




# Analysis of the impact of broad absorption lines on quasar redshift measurements with synthetic observations

Luz Ángela García <sup>1</sup>★, Paul Martini,<sup>2,3</sup> Alma X. Gonzalez-Morales,<sup>4,5</sup> Andreu Font-Ribera <sup>6,7</sup>, Hiram K. Herrera-Alcantar,<sup>5</sup> Jessica Nicole Aguilar,<sup>8</sup> Steve Ahlen,<sup>9</sup> David Brooks,<sup>7</sup> Axel de la Macorra,<sup>10</sup> Peter Doel,<sup>7</sup> Jaime E. Forero-Romero <sup>11</sup>, Julien Guy,<sup>8</sup> Theodore Kisner,<sup>8</sup> Martin Landriau,<sup>8</sup> Ramon Miquel,<sup>6</sup> John Moustakas,<sup>12</sup> Jundan Nie,<sup>13</sup> Claire Poppett,<sup>8,14,15</sup> Gregory Tarlé<sup>16</sup> and Zhimin Zhou<sup>13</sup>

<sup>1</sup>Universidad ECCI, Cra. 19 no. 49-20, Bogotá, Código 111311, Colombia

<sup>2</sup>Center for Cosmology and AstroParticle Physics, The Ohio State University, 191 West Woodruff Avenue, Columbus, OH 43210, USA

<sup>3</sup>Department of Astronomy, The Ohio State University, 4055 McPherson Laboratory, 140 W 18th Avenue, Columbus, OH 43210, USA

<sup>4</sup>Consejo Nacional de Ciencia y Tecnología, Av. Insurgentes Sur 1582. Colonia Credito Constructor, Del. Benito Jurez C.P. 03940, México D.F., México

<sup>5</sup>Departamento de Física, División de Ciencias e Ingenierías, Campus Leon, Universidad de Guanajuato, León 37150, México

<sup>6</sup>Institut de Física d'Altes Energies (IFAE), The Barcelona Institute of Science and Technology, Campus UAB, E-08193 Bellaterra Barcelona, Spain

<sup>7</sup>Department of Physics & Astronomy, University College London, Gower Street, London, WC1E 6BT, UK

<sup>8</sup>Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA 94720, USA

<sup>9</sup>Physics Department, Boston University, 590 Commonwealth Avenue, Boston, MA 02215, USA

<sup>10</sup>Instituto de Física, Universidad Nacional Autónoma de México, Cd. de México C.P. 04510, México

<sup>11</sup>Departamento de Física, Universidad de los Andes, Cra. 1 No. 18A-10, Edificio Ip, CP 111711, Bogotá, Colombia

<sup>12</sup>Department of Physics and Astronomy, Siena College, 515 Loudon Road, Loudonville, NY 12211, USA

<sup>13</sup>National Astronomical Observatories, Chinese Academy of Sciences, A20 Datun Rd., Chaoyang District, Beijing 100012, P. R. China

<sup>14</sup>Space Sciences Laboratory, University of California, Berkeley, 7 Gauss Way, Berkeley, CA 94720, USA

<sup>15</sup>University of California, Berkeley, 110 Sproul Hall #5800 Berkeley, CA 94720, USA

<sup>16</sup>University of Michigan, Ann Arbor, MI 48109, USA

Accepted 2023 September 28. Received 2023 September 6; in original form 2023 April 14

## ABSTRACT

Accurate quasar classifications and redshift measurements are increasingly important to precision cosmology experiments. Broad absorption line (BAL) features are present in 15–20 per cent of all quasars, and these features can introduce systematic redshift errors, and in extreme cases produce misclassifications. We quantitatively investigate the impact of BAL features on quasar classifications and redshift measurements with synthetic spectra that were designed to match observations by the Dark Energy Spectroscopic Instrument (DESI) survey. Over the course of 5 yr, DESI aims to measure spectra for 40 million galaxies and quasars, including nearly three million quasars. Our synthetic quasar spectra match the signal-to-noise ratio and redshift distributions of the first year of DESI observations, and include the same synthetic quasar spectra both with and without BAL features. We demonstrate that masking the locations of the BAL features decreases the redshift errors by about 1 per cent and reduces the number of catastrophic redshift errors by about 80 per cent. We conclude that identifying and masking BAL troughs should be a standard part of the redshift determination step for DESI and other large-scale spectroscopic surveys of quasars.

**Key words:** methods: numerical – techniques: spectroscopic – (galaxies:) quasars: absorption.

## 1 INTRODUCTION

The Dark Energy Spectroscopic Instrument (DESI) is an ongoing Stage IV ground-based facility focused on studying dark energy and the evolution of structure through baryon acoustic oscillations (BAO) and redshift-space distortions techniques (Levi et al. 2013). DESI will provide us with the most extensive redshift map of galaxies and quasars to date (DESI Collaboration 2016a, b, 2023a, b; Zou et al. 2017; Dey et al. 2019; Raichoor et al. 2020, 2023; Raichoor

et al. in preparation, Ruiz-Macias et al. 2020; Yèche et al. 2020; Zhou et al. 2020; Hahn et al. 2022; Guy et al. 2023; Lan et al. 2023; Miller et al. in preparation; Myers et al. in preparation; Schlafly et al. in preparation; Schlegel et al. in preparation; Silber et al. 2023; Zhou et al. 2023). The DESI survey successfully measures about 205 quasars per square degree, including 60 deg<sup>-2</sup> about  $z > 2.1$  that will include measurements of the Ly  $\alpha$  forest (Chaussidon et al. 2023; Moustakas et al. in preparation). This corresponds to nearly three million quasars, including over 0.8 million at  $z > 2.1$  within the 14 500 deg<sup>2</sup> survey footprint.

The Sloan Digital Sky Survey (SDSS), and especially BOSS (the Baryon Oscillation Spectroscopic Survey) and eBOSS (extended

\* E-mail: [lgarciap@eccci.edu.co](mailto:lgarciap@eccci.edu.co)

BOSS) surveyed 6000 deg<sup>2</sup> and measured more than 500 000 quasars in the redshift range of 0.8–3.5. As part of the analysis, different contaminants to the quasar spectra, such as DLAs (damped Ly  $\alpha$  systems), BALs (broad absorption lines), and other metal absorption lines were identified, and different strategies to mitigate their impact on clustering measurements were implemented (du Mas des Bourboux et al. 2020). The last SDSS data release with new eBOSS data (DR16) catalogued more than 750 000 quasars, including nearly 100 000 BALs (Lyke et al. 2020).

BALs are high column density features in the spectra of quasars produced by clouds of gas moving at high velocities in the quasar host galaxy. Their distance from the black hole is somewhat disputed, as is the mechanism that launches them (Goodrich 1997; Ganguly et al. 2007; Capellupo et al. 2011; De Cicco et al. 2017). Quasars with BAL troughs have been primarily associated with C IV and Si IV, nearly always exhibiting absorption on the blue side of the emission lines’ systemic redshift, with outflow velocities up to 0.1–0.2 of the speed of light (Gibson et al. 2010; Rodriguez Hidalgo et al. 2012; Guo & Martini 2019; Hamann et al. 2019). None the less, BAL-QSOs are also identified with other features such as Al III, Fe II N V, and O VI (among other ionization metals) and the Ly  $\alpha$  emission (Turnshek 1997; Hall et al. 2012; Capellupo et al. 2017).

Different works focus on understanding BAL through other associated transition lines, claiming that C IV only provides a lower limit of the BAL’s effects. For instance, Hall et al. (2012) explores Si IV and N V emission lines in the quasar spectra, in addition to the typical search for C IV to inform about the properties of BAL-quasars. On the other hand, Capellupo et al. (2017) diagnose BALs through the description of powerful outflows associated with P v. Finally, Chen et al. (2020) examine the correlation of BALs in SDSS DR12 quasar spectra with absorption lines environments.

Importantly, these absorption lines generate several issues when present in the Ly  $\alpha$  forest: they add noise to the intrinsic signal of the spectrum, thus, induce an incorrect redshift estimate compared with the quasar systemic redshift up to  $dz \sim 0.01$ . Also, they absorb a significant amount of the flux blueshifted from the emission line counterpart, and consequently, less secure lines can be used to classify their spectra, leading to wrong spectral diagnostics. Finally, the presence of these BALs increases the complexity level when estimating the spectra continuum.

