by Alpaslan et al. (2015) investigated in some detail the various trends that galaxy properties have with different definitions of structure, including the same group definition investigated here. Figure 9 in that work demonstrates the enhancement of the $u-r$ colour bimodality for high mass groups, and in particular for satellite galaxies. In principle, this information can be used to better inform the galaxy tagging probabilities. For example, a 'red' galaxy in non-refined central or a refined satellite of a cluster (given the current *galtag* assignment method) is more likely to belong to the cluster. Given the lack of a clear bimodality for the halo mass range that dominates the GAMA group catalogue ($10^{13}M_{\odot}/h$, see Robotham et al. 2011) we have chosen not to utilise the information in our application to GAMA group refinement, but we note that potential future work based on higher mass cluster refinement might benefit from using colour (and possibly morphology and/or size etc) probability distribution functions.

ACKNOWLEDGEMENTS

PRK is funded through Australian Research Council (ARC) grant DP140100395 and The University of Western Australia Research Collaboration Award PG12104401 and PG12105203. We like to thank Violeta Gonzalez-perez for providing us DESI light-cones, Gary Mamon (IAP) for MAGGIE related Q&A, Maciej Bilicki for comments on the photometric redshift aspects, and Rob Finnegan and Dylan

Cusack-Paquelet for their supports during the earlier stage of the project.

GAMA is a joint European-Australasian project based around a spectroscopic campaign using the Anglo-Australian Telescope. The GAMA input catalogue is based on data taken from the Sloan Digital Sky Survey and the UKIRT Infrared Deep Sky Survey. Complementary imaging of the GAMA regions is being obtained by a number of independent survey programmes including GALEX MIS, VST KiDS, VISTA VIKING, WISE, Herschel-ATLAS, GMRT and ASKAP providing UV to radio coverage. GAMA is funded by the STFC (UK), the ARC (Australia), the AAO, and the participating institutions. The GAMA website is `http://www.gama-survey.org/`.

To construct the GAMA-Mock the DiRAC Data Centric system at Durham University, operated by the Institute for Computational Cosmology on behalf of the STFC DiRAC HPC Facility (`www.dirac.ac.uk`), was used. This equipment was funded by BIS National E-infrastructure capital grant ST/K00042X/1, STFC capital grant ST/H008519/1, and STFC DiRAC Operations grant ST/K003267/1 and Durham University. DiRAC is part of the National E-Infrastructure. The development of the GAMA-Mock was supported by a European Research Council Starting grant (DEGAS-259586) and the Royal Society.

Software credit: We like to thank the developers and curators of the following software that this paper benefits from: IPYTHON (Pérez & Granger 2007), MATPLOTLIB (Hunter 2007), NUMPY (van der Walt et al. 2011), PANDAS (McKinney 2012) and SCIPY (Jones et al. 2001).

