



**HAL**  
open science

# The effect of dust bias on the census of neutral gas and metals in the high-redshift Universe due to SDSS-II quasar colour selection

Jens-Kristian Krogager, Johan P.U. Fynbo, Palle Møller, Pasquier Noterdaeme, Kasper E. Heintz, Max Pettini

## ► To cite this version:

Jens-Kristian Krogager, Johan P.U. Fynbo, Palle Møller, Pasquier Noterdaeme, Kasper E. Heintz, et al.. The effect of dust bias on the census of neutral gas and metals in the high-redshift Universe due to SDSS-II quasar colour selection. *Monthly Notices of the Royal Astronomical Society*, 2019, 486 (3), pp.4377-4397. 10.1093/mnras/stz1120 . hal-02136774

**HAL Id: hal-02136774**

**<https://hal.science/hal-02136774v1>**

Submitted on 30 May 2024

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# The effect of dust bias on the census of neutral gas and metals in the high-redshift Universe due to SDSS-II quasar colour selection

Jens-Kristian Krogager <sup>1</sup>★, Johan P. U. Fynbo,<sup>2,3</sup> Palle Møller <sup>4</sup>,  
Pasquier Noterdaeme,<sup>1</sup> Kasper E. Heintz <sup>3,5</sup> and Max Pettini<sup>6</sup>

<sup>1</sup>*Institut d'Astrophysique de Paris, CNRS-SU, UMR7095, 98bis bd Arago, F-75014 Paris, France*

<sup>2</sup>*Dark Cosmology Centre, Niels Bohr Institute, University of Copenhagen, Juliane Maries Vej 30, DK-2100 Copenhagen Ø, Denmark*

<sup>3</sup>*The Cosmic Dawn Center, Niels Bohr Institute, University of Copenhagen, Juliane Maries Vej 30, DK-2100 Copenhagen Ø, Denmark*

<sup>4</sup>*European Southern Observatory, Karl-Schwarzschildstrasse 2, D-85748 Garching bei München, Germany*

<sup>5</sup>*Centre for Astrophysics and Cosmology, Science Institute, University of Iceland, Dunhagi 5, 107 Reykjavík, Iceland*

<sup>6</sup>*Institute of Astronomy, Kavli Institute for Cosmology, Madingley Road, Cambridge CB3 0HA, UK*

Accepted 2019 March 29. Received 2019 March 22; in original form 2018 August 28

## ABSTRACT

We present a systematic study of the impact of a dust bias on samples of damped Ly  $\alpha$  absorbers (DLAs). This bias arises as an effect of the magnitude and colour criteria utilized in the Sloan Digital Sky Survey (SDSS) quasar target selection up until data release 7 (DR7). The bias has previously been quantified assuming only a contribution from the dust obscuration. In this work, we apply the full set of magnitude *and* colour criteria used up until SDSS-DR7 in order to quantify the full impact of dust biasing against dusty and metal-rich DLAs. We apply the quasar target selection algorithm on a modelled population of intrinsic colours, and by exploring the parameter space consisting of redshift, ( $z_{\text{QSO}}$  and  $z_{\text{abs}}$ ), optical extinction, and HI column density, we demonstrate how the selection probability depends on these variables. We quantify the dust bias on the following properties derived for DLAs at  $z \approx 3$ : the incidence rate, the mass density of neutral hydrogen and metals, and the average metallicity. We find that all quantities are significantly affected. When considering all uncertainties, the mass density of neutral hydrogen is underestimated by 10–50 per cent, and the mass density in metals is underestimated by 30–200 per cent. Lastly, we find that the bias depends on redshift. At redshift  $z = 2.2$ , the mass density of neutral hydrogen and metals might be underestimated by up to a factor of 2 and 5, respectively. Characterizing such a bias is crucial in order to accurately interpret and model the properties and metallicity evolution of absorption-selected galaxies.

**Key words:** galaxies: high-redshift – quasars: absorption lines – cosmology: observations.

## 1 INTRODUCTION

Quasar absorption systems are invaluable tools for our understanding of neutral gas at high redshift, as this is currently the only way to study the neutral gas phases beyond the local environment where emission studies are sensitive. One such class of quasar absorption systems, the so-called damped Ly  $\alpha$  absorbers (DLAs), are of particular importance since the high column density of HI ( $N_{\text{HI}} > 2 \times 10^{20} \text{ cm}^{-2}$ ; Wolfe et al. 1986) ensures that the hydrogen is predominantly neutral. Due to the low fraction of ionized gas in DLAs, we can infer very precise measurements of the abundances of metal species. Moreover, the neutral gas mass density in the Universe is dominated by these high column density

systems (Lanzetta, Wolfe & Turnshek 1995; Prochaska, Herbert-Fort & Wolfe 2005; Noterdaeme et al. 2009b). Since the neutral gas phase is needed in order to form molecules and subsequently stars, DLAs therefore provide important clues about the gas reservoir from which stars form over most of cosmic time.

Since certain elements (e.g. Fe, Cr, and Mn) tend to deplete strongly into dust grains, their gas-phase abundances (as measured from absorption lines) are lowered according to the strength of the depletion. By studying abundance ratios in DLAs, it is therefore possible to conclude that some amount of dust must be present in most DLAs (e.g. Ledoux, Petitjean & Srianand 2003; De Cia et al. 2016). Moreover, it is found that the optical extinction,  $A_V$ , scales with  $N_{\text{HI}}$  and metallicity (Vladilo & Péroux 2005; Zafar & Watson 2013). The more metal-rich and high-column density systems will therefore have a higher dust opacity. This will lead to a bias against metal-rich and high  $N_{\text{HI}}$  systems as these are

\* E-mail: krogager@iap.fr



more likely to be missed in optically selected, flux-limited surveys (Pei, Fall & Bechtold 1991). The effect of obscuration has been studied statistically throughout the last decades, reaching conflicting conclusions about the average reddening from DLAs. Most studies agree that the average reddening is low but measurements still present a rather large scatter ranging from around 0.002 to 0.01 mag, with the strength of the effect varying from weak to no bias in optically selected DLA samples (e.g. Pei et al. 1991; Murphy & Liske 2004; Vladilo, Prochaska & Wolfe 2008; Pontzen & Pettini 2009; Frank & Péroux 2010; Kaplan et al. 2010; Khare et al. 2012). Alternatively, by selecting quasars purely based on their radio properties, the effect of a dust bias can be largely overcome as electromagnetic radiation at radio wavelengths does not suffer from dust obscuration (Ellison et al. 2001; Jorgenson et al. 2006). Yet, an optical counterpart must still be detectable to spectroscopically confirm the source as being a quasar. The largest analysis of radio selected quasars has put an upper limit on the average reddening from DLAs of  $\langle E(B - V) \rangle < 0.040$  mag (Ellison, Hall & Lira 2005). However, the study is limited by the small sample size of only 14 DLAs from a total of 42 radio-selected quasars with optical and near-infrared data available. Combining optical and radio constraints, Pontzen & Pettini (2009) estimate that a small fraction of absorbers (less than around 5 per cent) will be missed due to dust obscuration, but that the mass density of metals,  $\Omega_Z$ , might be underestimated by up to a factor of 2. Recently, Murphy & Bernet (2016) reanalysed the quasar sample of the Sloan Digital Sky Survey (SDSS; York et al. 2000) data release 7 (DR7). The authors find a low average reddening of  $\langle E(B - V) \rangle = 0.003 \pm 0.001$  mag and conclude that the optical criteria (both in terms of flux and colours) used for target selection in SDSS prior to DR7 have a negligible effect on their estimated reddening. While the authors replicate the complex selection function of SDSS-DR7 to support their conclusion, they do not provide a full analysis of the impact of the selection effects on the observed metallicity and  $N_{\text{HI}}$  distributions.

While the number of systems missed due to a dust bias might be small (less than 5 per cent in number, Pontzen & Pettini 2009), these systems are preferentially metal-rich and dusty and thus more likely to harbour cold neutral gas and molecules (Ledoux et al. 2003; Srianand et al. 2005, 2008; Noterdaeme et al. 2009a, 2018). It is therefore important to quantify the full implications of dust reddening and obscuration for our understanding of the cold ( $T \sim 100$  K), dust-rich gas phases in DLAs. In an attempt to identify this obscured population of dusty DLAs, we have applied optical and near-infrared photometric criteria to observe reddened quasars that are not targeted by the SDSS selection algorithm (Fynbo et al. 2013; Krogager et al. 2015, 2016b). Based on these tailored surveys, we have identified three dusty DLAs at high redshift ( $z \approx 2$ ) towards quasars that were not flagged as quasar candidates in SDSS (Krogager et al. 2016a; Fynbo et al. 2017; Heintz et al. 2018b). All of these absorbers have high metallicity, and while cold ( $T \sim 100$  K), neutral gas is only securely detected in the absorber HAQ2225+0527 (Krogager et al. 2016a), it is highly probable that higher resolution spectroscopy of the remaining DLAs will reveal similarly cold, neutral gas.

In this work, we study the combined effects of dust obscuration and reddening on the quasar selection in SDSS-DR7 (Richards et al. 2002). In order to assess the selection probability of quasars, we first simulate intrinsic quasar colours and magnitudes as a function of redshift. We then introduce various absorber properties in front of the intrinsic quasar photometry and replicate the SDSS selection algorithm in order to quantify the effect of various parameters. Following the analysis of Pontzen & Pettini (2009), we use our

calculated selection probabilities to infer the intrinsic distribution functions for metallicity and  $N_{\text{HI}}$ , and find that the full, combined optical selection (using both colour and flux-limit) leads to a significantly stronger bias than assuming a flux-limit alone. This is somewhat alleviated if assuming that a small fraction of quasars are selected irrespective of their optical properties (similar to the cross-matching to radio sources performed in the SDSS). We stress that all references to SDSS throughout this work refers solely to the first epochs of SDSS (I and II up until DR7) before the onset of the Baryon Oscillation Spectroscopic Survey (BOSS; also referred to as SDSS-III), since the spectroscopic target selection changed completely from the algorithm described in Richards et al. (2002) for SDSS I/II to the more complex algorithms used for BOSS (Richards et al. 2009; Yèche et al. 2010; Bovy et al. 2011; Kirkpatrick et al. 2011; Ross et al. 2012). While the spectroscopic target selection for quasars is more complex for the BOSS sample, the selection is still based on optical information and is therefore still susceptible to biases. Moreover, the average reddening effect by DLAs has only been studied for DLAs selected from the quasar sample of SDSS-DR7 (G74). We therefore do not consider the quasar samples of later data releases in the current work. The effect of the BOSS quasar selection will be studied in a forthcoming paper.

This paper is organized as follows. In Section 2, we describe our simulated photometry and the calculation of selection probabilities. In Section 3, we present our calculation of fractional completeness of absorption-derived quantities such as the gas mass density of neutral gas and metals in DLAs. The systematic uncertainties related to our modelling are investigated in Section 4. In Section 5, we discuss the results and implications of our work as well as the related caveats, and lastly in Section 6, we summarize our results.

Throughout this work, we will assume a flat  $\Lambda$ CDM cosmology with  $H_0 = 68 \text{ kms}^{-1} \text{ Mpc}^{-1}$ ,  $\Omega_\Lambda = 0.69$ , and  $\Omega_M = 0.31$  (Planck Collaboration 2014).

## 2 QUASAR COLOUR SELECTION

We investigate the effect of dust reddening by replicating the colour and magnitude selection algorithm of SDSS. As input to the selection algorithm, we simulate an intrinsic population of quasars at different redshifts ( $z_{\text{QSO}}$ ) with varying amounts of foreground dust ( $A_V$ ) at various absorption redshifts ( $z_{\text{abs}}$ ) for two broad classes of extinction laws (SMC and LMC; as quantified by Gordon et al. 2003). Moreover, we include the reddening effect of strong HI absorption in our model. In the following section, we will describe the details of how we simulate the intrinsic quasar properties, how we apply the SDSS quasar selection algorithm, and how we simulate the absorption properties and calculate the selection probabilities. Using a physically motivated model in which metallicity ( $Z$ ) and  $N_{\text{HI}}$  scale with  $A_V$ , we then calculate the selection probabilities as a function of  $Z$  and  $N_{\text{HI}}$ . We use this two-dimensional (2D) selection probability to fit intrinsic distributions of  $Z$  and  $N_{\text{HI}}$  in order to reproduce the observed distributions of  $Z$  by Rafelski et al. (2012) and Jorgenson, Murphy & Thompson (2013) and of  $N_{\text{HI}}$  by Noterdaeme et al. (2009b) as well as the observed average reddening measured by Murphy & Bernet (2016). Lastly, we calculate the fractional completeness in various observable quantities for DLAs following Pontzen & Pettini (2009).

### 2.1 Constructing the statistical samples

In order to make a meaningful statistical comparison between our model and the data, we only consider quasars brighter than the

magnitude-limit used for the complete sample of quasars in SDSS ( $i \leq 19.1$ ). The quasar sample of SDSS does include a significant number of targets fainter than this limit, but the completeness is low ( $\sim 30$  per cent for  $z_{\text{QSO}} \approx 3$  based on the luminosity function at this redshift). It is therefore not possible for us to replicate this deeper selection (referred to as ‘high- $z$  quasar selection’ or ‘*griz*’-selection, see Section 2.3) in a statistically meaningful way.

Since the data by Rafelski et al. (2012), Noterdaeme et al. (2009b), and Murphy & Bernet (2016) are based on the full sample of DLAs in SDSS up until DR7 irrespective of the brightness of the background quasar, we first have to clean the samples and remove DLAs observed towards quasars that do not meet the  $i$ -band flux limit. The sample by Rafelski et al. (2012) is composed of a purely SDSS-selected sub-sample from DR5 (their table 2) together with a sample compiled from the literature (their table 3). Since a large part of the literature sample is not covered by the SDSS foot-print, we are not able to obtain SDSS  $i$ -band magnitudes. Instead, we have used available  $i$ -band photometry from the Pan-STARRS1 (Chambers et al. 2016) as this survey uses a very similar  $i$ -band to that of SDSS. When no photometry is available from neither SDSS nor Pan-STARRS1, we have used available broad-band photometry in  $B$ ,  $V$ , and  $R$  bands to infer an  $i$ -band magnitude based on colour transformations between the various magnitude systems. In those cases, the targets are several magnitudes brighter than the  $i \leq 19.1$  limit, and the large uncertainty in the colour transformations is therefore not important. For one target (J0242–2917) we have not been able to obtain any photometric information. This object has been excluded from the final sample. Based on the photometry, we discard all DLAs for which the background quasar is fainter than  $i = 19.1$ . Moreover, we exclude targets where the metallicity is based on a combination of limits or on refractory elements alone (Fe, Ni, and Cr).

We furthermore include the DLA sample of metallicities by Jorgenson et al. (2013) homogeneously selected from SDSS DR5 with  $i < 19$ . However, we remove duplicates between the two samples giving preference to measurements based on sulphur over those based on silicon and similarly we give preference to higher resolution measurements. As for the ‘Rafelski sample’, we remove measurements based on refractory elements or limits. This results in a total sample of 202 DLAs with good-quality metallicity measurements.

We highlight that the quasar selection algorithm did not change up until DR7 and hence all SDSS data releases before and including DR7 are based on the selection algorithm as presented by Richards et al. (2002). The literature sample included by Rafelski et al. (2012) is not strictly selected based on the same criteria as the SDSS selection algorithm; however, since quasars before SDSS were mainly identified either via radio observations or the UV excess method, both of which are explicitly included in the SDSS algorithm, we argue that the literature sample of Rafelski et al. would very likely have been targeted by the SDSS selection algorithm that we study in this work. This claim is further bolstered by the fact that such ‘pre-SDSS’ quasars were incorporated in the sample of quasars used to optimize the SDSS algorithm (Richards et al. 2002).

The samples by Noterdaeme et al. (2009b) and Murphy & Bernet (2016) are derived purely from SDSS-DR7 and we can therefore easily clean out the DLAs in front of quasars fainter than  $i = 19.1$ . The average reddening measurement by Murphy & Bernet (2016) is  $\langle E(B - V) \rangle = 3 \pm 1$  mmag. We find a consistent average reddening for the flux-limited sub-sample yet with a slightly larger uncertainty:  $\langle E(B - V) \rangle = 3.0 \pm 1.5$  mmag.

Similarly, we find no effect on the  $N_{\text{H I}}$  distribution function derived by Noterdaeme et al. (2009b) when restricting the sample to quasars with  $i < 19.1$ .