From early in the SDSS project, the BOSS team created a pipeline to identify BALs by visual inspection, characterize and archive them. For instance, SDSS-III labelled BAL-quasars and removed them from their catalogue for Ly  $\alpha$  forest analysis (Slosar et al. 2011). A similar approach was decided for SDSS DR9 (Pâris et al. 2012) and SDSS DR12Q (Bautista et al. 2017; Pâris et al. 2017), due to the large uncertainties in the quasar systemic redshift caused by BAL. Finally, Lyke et al. (2020) explores an algorithm to identify and analyse C IV- and Si IV-BALs in the SDSS DR16 catalogue. The latter approach opens the possibility of treating these lines instead of eliminating spectra with BALs. More quasars are observed with progressively larger spectroscopic facilities; thus, more BAL-QSOs are also being detected, and discarding valid data is not a smart strategy. In particular, Guo & Martini (2019) find 16.8 per cent of BAL-quasars in SDSS DR14. Thus, we seek optimal ways to treat BAL quasars in DESI. For instance, Ennesser et al. (2022) explore masking BALs in eBOSS (DR14) and find that the procedure returns up to 95 per cent of the total forest pathlength lost in previous surveys and discuss how this strategy impacts the Lyman- $\alpha$  autocorrelation functions. However, even when one can mask the BALs from the spectra, there is another effect to consider: the error in the redshift estimation due to the BAL presence. This is

important for the measurements of the quasar autocorrelation function. In addition, Youles et al. (2022) showed that quasar redshifts uncertainties impact the Lyman  $\alpha$  forest auto- and cross-correlation functions.

This work aims to use synthetic spectra to: (i) determine the impact of BAL features on quasars redshifts, in particular, those used for Ly  $\alpha$  forest studies; (ii) quantify the gain on redshift precision by masking the BAL features; (iii) determine if masking BALs at the redshift fitting stage is a viable strategy for an experiment like DESI. Throughout this paper, we focus specifically on quasars at  $z > 1.8$ . This lower limit is necessary to identify BALs that may be present up to 25 000 km s<sup>-1</sup> on the blue side of the C IV line.

The paper is structured as follows: Section 2 describes the simulated spectra and focuses on the introduction and description of BAL features in the mocks. In Section 3, we discuss the pipeline and main assumptions. Finally, Section 4 builds on this work’s results to offer insight on how to treat future DESI data containing BAL features.

## 2 SIMULATED DATA SETS

The simulated spectra, also referred to as mocks, used in this work were produced in two stages:

### 2.1 Raw mocks

The raw mocks assume a spatially flat  $\Lambda$ CDM Planck 2015 cosmology (Planck Collaboration 2016), the corresponding mass power spectrum, and a given observed number density of quasars. A quasar catalogue is generated by identifying the high-density regions in a Gaussian random field realization and locating quasars in such regions. The set of sightlines from the position of each quasar to an observer’s position, what we call skewers, are also drawn from the same gaussian realizations. Subsequently, the skewers are post-processed with LYACOLORE, as discussed in Farr et al. (2020). LYACOLORE adds small-scale fluctuations to a gaussian field, then turns this into a physical density used to calculate the optical depth and, finally, the transmitted flux. As a result, for each skewer in the catalogue, we have the transmitted flux fraction as a function of wavelength. The position of high-column density lines, such as DLAs, has been identified at this stage. However, for the purpose of this study, we do not take them into account hereafter.

### 2.2 Synthetic spectra

The raw mocks are processed with the code QUICKQUASARS (Herrera-Alcantar in preparation), that is based on the DESISIM and SPECSIM repositories (Kirkby et al. 2016). This implementation generates a distinct realistic representation of each quasar spectral energy distribution – for a detailed description of QUICKQUASARS, we refer the reader to Herrera-Alcantar (in preparation). To increase the level of accuracy of the mocks, the code also adds a background QSO continuum, noise and some smoothing of the forest to mimic instrumental resolution to the transmissions. Absorption by BALs, DLAs, other metals in the intergalactic medium (IGM), and the Lyman  $\alpha$  forest are then applied to these spectra. However, we prepare mock spectra containing only BALs and no other contaminant. BAL features are created on a library of 1500 empirical templates constructed from BAL in the SDSS DR14 quasar catalogue (Niu 2020). To construct a template from a BAL QSO, we fit a set of Principal Components Analysis (PCA) components to a BAL in the C IV and Si IV region following the procedure described in

Guo & Martini (2019). We shifted the QSO spectrum to the rest frame and masked the BAL features before we performed this fit, and then divided the BAL-QSO by this best-fitting template to produce a map of the fractional absorption. This procedure works best for the C IV and Si IV region because it is unaffected by Ly $\alpha$  absorption or strong emission lines. However, BAL features are known to be associated with many other lines in regions where it is not possible to measure them directly, such as Ly $\alpha$ , N V, and O VI. We consequently synthesized absorption troughs associated with these other lines based on the absorption versus velocity of the C IV line. We used this procedure to create 1500 BAL templates from a randomly selected subset of the BAL in the Guo & Martini (2019) catalogue with high signal-to-noise ratio (SNR) spectra. The high SNR is necessary because it is more straightforward to identify weaker BAL features in higher SNR data and to measure the structure of the absorption line. We calculated the distributions of absorption index (AI), balnicity index (BI), and velocity offset for the 1500 template BAL and confirmed that their distributions are consistent with the distribution of these parameters in the DR14 BAL catalogue.

Two important properties to characterize BAL quasars are the AI described by Hall et al. (2002) and the BI proposed by Weymann et al. (1991). Both of these parameters are computed in the vicinity of the C IV emission line and are defined as follows:

$$AI_{CIV} = - \int_{25000}^0 \left[ 1 - \frac{f(v)}{0.9} \right] C(v) dv. \quad (1)$$

AI<sub>CIV</sub> is computed from 25000 to 0 km s<sup>-1</sup> bluewards the C IV emission line. The term  $f(v)$  is the normalized flux density of the quasar measured with the C IV line's velocity shift. On the other hand,  $C(v)$  is a parameter set to one if the trough extends for more than 450 km s<sup>-1</sup>, and zero otherwise. Finally, the factor 0.9 captures the fact that BAL troughs absorb at least 10 per cent of the continuum.

We adopt the associated error to the AI<sub>CIV</sub> parameter,  $\sigma_{AI}^2$ , as described by Guo & Martini (2019):

$$\sigma_{AI}^2 = - \int_{25000}^0 \left( \frac{\sigma_{f(v)}^2 + \sigma_{PCA}^2}{0.9^2} \right) C(v) dv, \quad (2)$$

where  $\sigma_{f(v)}$  corresponds to the flux error in each pixel of the normalized flux density  $f(v)$  and  $\sigma_{PCA}$  is the uncertainty found by Guo & Martini (2019) in their PCA fitting.

The definition of BI<sub>CIV</sub> and its error  $\sigma_{BI}^2$  differs from equations (1) and (2) by the fact that it only extends to within 3000 km s<sup>-1</sup> of the line centre and the trough has to extend for at least 2000 km s<sup>-1</sup>, rather than for just 450 km s<sup>-1</sup> as for AI<sub>CIV</sub>.

$$BI_{CIV} = - \int_{25000}^{3000} \left[ 1 - \frac{f(v)}{0.9} \right] C(v) dv. \quad (3)$$

The error for BI was introduced by Trump et al. (2006); however, Guo & Martini (2019) included the additional term  $\sigma_{PCA}$  to account for the error associated with the PCA fitting in their pipeline.

$$\sigma_{BI}^2 = - \int_{25000}^{3000} \left( \frac{\sigma_{f(v)}^2 + \sigma_{PCA}^2}{0.9^2} \right) C(v) dv. \quad (4)$$

Both AI<sub>CIV</sub> and BI<sub>CIV</sub> parameterize the equivalent width of BAL troughs. The fraction of BAL quasars identified with the AI<sub>CIV</sub>  $\neq$  0 criterion is larger than the fraction identified with the BI<sub>CIV</sub> criterion because the AI criterion is sensitive to narrower BAL troughs and it extends closer to the line centre.

The balnicity and absorption index distributions of the templates used for the mocks in this work are representative of the full

distributions found in Guo & Martini (2019). BAL templates are added multiplicatively to the model quasar spectra before adding the Ly $\alpha$  forest absorption and other features related to the IGM. These simulated spectra are then convolved with a model for the instrument resolution. Lastly, QUICKQUASARS adds noise appropriate to the apparent magnitude of the source and the integration time.