REFERENCES

Alpaslan M., et al., 2015, MNRAS, 451, 3249
Aragon-Calvo M. A., van de Weygaert R., Jones B. J. T., Mobasher B., 2015, MNRAS, 454, 463
Baldry I. K., et al., 2010, MNRAS, 404, 86
Baldry I. K., et al., 2014, MNRAS, 441, 2440
Baum W. A., 1962, in McVittie G. C., ed., IAU Symposium Vol. 15, Problems of Extra-Galactic Research. p. 390
Beers T. C., Flynn K., Gebhardt K., 1990, AJ, 100, 32
Benítez N., 2000, ApJ, 536, 571
Bilicki M., et al., 2017, preprint, (`arXiv:1709.04205`)
Bolzonella M., Miralles J.-M., Pelló R., 2000, A&A, 363, 476
Bonfield D. G., Sun Y., Davey N., Jarvis M. J., Abdalla F. B., Banerji M., Adams R. G., 2010, MNRAS, 405, 987
Boylan-Kolchin M., Bullock J. S., Kaplinghat M., 2011, MNRAS, 415, L40
Boylan-Kolchin M., Bullock J. S., Kaplinghat M., 2012, MNRAS, 422, 1203
Brammer G. B., van Dokkum P. G., Coppi P., 2008, ApJ, 686, 1503
Brown M. J. I., et al., 2014, ApJS, 212, 18
Brunner R. J., Connolly A. J., Szalay A. S., Bershady M. A., 1997, ApJ, 482, L21
Budavári T., 2009, ApJ, 695, 747
Budavári T., 2012, Photometric Redshifts: 50 Years After. pp 323–335
Capak P., et al., 2007, ApJS, 172, 99
Carliles S., Budavári T., Heinis S., Priebe C., Szalay A. S., 2010, ApJ, 712, 511
Cavuoti S., et al., 2017, MNRAS, 466, 2039
Collister A. A., Lahav O., 2004, PASP, 116, 345
Connolly A. J., Csabai I., Szalay A. S., Koo D. C., Kron R. G., Munn J. A., 1995, AJ, 110, 2655
Dahlen T., et al., 2013, ApJ, 775, 93
Driver S. P., et al., 2011, MNRAS, 413, 971
Duarte M., Mamon G. A., 2015, MNRAS, 453, 3848
Duarte M., Mamon G. A., 2016, MNRAS, 458, 1301
Duffy A. R., Schaye J., Kay S. T., Dalla Vecchia C., 2008, MNRAS, 390, L64
Edge A., Sutherland W., Kuijken K., Driver S., McMahon R., Eales S., Emerson J. P., 2013, The Messenger, 154, 32
Firth A. E., Lahav O., Somerville R. S., 2003, MNRAS, 339, 1195
Fontana A., D'Odorico S., Poli F., Giallongo E., Arnouts S., Cristiani S., Moorwood A., Saracco P., 2000, AJ, 120, 2206
Furusawa H., Shimasaku K., Doi M., Okamura S., 2000, ApJ, 534, 624
Geha M., et al., 2017, preprint, (`arXiv:1705.06743`)
Gonzalez-Perez V., Lacey C. G., Baugh C. M., Lagos C. D. P., Helly J., Campbell D. J. R., Mitchell P. D., 2014, MNRAS, 439, 264
Graham M. L., Connolly A. J., Ivezić Ž., Schmidt S. J., Jones R. L., Jurić M., Daniel S. F., Yoachim P., 2017, preprint, (`arXiv:1706.09507`)
Guo Q., et al., 2011, MNRAS, 413, 101
Hildebrandt H., et al., 2012, MNRAS, 421, 2355
Hildebrandt H., et al., 2017, MNRAS, 465, 1454
Hopkins A. M., et al., 2013, MNRAS, 430, 2047
Hunter J. D., 2007, Computing In Science & Engineering, 9, 90
Ilbert O., et al., 2009, ApJ, 690, 1236
Jones E., Oliphant T., Peterson P., et al., 2001, SciPy: Open source scientific tools for Python, `http://www.scipy.org/`
Kafle P. R., Sharma S., Lewis G. F., Bland-Hawthorn J., 2014, ApJ, 794, 59
Kafle P. R., Sharma S., Lewis G. F., Robotham A. S. G., Driver S. P., 2018, MNRAS, 475, 4043
Klypin A., Kravtsov A. V., Valenzuela O., Prada F., 1999, ApJ, 522, 82
Kovač K., et al., 2010, ApJ, 708, 505
Kuijken K., et al., 2015, MNRAS, 454, 3500
Laigle C., et al., 2016, ApJS, 224, 24
Le Borgne D., Rocca-Volmerange B., 2002, A&A, 386, 446
Leistedt B., Hogg D. W., 2017, ApJ, 838, 5
Libeskind N. I., Guo Q., Tempel E., Ibata R., 2016, ApJ, 830, 121
Liske J., et al., 2015, MNRAS, 452, 2087
Loh E. D., Spillar E. J., 1986, ApJ, 303, 154
Mamon G. A., Łokas E. L., 2005, MNRAS, 363, 705
Mamon G. A., Biviano A., Murante G., 2010, A&A, 520, A30
Mamon G. A., Biviano A., Boué G., 2013, MNRAS, 429, 3079
Matthews D. J., Newman J. A., 2010, ApJ, 721, 456
McKinney W., 2012, Python for data analysis: Data wrangling with Pandas, NumPy, and IPython. O'Reilly Media, Inc.
Merson A. I., et al., 2013, MNRAS, 429, 556
Morrison C. B., Hildebrandt H., Schmidt S. J., Baldry I. K., Bilicki M., Choi A., Erben T., Schneider P., 2017, MNRAS, 467, 3576
Murray S. G., Power C., Robotham A. S. G., 2013, MNRAS, 434, L61
Navarro J. F., Frenk C. S., White S. D. M., 1996, ApJ, 462, 563
Navarro J. F., et al., 2004, MNRAS, 349, 1039
Pawlowski M. S., Ibata R. A., Bullock J. S., 2017, preprint, (`arXiv:1710.07639`)
Pérez F., Granger B. E., 2007, Comput. Sci. Eng., 9, 21
Rahman M., Mendez A. J., Ménard B., Scranton R., Schmidt S. J., Morrison C. B., Budavári T., 2016, MNRAS, 460, 163
Robotham A., et al., 2010, Publ. Astron. Soc. Australia, 27, 76
Robotham A. S. G., et al., 2011, MNRAS, 416, 2640
Robotham A. S. G., et al., 2012, MNRAS, 424, 1448
Sadeh I., Abdalla F. B., Lahav O., 2016, PASP, 128, 104502
Sheth R. K., Mo H. J., Tormen G., 2001, MNRAS, 323, 1

Wojtak R., Łokas E. L., Gottlöber S., Mamon G. A., 2005, MNRAS, 361, L1
Wolf C., 2009, MNRAS, 397, 520
Wright A. H., et al., 2016, MNRAS, 460, 765
van der Walt S., Colbert S. C., Varoquaux G., 2011, Computing in Science & Engineering, 13