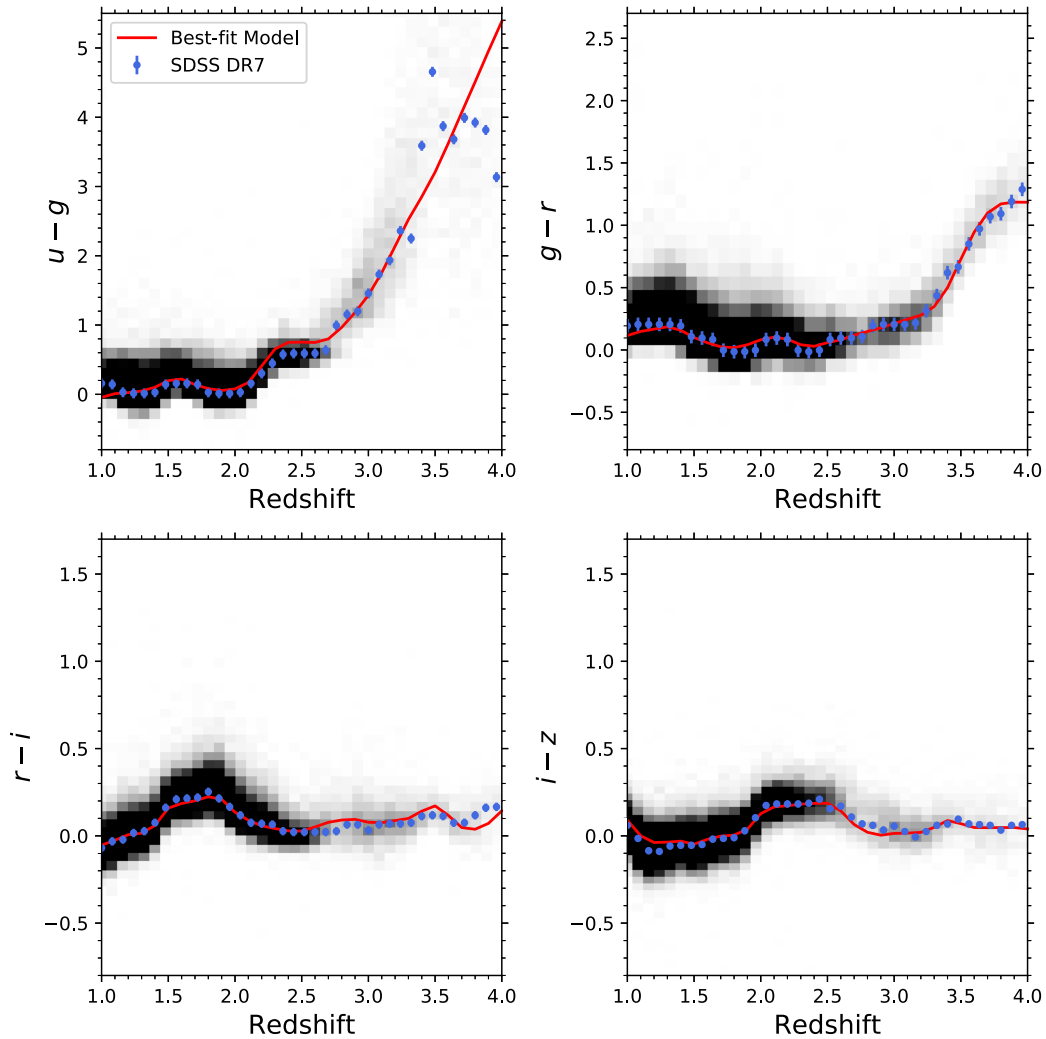
## 2.2 Simulating quasar photometry

In order to generate proper input for the SDSS quasar selection algorithm, we need not only the colour information but also the magnitudes in the 5 filters of SDSS ( $u$ ,  $g$ ,  $r$ ,  $i$ , and  $z$ ). We therefore draw samples from the quasar luminosity function as a function of quasar redshift (Manti et al. 2017) down to a limiting magnitude of  $i < 19.1$ . To get the right distribution of colours, we construct a fiducial quasar model based on available constraints from the literature. For this purpose, we use the quasar template by Selsing et al. (2016) as a basis for the spectral energy distribution; however, as the template only covers rest-frame wavelengths down to  $\sim 1000$  Å, we extend the template bluewards of  $\lambda_{\text{rest}} = 1000$  Å by assuming a power law with a spectral index of  $\alpha_\lambda = -0.3$  (Lusso et al. 2015). This break wavelength corresponds fairly well to the break in the power-law spectral shape observed by Lusso et al. (2015), who infer  $\lambda_{\text{break}} \sim 912$  Å. The template is redshifted to the given  $z_{\text{QSO}}$  and scaled to the absolute flux as determined from the luminosity function. This results in a five-element vector of magnitudes per band denoted  $\vec{m}_{0,i}$  referring to the  $i$ th random realization of quasar photometry drawn from the luminosity function.

Since the intrinsic properties of quasars are known to vary from object to object, we include variations in the intrinsic power-law spectrum. The change in power-law index is applied as a relative offset ( $\Delta\alpha$ ) with respect to the intrinsic value of  $\alpha_\lambda = -1.70$  (Selsing et al. 2016). For the fiducial model,  $\Delta\alpha$  has a value of 0. Following Krawczyk et al. (2015), we assume that the power-law index is distributed as a Gaussian centred at the intrinsic value of the template with a width of  $\sigma_\alpha = 0.186 \pm 0.001$ . Rather than randomly assigning a value of the power-law index for each realization, we calculate the offset in magnitudes per band as a function of redshift relative to the intrinsic template for a change of  $\sigma_\alpha$  in power-law index. This allows us to evaluate the mean colours at a given redshift without having to evaluate a large ensemble. The pre-evaluated offsets per band is denoted as the five-element vector,  $\vec{\Delta}_\alpha(z)$ . For each quasar realization, we then draw a random number from a Normal distribution  $X_i \sim \mathcal{N}(\mu = 0; \sigma = 1)$  and multiply  $\vec{\Delta}_\alpha(z)$  by  $X_i$ . Lastly, in order to take into account variations of emission lines, we scale  $\vec{\Delta}_\alpha(z)$  by the parameter  $w_\alpha$ , which for the fiducial model has the value  $w_\alpha = 1$ .

The last effect we include in the fiducial model is the attenuation bluewards of the quasar Ly  $\alpha$  emission line due to hydrogen Lyman-series absorption from the intergalactic medium. We use the theoretically derived average attenuation for different redshifts as a function of wavelength by Meiksin (2006) to calculate the redshift-dependent average attenuation per band denoted  $\vec{A}_{\text{IGM}}(z)$ .<sup>1</sup> The attenuation is assigned a randomly drawn weight from a Normal distribution:  $Y_i \sim \mathcal{N}(w_{\text{IGM}}, \sigma_{\text{IGM}})$ , where  $w_{\text{IGM}}$  is an overall normalization of the theoretical prediction in order to match the data, and  $\sigma_{\text{IGM}}$  determines the dispersion in IGM absorption from one object to another. This mimics effects of random sightlines through the Ly  $\alpha$  forest. For the fiducial model, the parameters of this distribution are  $(w_{\text{IGM}}, \sigma_{\text{IGM}}) = (1.0, 0.2)$ .

<sup>1</sup>The average attenuation is calculated by interpolating between the values given by Meiksin (2006).



**Figure 1.** Colour–redshift relations for quasars in SDSS-DR7. The underlying grey-scale distribution indicates the number of quasars. The mode of the observed distribution as a function of redshift is shown as blue dots with error bars corresponding to half the bin-size. The red curve shows the best-fitting quasar model for  $\Delta\alpha = -0.01 \pm 0.06$  and  $w_{\text{IGM}} = 0.67 \pm 0.01$ .

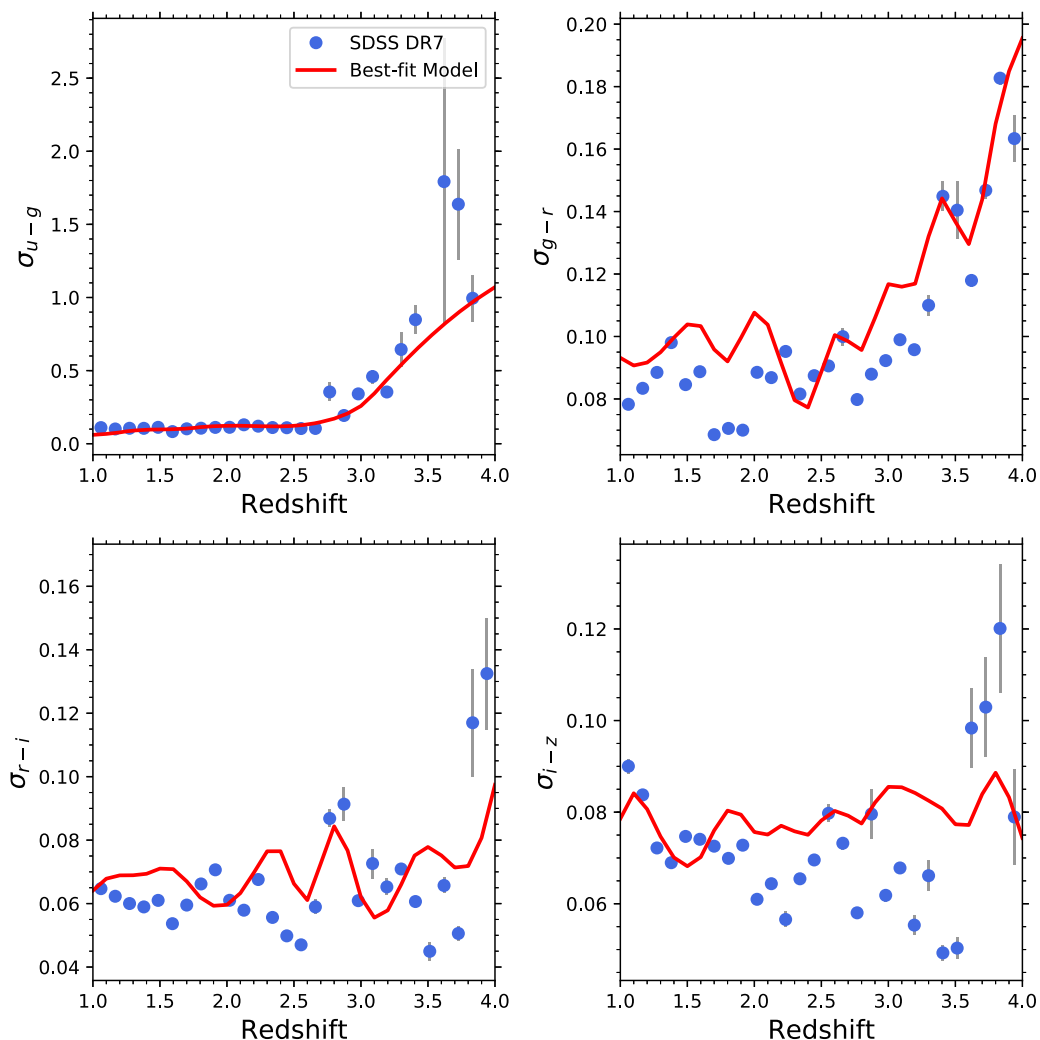
The final five-band photometry for a given realisation,  $\bar{\mathbf{m}}_i$ , is then calculated as

$$\bar{\mathbf{m}}_i = \bar{\mathbf{m}}_{0,i}(\Delta\alpha, z) + w_\alpha X_i \bar{\Delta}_\alpha(z) + Y_i(w_{\text{IGM}}, \sigma_{\text{IGM}}) \bar{\mathbf{A}}_{\text{IGM}}(z). \quad (1)$$

For each of the realizations of quasar magnitudes in the five SDSS filters, we assign photometric uncertainties following an empirically derived model between magnitude and uncertainty inferred from point source data in SDSS-DR7. We fit the distribution of photometric uncertainties as a function of observed magnitudes for each of the five bands using a fourth-order polynomial. For sources with magnitudes larger than 22.5 mag, we assign a constant photometric uncertainty. This does not significantly affect our simulations, as targets this faint will not meet the  $i$ -band criterion and hence will not be considered for target selection. Following Richards et al. (2002), we increase the photometric uncertainties by 0.0075 mag (added in quadrature) to account for systematic uncertainties in the flux calibration. The authors moreover add an error contribution from the correction for Galactic extinction, for which they add in quadrature 15 per cent of the extinction value to the photometric uncertainty. Since we do not simulate the projected sky position of our targets, we simply assume a median correction in the five filters of  $A_{\text{med}} = 0.184, 0.135, 0.098, 0.074, 0.053$  for  $u, g, r,$

$i, z,$  respectively. The median Galactic correction values are derived from 10 000 quasars in SDSS-DR7. We then add in quadrature 15 per cent of the median for each filter to the uncertainty of the given filter.

In order to optimize this fiducial model, we fit the four colour relations of the SDSS colour-space ( $u - g, g - r, r - i, i - z$ ) as a function of redshift. We parametrize the colour–redshift relations by measuring the ‘mode’ and standard deviation of the four colour distributions in redshift bins of  $\Delta z \approx 0.1$ . In order to reproduce the mode of the observed colour–redshift distribution, we only need to vary the parameters:  $\Delta\alpha$  and  $w_{\text{IGM}}$ . The widths of the distributions are not fitted during the same process, since they do not alter the mode of the distributions. We will instead constrain the width of the distributions in a subsequent step. We note that we fit the mode of the observed distributions since this quantity will be less affected by dust reddening compared to the mean and the median of the colour distributions. The best-fitting model parameters are  $\Delta\alpha = -0.01 \pm 0.06$  and  $w_{\text{IGM}} = 0.67 \pm 0.01$ . However, we find that in order to fit the  $u - g$  colour properly we need to increase the IGM absorption in the  $u$  band by a factor  $w_{\text{IGM}}^u = 1.5$ . The resulting colour–redshift relations are shown in Fig. 1.



**Figure 2.** Width of colour–redshift relations for quasars in SDSS-DR7. The red curve shows the best-fitting quasar model for  $w_\alpha = 1.5 \pm 0.2$  and  $\sigma_{\text{IGM}} = 0.13 \pm 0.01$ .

In order to constrain the parameters  $w_\alpha$  and  $\sigma_{\text{IGM}}$ , we fit the width of the colour–redshift distributions while keeping the parameters  $\Delta\alpha$  and  $w_{\text{IGM}}$  fixed to their best-fitting values. In Fig. 2, we show the model for the best-fitting parameters  $w_\alpha = 1.5 \pm 0.2$  and  $\sigma_{\text{IGM}} = 0.13 \pm 0.01$ . Although we include uncertainties in Fig. 2, these are not taken into account in the fit, since this would give much stronger weight to the low-redshift bins where there are many more quasars. In order to give equal weight to all redshifts during the fit, we use an average uncertainty of 0.01 for all measurements.

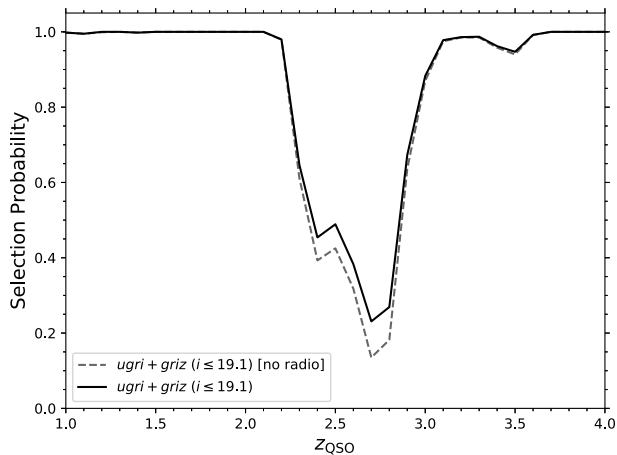
The intrinsic properties of the simulated data match well the observed properties of spectroscopically confirmed quasars as shown in Figs 1 and 2. This allows us to probe the overall redshift dependence of the quasar selection algorithm (see Section 2.3). Yet, there are still significant discrepancies for certain redshift intervals. While it is beyond the scope of the current work to reproduce in detail the quasar properties over the full redshift range, we do attempt to optimize the model in certain redshift ranges of interest. Specifically for quasars at  $z_{\text{QSO}} = 3$  and 2.5, which we use for the detailed analysis of dust bias against absorption systems in Section 3, we optimize the quasar model. The additional optimization and robust evaluation of model uncertainties are described in Appendix A.

In the quasar model described above, we have neglected the contribution of BAL quasars as these are anyways not included in the sample of quasars used for absorption analyses. Other possible effects that could affect our intrinsic quasar photometry model would be systematic variations in the emission lines, e.g. the Baldwin effect (Baldwin 1977), blazars with no apparent emission lines, or quasars with abnormal spectral shapes (Meusinger et al. 2012). However, as blazars and quasars with abnormal spectral shapes contribute very little to the overall population of quasars (Meusinger et al. 2012; Li et al. 2015), we have neglected the contribution of these in order to keep the model as simple as possible.

### 2.3 SDSS quasar candidate selection

The simulated photometry and errors described above are then passed through the steps of the selection algorithm as outlined in fig. 1 of Richards et al. (2002). The calculation is split into two selection ‘branches’: *ugri*, selected in the three-dimensional (3D)  $u - g, g - r$ , and  $r - i$  colour-space, and *griz*, selected in the 3D  $g - r, r - i, i - z$  colour-space. We have re-implemented the definition of the ‘stellar locus’ in *ugri* and *griz* colour-spaces by Richards et al.