Once QUICKQUASARS is run, three main files are produced for each HEALPIX pixel: a truth-, a zbest-, and a spectra-file.<sup>1</sup> The first one contains the most relevant information about the quasar and includes the *true* redshift, the number of exposures, and the fluxes and magnitudes used to produce that particular data set. The segment of this file devoted to BALs contains information about the BAL templates that were used to generate the BAL features, including the BAL template ID, the BAL redshift, the AI<sub>CIV</sub> and BI<sub>CIV</sub> parameters (and their corresponding errors), in addition to the number of distinct components of C IV with absorption width larger than 450 km s<sup>-1</sup>, N<sub>CIV450</sub>, and the minimum and maximum velocities of the C IV troughs defined for each one of the N<sub>CIV450</sub> components,  $v_{min450}$ , and  $v_{max450}$ , respectively.

The zbest file contains a catalogue with modified redshifts, including the finger-of-god effect. This intends to emulate the redshift changes that would be introduced when using a redshift fitter. Finally, the spectra file contains the simulated flux for each camera (b, r, z), the inverse variance of the flux,  $\sigma^{-2}$ , the resolution, and other metadata that are not used in this work. The three files are related through the TARGETID, a unique identifier for each simulated quasar.

In this work, we use two mock realizations:

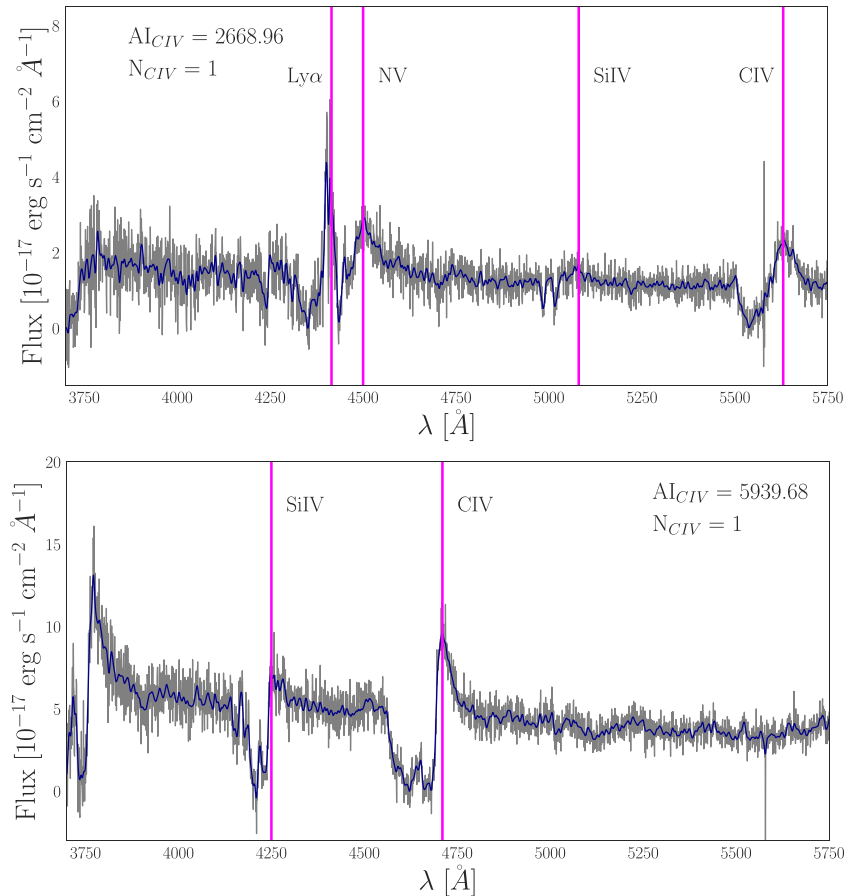
(i) **No BAL mock:** Spectra with continuum and Ly $\alpha$  absorption only. These are simulations where the spectra have only the quasar continuum and Ly $\alpha$  absorption but no BALs or other astrophysical effects. We have generated the same realization (same set of spectra) at several multiples of the standard DESI exposure time of 1000 s to simulate similar SNR as in the DESI Y1 survey: 1000, 2000, 3000, and 4000 s. Hereafter, we identify this realization with the subscript  $noBAL$ .

(ii) **BAL mock:** Same as the 'No BAL mocks' except with 16 per cent of the quasars with BALs. These spectra are identical to the previous case, just with the BAL absorption added. We use these to investigate redshift changes due to the presence of BAL absorption on the same underlying quasar spectra. We have also produced these spectra at several multiples of the standard DESI exposure time (see cases above). We label this realization with the subscript  $BAL$ .

The simulated spectra cover the redshift range from 1.8 to 3.8, and the quasar density of each HEALPIX pixel follows the expected density of quasars for DESI when we began this work (50 quasars per deg<sup>2</sup>, rather than the 60 deg<sup>-2</sup> DESI presently achieves). We used a total of 116 750 quasar spectra, which constitute the 'No BAL mock'. These same quasars are repeated in the 'BAL mock', with the exception that 16 per cent or 18 555 have BAL features. We compute the analysis in a subsample of the DESI Y1 expected footprint because running the redshift classifier REDROCK is computationally expensive. None the less, the subset of spectra is representative of the overall sample.

Fig. 1 shows two examples of synthetic spectra in our catalogue with BALs blueshifted from the C IV emission line. Once these troughs are identified, we masked them following the pipeline discussed above.

<sup>1</sup>The simulated data is organized by healpy pixels, following the data model for DESI data.

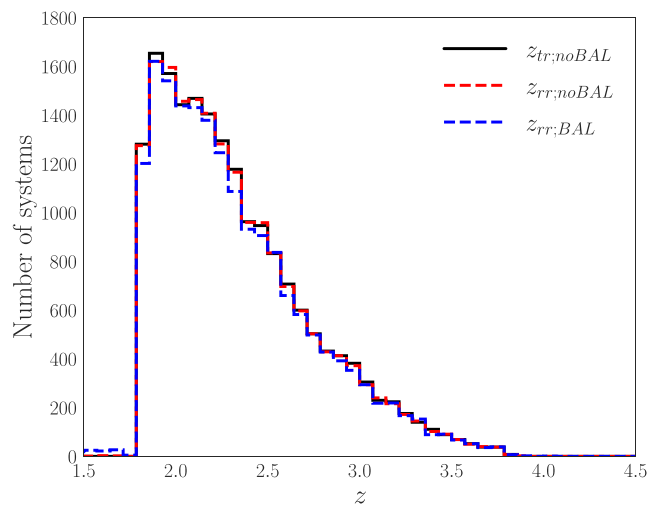


**Figure 1.** Examples of synthetic quasar spectra from DESI Y1 mocks as a function of the observed wavelength. The grey region corresponds to the total flux, whereas the dark blue line shows the smoothed flux. The vertical lines show the main lines used to identify the BAL features. In each panel, BAL troughs are exhibited blueward of C IV emission lines. These spectra’s ‘true’ synthetic redshifts are 2.632 and 2.039 – upper and lower panels, respectively. These redshifts are extracted from the truth files of each spectrum.

### 3 THE EFFECT OF MASKING BALs IN QUASAR SPECTRA

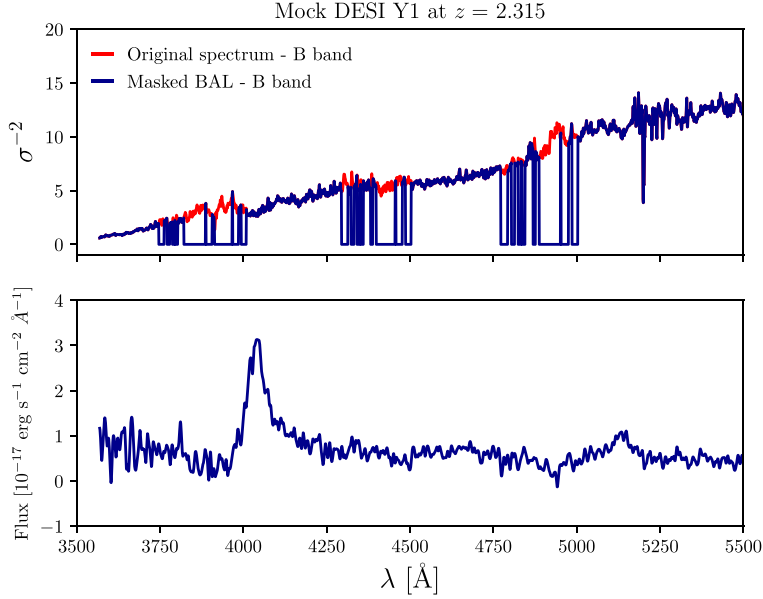
We quantify the impact of BALs by measuring redshifts on both mock realizations, one with BALs and one without. Spectral classification and redshift fits for DESI are calculated with the spectral template–redshift fitting code REDROCK, which was developed by members of the DESI collaboration<sup>2</sup> (Bailey et al. in preparation). The REDROCK redshift fitter compares each spectrum against a set of templates for stars, galaxies, and quasars and returns the best match for the input spectrum based on the fit with the minimal  $\chi^2$ . The return values are a redshift estimate, a class for the type of spectrum fit, and the ZWARNING flag which is primarily different from 0 for a contaminant (any absorption or skyline that could cause an error in the classification of the spectrum). We run REDROCK on both realization and study the differences between the output redshifts. We then assess the effect of masking BALs on the redshift fitting to see if there is an improvement with respect to the no-masked BAL case (see details below).

Fig. 2 shows three redshift distributions for the quasars: (1) the true redshifts of the quasars  $z_{\text{tr,noBAL}}$ ; (2) the REDROCK redshift distribution



**Figure 2.** Redshift distributions of the 18 555 simulated spectra that contain BALs. The histograms show the redshift from the truth file ( $z_{\text{tr,noBAL}}$ , black line), the measurements from REDROCK ( $z_{\text{rr,noBAL}}$ , dashed red lines) for the quasars without the BAL templates; 3) the Ly  $\alpha$  + BAL mocks in (blue dashed). The distributions peak at  $z \sim 2$  and decrease at higher redshift.

<sup>2</sup><https://redrock.readthedocs.io/en/latest/api.html>



**Figure 3.** Inverse variance  $\sigma^{-2}$  and smoothed flux of a DESI Y1 synthetic quasar spectra as a function of observed wavelength. The upper panel shows the masked (blue) and the original inverse variance  $\sigma^{-2}$  (red). When  $\sigma^{-2}$  is set to zero, REDROCK does not fit that part of the spectrum. The lower panel displays the corresponding smoothed flux for each spectrum. The inverse variance set to zero for the eight BAL absorption components on the blue side of C IV, Si IV, N V, and Ly  $\alpha$ .

for quasars without BALs  $z_{\text{tr,noBAL}}$ ; (3) the REDROCK distribution for quasars with BALs  $z_{\text{tr,BAL}}$ . The  $z_{\text{tr,noBAL}}$  redshift is not used in our analysis. However, we draw the comparison here to demonstrate that running REDROCK introduces a variation in the intrinsic redshift, in addition to the shift caused by the presence of BAL features that induce errors in the redshift estimation. We note that the histograms in Fig. 2 follow the distribution of the quasars in our catalogue, with BAL-quasars being 16 per cent of the total number of quasar spectra. While the true redshift range only extends from  $1.8 < z < 3.8$ , the redshift ranges after running REDROCK are  $0.009 < z_{\text{tr,noBAL}} < 5.907$ , and  $-0.003 < z_{\text{tr,BAL}} < 5.907$ . This is because REDROCK classifies some of the quasars as stars and in other cases overestimates the redshift by a substantial amount. This problem is more significant for the quasars that are BALs.

We next reran REDROCK after masking the locations of the BAL features. The information about the locations of the BAL features is stored in a truth catalogue that includes the number of troughs associated with the C IV line that meet the AI criterion  $N_{\text{CIV450}}$  along with the minimum and maximum velocities of each BAL component,  $v_{\text{minCIV450}}$  and  $v_{\text{maxCIV450}}$ , respectively. Specifically, we determine the observed frame wavelengths that contain each absorption trough based on  $v_{\text{minCIV450}}$  and  $v_{\text{maxCIV450}}$  and set the inverse variance of the pixels that contain those wavelengths equal to zero (the catalogue also includes similar information based on the BI criterion, although we do not use that for this study as the trough information based on the AI criterion is more complete). We also apply the mask to the equivalent wavelength ranges associated with the Ly  $\alpha$ , Si IV (1394 Å), and N V (1239 Å) lines. Fig. 3 shows the flux and inverse variance of a quasar with and without the BAL features.

Fig. 4 shows the difference between the estimated redshift in the BAL mock, characterized by the absorption index,  $\text{AI}_{\text{CIV}} > 0$ , and the no BAL mock. We use  $z_{\text{tr,noBAL}}$  as our *true* redshift, meaning it has not been affected by BALs or other contaminants, and  $z_{\text{tr,BAL}}$  is the measured redshift in the sample with BALs (or  $z_{\text{tr,mas}}$  for masked BAL-QSO). The red line shows the difference for cases where the

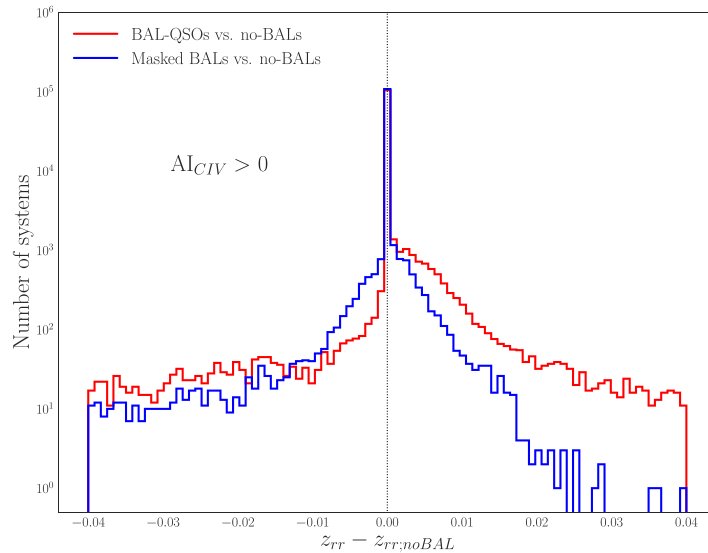
BAL features are not masked. There are significant redshift changes that indicate the presence of BALs increases the redshift errors. Furthermore, the distribution is asymmetric because the redshifts for the BALs are overestimated relative to the no-BAL sample. This is because the BAL features impact the blue side of the strong emission lines. The blue line in Fig. 4 shows the redshift difference  $z_{\text{tr}} - z_{\text{tr,noBAL}}$  distribution after masking the BAL features. In this case, the distribution is nearly symmetric, and the negative  $z_{\text{tr}} - z_{\text{tr,noBAL}}$  tail is very weak, but still visible in the plot because of the logarithmic scale in the y-axis.

There is an improvement when masks are applied to the BAL features, although a few outliers with large  $z_{\text{tr,mas}} - z_{\text{tr,noBAL}}$  persist. Only 557 out of the 18 555 BALs have  $|z_{\text{tr,BAL}} - z_{\text{tr,noBAL}}| > 0.01$  before masking. After masking this number reduces to 103, which corresponds to a reduction of the catastrophic error rate by more than 80 per cent. These numbers indicate that masking is an excellent approach to reducing the number of catastrophic errors due to the BALs.

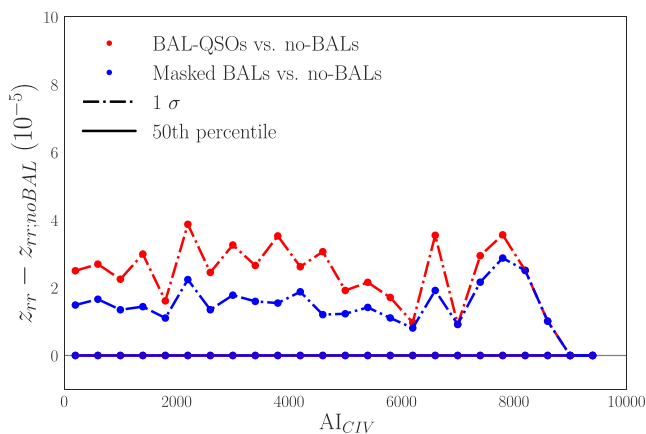
In Fig. 5, we display the distribution of  $z_{\text{tr}} - z_{\text{tr,noBAL}}$  as a function of the absorption index  $\text{AI}_{\text{CIV}}$ . The DESI redshift requirements (Abareshi et al. 2022) for tracer quasars are (1) the tracer quasar redshift accuracy should be  $\sigma_z = 0.0025(1+z)$  and (2) a systematic offset on the redshift should be less than  $\sigma_z = 0.0004(1+z)$ . Figs 5 and 6 show that  $z_{\text{tr}} - z_{\text{tr,noBAL}}$  is well within the DESI science requirements for both the masked and non-masked BALs.