**Figure 3.** Selection probability as function of  $z_{\text{QSO}}$  for 1000 simulated quasars in each redshift bin. The solid line shows the full selection probability ( $ugri + griz$ ) with radio selection for quasars brighter than  $i \leq 19.1$ . The grey, dashed line shows the effect of turning off the radio selection.

(2002). The details of our new implementation in PYTHON are given in Appendix C. Since we are working on simulated photometry, we do not implement the checks for photometric data-quality flags. Moreover, these photometric impurities will not introduce any significant bias but merely lower the completeness. Lastly, we apply the same magnitude cut as Richards et al. (2002), namely,  $i < 19.1$  mag and  $i < 20.2$  mag for targets in the  $ugri$  and  $griz$  colour-spaces, respectively. However, for our statistical analysis, we use a single magnitude cut of  $i < 19.1$ . The resulting fraction of simulated quasars identified through the algorithm will hereafter be referred to as the selection probability,  $P$ .

Since we only consider quasars brighter than  $i < 19.1$  for the statistical bias calculation in this work, the total selection probability is then given by the union of the two selection branches, and simultaneously fulfilling the  $i$ -band criterion:

$$P_{\text{tot}} = \frac{N(ugri \cup griz \ \& \ i < 19.1)}{N(i < 19.1)} \quad (2)$$

where  $N(ugri \cup griz)$  denotes the number of quasars that are selected either by the  $ugri$  or  $griz$  criteria.

Lastly, we implement a mock radio selection corresponding to the cross-matching between FIRST (Becker, White & Helfand 1995) and SDSS by simply allowing a constant fraction of quasars to be selected irrespective of their optical colours. Here, we assume a fraction of radio detections of 10 per cent inferred from FIRST cross-matching in independent surveys (Hewett, Foltz & Chaffee 2001; Krogager et al. 2016b). Thus, 10 per cent of the simulated quasars will be allowed to pass the quasar selection irrespective of the colour criteria. Nevertheless, the quasars still have to pass the  $i$ -band magnitude limit. The result of passing the intrinsic quasar photometry as described in Section 2.2 is shown in Fig. 3. We recover the strong drop in selection probability for quasars with redshift  $z \sim 2.5$  as demonstrated in the original work by Richards et al. (2002). This low sensitivity for quasars in this redshift range is caused by the fact that the quasar colours cross the stellar locus in the optical colour-spaces utilized by the SDSS. For the redshift range  $2.3 \leq z_{\text{QSO}} \leq 3.0$ , we have carried out a similar optimization of the quasar model as described in Appendix A.

## 2.4 Absorption properties

We here investigate the effect of applying intervening dust reddening and strong H I absorption to the quasar photometry. In the first part, we will simply calculate the selection probability for the various parameters in order to visualize the dependence of the selection probability on these variables. We therefore sample all properties uniformly and calculate the selection probability in this parameter space. In the second part, we will use a physically motivated model to calculate how the selection probability affects the observed distributions of metallicity and column density of H I.

When simulating the absorber properties, we set up a grid of quasar redshifts in the interval  $1 < z_{\text{QSO}} < 3.6$  (in steps of 0.2), with absorber redshifts in the range  $1.0 < z_{\text{abs}} < 3.5$  in steps of 0.1 and requiring that  $z_{\text{abs}} \leq z_{\text{QSO}}$ , with  $A_V$  in the interval  $0 < A_V < 2$  mag in steps of 0.1, and with  $\log(N_{\text{HI}})$  in the range  $20 < \log(N_{\text{HI}} / \text{cm}^{-2}) < 23$  in steps of 0.3. For each point in this four-dimensional (4D) parameter space  $[z_{\text{QSO}}, z_{\text{abs}}, A_V, \log(N_{\text{HI}})]$ , we then simulate an ensemble of 1000 quasars as described in Section 2.2 and apply the properties  $A_V$  and  $\log(N_{\text{HI}})$  to the simulated quasar photometry. The ensemble is then passed through the SDSS quasar selection algorithm and the selection probability is calculated. We perform this calculation twice assuming different reddening laws: either SMC- or LMC-type as parametrized by Gordon et al. (2003).

In Fig. 4, we show the resulting  $P_{ugri}$  and  $P_{griz}(i < 19.1)$  in the  $(z_{\text{abs}}, A_V)$ -grid averaged over all quasar redshifts and all values of  $\log(N_{\text{HI}})$ . For comparison, we show the selection probabilities of quasars for a purely flux-limited selection of  $i < 19.1$ . These flux-limited-only calculations were performed on the same simulated data simply by skipping the colour selection criteria and only applying the  $i$ -band magnitude cut. In Appendix B, we show the averaged  $P_{ugri}$  and  $P_{griz}(i < 19.1)$  in the  $(z_{\text{abs}}, z_{\text{QSO}})$  and  $(A_V, z_{\text{QSO}})$  parameter spaces respectively.

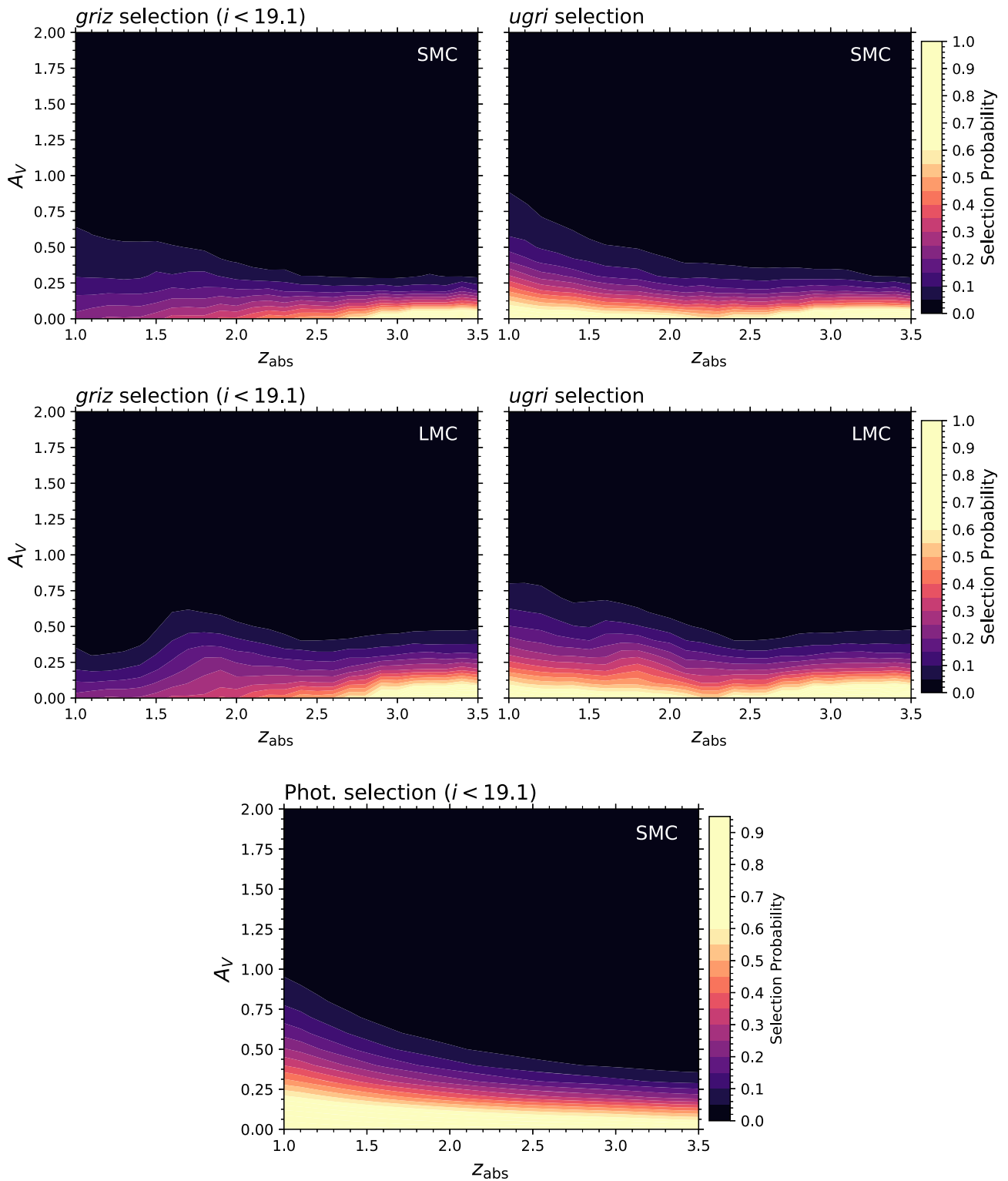
In the above analysis, we have not assumed any physical relation between absorber properties and  $A_V$ ; however, it is well known that  $A_V$  scales with both  $\log(N_{\text{HI}})$  and metallicity (e.g. Vladilo & Péroux 2005; Zafar & Watson 2013). We therefore run a similar set of models as described above assuming a fixed dust-to-metals ratio ( $\kappa_z$ ) to investigate the effect of the colour selection on the  $Z$  and  $\log(N_{\text{HI}})$  distribution of absorbers. We use the relation by Zafar & Watson (2013) to quantify the  $V$ -band dust extinction in terms of  $Z$  and  $\log(N_{\text{HI}})$ :

$$\log(A_V) = \log(N_{\text{HI}}) + \log(Z/Z_{\odot}) + \log(\kappa_z), \quad (3)$$

where  $\kappa_z$  is the dust-to-metal ratio.

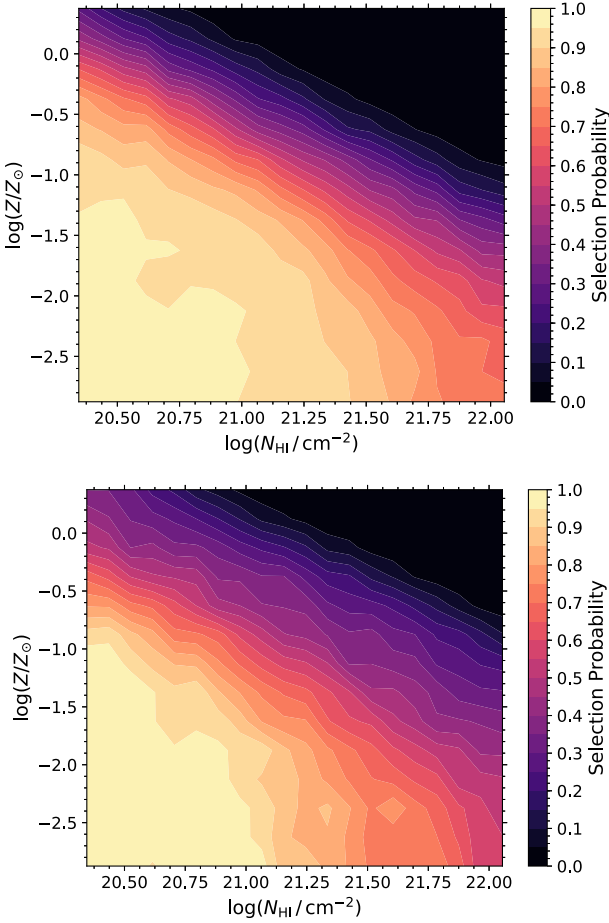
In order to compare our results to the previous work by Pontzen & Pettini (2009), we will assume an absorption redshift of  $z_{\text{abs}} = 3.0$  that corresponds to the median redshift of absorbers in the analysis by Pontzen & Pettini (2009). While Pontzen & Pettini (2009) do not worry about the redshift of the background quasar, we must take this into account when estimating the bias of the full SDSS selection criteria, since the colours depend on  $z_{\text{QSO}}$ . In the following, we assume that the background quasars are at an average  $z_{\text{QSO}} = 3.2$ . We will also investigate the selection probability at lower redshift ( $z \approx 2$ ) to demonstrate the redshift dependence of the bias.

We evaluate the selection probability at the specific redshift as a function of  $Z$  and  $N_{\text{HI}}$ , denoted  $P(Z, N_{\text{HI}})$ , in the ranges of  $-3 < \log(Z/Z_{\odot}) < 0.5$  and  $20.3 < \log(N_{\text{HI}}) < 22$ , by uniformly drawing 200 000 sets of  $Z$  and  $N_{\text{HI}}$ . For each draw, we randomly assign a realization of intrinsic quasar photometry in the five SDSS bands for the given  $z_{\text{QSO}}$  following the same method as described in Section 2.2. The optical extinction is then calculated



**Figure 4.** Selection probability for the *ugri* and *griz* selections as function of absorber redshift and  $A_V$ . The colour coding represents the selection probability based on the *griz* (only for  $i < 19.1$ ; left-hand panels) and *ugri* (right-hand panels) selection criteria. The two upper rows show the probabilities for the colour and magnitude selection for SMC- and LMC-type extinction curves, respectively. The bottom row shows the probability for a purely flux-limited selection assuming SMC-type extinction.





**Figure 5.** Selection probability as function of metallicity and H I column density assuming  $z_{\text{QSO}} = 3.2$  and  $z_{\text{abs}} = 3.0$  (top), and  $z_{\text{QSO}} = 2.5$  and  $z_{\text{abs}} = 2.2$  (bottom). Both models assume the SMC extinction curve and a dust-to-metals ratio of  $\log \kappa_Z = -21.5$ .

using equation (3), and the quasar fluxes are attenuated according to the SMC extinction law applied in the absorber rest frame. Here, we only focus on the SMC extinction curve as this the most commonly observed type of extinction for DLAs (see also Section 5.2). Furthermore, we apply the extinction from the DLA absorption itself. As before, the quasar photometry is passed through the selection algorithm and the selection probability in the 2D  $(Z, N_{\text{HI}})$ -space is calculated. In Fig. 5, we show the selection probability for two models of  $z_{\text{QSO}} = 3.2$  and  $z_{\text{abs}} = 3.0$ , and  $z_{\text{QSO}} = 2.5$  and  $z_{\text{abs}} = 2.2$ . Both models are calculated assuming a value of  $\log \kappa_Z = -21.5$  for the SMC extinction law. It is clearly seen that the selection probability drops for high metallicity and high  $\log(N_{\text{HI}})$  due to the increasing reddening moving diagonally from the bottom left corner to the top right corner.

### 3 DUST BIAS IN DAMPED LY $\alpha$ ABSORBERS

We are now able to use the inferred selection probability,  $P(Z, N_{\text{HI}})$ , to infer the completeness in metallicity and  $N_{\text{HI}}$  distributions of DLAs at  $z \approx 3$ . Since the colour selection is very redshift dependent, we also perform a calculation of the completeness for lower redshift DLAs at  $z \approx 2$  towards quasars at  $z_{\text{QSO}} \approx 2.5$  where the SDSS selection algorithm is very inefficient. For this purpose, we assume analytic, parametric functional forms for the intrinsic distributions

of  $Z$  and  $N_{\text{HI}}$ . For a given set of parameters for intrinsic distributions, we apply the selection probability to infer the corresponding biased distributions (i.e. the distributions that would be observed). The parameters for the intrinsic distributions are then varied in order to match the observed distributions inferred from SDSS data up until DR7. Specifically, we use a lognormal metallicity distribution:

$$f(Z, \mu_Z, \sigma_Z) \propto \exp\left(-\frac{(Z - \mu_Z)^2}{2\sigma_Z^2}\right), \quad (4)$$

with mean and standard deviation denoted as  $\mu_Z$  and  $\sigma_Z$ , respectively. This functional form accurately describes the observed distribution (Rafelski et al. 2014) as well as simulations (Pontzen et al. 2008; Bird et al. 2015).

For the  $N_{\text{HI}}$  distribution, we use a  $\Gamma$ -function:

$$f(N_{\text{HI}}, \alpha, N_0) \propto \left(\frac{N_{\text{HI}}}{N_0}\right)^\alpha \exp\left(-\frac{N_{\text{HI}}}{N_0}\right), \quad (5)$$

as argued by Pei & Fall (1995) and further supported by observations (Noterdaeme et al. 2009b, 2012).

We assume that the two distributions are independent since simulations of DLAs in cosmological contexts show no significant correlation between  $N_{\text{HI}}$  and halo-properties such as mass or metallicity (Pontzen et al. 2008), see also discussion in Section 5.3. We can therefore construct a 2D distribution,  $F(\theta)$ , as the product of the two distributions in the  $(Z, N_{\text{HI}})$  parameter space, such that

$$F(Z, N_{\text{HI}}, \theta) = f(Z, \mu_Z, \sigma_Z) \times f(N_{\text{HI}}, \alpha, N_0), \quad (6)$$

where  $\theta$  refers to the set of intrinsic parameters  $\theta = \{\mu_Z, \sigma_Z, \alpha, N_0\}$  for the two distributions.