Fig. 5 shows that the 50th percentile has a null difference, and the dispersion is roughly constant regardless of the value of  $\text{AI}_{\text{CIV}}$ . This is because even though there is a trend for larger AI values to produce larger redshift errors, most of the BAL features are sufficiently blueshifted that they do not have an appreciable impact on the line profiles.

Fig. 6 presents  $z_{\text{tr}} - z_{\text{tr,noBAL}}$  as a function of  $z_{\text{tr,noBAL}}$ . There is no trend with redshift for the vast majority of the sample, with the exceptions at the limits of the redshift range where we identify BALs. At low redshift, there is more scatter due to misclassifications,



**Figure 4.** Distributions of redshift differences for the same quasars with  $z_{\text{tr}}$  and without BAL features  $z_{\text{tr,noBAL}}$ , both for the case where the BAL features are not masked (*red*) and where the BAL features are masked (*blue*). In both cases, the presence of BAL features changes the redshift estimate from REDROCK, although there are fewer such cases when the BAL features are masked.

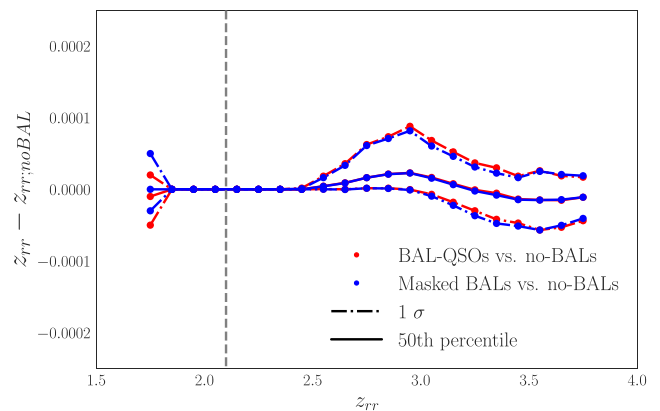


**Figure 5.** Difference between the estimated redshift and the true redshift versus  $AI_{CIV}$ . In this diagram, we present masked BAL with blue points and the original sample of BAL-quasar in red. The 16th (and 84th) and 50th percentiles in the two cases are dashed–dotted and solid lines. Redshift differences in the 16th percentile are overall covered by the 50th percentile.

while above  $z > 2.5$ , the Mg II line is no longer in the spectrograph bandpass, and therefore the C IV line is more critical for the redshift measurement.

### 3.1 Misclassifications and poor fits

The redshift fit performed by REDROCK also provides the classification corresponding to the minimum  $\chi^2$ . The output is reflected in Table 1. We show the percentage of spectra identified as a quasar, a galaxy, or a star. By construction, SPECSIM simulates only quasar spectra; thus, if it was perfect, REDROCK should have classified all the spectra in the input sample as quasars. With the no-BAL sample, REDROCK misidentifies 1.6 per cent of the spectra and tags them as galaxies. Although REDROCK does not always diagnose the spectrum as generated by a quasar, there are very few spotted *wrong* cases. None the less, when 16 per cent of the synthetic quasars contain



**Figure 6.** Difference between the redshift in the absence of BAL features  $z_{\text{tr,noBAL}}$  compared to when they are present but unmasked (*red*), and present and masked (*blue*). The samples at  $1\sigma$  and the 50th percentile are presented in dashed–dotted and solid lines. The dashed grey vertical line splits the mock sample between tracer and Ly  $\alpha$  quasars (left- and right-hand side of the line, respectively). Note that the range of the window is well inside the DESI science requirements.

**Table 1.** Spectral type returned by REDROCK, comparing synthetic spectra with no BALs, unmasked BALs, and masked BALs. All input spectra are quasars, so the percentage classified as galaxies represents misclassifications. There is an improvement in the number of spectra identified as quasar spectra when the BAL features are masked. A variation in the percentage of the misclassified spectra indicates that the masking procedure considerably improves the performance of REDROCK. Note the similar results for the no-BAL and masked-BAL samples.

	Quasar (per cent)	Galaxy (per cent)
<b>No BALs</b>	98.4	1.6
<b>Unmasked BALs</b>	96.4	3.6
<b>Masked BAL</b>	98.0	2.0

**Table 2.** Number of systems with absolute values of  $|dv_{\text{tr}}| > 15000 \text{ km s}^{-1}$ . We compare a sample of 18 555 quasars with unmasked and masked BAL troughs in their spectra. The results reveal that masking BAL troughs effectively reduces cases with  $z_{\text{tr}} < z_{\text{tr,noBAL}}$ , despite missing a few dozen cases with estimated redshifts that are greater than the no-BAL case.

	$dv_{\text{tr}} < -15000 \text{ km s}^{-1}$ (per cent)	$dv_{\text{tr}} > 15000 \text{ km s}^{-1}$ (per cent)
$z_{\text{tr,BAL}} - z_{\text{tr,noBAL}}$	2.9	0.1
$z_{\text{tr,mas}} - z_{\text{tr,noBAL}}$	0.5	0.1

BAL features, the number of spectra misidentified as galaxies rises to 3.6 per cent (the additional 2 per cent; therefore, originates with the 16 per cent that are BALs), and one object is classified as a star. In contrast, when the BALs are masked out, the quasar misclassified as a star is no longer present. The number of galaxy-type objects reported by REDROCK decreases to 2.0 per cent, close to the misclassification percentage in the absence of BALs. Thus, our masking process cuts down the spectral misclassifications.

We also investigate the rate of redshift warnings ZWARNING and the rate of spectral type misclassifications. For the mock spectra without BALs, we find that 2 per cent of the spectra have ZWARNING  $\neq 0$ , which indicates there is an error associated with the redshift. When BALs are present, this increases to 2.5 per cent, while after masking the percentage is 2.2 per cent, very close to the no-BAL value.

We quantify the improvement in the redshift measurements after masking the BAL features with:

$$dv_{\text{tr}} = c \left( \frac{z_{\text{tr}} - z_{\text{tr,noBAL}}}{1 + z_{\text{tr,noBAL}}} \right). \quad (5)$$

and define the catastrophic error rate as the fraction of systems with  $|dv_{\text{tr}}| > 15000 \text{ km s}^{-1}$  and report the values in Table 2.

The results in Table 2 draw three main insights: (i) cases with high redshift dispersion occur mostly if BAL features are present. (ii) Masking brings down 83 per cent of the errors due to the presence of BALs in the spectra in the science requirements window. (iii) Our masking strategy not only reduces the scatter in the overall quasar sample but also induces a significant decline in the redshift errors for quasars with the condition  $z_{\text{tr}} < z_{\text{tr,noBAL}}$ . The same conclusion does not hold for quasars otherwise. In such case, the numbers stay constant regardless of the masking (reflected in large positive values of  $dv_{\text{tr}}$  in Table 2). The redshift uncertainty for  $dv_{\text{tr}} > 15000 \text{ km s}^{-1}$  is not boosted by the presence of BAL in the spectrum, but instead, other systematics that are not affected by the masks.

We define four metrics to investigate further the impact of BALs on the redshift fitting and classification:

- (i) *Good fit*: difference in redshift is below a threshold (compared with the *true* redshift  $z_{\text{tr,noBAL}}$ ) and ZWARNING = 0;
- (ii) *Failed fit*: difference in redshift is above a given threshold (compared with the *true* redshift  $z_{\text{tr,noBAL}}$ ) and ZWARNING = 0 (catastrophic failures);
- (iii) *Missed opportunities*: difference in redshift is below a threshold (compared with the *true* redshift  $z_{\text{tr,noBAL}}$ ) and ZWARNING  $\neq 0$ ;
- (iv) *Lost*: difference in redshift is above a threshold (compared with the *true* redshift  $z_{\text{tr,noBAL}}$ ) and ZWARNING  $\neq 0$ .

The threshold that we use is  $\frac{|z_{\text{tr}} - z_{\text{tr,noBAL}}|}{z_{\text{tr,noBAL}}} = 0.05$ , which is comparable to the allowed tolerance for quasar observations with ground-based telescopes for tracer quasars ( $z < 2.1$ ). The latter is the

**Table 3.** Goodness levels of the fits achieved with REDROCK. The good fits increase by 1.0 per cent when BAL features are masked, mostly offset by a corresponding decrease in the percentage of catastrophic errors and lost cases.

Fit	Good (per cent)	Failed (per cent)	Missed (per cent)	Lost (per cent)
$z_{\text{tr,BAL}} - z_{\text{tr,noBAL}}$	97.29	0.11	2.14	0.46
$z_{\text{tr,mas}} - z_{\text{tr,noBAL}}$	98.10	0.05	1.83	0.02

same threshold assumed to compute results in Table 2, with the exception that  $dv_{\text{tr}}$  is expressed in velocity units (a factor of the speed of light,  $c$ ), and this threshold is dimensionless. It presents an error in percentage. Perfect fits in the code would give  $dv_{\text{tr}} = 0$  or equivalently,  $\frac{|z_{\text{tr}} - z_{\text{tr,noBAL}}|}{z_{\text{tr,noBAL}}} = 0$ .

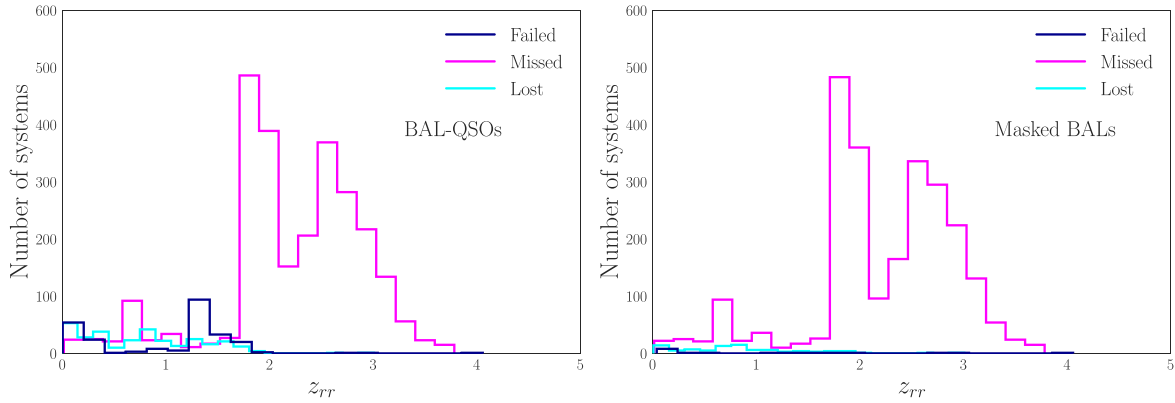
In Table 3, we report the percentages of quasars in each of these categories for unmasked BALs (top row) and masked BALs (bottom row). The good fits increase by  $\sim 1.0$  per cent when BAL troughs are covered up; consequently, failed fits and lost opportunities reduce in a similar proportion.

Fig. 7 shows the redshift distributions of the Failed, Missed, and Lost categories. The left panel shows the BAL mock without masking compared with the redshift from the mock without BALs, and the right panel shows the same distribution except the BALs are masked. We do not compare these fitting cases with those considered *good* for our pipeline because they outnumbered ‘bad’ fits by more than 97–98 per cent. Fig. 7 shows an interesting trend in that the number of failed and lost fits are mostly present around  $z_{\text{tr}} \sim 1.8$ . Once the synthetic mocks are extended to lower redshifts, we could investigate this effect in further detail.

When BALs are present in the mocks, the distribution in redshift for the catastrophic failures (dark blue histograms) is primarily seen in the lower-redshift ‘tracer’ quasars, with a few occurrences at redshifts above 2.5. On the other hand, missed opportunities (good redshift fitting but ZWARNING  $\neq 0$ ) in magenta lines are centred at  $\sim 2.0$ , and their distribution spans the redshift range of 1.8–3.7. Finally, lost chances (wrong redshift estimate and ZWARNING  $\neq 0$ ) are barely spotted in the masked sample. ‘Bad’ fits are largely spotted when BAL occurs in the spectra, and both lost and failed fits scale down when BALs are masked out. The latter results agree with the assumption that led us to run this test: masking the BALs will reduce the redshift errors in the quasar sample used to study the Ly  $\alpha$  forest.

Finally, we briefly revisit the impact of masking the BAL troughs if we only focus on the BAL subsample (this is, of course, not realistic to observations since BAL-QSOs only occur in less than 20 per cent of the detected quasars). We consider 18 555 BAL-QSOs and the same metrics presented in this section and find the following results in Table 4.

Two main conclusions are drawn from this part of the analysis: (i) good fits increase by more than three per cent when the BAL troughs are masked, compared with a rough one per cent if the entire QSO catalogue is considered (Table 3). Conversely, wrong classifications (catastrophic errors and lost opportunities) cut down by  $\sim 19$  per cent with the masking procedure when we limit the sample to BAL-QSOs; (ii) missed opportunities remain constant even if masking is implemented (a trend already seen for the overall sample of QSOs in Table 3), with a modest decrease of 0.8 per cent. This effect is explained by the warnings displayed by the redshift classifier: the fit made by REDROCK shows a chi-squared best fit too close to the second best, the chi-squared minimum is at the edge of the redshift fitting range, or a poor parabola is proposed to compute the  $\chi^2$  minimum.



**Figure 7.** Distribution of failed fits, missed and lost opportunities with redshift (we skip good fits here since they account for more than 97 per cent in all realizations) versus the redshift of the fit computed by REDROCK. We compare 116 750 quasar realizations with unmasked and masked BAL troughs in the left and right panels.

**Table 4.** Goodness levels of the fits achieved with REDROCK. Here, we only consider the BAL-QSO subsample (i.e. 18 555 quasars).

Fit	Good (per cent)	Failed (per cent)	Missed (per cent)	Lost (per cent)
$z_{rr,BAL} - z_{rr,noBAL}$	93.0	1.3	3.8	1.9
$z_{rr,mas} - z_{rr,noBAL}$	96.4	0.1	3.0	0.5

**Table 5.** Effective number of exposures for spectra in the quasar mock sample with BALs. This sample has a total of 116 750 quasars, distributed as 41 366 tracers quasars ( $z < 2.1$ ) and 75 384 Ly  $\alpha$  quasars ( $z > 2.1$ ).

Number of exposures	1000 s	2000 s	3000 s	4000 s
<b>Tracer quasars</b>	41 366	0	0	0
<b>Ly <math>\alpha</math> quasars</b>	33 371	16 935	11 372	13 706

Combinations of the latter flags are also raised, most likely due to other absorption lines in the spectra that turned unaffected by BAL masking.

### 3.2 Exposure time dependence

The results presented so far considered mocks designed to represent the data quality for the DESI Year 1 data set, which corresponds to a nominal exposure time of 1000 s for all spectra. At the conclusion of the 5-yr survey, DESI will observe the  $z > 2.1$  quasars up to four times (for a nominal exposure time of 4000 s) to improve the SNR of the Ly  $\alpha$  forest measurement. In this subsection, we investigate the impact of that greater exposure time on the redshift performance. Only quasars with  $z > 2.1$  are candidates for multiple observations, and only a subset of those in our mock data have such longer exposure times. Specifically, the mock data have 41 366 and 75 384 quasars at  $z < 2.1$  and  $z \geq 2.1$ , respectively. The number of quasars with each exposure time are listed in Table 5.

Fig. 8 presents the distribution  $z_{rr} - z_{rr,noBAL}$  for quasars with different exposure times. The histograms compare the original sample of BALs in solid lines and masked BALs in dashed lines. The plots exhibit the overall distribution for Ly  $\alpha$  quasars ( $z_{rr,noBAL} > 2.1$ ) with  $AI_{CIV} > 0$  – only BAL quasars, i.e. 18 555 in total; thus, it is clear why many systems have a single exposure, according to Table 5.

Fig. 8 reinforces the hypothesis of this work: masking out BALs reduces the discrepancy between the estimated redshift for realizations without and with BALs ( $z_{rr,noBAL}$  and  $z_{rr,BAL}$ ). The benefit of masking BALs is also seen for longer exposure times of 2000–4000 s. This effect is also seen in Fig. A1 and Table A3, with a significant reduction in catastrophic failure rate and the fraction of lost opportunities with an increasing exposure time.