We then construct the biased distributions:

$$f'(Z, \theta) = \int F(Z, N_{\text{HI}}, \theta) P(Z, N_{\text{HI}}) dN_{\text{HI}} \quad (7)$$

and

$$f'(N_{\text{HI}}, \theta) = \int F(Z, N_{\text{HI}}, \theta) P(Z, N_{\text{HI}}) dZ, \quad (8)$$

which we can compare to the actual observed distributions  $f_{\text{obs}}(Z)$  (Rafelski et al. 2012; Jorgenson et al. 2013) and  $f_{\text{obs}}(N_{\text{HI}})$  (Noterdaeme et al. 2009b). Although more recent and larger compilations of metallicity and H I column density measurements are available (e.g. Noterdaeme et al. 2012; Quiret et al. 2016), we use the data by Rafelski et al. (2012), Jorgenson et al. (2013), and Noterdaeme et al. (2009b) as these are, to our knowledge, the largest, homogeneous samples selected purely from SDSS-II up until DR7 (see Section 2.1). For later data releases, the quasars were selected based on different criteria than the colour selection algorithm that we investigate in this work. We highlight that the data presented by Rafelski et al. (2012) and Jorgenson et al. (2013) are compiled using various elements; sulphur is used whenever possible. If sulphur is not available, silicon or zinc is used, and lastly, iron is used at low metallicities as the authors argue that dust depletion does not significantly affect the abundances at low metallicity (Rafelski et al. 2012; Jorgenson et al. 2013). For the high-redshift absorbers ( $z_{\text{abs}} = 3.0$ ), we compare to metallicity measurements of DLAs in the redshift range  $2.5 < z_{\text{abs}} < 3.2$  in order to have a large enough sample. For the low-redshift absorbers ( $z_{\text{abs}} = 2.2$ ), we use metallicity measurements of DLAs in the redshift range  $1.8 < z_{\text{abs}} < 2.5$ .

**Table 1.** Fractional completeness and average reddening of  $z \approx 3$  DLA properties for different dust-to-metals ratios,  $\kappa_Z$ .

$\log \kappa_Z$	-21.1	-21.2	-21.3	-21.4	-21.5	-21.7	-21.9
$C [l_{\text{DLA}}]$	$0.72 \pm 0.03$	$0.77 \pm 0.02$	$0.81 \pm 0.02$	$0.84 \pm 0.02$	$0.87 \pm 0.01$	$0.91 \pm 0.01$	$0.93 \pm 0.01$
$C [\Omega_{\text{DLA}}]$	$0.61 \pm 0.03$	$0.68 \pm 0.03$	$0.73 \pm 0.02$	$0.77 \pm 0.02$	$0.81 \pm 0.02$	$0.86 \pm 0.01$	$0.89 \pm 0.01$
$C [Z]$	$0.48 \pm 0.05$	$0.55 \pm 0.05$	$0.62 \pm 0.05$	$0.68 \pm 0.05$	$0.74 \pm 0.04$	$0.83 \pm 0.03$	$0.89 \pm 0.02$
$C [Z]_{\text{N}_{\text{H}1}}$	$0.41 \pm 0.05$	$0.48 \pm 0.05$	$0.55 \pm 0.05$	$0.61 \pm 0.05$	$0.67 \pm 0.04$	$0.76 \pm 0.04$	$0.84 \pm 0.03$
$C [\Omega_Z]$	$0.25 \pm 0.04$	$0.32 \pm 0.04$	$0.40 \pm 0.05$	$0.47 \pm 0.05$	$0.54 \pm 0.04$	$0.66 \pm 0.04$	$0.75 \pm 0.03$
$\langle E(B - V) \rangle_{\text{obs}} / \text{mmag}$	$9.7^{+1.0}_{-0.9}$	$8.0^{+0.7}_{-0.7}$	$6.5^{+0.6}_{-0.5}$	$5.3^{+0.6}_{-0.4}$	$4.3^{+0.4}_{-0.4}$	$2.9^{+0.3}_{-0.3}$	$1.9^{+0.2}_{-0.2}$
$\langle E(B - V) \rangle_{\text{int}} / \text{mmag}$	$28.0^{+5.3}_{-5.1}$	$19.1^{+3.5}_{-3.1}$	$13.2^{+2.4}_{-1.9}$	$9.6^{+1.7}_{-1.4}$	$7.0^{+1.1}_{-1.0}$	$4.0^{+0.6}_{-0.6}$	$2.3^{+0.4}_{-0.3}$

The best-fitting intrinsic parameters,  $\theta$ , for the two distributions can then be found by  $\chi^2$  minimization:

$$\chi^2 = \sum \frac{(\tilde{f}_{\text{obs}}(Z) - \tilde{f}'(Z, \theta))^2}{\tilde{f}_{\text{obs}}(Z)} \quad (9)$$

$$+ \sum \frac{(\tilde{f}_{\text{obs}}(\text{N}_{\text{H}1}) - \tilde{f}'(\text{N}_{\text{H}1}, \theta))^2}{\tilde{f}_{\text{obs}}(\text{N}_{\text{H}1})}, \quad (10)$$

where the  $\tilde{f}'$  and  $\tilde{f}_{\text{obs}}$  refer to the normalized distributions, e.g.  $\tilde{f}_{\text{obs}}(Z) = f_{\text{obs}}(Z) / \int dZ f_{\text{obs}}(Z)$ .

In order to make a direct comparison to the work by Pontzen & Pettini (2009), we calculate the fractional completeness in the same quantities as they do. These are the incidence rate:

$$l_{\text{DLA}} \propto \int F(Z, \text{N}_{\text{H}1}, \theta) d\text{N}_{\text{H}1} dZ \equiv \phi_0; \quad (11)$$

the total mass density of neutral hydrogen in DLAs:

$$\Omega_{\text{DLA}} \propto \int \text{N}_{\text{H}1} F(Z, \text{N}_{\text{H}1}, \theta) d\text{N}_{\text{H}1} dZ; \quad (12)$$

the mean metallicity of DLAs:

$$\langle Z \rangle \propto \int Z F(Z, \text{N}_{\text{H}1}, \theta) d\text{N}_{\text{H}1} dZ / \phi_0; \quad (13)$$

the total mass density of metals in DLAs:

$$\Omega_Z \propto \int \text{N}_{\text{H}1} Z F(Z, \text{N}_{\text{H}1}, \theta) d\text{N}_{\text{H}1} dZ; \quad (14)$$

and the column density weighted mean metallicity of DLAs:

$$\langle Z \rangle_{\text{N}_{\text{H}1}} = \frac{\Omega_Z}{\Omega_{\text{DLA}}}. \quad (15)$$

The fractional completeness,  $C$ , is then calculated for the above quantities as the ratio of the biased over the intrinsic quantity, e.g.:

$$C [l_{\text{DLA}}] = \frac{\int F(Z, \text{N}_{\text{H}1}, \theta) P(Z, \text{N}_{\text{H}1}) d\text{N}_{\text{H}1} dZ}{\int F(Z, \text{N}_{\text{H}1}, \theta) d\text{N}_{\text{H}1} dZ}. \quad (16)$$

Moreover, we calculate the mean biased and intrinsic reddening as

$$\langle E(B - V) \rangle_{\text{obs}} = \frac{\kappa_Z}{R_V \phi_0} \int \text{N}_{\text{H}1} Z F(\theta) P d\text{N}_{\text{H}1} dZ \quad (17)$$

and

$$\langle E(B - V) \rangle_{\text{int}} = \frac{\kappa_Z}{R_V \phi_0} \int \text{N}_{\text{H}1} Z F(\theta) d\text{N}_{\text{H}1} dZ, \quad (18)$$

where  $F(\theta)$  and  $P$  are shorthand notations for  $F(Z, \text{N}_{\text{H}1}, \theta)$  and  $P(Z, \text{N}_{\text{H}1})$ , respectively.

In Table 1, we summarize the results from our analysis for different dust-to-metals ratios,  $\kappa_Z$ , in order to match the range of probable values of the average observed reddening:  $2 < \langle E(B - V) \rangle < 10$  mmag (Murphy & Bernet 2016). The best-fitting intrinsic

distributions of  $Z$  and  $\text{N}_{\text{H}1}$  are shown in Fig. 6 as the dotted blue lines. The corresponding biased distributions are shown as solid lines together with the observational data, to which the biased distributions are fitted. The red lines result from a purely flux-limited selection as described below. In Table 2, we summarize the results for the low-redshift configuration of  $z_{\text{QSO}} = 2.5$  and  $z_{\text{abs}} = 2.2$ . As the metallicity distribution is very sparsely sampled in this redshift range, we advise that these results should be interpreted with caution. Moreover, there are no direct measurements of the average reddening induced by DLAs in the SDSS-DR7 at this redshift. We therefore simply assume the same range of  $\kappa_Z$  as used to match the observed range of  $\langle E(B - V) \rangle$  for the high-redshift absorbers.

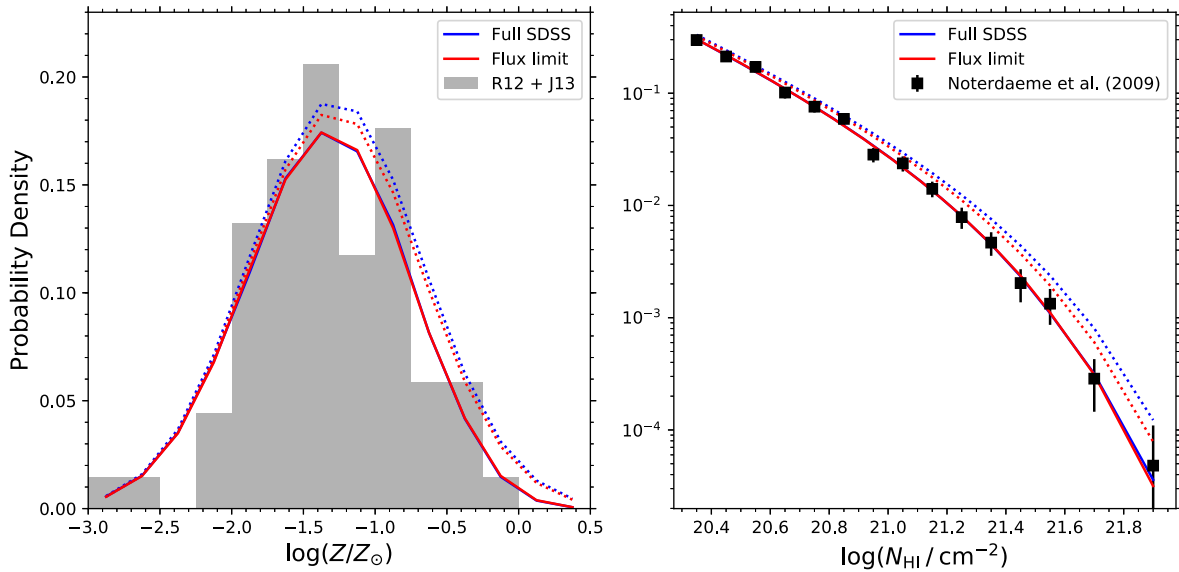
Lastly, we calculate the fractional completeness assuming a purely flux-limited selection in order to evaluate the separate contributions of the colour-selection and the magnitude limit and to compare our results to the previous results by Pontzen & Pettini (2009). For the comparison to be valid, we only compare models with a similar dust-to-metals ratio as used by the authors. Pontzen & Pettini (2009) infer a most likely value of the dust opacity at  $\lambda \sim 1900 \text{ \AA}$  of  $\log \tau_0 \approx -21.8$ , which corresponds to a value of  $\kappa_Z$  of  $\log \kappa_Z \approx -21.5$ .<sup>2</sup> For the comparison, we therefore focus on models with  $\log \kappa_Z = -21.5$ , see the red lines in Fig. 6. The calculated fractional completeness for a purely flux-limited sample is given in Table 3. We highlight that only the high-redshift results should be compared to the results by Pontzen & Pettini (2009), as this is the only redshift range considered by these authors.

We find that the main contribution to the incompleteness at  $z \approx 3$  is due to the flux-limit imposed on the quasar selection. This is also evident from the close resemblance between the dotted blue and red lines in Fig. 6. However, the fractional completeness is systematically lowered when taking into account the full quasar selection algorithm. The effect on the fractional completeness of  $\Omega_Z$  is 0.05, corresponding to a decrease in completeness by  $\sim 10$  per cent. For the lower redshift model, the fractional completeness is significantly affected by the additional colour selection:  $\Omega_{\text{DLA}}$  and  $\Omega_Z$  are lowered by 0.15 (22 per cent) and 0.13 (37 per cent), respectively, when taking the full selection algorithm into account.

#### 4 SYSTEMATIC UNCERTAINTIES

Our modelling of the dust bias in Section 3 is based on our best-fitting intrinsic quasar model (Section 2.2). It is therefore important

<sup>2</sup>We note that Pontzen & Pettini use a metallicity dependent dust model where low metallicity absorbers below a threshold of roughly  $Z_{\text{lim}} < 0.1 Z_{\odot}$  have a lower dust-to-metals ratio. This does not significantly alter the calculation of fractional completeness as this simply lowers the dust extinction at low metallicities where the completeness is already roughly unity. The converted value of  $\kappa_Z$  is thus quoted for high metallicities,  $Z \sim Z_{\odot}$ .



**Figure 6.** Metallicity and  $N_{\text{HI}}$  distribution functions from literature. The dotted lines show the intrinsic, best-fitting functions, while the solid lines show the corresponding biased functions assuming either the full SDSS colour selection plus radio (blue) or the flux-limited selection only (red). The red and blue solid lines overlap making it hard to distinguish the two. Here, we assume  $z_{\text{QSO}} = 3.2$  and  $z_{\text{abs}} = 3.0$  for SMC type dust with a single dust-to-metals ratio of  $\log(\kappa_Z) = -21.5$ .

**Table 2.** Fractional completeness for  $z \approx 2$  DLAs for different values of  $\kappa_Z$ .

$\log \kappa_Z$	-21.1	-21.5	-21.9
$C [l_{\text{DLA}}]$	$0.58 \pm 0.05$	$0.77 \pm 0.03$	$0.91 \pm 0.01$
$C [\Omega_{\text{DLA}}]$	$0.49 \pm 0.05$	$0.68 \pm 0.03$	$0.84 \pm 0.02$
$C [ \langle Z \rangle ]$	$0.40 \pm 0.04$	$0.57 \pm 0.03$	$0.71 \pm 0.02$
$C [ \langle Z \rangle_{N_{\text{HI}}} ]$	$0.34 \pm 0.04$	$0.51 \pm 0.03$	$0.67 \pm 0.03$
$C [\Omega_Z]$	$0.16 \pm 0.03$	$0.35 \pm 0.03$	$0.56 \pm 0.03$
$\langle E(B - V) \rangle_{\text{obs}} / \text{mmag}$	$17.1^{+3.5}_{-3.2}$	$8.0^{+1.3}_{-1.2}$	$3.3^{+0.4}_{-0.4}$
$\langle E(B - V) \rangle_{\text{int}} / \text{mmag}$	$61.0^{+12.1}_{-12.2}$	$17.7^{+3.6}_{-3.1}$	$5.4^{+0.9}_{-0.8}$

**Table 3.** Fractional completeness for DLA statistics assuming only flux-limited selection ( $i < 19.1$  mag).

Model	PP09	$z_{\text{QSO}} = 3.2^a$	$z_{\text{QSO}} = 2.5^a$
$C [l_{\text{DLA}}]$	$0.93^{+0.07}_{-0.03}$	$0.90 \pm 0.01$	$0.88 \pm 0.02$
$C [\Omega_{\text{DLA}}]$	$0.87^{+0.10}_{-0.06}$	$0.85 \pm 0.01$	$0.83 \pm 0.03$
$C [ \langle Z \rangle ]$	$0.75^{+0.25}_{-0.09}$	$0.76 \pm 0.04$	$0.67 \pm 0.05$
$C [ \langle Z \rangle_{N_{\text{HI}}} ]$	$0.63^{+0.19}_{-0.20}$	$0.69 \pm 0.04$	$0.59 \pm 0.06$
$C [\Omega_Z]$	$0.56^{+0.19}_{-0.26}$	$0.59 \pm 0.04$	$0.48 \pm 0.06$
$\langle E(B - V) \rangle_{\text{obs}} / \text{mmag}$	$4^{+12}_{-2}$	$4.4^{+0.4}_{-0.4}$	$7.2^{+0.9}_{-0.8}$
$\langle E(B - V) \rangle_{\text{int}} / \text{mmag}$	$8^{+24}_{-5}$	$6.7^{+1.1}_{-0.8}$	$13.1^{+2.1}_{-2.2}$

<sup>a</sup>The models reported here assume a fixed value of  $\log \kappa_Z = -21.5$ .

to understand to what degree the quasar model affects the inferred fractional completeness. In the following section, we study the systematic uncertainty introduced by variations in the intrinsic quasar properties. Furthermore, we assess the possibility of a non-Gaussian metallicity distribution.

#### 4.1 Intrinsic quasar model

In order to test the effect of the intrinsic quasar model, we recalculate the selection probability  $P(Z, N_{\text{HI}})$  for two extreme ranges of

the best-fitting model parameters. For the first model, referred to as ‘199’, we offset the parameters by  $-3\sigma$  taking into account parameter correlations (see details in Appendix A). This results in a *bluer* intrinsic quasar shape. Similarly for the second model, referred to as ‘u99’, we offset the parameters by  $+3\sigma$ , resulting in a *redder* intrinsic quasar model. We then recalculate the fractional completeness in the various properties listed in Table 1 as a function of  $\kappa_Z$ . The  $3\sigma$  difference in the inferred fractional completeness is at most 0.02, thus well within the statistical uncertainty. Similarly, the systematic uncertainty on  $\langle E(B - V) \rangle$  is  $\sim 0.5$  mmag and  $< 0.1$  mmag for, respectively, the highest and lowest value of  $\kappa_Z$  considered in this work, both significantly smaller than the statistical uncertainty.

The systematic uncertainty might be more significant for the lower-redshift models, however, due to the complexity of fitting the intrinsic quasar colour distributions when these are heavily affected by the selection algorithm, it is not possible to formally address the error in the same way as above. Instead, we assume the same range of parameters as used for the high-redshift model and offset the quasar model similarly. We then recalculate the selection probability and the fractional completeness. The systematic uncertainties are significantly stronger for the low-redshift model, with the quantities  $l_{\text{DLA}}$ ,  $\Omega_{\text{DLA}}$ , and  $\Omega_Z$  being mostly affected. For these quantities, the  $1\sigma$  systematic uncertainty is of the same order as the statistical uncertainty yet slightly smaller ( $\sim 0.01$ ).

The precise systematic uncertainties for each quantity are given in Table A2 in Appendix A.

#### 4.2 Non-Gaussian metallicity distribution

In the bias analysis above, we have assumed a lognormal distribution of metallicity as the intrinsic shape. However, given the strong bias against high-metallicity absorbers, the observed distribution might equally well arise from a skewed distribution where the high-metallicity tail has been suppressed by the dust bias. In the following, we will investigate the effect of an intrinsically skewed lognormal distribution. We use the following parametrization of the

skewed normal distribution (O’Hagan & Leonard 1976):

$$f(Z, \mu_z, \sigma_z, \zeta) = 2\phi\left(\frac{Z - \mu_z}{\sigma_z}\right) \Phi\left(\zeta \frac{Z - \mu_z}{\sigma_z}\right), \quad (19)$$

where  $\phi$  and  $\Phi$  are the probability density and cumulative probability of the normal distribution, respectively. The skewness is controlled by the parameter  $\zeta$ , which for positive (negative) values will skew the distribution towards higher (lower) values than the mean. For  $\zeta = 0$ , one recovers the symmetrical normal distribution.

We then perform the same analysis as in Section 3; however, we use the skewed normal distribution of metallicity instead of the standard normal distribution. For high-redshift configuration ( $z_{\text{abs}} = 3$ ), we find that the skewed distribution is poorly constrained:  $\zeta = 1.4 \pm 1.3$ . For the lower redshift model ( $z_{\text{abs}} = 2.2$ ), we are not able to constrain  $\zeta$  at all. Hence, in both cases we find no evidence for a skewed intrinsic distribution providing a better fit to the data.

## 5 DISCUSSION

In this work, we have calculated the fractional completeness of DLA statistics in terms of the incidence rate, the mass density of neutral hydrogen, the mean metallicity, the column density weighted mean metallicity, and the mass density of metals assuming a wide range of dust-to-metals ratios,  $\kappa_z$ . In the following discussion, we focus on the high-redshift results, as these are better constrained by data (for a discussion of the redshift dependence, see Section 5.1). Murphy & Bernet (2016) provide the latest measurement of  $\langle E(B - V) \rangle$  from SDSS-DR7 of  $3.0 \pm 1.5$  mmag (when restricting to  $i < 19.1$ ). Moreover, the authors provide a review of recent measurements in the literature. The various studies reach different results concerning  $\langle E(B - V) \rangle$ ; none the less, Murphy & Bernet (2016) argue that the most probable range is  $2 < \langle E(B - V) \rangle < 10$  mmag. This range of  $\langle E(B - V) \rangle$  corresponds to the full range of  $\kappa_z$  studied in this work. As observed from the results in Table 1, the fractional completeness depends strongly on  $\kappa_z$ . For the largest assumed value of  $\kappa_z = -21.1$ , we infer that  $\Omega_{\text{DLA}}$  as measured in SDSS-DR7 may be underestimated by 60 per cent, and  $\Omega_z$  by a factor of 4. This high value of  $\log \kappa_z$  leads to an intrinsic average reddening of 28 mmag which is consistent with the limit on the unbiased  $\langle E(B - V) \rangle$  as inferred from a purely radio-selected sample of  $\langle E(B - V) \rangle < 40$  mmag (Ellison et al. 2005). However, such a large value of  $\kappa_z$  is slightly at odds with the recent results obtained by Zafar & Møller (2019) of  $\log \kappa_z = -21.4 \pm 0.1$ , yet consistent at the  $3\sigma$  level. Instead, taking the Murphy & Bernet value of  $\langle E(B - V) \rangle = 3.0 \pm 1.5$  mmag at face value, we find a most likely value of  $\kappa_z$  of  $\log \kappa_z = -21.7 \pm 0.2$ . This is equally at odds with the constraint by Zafar & Møller (2019). Murphy & Bernet (2016), however, note that their measurement of  $\langle E(B - V) \rangle$  might be underestimated by up to 3 mmag (see section 6.1 and their fig. 9). A correction of this size would bring their result in agreement with the measurement by Vladilo et al. (2008) ( $\langle E(B - V) \rangle = 6 \pm 2$  mmag), and instead favour a dust-to-metals ratio between  $\log \kappa_z = -21.4$  and  $-21.3$ , fully consistent with the result by Zafar & Møller (2019).

Given the lack of strong evidence as to the true value of  $\langle E(B - V) \rangle$  for DLAs from SDSS-DR7, we instead use the constraint on  $\kappa_z$  from the analysis by Zafar & Møller (2019) of  $\log \kappa_z = -21.4 \pm 0.1$ . Based on this value, we infer a fractional completeness for  $\Omega_{\text{DLA}}$  and  $\Omega_z$  in DLAs of  $0.77 \pm 0.04$  and  $0.47 \pm 0.09$ , respectively, and we obtain an independent determination of the average reddening in optically selected DLA samples of  $\langle E(B - V) \rangle = 5_{-1}^{+2}$  mmag. We further note that these results do not significantly depend on the

intrinsic quasar model as the systematic uncertainty is much smaller than the statistical uncertainty (Section 4).

One last aspect to consider is the fraction of the population with supersolar metallicities, since the results must agree with constraints on the observed fraction of high-metallicity systems. For  $z_{\text{abs}} = 3.0$  DLAs, we find that the intrinsic fraction of supersolar DLAs is 1–2 per cent. This is consistent with the observed fraction of galaxies with supersolar metallicity as probed by gamma-ray bursts (GRBs) is  $4_{-2}^{+5}$  per cent (Krühler et al. 2015). However, as GRBs trace the galaxy population differently than DLAs (Fynbo et al. 2008) it is not straightforward to directly compare the fraction of supersolar GRB-hosts to that of DLAs. None the less, we argue that the inferred fraction of 1–2 per cent in an unbiased sample of DLAs is not in stark contrast with observations, since these highly enriched DLAs would be completely missed in the DR7 samples due to the strong dust bias.

### 5.1 Redshift evolution of $\Omega_z$

Due to the redshift dependent nature of the colour criteria used in SDSS-DR7, the magnitude of the dust bias depends on redshift. Based on our calculations of the fractional completeness at  $z_{\text{abs}} = 2.2$ , we find that  $\Omega_z$  at this redshift is underestimated by a factor of  $\sim 3$  assuming  $\log \kappa_z = -21.5$  as argued above. The corresponding  $\langle E(B - V) \rangle$  inferred for an optically selected sample is  $\langle E(B - V) \rangle = 8$  mmag. While there are no direct constraints on  $\langle E(B - V) \rangle$  in this redshift range, a value of 8 mmag at  $z_{\text{abs}} = 2.2$  is consistent with the results by Murphy & Bernet (2016). For this value of  $\kappa_z$ , we find a fractional completeness of  $\Omega_{\text{DLA}}$  and  $\Omega_z$  of  $0.68 \pm 0.03$  and  $0.35 \pm 0.03$ .

Taking the uncertainties on  $\langle E(B - V) \rangle$  into account, we infer a lower estimate for the fractional completeness on  $\Omega_{\text{DLA}}$  and  $\Omega_z$  of  $0.55 \pm 0.04$  and  $0.22 \pm 0.03$ , respectively. These values indicate a significant redshift evolution of both  $\Omega_{\text{DLA}}$  and  $\Omega_z$ .

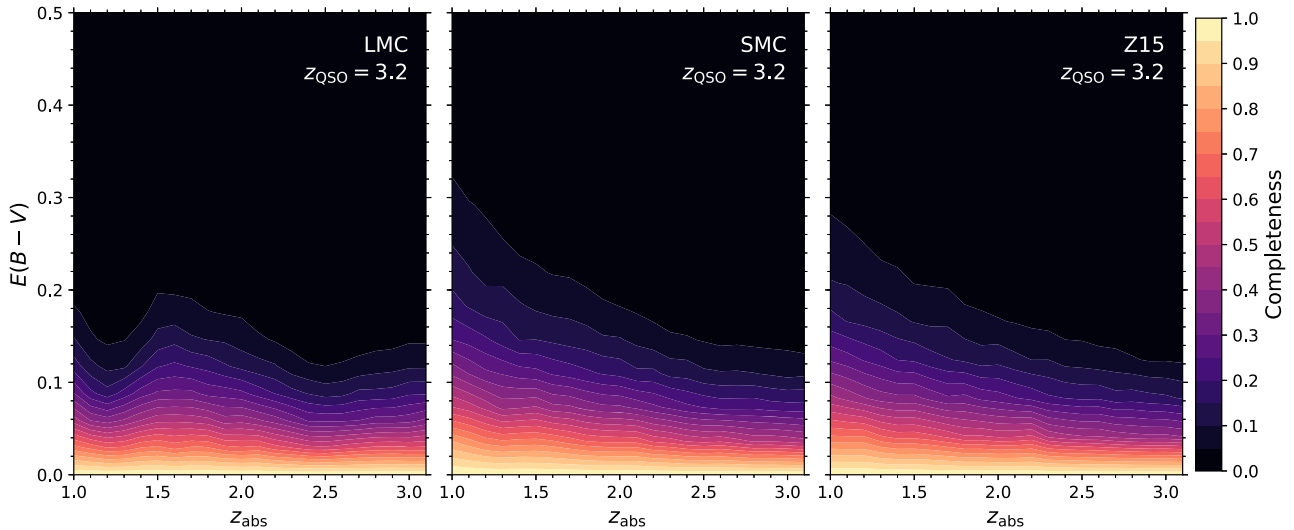
Such a redshift evolution of the fractional completeness of  $\Omega_z$  would significantly affect the measurements presented by Rafelski et al. (2014), and bring the redshift evolution of  $\Omega_z$  for DLAs in closer agreement with that inferred for star-forming galaxies (as traced by LBGs), albeit with a lower fraction of metals in DLAs. Still, more detailed calculations and larger samples of precise metallicity and reddening measurements are needed in order to firmly address all the redshift dependencies in the DR7 (and later) samples of absorbers taking into account both absorber and quasar redshifts.

Lastly, as larger samples of DLAs are being compiled on the basis of more complex and varying selection criteria, the effect of a dust-bias on the metallicity distribution should be studied in detail. We have restricted our analysis to the magnitude limit of  $i < 19.1$  in order to make a statistically meaningful comparison. However, the parent sample of DLAs include many more targets that are selected down to a much deeper magnitude limit ( $i < 20.2$  and even deeper for radio selected quasars and later SDSS surveys). One might therefore naïvely expect that the bias would be alleviated by the inclusion of such targets. However, since the completeness of quasars in SDSS-DR7 for  $19.1 < i < 20.2$  is only around 30 per cent, the direct impact on the bias is not easily quantified.

### 5.2 Different extinction curves

Recently, it has been shown that dust in various environments follows a steeper reddening law than the SMC reddening law (e.g. Fitzpatrick & Massa 2007; Zafar et al. 2012, 2015; Fynbo et al.





**Figure 7.** Completeness as a function of absorber redshift and optical reddening for three different reddening laws: LMC (left) and SMC (middle) from Gordon et al. (2003) and the average reddening law from Zafar et al. (2015) (right). The plots assume a quasar redshift of  $z_{\text{QSO}} = 3.2$  and  $\log(N_{\text{HI}} / \text{cm}^{-2}) = 20.3$ .

2014; Heintz et al. 2017; Noterdaeme et al. 2017). In such cases, the reddening of the background quasar would be even larger than that of the SMC for the same  $A_V$ . We have tested the impact of inserting a steeper extinction curve in our simulations. For this purpose, we use the reddening law inferred by Zafar et al. (2015, hereafter Z15) and in Fig. 7 we show a comparison of the absorber completeness as a function of absorber redshift and  $E(B - V)$  for three different extinction laws. We highlight that this figure, in contrast with similar figures in this work, shows the selection *completeness* not the selection probability; i.e. the selection probability has been corrected for the intrinsic quasar selection probability assuming no foreground dust.

Based on Fig. 7, it is seen that the smooth extinction curves (SMC and Z15) lead to a similar pattern of completeness as a function of  $z_{\text{abs}}$  and  $E(B - V)$ , yet with a slightly lower completeness for the steeper reddening law (Z15). In contrast, the characteristic 2175 Å bump for the LMC-type extinction curve leads to a significant drop in completeness of absorbers in the redshift range  $1 < z_{\text{abs}} < 1.5$  (see Fig. 7).

Since it is not yet possible to constrain the relative frequency of such steep reddening laws in DLA samples we cannot model the effects of a mixed population of extinction properties on the metallicity distribution. We are merely able to draw the qualitative conclusion that if a non-negligible part of the DLA population arises in galaxies with steeper reddening laws than the SMC, the bias on  $\Omega_{\text{DLA}}$  and  $\Omega_Z$  will be larger than we estimate in this work. DLAs with such steep reddening laws would preferentially be missed in the SDSS sample (due to the larger bias against these) and hence would not significantly alter the average reddening measurements that generally show good agreement with the SMC reddening law (e.g. Vladilo et al. 2008; Murphy & Bernet 2016).

Similarly, there are not many quantitative estimates of the frequency of the 2175 Å bump in DLA extinction curves. The largest sample is presented by Zhao et al. (2016), who identify around 400 absorbers (with  $0.7 < z_{\text{abs}} < 2.7$ ) with LMC-type extinction curves out of roughly 40 000 absorbers. However, as these are all preselected based on the Mg II line, it is not clear what fraction of them would be DLAs. A proper comparison is further hampered

by the fact that their full sample and the detection method has not yet been published. At face value though, their study implies that the fraction of DLAs exhibiting LMC-type extinction is less than 1 per cent. This is consistent with the C I-selected sample by Ledoux et al. (2015) who find that roughly 1 per cent of DLAs are selected as C I absorbers and roughly one third of them have LMC-type extinction. Given the very low frequency of LMC-type extinction in DLAs, we have focused on the SMC extinction law in this work.

### 5.3 Correlation between $N_{\text{HI}}$ and $Z$ ?

In Section 3, we explicitly assume that the distribution of metallicity and HI are independent, i.e. there are no intrinsic correlation between the two quantities. However, if large  $N_{\text{HI}}$  systems systematically arise in higher mass haloes one would expect a positive correlation between  $N_{\text{HI}}$  and  $Z$ . If such a correlation indeed exists, then the dust bias would be enhanced, since there would be relatively more high- $Z$  and high- $N_{\text{HI}}$  absorbers for which the bias is stronger.

Alternatively, there might be an anticorrelation between  $Z$  and  $N_{\text{HI}}$  at the highest column densities, since the high metallicity leads to very efficient cooling which in turns leads to efficient molecule formation thereby depleting the HI into  $\text{H}_2$ . This effect would slightly alleviate the bias, since the part of the parameter space with high metal column (high  $Z$  and high  $N_{\text{HI}}$ ) would be intrinsically less populated and hence less affected by a dust bias.

None the less, no significant correlation between  $N_{\text{HI}}$  and  $Z$  is observed in simulations where  $\text{H}_2$  formation is included at the highest  $N_{\text{HI}}$  (e.g. Rahmati & Schaye 2014) nor is such a correlation reported in the observed data. The combination of both effects discussed above could explain the lack of any strong correlation between  $N_{\text{HI}}$  and  $Z$  in simulations (Pontzen et al. 2008; Rahmati & Schaye 2014). We therefore do not include any intrinsic correlations in our dust bias analysis, and hence assume that the two distributions are fully separable.

## 5.4 DLA identification bias

Throughout this analysis, we have made the assumption that all DLAs would be identified correctly if the background quasars were indeed selected as quasars based on their photometric properties alone. However, as DLAs are identified and searched for in the resulting spectra after quasar target selection, the spectral quality also impacts the bias. Specifically, the noise level in the Ly $\alpha$ -forest continuum is important since DLAs are found in this wavelength range. The largest catalogue of DLAs from SDSS-II DR7 is based on an algorithm which only considers spectra with a continuum-to-noise (CNR) ratio larger than 4 in the appropriate wavelength range (Noterdaeme et al. 2009b). As reddened quasars with dusty foreground absorbers will have systematically lower CNR in the blue part of the spectrum where the DLAs are found, it is indeed possible that even though the quasar was correctly identified and observed, the DLA would not be identified in the algorithm due to the CNR being too low. This will therefore lead to an increase in the bias against dusty and metal-rich DLAs. Quantifying the additional effect would require a complete model of the spectral data which is beyond the scope of this paper.