In the Appendices, we present additional assessments of the impact of longer exposure times. In Appendix A, we perform two additional tests to assess if masking effects in the spectra are related to the specific exposure times. The latter quantity is an indirect measurement of average SNR gain. The results presented in Tables A1–A3 show an improvement in the redshift classification with increasing exposure times when BALs are masked. Yet those results have a very uneven distribution of exposure times, with four times more single-exposure mocks than longer exposure times. We evaluate if this distinction has an impact with a second analysis, presented in Appendix B, with the same number of spectra (11546) for each exposure time: 1000, 2000, 3000, and 4000 s with no BALs and BALs. The main distinction between the results in the Appendices are that only 16 per cent of the spectra exhibit BAL features in Appendix A, whereas in Appendix B, the percentage of BALs in the spectra is set to a hundred per cent.

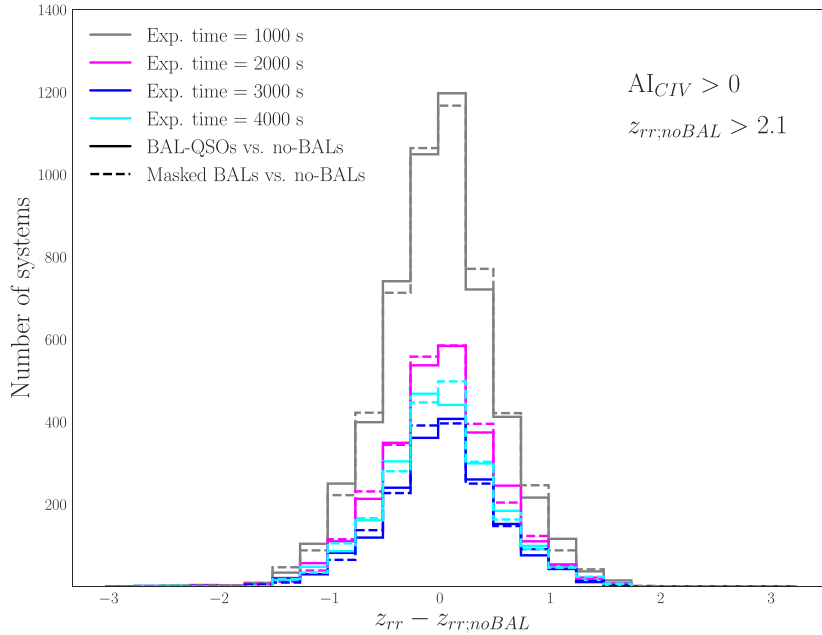
We draw two important conclusions from these tests: (i) the relative distribution of good and ‘bad’ fits remains unchanged regardless of the number of exposure times distribution. The most critical cases have a single exposure – an indirect measure of a low SNR in the spectra – that mainly affects tracer quasars. (ii) Masking BAL contaminants makes a difference in our synthetic results since this strategy improves the success rate achieved by REDROCK, regardless of the exposure time of the mock spectra.

## 4 DISCUSSION AND CONCLUSIONS

We have used synthetic quasar spectra to understand the impact of BAL features on quasar redshifts calculated with REDROCK, the main redshift fitter and object classifier used in DESI.

The first part of the study was devoted to understanding how BAL in the Ly  $\alpha$  spectra affect the redshift estimation with the software REDROCK. Like other absorption lines with large equivalent widths, BAL troughs distort the shape of the spectrum and add noise in the regions with large absorption, which reduces the redshift success rate.





**Figure 8.** Distribution  $z_{rr} - z_{rr,noBAL}$  for Ly  $\alpha$  quasars. The number of exposures displayed is 1000 s (in grey), 2000 s (in magenta), 3000 s (in blue), and 4000 s (in cyan). The original BAL and masked BAL samples are presented in solid and dashed lines, respectively.

We find that the performance of REDROCK decreases in several different ways when BALs are present: a small percentage of the synthetic spectra are misclassified (primarily as galaxies, although one as a star), the velocity error  $dv$  increases, there are more redshift warnings indicated with the ZWARNING flag, and the percentage of good fits decreases by  $\sim 0.5$  per cent.

As discussed in Chaussidon et al. (2023) and Alexander et al. (2023), there is room for improvement with the fitting procedure performed by REDROCK. This work demonstrates that masking the BAL regions and re-running the redshift fitter reduces the error in redshift estimation to be nearly comparable to the non-BAL quasars. Specifically, the masking process reduces the number of misidentified objects by more than half and decreases the incidence of redshift warnings ZWARNING  $> 0$  to only 0.2 per cent above the 2 per cent incidence for the non-BAL sample. The redshift efficiency reflected in Table 3 shows an improvement at about 1 per cent for good fits, which results from the combination of fewer misidentified spectra and fewer ZWARNING flags, and a level of scatter that approximates the non-BAL mocks, in particular at low  $z$ . However, very large redshift dispersions of  $dv_{rr} > 15\,000$  km s $^{-1}$  are not corrected even with the masking.

In summary, the BALs troughs exhibited by  $\sim 16$  per cent of quasars introduce redshift errors and contaminate the Ly  $\alpha$  forest region. We have used mock DESI quasar spectra both with and without BAL features to quantify the magnitude of the redshift errors measured with REDROCK and spectral type misclassifications and catastrophic errors. We have also shown that masking the BAL troughs at the wavelengths of C IV, Si IV, N V, and Lyman  $\alpha$  substantially reduces all of these sources of uncertainty and advocate for the automatic identification and masking of the BAL features as part of the quasar identification and redshift fitting process.

## ACKNOWLEDGEMENTS

LA García thanks the Ly  $\alpha$  WG for allowing her to carry out this project, Universidad ECCI for contributing with funding through the

internal allocation v.05–2019, and Jaime Forero-Romero for presenting her into the collaboration. AFR acknowledges support from the Spanish Ministry of Science and Innovation through the program Ramon y Cajal (RYC-2018-025210) and from the European Union’s Horizon Europe research and innovation programme (COSMO-LYA, grant agreement 101044612). IFAE is partially funded by the CERCA program of the Generalitat de Catalunya. All calculations presented in this work, including the production and storage of the simulated spectra, were done in the supercomputer CORI from the National Energy Research Scientific Computing Center (NERSC) facilities.

This research is supported by the Director, Office of Science, Office of High Energy Physics of the U.S. Department of Energy under contract no. DE-AC02-05CH11231, and by the NERSC, a DOE Office of Science User Facility under the same contract; additional support for DESI is provided by the U.S. National Science Foundation, Division of Astronomical Sciences under contract no. AST-0950945 to the NSF’s National Optical-Infrared Astronomy Research Laboratory; the Science and Technologies Facilities Council of the United Kingdom; the Gordon and Betty Moore Foundation; the Heising-Simons Foundation; the French Alternative Energies and Atomic Energy Commission (CEA); the National Council of Science and Technology of Mexico (CONACYT); the Ministry of Science and Innovation of Spain (MICINN), and by the DESI Member Institutions: <https://www.desi.lbl.gov/collaborating-institutions>.

The authors are honoured to be permitted to conduct scientific research on Iolkam Du’ag (Kitt Peak), a mountain with particular significance to the Tohono O’odham Nation.

## DATA AVAILABILITY

Mock data will be released as part of the DESI data releases. All data points shown on the figures are available in a machine-readable form on <https://zenodo.org/record/7799198#.ZDIPZuzMK3V>.