## 6 SUMMARY

In this work, we have analysed the impact of a bias against dusty absorption systems in SDSS-DR7 on the measured cosmic mass density of neutral hydrogen and metals as inferred by DLAs. For this purpose, we develop a model to describe the intrinsic photometric properties of quasars as a function of redshift (Section 2.2). This model is calibrated to match the observed colour distributions in SDSS-DR7. Next, we implement the colour selection algorithm used by SDSS up until DR7 (Richards et al. 2002) to calculate the selection probability as a function of quasar redshift, details are provided in Appendix C.

Based on the intrinsic quasar model, we furthermore calculate the selection probability of quasars for various foreground absorber properties (redshift, visual extinction and  $N_{\text{H I}}$ ). The selection probability of quasars and absorbers is highly redshift dependent and the final probability depends on both the absorber and the quasar redshift. In the remaining of the analysis, we focus our attention on the  $z \approx 3$  absorbers analysed by Pontzen & Pettini (2009) and Murphy & Bernet (2016), though we also explore a lower redshift model. We then calculate the fractional completeness of the following absorption inferred quantities: the DLA incidence rate,  $I_{\text{DLA}}$ , the cosmic mass density of neutral hydrogen in DLAs,  $\Omega_{\text{DLA}}$ , the average metallicity,  $\langle Z \rangle$ , the hydrogen weighted average metallicity,  $\langle Z \rangle_{\text{NHI}}$ , and the cosmic mass density of metals,  $\Omega_{\text{Z}}$ .

By assuming a constant dust-to-metals ratio of  $\log \kappa_{\text{Z}} = -21.5$  and a fixed flux-limited selection criterion, we are able to replicate the results by Pontzen & Pettini (2009). However, when we include the full set of colour criteria of the SDSS-DR7 selection algorithm including radio selection, we find that the fractional completeness is lower compared to a single flux-limited sample (see Tables 1 and 3).

The extent of the incompleteness varies with the assumed  $\kappa_{\text{Z}}$ . For a probable range in  $\kappa_{\text{Z}}$  (reproducing the observed  $\langle E(B - V) \rangle$ ), we find that  $\Omega_{\text{DLA}}$  and  $\Omega_{\text{Z}}$  are underestimated by, respectively, 10–50 per cent and 30–200 per cent.

At lower redshifts ( $z_{\text{abs}} = 2.2$ ) assuming the same  $\kappa_{\text{Z}}$ , we find a fractional completeness of  $\Omega_{\text{DLA}}$  and  $\Omega_{\text{Z}}$  of  $0.68 \pm 0.03$  and  $0.35 \pm 0.03$ , respectively. However, taking all uncertainties into account, the fractional completeness of  $\Omega_{\text{DLA}}$  and  $\Omega_{\text{Z}}$  may be as

low as 0.55 and 0.22, respectively. We advise that these results at  $z = 2.2$  should be interpreted with caution as the metallicity distribution and the dust measurements in DLAs at this redshift are not well constrained; hence, the resulting bias calculation is only preliminary.

We conclude that while the main driver of the dust bias against metal-rich absorption systems at  $z \approx 3$  is the flux-limit used in SDSS up until DR7, the colour selection adds additional incompleteness and makes the incompleteness a strong function of redshift. Such selection effects must be kept in mind (and preferably modelled) when interpreting absorber samples in any large quasar survey. Alternatively, the absorber statistics should be inferred using a purely unbiased selection such as larger radio-selected samples or *Gaia*-assisted selection (Heintz et al. 2018a, see also Geier et al. 2019, A&A in production).

## ACKNOWLEDGEMENTS

The authors thank the anonymous referee for the thorough and constructive comments. JK acknowledges financial support from the Danish Council for Independent Research (EU-FP7 under the Marie-Curie grant agreement no. 600207) with reference DFF-MOBILEX-5051-00115. The research leading to these results has received funding from the French *Agence Nationale de la Recherche* under grant no ANR-17-CE31-0011-01 (project ‘HIH2’ – PI Noterdaeme). The Cosmic Dawn Center is funded by the DNRF. KEH acknowledges support by a Project Grant (162948–051) from the Icelandic Research Fund. The Pan-STARRS1 Surveys (PS1) and the PS1 public science archive have been made possible through contributions by the Institute for Astronomy, the University of Hawaii, the Pan-STARRS Project Office, the Max-Planck Society and its participating institutes, the Max Planck Institute for Astronomy, Heidelberg and the Max Planck Institute for Extraterrestrial Physics, Garching, The Johns Hopkins University, Durham University, the University of Edinburgh, the Queen’s University Belfast, the Harvard-Smithsonian Center for Astrophysics, the Las Cumbres Observatory Global Telescope Network Incorporated, the National Central University of Taiwan, the Space Telescope Science Institute, the National Aeronautics and Space Administration under Grant No. NNX08AR22G issued through the Planetary Science Division of the NASA Science Mission Directorate, the National Science Foundation Grant No. AST-1238877, the University of Maryland, Eotvos Lorand University (ELTE), the Los Alamos National Laboratory, and the Gordon and Betty Moore Foundation.

## REFERENCES

- Baldwin J. A., 1977, *ApJ*, 214, 679  
 Becker R. H., White R. L., Helfand D. J., 1995, *ApJ*, 450, 559  
 Bird S., Haehnelt M., Neeleman M., Genel S., Vogelsberger M., Hernquist L., 2015, *MNRAS*, 447, 1834  
 Bovy J. et al., 2011, *ApJ*, 729, 141  
 Chambers K. C. et al., 2016, preprint (arXiv:1612.05560)  
 De Cia A., Ledoux C., Mattsson L., Petitjean P., Srianand R., Gavignaud I., Jenkins E. B., 2016, *A&A*, 596, A97  
 Ellison S. L., Hall P. B., Lira P., 2005, *AJ*, 130, 1345  
 Ellison S. L., Yan L., Hook I. M., Pettini M., Wall J. V., Shaver P., 2001, *A&A*, 379, 393  
 Fitzpatrick E. L., Massa D., 2007, *ApJ*, 663, 320  
 Foreman-Mackey D., Hogg D. W., Lang D., Goodman J., 2013, *PASP*, 125, 306  
 Frank S., Péroux C., 2010, *MNRAS*, 406, 2235



- Fynbo J. P. U., Krogager J.-K., Venemans B., Noterdaeme P., Vestergaard M., Møller P., Ledoux C., Geier S., 2013, *ApJS*, 204, 6
- Fynbo J. P. U., Prochaska J. X., Sommer-Larsen J., Dessauges-Zavadsky M., Møller P., 2008, *ApJ*, 683, 321
- Fynbo J. P. U. et al., 2014, *A&A*, 572, A12
- Fynbo J. P. U. et al., 2017, *A&A*, 606, A13
- Geier S. et al., 2019, *A&A*, preprint ([arXiv:1904.01686](https://arxiv.org/abs/1904.01686))
- Gordon K. D., Clayton G. C., Misselt K. A., Landolt A. U., Wolff M. J., 2003, *ApJ*, 594, 279
- Heintz K. E., Fynbo J. P. U., Høg E., Møller P., Krogager J. K., Geier S., Jakobsson P., Christensen L., 2018a, *A&A*, 615, L8
- Heintz K. E. et al., 2017, *A&A*, 601, A83
- Heintz K. E. et al., 2018b, *A&A*, 615, A43
- Hewett P. C., Foltz C. B., Chaffee F. H., 2001, *AJ*, 122, 518
- Jorgenson R. A., Murphy M. T., Thompson R., 2013, *MNRAS*, 435, 482
- Jorgenson R. A., Wolfe A. M., Prochaska J. X., Lu L., Howk J. C., Cooke J., Gawiser E., Gelino D. M., 2006, *ApJ*, 646, 730
- Kaplan K. F., Prochaska J. X., Herbert-Fort S., Ellison S. L., Dessauges-Zavadsky M., 2010, *PASP*, 122, 619
- Khare P., vanden Berk D., York D. G., Lundgren B., Kulkarni V. P., 2012, *MNRAS*, 419, 1028
- Kirkpatrick J. A., Schlegel D. J., Ross N. P., Myers A. D., Hennawi J. F., Sheldon E. S., Schneider D. P., Weaver B. A., 2011, *ApJ*, 743, 125
- Krawczyk C. M., Richards G. T., Gallagher S. C., Leighly K. M., Hewett P. C., Ross N. P., Hall P. B., 2015, *AJ*, 149, 203
- Krogager J.-K., Fynbo J. P. U., Noterdaeme P., Zafar T., Møller P., Ledoux C., Krühler T., Stockton A., 2016a, *MNRAS*, 455, 2698
- Krogager J.-K. et al., 2015, *ApJS*, 217, 5
- Krogager J.-K. et al., 2016b, *ApJ*, 832, 49
- Krühler T. et al., 2015, *A&A*, 581, A125
- Lanzetta K. M., Wolfe A. M., Turnshek D. A., 1995, *ApJ*, 440, 435
- Ledoux C., Noterdaeme P., Petitjean P., Srianand R., 2015, *A&A*, 580, A8
- Ledoux C., Petitjean P., Srianand R., 2003, *MNRAS*, 346, 209
- Li H.-Z., Chen L.-E., Jiang Y.-G., Yi T.-F., 2015, *Res. Astron. Astrophys.*, 15, 929
- Lusso E., Worseck G., Hennawi J. F., Prochaska J. X., Vignali C., Stern J., O'Meara J. M., 2015, *MNRAS*, 449, 4204
- Manti S., Gallerani S., Ferrara A., Greig B., Feruglio C., 2017, *MNRAS*, 466, 1160
- Meiksin A., 2006, *MNRAS*, 365, 807
- Meusinger H., Schalldach P., Scholz R.-D., in der Au A., Newholm M., de Hoon A., Kaminsky B., 2012, *A&A*, 541, A77
- Murphy M. T., Bernet M. L., 2016, *MNRAS*, 455, 1043
- Murphy M. T., Liske J., 2004, *MNRAS*, 354, L31
- Noterdaeme P., Ledoux C., Srianand R., Petitjean P., Lopez S., 2009a, *A&A*, 503, 765
- Noterdaeme P., Ledoux C., Zou S., Petitjean P., Srianand R., Balashev S., López S., 2018, *A&A*, 612, A58
- Noterdaeme P., López S., Dumont V., Ledoux C., Molaro P., Petitjean P., 2012, *A&A*, 542, L33
- Noterdaeme P., Petitjean P., Ledoux C., Srianand R., 2009b, *A&A*, 505, 1087
- Noterdaeme P. et al., 2017, *A&A*, 597, A82
- O'Hagan A., Leonard T., 1976, *Biometrika*, 63, 201
- Pei Y. C., Fall S. M., 1995, *ApJ*, 454, 69
- Pei Y. C., Fall S. M., Bechtold J., 1991, *ApJ*, 378, 6
- Planck Collaboration et al., 2014, *A&A*, 571, A16
- Pontzen A., Pettini M., 2009, *MNRAS*, 393, 557
- Pontzen A. et al., 2008, *MNRAS*, 390, 1349
- Pope S. B., 2008, Technical report, Algorithms for Ellipsoids. Cornell University Report FDA 08-01
- Prochaska J. X., Herbert-Fort S., Wolfe A. M., 2005, *ApJ*, 635, 123
- Quiret S. et al., 2016, *MNRAS*, 458, 4074
- Rafelski M., Neeleman M., Fumagalli M., Wolfe A. M., Prochaska J. X., 2014, *ApJ*, 782, L29
- Rafelski M., Wolfe A. M., Prochaska J. X., Neeleman M., Mendez A. J., 2012, *ApJ*, 755, 89
- Rahmati A., Schaye J., 2014, *MNRAS*, 438, 529
- Richards G. T. et al., 2002, *AJ*, 123, 2945
- Richards G. T. et al., 2009, *ApJS*, 180, 67
- Ross N. P. et al., 2012, *ApJS*, 199, 3
- Schneider D., 2010, *AJ*, 139, 2360
- Selsing J., Fynbo J. P. U., Christensen L., Krogager J.-K., 2016, *A&A*, 585, A87
- Srianand R., Noterdaeme P., Ledoux C., Petitjean P., 2008, *A&A*, 482, L39
- Srianand R., Petitjean P., Ledoux C., Ferland G., Shaw G., 2005, *MNRAS*, 362, 549
- Vladilo G., Prochaska J. X., Wolfe A. M., 2008, *A&A*, 478, 701
- Vladilo G., Péroux C., 2005, *A&A*, 444, 461
- Wolfe A. M., Turnshek D. A., Smith H. E., Cohen R. D., 1986, *ApJS*, 61, 249
- York D. G. et al., 2000, *AJ*, 120, 1579
- Yèche C. et al., 2010, *A&A*, 523, A14
- Zafar T., Møller P., 2019, *MNRAS*, 482, 2731
- Zafar T., Watson D., 2013, *A&A*, 560, A26
- Zafar T. et al., 2012, *ApJ*, 753, 82
- Zafar T. et al., 2015, *A&A*, 584, A100 (Z15)
- Zhao Y. et al., 2016, American Astronomical Society Meeting Abstracts #227. p. 339.07

## APPENDIX A: OPTIMIZATION OF INTRINSIC QUASAR MODEL

We optimize the quasar model to fit more accurately the intrinsic properties of quasars in the redshift ranges used for the bias analysis, i.e.  $z_{\text{QSO}} = 3.2$  and  $z_{\text{QSO}} = 2.5$ . For the high- $z$  solution we do not need to take selection effects into account, since the completeness is rather high (95 per cent). However, for the low- $z$  model, the model strongly depends on the colour-selection algorithm, which increases computation time beyond what is feasible for the scope of this paper. We therefore have to manually tweak the low-redshift model in order to match the observed properties. The details for the two models are presented below.

In order to match the correlated widths of the colour distributions, we introduce an additional parameter:  $w_\alpha^r$ , the weight of the  $r$ -band dispersion due to intrinsic variations in the quasar spectral shape. The remaining four bands are all scaled using the parameter  $w_\alpha$  as before. This can be expressed in vector notation as the five-element vector:  $\vec{w}_\alpha = w_\alpha \times \{1, 1, w_\alpha^r/w_\alpha, 1, 1\}$ . Hence, the parameter  $w_\alpha^r$  is strongly correlated with  $w_\alpha$  by construction. The randomly assigned intrinsic variation (the second term of equation 1) is then calculated as

$$\vec{m}_{\alpha,i} = X_i(\vec{w}_\alpha \circ \vec{\Delta}_\alpha),$$

where ‘ $\circ$ ’ denotes the Hadamard product, i.e. component-wise multiplication.

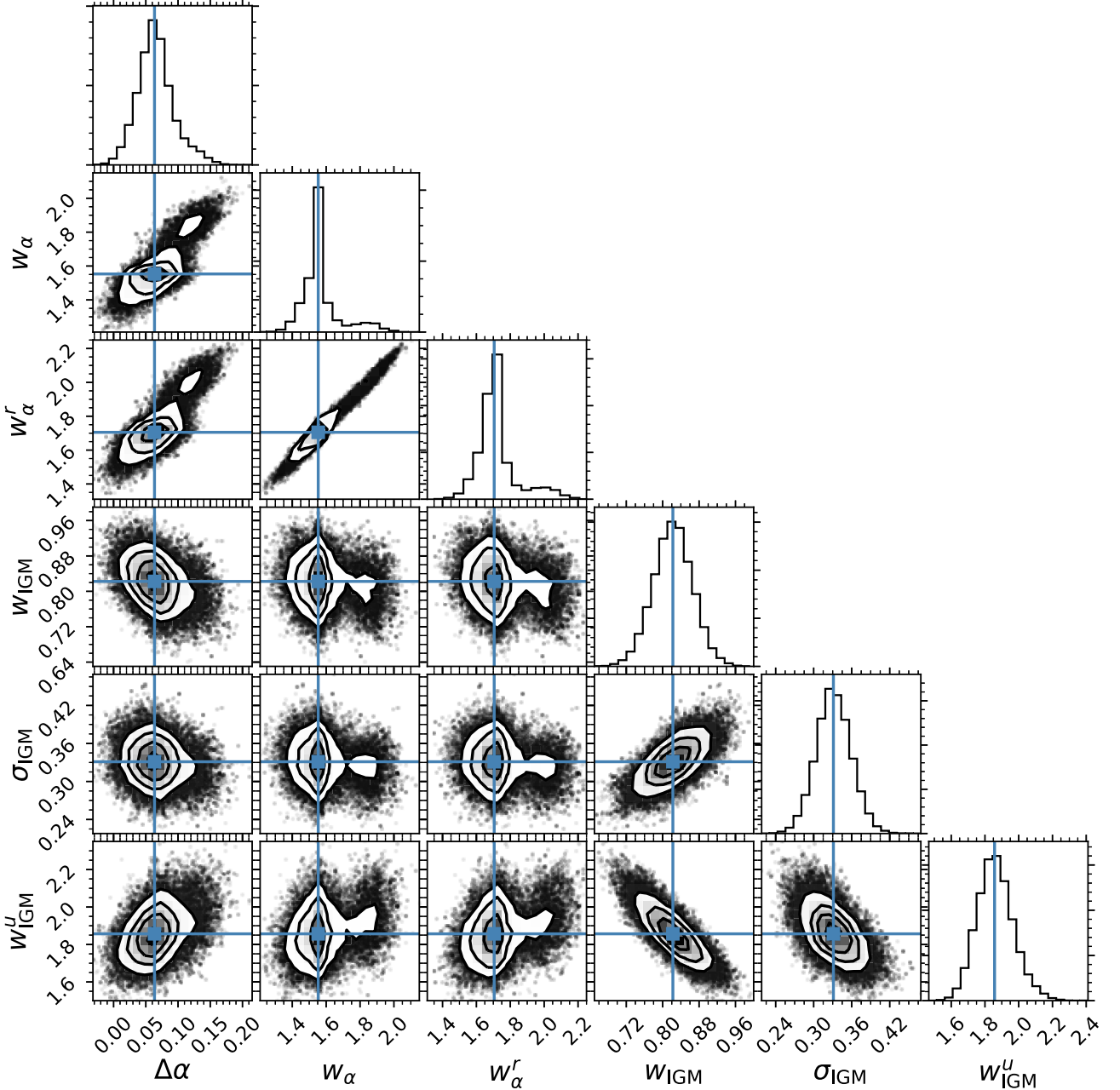


Figure A1. Posterior probability distributions and correlations for best-fitting intrinsic quasar model for  $z_{\text{QSO}} = 3.2$ .

We fit the quasar model to the four colour distributions by minimizing the squared residual of the median, the standard deviation, the skewness (third moment), and the kurtosis (fourth moment) for each of the distributions. The four quantities are weighted by the uncertainty on each of them that are determined by bootstrapping. For the high-redshift model, we optimize the six model parameters using a simple max-likelihood estimator. The fit is then refined using a Monte Carlo approach in order to properly study parameter correlations. For this purpose, we use the PYTHON package *emcee* (Foreman-Mackey et al. 2013) with 100 ‘walkers’ initiated around the max-likelihood solution. We run the chain for 1000 iterations discarding the first 100 as ‘burn-in’ and using ‘flat’ priors for all parameters.

The best-fitting parameters are given in Table A1, and the posterior probability distributions and parameter correlations are shown in Fig. A1. We furthermore a comparison between the observed SDSS photometry and the model prediction in Fig. A2. As indicated by the  $p$ -values of the two-sample Kolmogorov–Smirnov (KS) test, all the modelled colour distributions are fully consistent with the data.

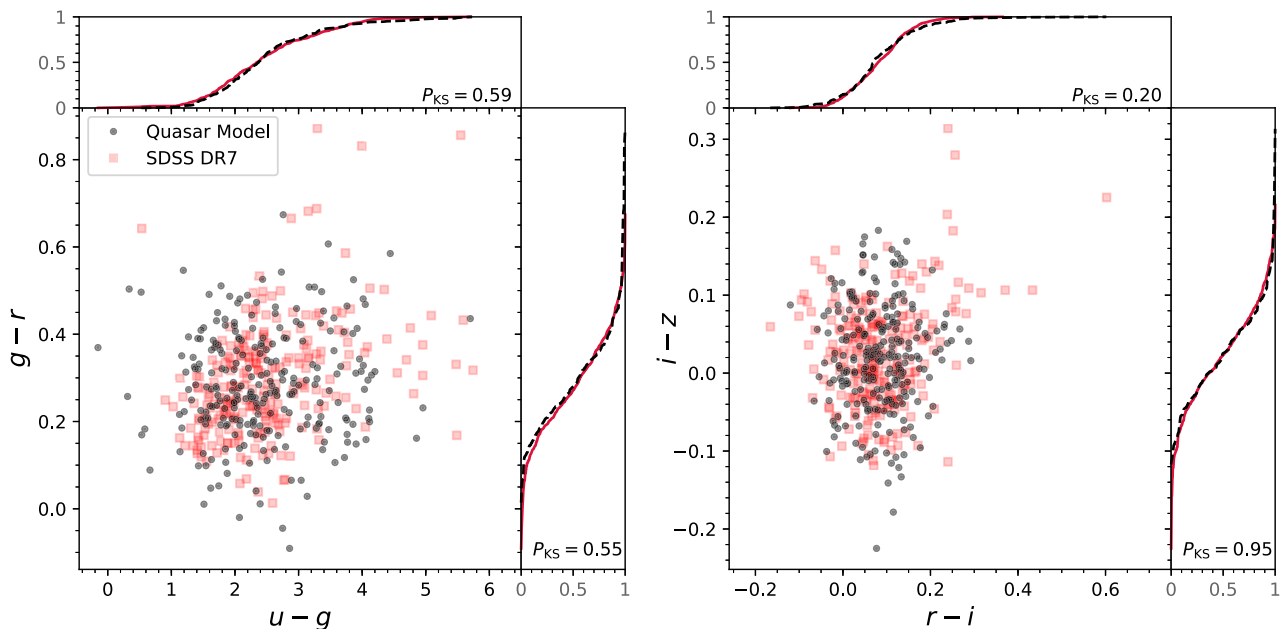
For the low-redshift model, as explained above, the colour selection algorithm strongly affects the quasar colour distributions and it is therefore not possible to do a robust Monte Carlo analysis. Moreover, we find that the six model parameters are not able to provide a good fit based on KS tests of the four distributions. Instead, we manually match the distributions by fixing the best-fitting parameters and applying ad hoc shifts to the model photometry in order to match the distributions. The applied shift per band is 0.02, 0.0, 0.088, 0.088, 0.092 mag for the  $u$ ,  $g$ ,  $r$ ,  $i$ , and  $z$  bands, respectively. A comparison between the observed SDSS colours and the model predictions are shown in Fig. A3. The most problematic distribution to match is for the  $u - g$  colour, where we are not able to fully match the shape of the distribution. However, the median is matched (by design) and the width of the distribution fits well, only the asymmetries in the distribution are not well reproduced. Since our main analysis is focused on the high-redshift model, we make no further attempts of improving the quasar model for this particular redshift. Moreover, due to uncertainties in the  $griz$  selection (Section C2), which is very important for this redshift range, the analysis at this redshift is inherently uncertain.

### A1 Quasar model systematics

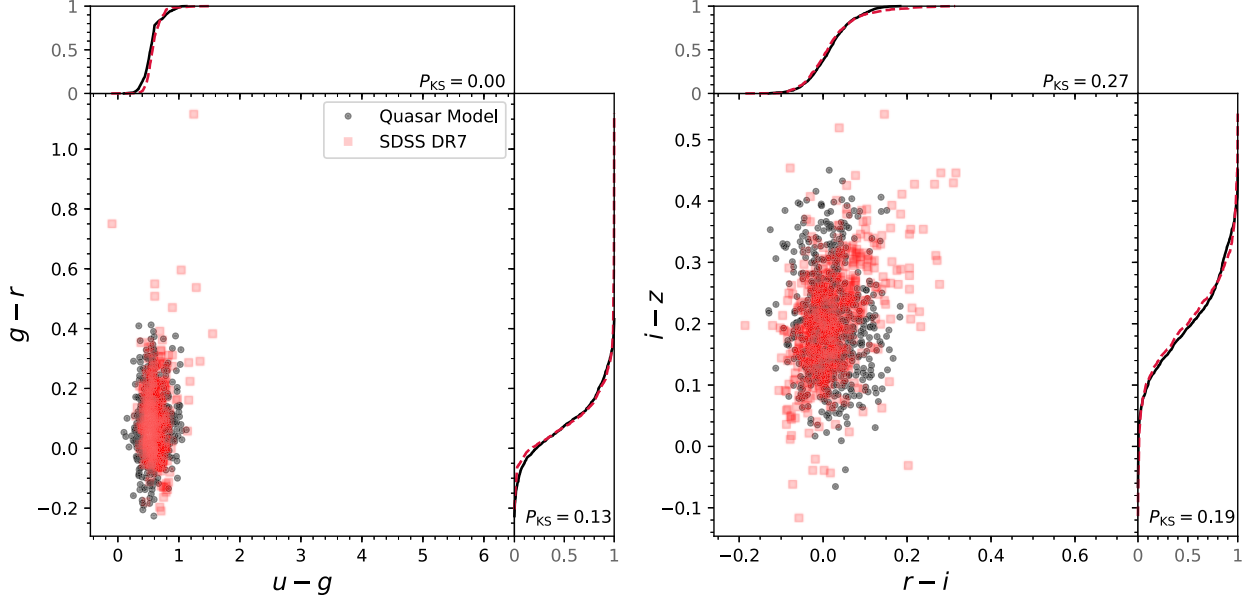
In order to assess the systematic effect of varying the quasar model within the parameter uncertainties, we perform the calculation of the fractional completeness as outlined in Section 3 for quasar models using the parameters offset by  $3\sigma$  towards bluer (l99) and redder (u99)

**Table A1.** Best-fit parameters for optimized quasar model.

Parameter	$z_{\text{QSO}} = 2.5$	$z_{\text{QSO}} = 3.2$
$\Delta\alpha$	$0.06 \pm 0.02$	$0.06 \pm 0.02$
$w_\alpha$	$1.8 \pm 0.1$	$1.55 \pm 0.08$
$w_\alpha^r$	$1.7 \pm 0.1$	$1.71 \pm 0.08$
$w_{\text{IGM}}$	$0.2 \pm 0.1$	$0.82 \pm 0.04$
$\sigma_{\text{IGM}}$	$0.13 \pm 0.03$	$0.33 \pm 0.03$
$w_{\text{IGM}}^u$	$1.9 \pm 0.1$	$1.9 \pm 0.1$



**Figure A2.** Colour–colour distributions for best-fitting intrinsic quasar model for  $z_{\text{QSO}} = 3.2$ . The top and right panels of each figure show the cumulative distribution of the given colour and the  $p$ -value derived from a two-sample KS test.



**Figure A3.** Colour–colour distributions for best-fitting intrinsic quasar model for  $z_{\text{QSO}} = 2.5$ . Same as Fig. A2.

**Table A2.** Systematic variation in fractional completeness of  $z \approx 3$  DLA statistics and average reddening.

Model	199	u99
$C [ I_{\text{DLA}} ]$	$0.89 \pm 0.01$	$0.86 \pm 0.01$
$C [ \Omega_{\text{DLA}} ]$	$0.83 \pm 0.02$	$0.80 \pm 0.02$
$C [ \langle Z \rangle ]$	$0.73 \pm 0.04$	$0.74 \pm 0.04$
$C [ \langle Z \rangle_{\text{NH}_1} ]$	$0.66 \pm 0.04$	$0.67 \pm 0.05$
$C [ \Omega_Z ]$	$0.55 \pm 0.05$	$0.53 \pm 0.05$
$\langle E(B - V) \rangle_{\text{obs}} / \text{mmag}$	$4.3^{+0.4}_{-0.4}$	$4.4^{+0.4}_{-0.4}$
$\langle E(B - V) \rangle_{\text{int}} / \text{mmag}$	$7.0^{+1.1}_{-1.2}$	$7.1^{+1.2}_{-1.1}$

*Note.* Both models are calculated for  $\log \kappa_Z = -21.5$ .

**Table A3.** Systematic variation in fractional completeness of  $z \approx 2$  DLA statistics and average reddening.

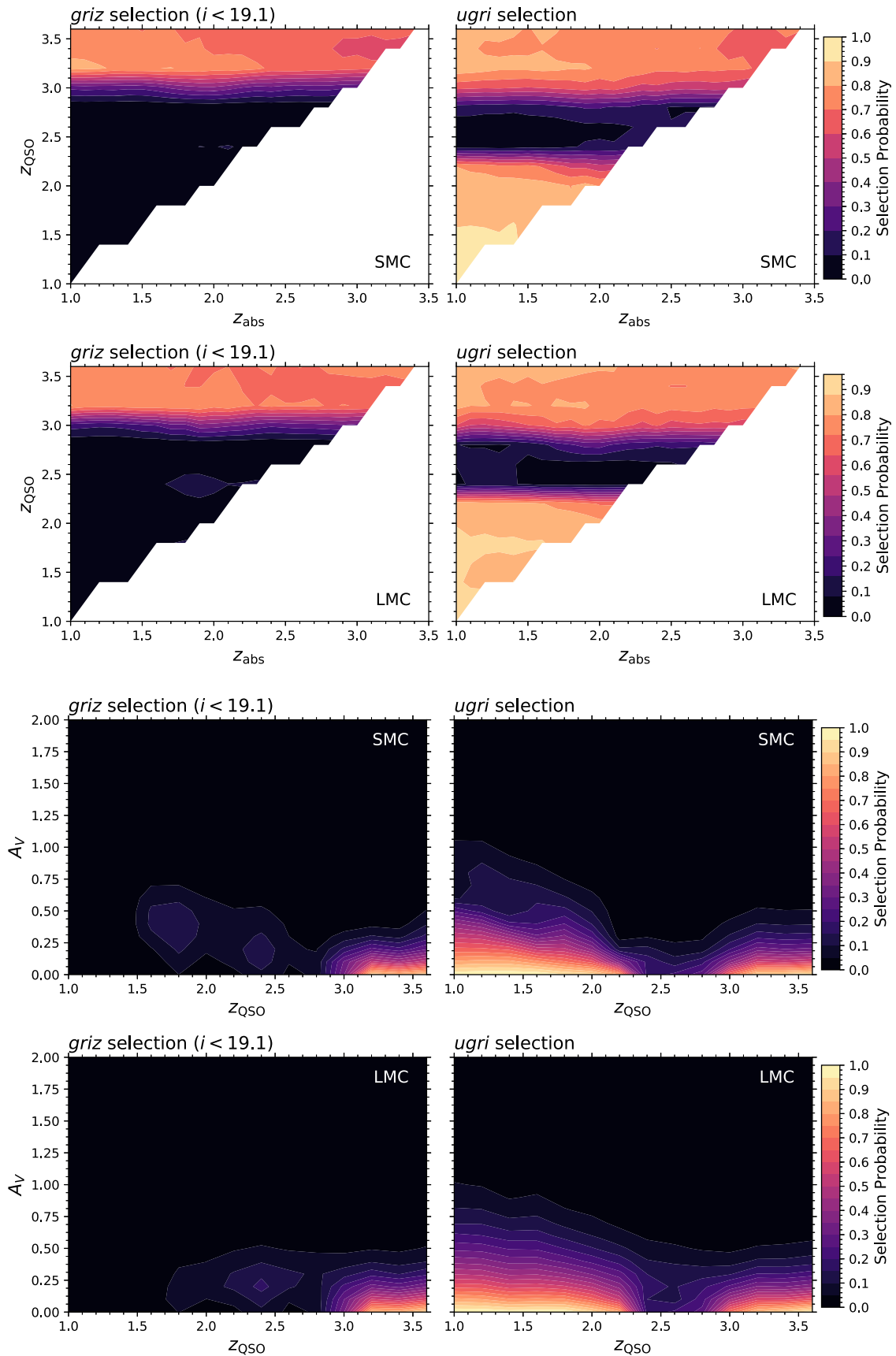
Model	199	u99
$C [ I_{\text{DLA}} ]$	$0.72 \pm 0.04$	$0.86 \pm 0.03$
$C [ \Omega_{\text{DLA}} ]$	$0.64 \pm 0.04$	$0.76 \pm 0.03$
$C [ \langle Z \rangle ]$	$0.53 \pm 0.03$	$0.58 \pm 0.03$
$C [ \langle Z \rangle_{\text{NH}_1} ]$	$0.47 \pm 0.03$	$0.52 \pm 0.03$
$C [ \Omega_Z ]$	$0.30 \pm 0.03$	$0.39 \pm 0.03$
$\langle E(B - V) \rangle_{\text{obs}} / \text{mmag}$	$8.1^{+1.5}_{-1.3}$	$8.1^{+1.1}_{-1.3}$
$\langle E(B - V) \rangle_{\text{int}} / \text{mmag}$	$19.5^{+3.6}_{-3.9}$	$17.6^{+2.6}_{-3.4}$

*Note.* Both models are calculated for  $\log \kappa_Z = -21.5$ .

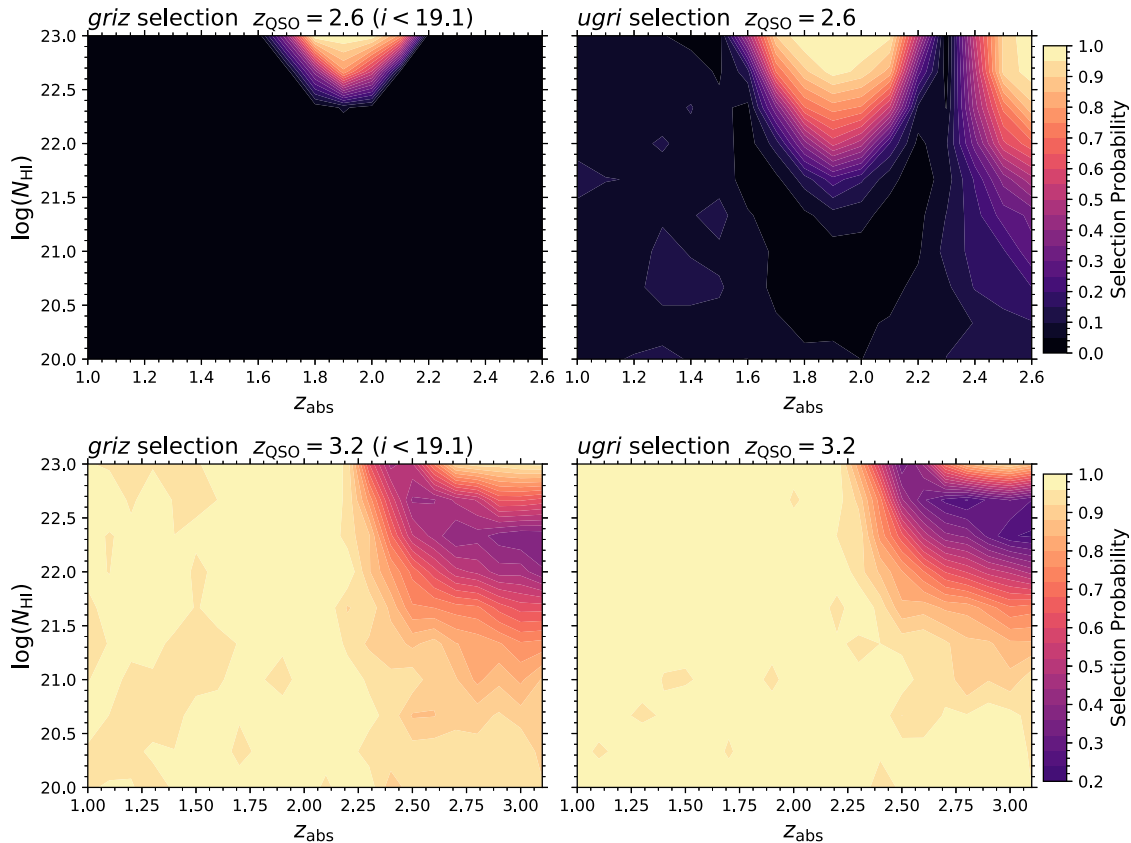
spectral slopes. We take the parameter correlations into account when offsetting the model parameters. The resulting fractional completeness for the two redshift configurations are shown in Tables A2 and A3 for a fixed value of  $\log \kappa_Z = -21.5$ .

## APPENDIX B: SUPPORTING FIGURES

In Fig. B1, we show the calculated selection probability as a function of  $z_{\text{abs}}$  and  $z_{\text{QSO}}$  for  $A_V = 0.2$  mag in the top two rows. In the bottom two rows of Fig. B1, we show the selection probability as a function of  $z_{\text{QSO}}$  and  $A_V$  averaged over all absorption redshifts and  $\text{NH}_1$ . In Fig. B2, we show the quasar selection probability as a function of  $\text{NH}_1$  and  $z_{\text{abs}}$  for fixed visual extinction ( $A_V = 0$ ) and two different quasar redshifts.



**Figure B1.** Selection probability as function of  $z_{\text{abs}}$  versus  $z_{\text{QSO}}$  with  $A_V = 0.2$  mag (top two rows) and  $z_{\text{QSO}}$  versus  $A_V$  (bottom two rows). The left-hand and right-hand panels show the probabilities for the criteria of *griz* (only for  $i < 19.1$ ) and *ugri*, respectively. The extinction curve used for the given model is indicated in the top right corner of each panel.



**Figure B2.** Selection probability as function of  $z_{\text{abs}}$  versus  $N_{\text{HI}}$  for  $A_V = 0.0$  assuming  $z_{\text{QSO}} = 2.6$  (top row) and  $z_{\text{QSO}} = 3.2$  (bottom row) for the SMC extinction law. The left-hand and right-hand panels show the probabilities for the criteria of *griz* (only for  $i < 19.1$ ) and *ugri*, respectively.

## APPENDIX C: IMPLEMENTATION OF SDSS COLOUR SELECTION ALGORITHM

### C1 Implementing the selection algorithm in PYTHON

We follow the outline of the selection algorithm as presented by Richards et al. (2002) using their tabulated definition of the stellar locus in the two 3D colour-spaces (*ugri* for the three colours:  $u - g$ ,  $g - r$ , and  $r - i$ ; and *griz* for the three colours:  $g - r$ ,  $r - i$ , and  $i - z$ ). The coordinates in 3D colour-space is denoted  $(x, y, z)$ . The stellar locus is defined as a sequence of ellipsoids and elliptical cylinders centred on so-called locus points in the 3D colour-space. For each locus point, there is an associated major ( $a_l$ ) and minor ( $a_m$ ) axis of the ellipse forming the base of the cylinder, a cylinder length  $k$ , an angle  $\theta$  defining the orientation of the major axis in the ellipse-plane<sup>3</sup> denoted by coordinates  $(i, j)$ , and a normal vector  $\hat{k}$  perpendicular to the ellipse-plane indicating the direction of the ellipse along the stellar locus in the 3D colour-space. Any target with observed magnitudes in the five-band photometric system describes a point in each of the two 3D colour-spaces denoted by a vector  $\mathbf{x}$ . The first step in order to assess whether an observed point lies within the locus is to identify the nearest locus point to the data point  $\mathbf{x}$ . This is calculated by the Euclidian distance in colour-space.

Since a given data point in colour-space has an associated uncertainty, any data point is represented by an error ellipsoid determined by its centre ( $\mathbf{x}$ ) and covariance matrix. We follow the definition of the covariance matrix given by Richards et al. (2002) who assume no correlation between individual *filters*. This translates to a covariance between two colours with the same filter occurring in both colours, e.g.  $\text{Cov}(g - r, r - i) = -V(r)$ . An example of the covariance matrix for the *ugri* colour-space is as follows:

$$\mathbf{S} = \begin{bmatrix} \sigma_u^2 - \sigma_g^2 & -\sigma_g^2 & 0 \\ -\sigma_g^2 & \sigma_g^2 - \sigma_r^2 & -\sigma_r^2 \\ 0 & -\sigma_r^2 & \sigma_r^2 - \sigma_i^2 \end{bmatrix}, \quad (\text{C1})$$

where  $\sigma$  denotes the  $1\sigma$  uncertainty in a given filter.

In order to take the measurement uncertainty into account when testing whether a point falls inside the locus, the base ellipse of the nearest locus-point cylinder is convolved with the error ellipse that arises from the projection of the error ellipsoid on to the  $(i, j)$  ellipse-plane. The

<sup>3</sup>The angle is given with respect to the basis vector  $\hat{i} = (\hat{k} \times \hat{z}) \times \hat{k}$  in the ellipse-plane.



data points are required to fall  $4\sigma$  away from the stellar locus. The covariance matrix is therefore multiplied by  $N_\sigma^2 = 16$ . This projection and convolution is implemented in terms of linear algebra and this is the only place where we do not directly follow Richards et al. (2002). However, we have tested that the two methods yield the exact same results. The projection is based on linear algebra equations from Pope (2008) and is performed as follows:

(i) Define a projection matrix,  $\mathbf{P}$ , which has the dimension  $(2 \times 3)$ , containing the two basis vectors of the locus ellipse-plane:  $\hat{\mathbf{j}} = \hat{\mathbf{k}} \times \hat{\mathbf{z}}$  and  $\hat{\mathbf{i}} = \hat{\mathbf{j}} \times \hat{\mathbf{k}}$ .

(ii) The inverse covariance matrix of the error ellipse in the  $(i, j)$  ellipse-plane is thus  $\mathbf{A}^{-1} = \mathbf{G}^{-1} \mathbf{P} \mathbf{B} \mathbf{P}^T \mathbf{G}^{-1}$ , where  $\mathbf{B}$  and  $\mathbf{G}$  are given by

$$\mathbf{B} = \mathbf{S}^{-1} - \frac{\mathbf{S}^{-1} \hat{\mathbf{k}} \hat{\mathbf{k}}^T \mathbf{S}^{-1}}{\hat{\mathbf{k}}^T \mathbf{S}^{-1} \hat{\mathbf{k}}} \quad \text{and} \quad \mathbf{G} = \begin{bmatrix} \hat{\mathbf{i}}^T \hat{\mathbf{i}} & \hat{\mathbf{i}}^T \hat{\mathbf{j}} \\ \hat{\mathbf{j}}^T \hat{\mathbf{i}} & \hat{\mathbf{j}}^T \hat{\mathbf{j}} \end{bmatrix}.$$

(iii) Construct the covariance matrix defining the base of the locus point cylinder in the  $(i, j)$  plane by applying a rotation to the diagonal matrix,  $\mathbf{D}$ , by the angle  $\theta$ :

$$\mathbf{E} = \mathbf{R} \mathbf{D} \mathbf{R}^T,$$

with

$$\mathbf{D} = \begin{bmatrix} a_l^2 & 0 \\ 0 & a_m^2 \end{bmatrix} \quad \text{and} \quad \mathbf{R} = \begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix}.$$

(iv) The convolution of the error ellipse, defined by  $\mathbf{A}$ , and the cylinder base in the ellipse-plane, defined by  $\mathbf{E}$ , is then simply obtained as the sum of the two covariance matrices:  $\mathbf{C} = \mathbf{A} + \mathbf{E}$ .

After convolving the base ellipse of the locus point cylinder, we project the data point on to the ellipse-plane, i.e. from colour-space  $(x, y, z)$  to ellipse-plane  $(i, j)$ . We then check whether the projected point,  $p_{ij}$ , lies within the convolved cylinder defined by the covariance matrix  $\mathbf{C}$ . The last step is to calculate whether the point falls inside the cylinder length, which is similarly convolved by the projected variance along  $\hat{\mathbf{k}}$ .

The first locus point is enclosed by a half ellipsoid defined by the two axes of the cylinder ellipse together with a third axis pointing along  $-\hat{\mathbf{k}}$  of length  $a_k$  (which is taken to be 0.2 and 0.5 for the *ugri* and *griz* colour-spaces, respectively). If  $\mathbf{x}$  is closest to the first locus point, it is therefore necessary to test whether the data point falls within this locus ellipsoid. Again, in order to take the photometric error into account we convolve the base ellipse as specified above. Lastly, we convolve the third axis  $a_k$  by the error by projecting the error ellipsoid on to the line defined by  $\hat{\mathbf{k}}$  that gives the measure  $\sigma_k$ . The length of the expanded half-ellipsoid is then given as the quadratic sum of  $a_k$  and  $\sigma_k$ . We subsequently test whether  $\mathbf{x}$  lies within the expanded half-ellipsoid.

If a data point falls inside the stellar locus, it will not be considered any further as a quasar candidate. In addition to the locus-exclusion step, Richards et al. (2002) include a series of colour and magnitude criteria that we similarly implement. These are given by Richards et al. (2002) and we shall therefore not replicate those steps here.

Since we only use the code on simulated photometry, we do not implement the photometric flag criteria nor do we include the handling of non-detections in certain bands.

All the equations necessary for the projections and operations presented above are implemented in a PYTHON module, *SCSPY*, and the code is publicly available on GitHub.<sup>4</sup>

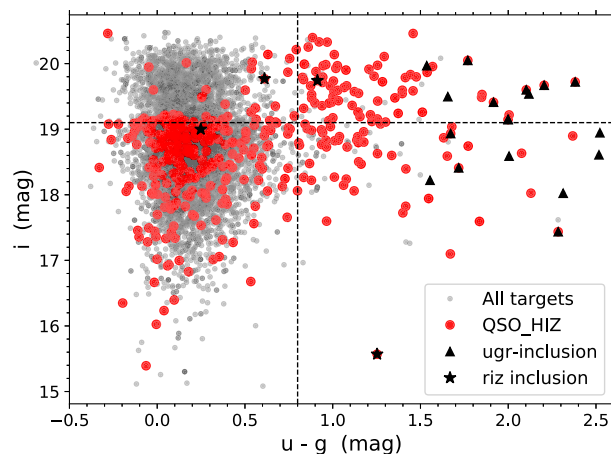
## C2 Testing the selection algorithm

In order to test the algorithm, we have compiled a random set of 10 000 quasars from SDSS-DR7 for which photometry is available in all five bands. For the tests here, we use the dereddened photometry as is the case in the actual target selection algorithm. We have obtained the target selection information from the SDSS data base<sup>5</sup> stored as the `prim_target` flag together with photometric flags indicating any possible errors with the photometry. Out of these 10 000 quasars, 8 417 quasars pass the photometric requirements imposed by Richards et al. (2002). Before testing the colour selection algorithm, we include the additional systematic uncertainties (Richards et al. 2002, Section 3.4.2), i.e. 0.0075 mag together with 15 per cent of the Galactic reddening value are added in quadrature to the photometric uncertainty. Of these 8 417 quasars, only 6 460 have been selected purely on the basis of colour selection, i.e. they have at least one of the primary target flags ‘QSO\_SKIRT’, ‘QSO\_CAP’, or ‘QSO\_HIZ’. When we pass the quasar photometry through our re-implementation of the algorithm, we recover 6 103 objects (94 per cent) as quasars based on colour selection.

We note that since it is not stated which photometric data set was used for the original classification, from which the `prim_target` flag stems, we are not able to reproduce the classification on a one-to-one basis. However, on a statistical basis we are able to recover the vast majority of quasars. This is further demonstrated by the fact that 280 of 6 101 quasars flagged as ‘QSO\_SKIRT’ or ‘QSO\_CAP’ by the SDSS *ugri* selection have  $i \geq 19.1$ . Yet the target selection algorithm only considers targets with  $i < 19.1$  as *ugri* candidates (Richards et al. 2002). When only regarding quasars which fulfil the *i*-band magnitude cuts for the *ugri* and *griz* selection branches, we are able to recover 99 per cent of the quasars.

<sup>4</sup><https://github.com/jkrogager/scspy>

<sup>5</sup><http://skyserver.sdss.org/CasJobs/>



**Figure C1.** Colour–magnitude diagram of quasars from the DR7 test data set with  $g - r < 1$ . The targets selected as ‘QSO\_HIZ’ from the *griz* branch are highlighted as large, red dots. Targets selected from one of the inclusion regions are shown as black stars and triangles. The dotted lines mark the limits ( $u - g \geq 0.8$  and  $i \geq 19.1$ ; see the text) used in the criteria for the low- $z$  rejection (see equation 1 of Richards et al. 2002).

In testing the algorithm, we discovered a mistake in equation (1) from the work by Richards et al. (2002) describing the low-redshift rejection criteria in the *griz* selection branch. When running the algorithm as specified in Richards et al. (2002), we find that nine times too many targets are classified as ‘QSO\_HIZ’. Investigating the colours and magnitudes of the ‘QSO\_HIZ’ candidates in the  $u - g$  versus  $i$ -band magnitude plane reveals that the rejection criteria of equation (1) do not match the distribution of the ‘QSO\_HIZ’ quasars in the test data set (see Fig. C1). Equation (1) of Richards et al. (2002) specify that quasars with  $g - r < 1$ ,  $u - g \geq 0.8$ , and  $i \geq 19.1$  or  $u - g < 2.5$  are *not* targeted as ‘QSO\_HIZ’. However, the distribution of  $u - g$  versus  $i$ -band magnitude for ‘QSO\_HIZ’ targets in the test data with  $g - r < 1$  does not follow the criteria laid out in equation (1). According to the criteria in equation (1), there should be no targets in the range  $0.8 \leq u - g < 2.5$  irrespective of their  $i$ -band magnitude, i.e. in Fig. C1 there should be no targets selected as ‘QSO\_HIZ’ (large red dots), only targets selected from inclusion regions (black stars and triangles) should be included. This is, however, clearly not the case. Since the criteria in equation (1) were designed to exclude low- $z$  quasars fainter than  $i \geq 19.1$  we find it more plausible that the criteria should read  $g - r < 1$ ,  $u - g \leq 0.8$ , and  $i \geq 19.1$ . This set of criteria would indeed discard the upper-left corner of the colour–magnitude diagram where the number of ‘QSO\_HIZ’ targets is very low.

Nevertheless, we have not been able to recover the actual criteria that were used in the classification of the SDSS target selection (Richards, private communication). Instead, we have come up with an approach that closely recovers the classification of the test data described above. However, we caution that the *griz* selection from our re-implemented algorithm does not provide fully robust results.

This error does not significantly affect our main results, since the magnitude limited ( $i < 19.1$ ) sample is selected almost entirely from the *ugri* branch (only 2 per cent are selected purely from the *griz* branch for our high-redshift modelling, i.e.  $z_{\text{QSO}} = 3.2$ ).

Similarly, the analysis by Murphy & Bernet (2016) is affected by the error in the paper describing the target selection.

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