## REFERENCES

- Abareshi B. et al., 2022, *AJ*, 164, 207  
Alexander D. M. et al., 2023, *AJ*, 165, 124  
Bautista J. E. et al., 2017, *A&A*, 603, A12  
Capellupo D. M., Hamann F., Shields J. C., Rodríguez Hidalgo P., Barlow T. A., 2011, *MNRAS*, 413, 908  
Capellupo D. M. et al., 2017, *MNRAS*, 469, 323  
Chaussidon E. et al., 2023, *ApJ*, 944, 107  
Chen C., Hamann F., Ma B., Lundgren B., York D., Nestor D., AlSayyad Y., 2020, *ApJ*, 902, 57  
DESI Collaboration, 2016a, preprint(arXiv:1611.00036)  
DESI Collaboration, 2016b, preprint(arXiv:1611.00037)  
DESI Collaboration, 2023a, preprint(arXiv:2306.06307)  
DESI Collaboration, 2023b, preprint(arXiv:2306.06308)  
De Cicco D., Brandt W. N., Grier C. J., Paolillo M., 2017, *Frontiers Astron. Space Sci.*, 4, 64  
Dey A. et al., 2019, *AJ*, 157, 168  
du Mas des Bourboux H. et al., 2020, *ApJ*, 901, 153  
Ennesser L., Martini P., Font-Ribera A., Pérez-Ràfols I., 2022, *MNRAS*, 511, 3514  
Farr J. et al., 2020, *J. Cosmol. Astropart. Phys.*, 2020, 068  
Ganguly R., Brotherton M. S., Cales S., Scoggins B., Shang Z., Vestergaard M., 2007, *ApJ*, 665, 990  
Gibson R. R., Brandt W. N., Gallagher S. C., Hewett P. C., Schneider D. P., 2010, *ApJ*, 713, 220  
Goodrich R. W., 1997, *ApJ*, 474, 606  
Guo Z., Martini P., 2019, *ApJ*, 879, 72  
Guy J. et al., 2023, *AJ*, 165, 144  
Hahn C. et al., 2022, *AJ*, 165, 24  
Hall P. B. et al., 2002, *ApJS*, 141, 267  
Hall P. B. et al., 2012, in Chartas G., Hamann F., Leighly K. M., eds, ASP Conf. Ser. Vol. 460, AGN Winds in Charleston. Astron. Soc. Pac., San Francisco, p.78  
Hamann F., Tripp T. M., Rupke D., Veilleux S., 2019, *MNRAS*, 487, 5041  
Kirkby D., Bailey S., Guy J., Weaver B. A., 2016, Zenodo Version 0.5, Quick simulations of fiber spectrograph response.  
Lan T.-W. et al., 2023, *ApJ*, 943, 68  
Levi M. et al., 2013, preprint(arXiv:1308.0847)  
Lyke B. W. et al., 2020, *ApJS*, 250, 8  
Niu W., 2020, in American Astronomical Society Meeting Abstracts #235.  
Pâris I. et al., 2012, *A&A*, 548, A66  
Pâris I. et al., 2017, *A&A*, 597, A79  
Planck Collaboration, 2016, *A&A*, 594, A13  
Raichoor A. et al., 2020, *Res. Notes Am. Astron. Soc.*, 4, 180  
Raichoor A. et al., 2023, *AJ*, 165, 126  
Rodríguez Hidalgo P., Hamann F., Eracleous M., Capellupo D., Charlton J., Shields J., 2012, in Chartas G., Hamann F., Leighly K. M., eds, ASP Conf. Ser. Vol. 460, AGN Winds in Charleston. Astron. Soc. Pac., San Francisco, p. 93  
Ruiz-Macias O. et al., 2020, *Res. Notes Am. Astron. Soc.*, 4, 187  
Silber J. H. et al., 2023, *AJ*, 165, 9  
Slosar A. et al., 2011, *J. Cosmol. Astropart. Phys.*, 2011, 001  
Trump J. R. et al., 2006, *ApJS*, 165, 1  
Turnshek D. A., 1997, in Arav N., Shlosman I., Weymann R. J., eds, ASP Conf. Ser. Vol. 128, Mass Ejection from Active Galactic Nuclei. Astron. Soc. Pac., San Francisco, p. 52  
Weymann R. J., Morris S. L., Foltz C. B., Hewett P. C., 1991, *ApJ*, 373, 23  
Yèche C. et al., 2020, *Res. Notes Am. Astron. Soc.*, 4, 179  
Youles S. et al., 2022, *MNRAS*, 516, 421  
Zhou R. et al., 2020, *Res. Notes Am. Astron. Soc.*, 4, 181  
Zhou R. et al., 2023, *AJ*, 165, 58  
Zou H. et al., 2017, *PASP*, 129, 064101

## APPENDIX A:

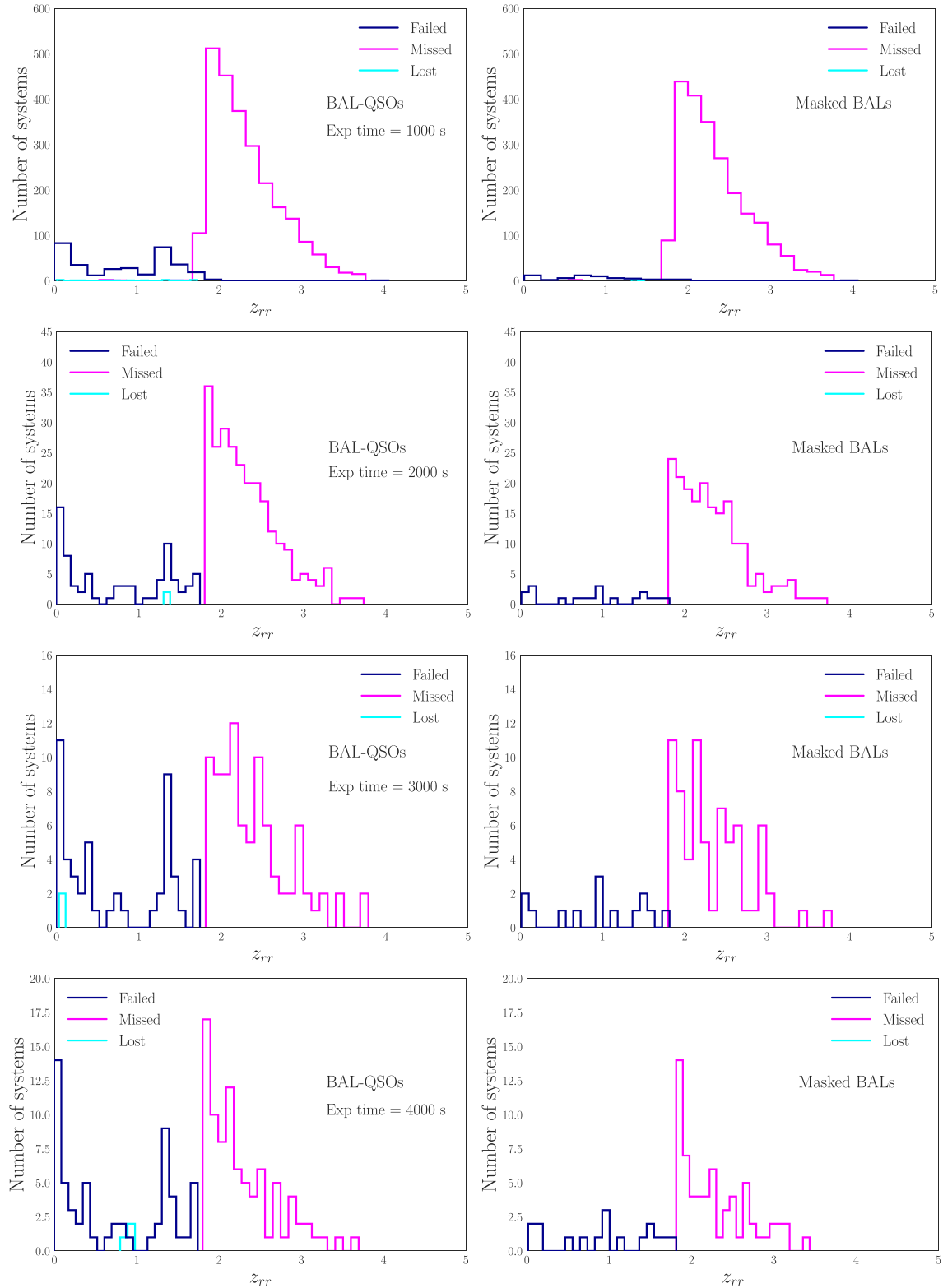
We complement the analysis shown in Section 3.2 to assess the performance of REDROCK by calculating the fits as a function of their time exposures. Table A1 lists the distribution of spectra identified by REDROCK in different exposure times. One main conclusion derived from this part of the analysis is that most errors in the spectra classifications occurs for a single exposure. The longer exposure

**Table A1.** Spectral type fitted by REDROCK, comparing synthetic spectra without BALs, unmasked BALs, and masked BALs. We split the entire sample of 116750 quasars in exposure times. Notably, we find that the spectrum fitted as a star (not shown in the table) has an exposure time of 2000 s and occurs when BALs are added to the synthetic spectra.

Exposure time = 1000 s		
	Quasar (per cent)	Galaxy (per cent)
<b>No BALs</b>	97.5	2.5
<b>Unmasked BALs</b>	95.4	4.6
<b>Masked BALs</b>	96.9	3.1
Exposure time = 2000 s		
<b>No BALs</b>	99.99	0.01
<b>Unmasked BALs</b>	98.1	1.9
<b>Masked BALs</b>	99.98	0.02
Exposure time = 3000 s		
<b>No BALs</b>	99.99	0.01
<b>Unmasked BALs</b>	98.0	2.0
<b>Masked BALs</b>	99.98	0.02
Exposure time = 4000 s		
<b>No BALs</b>	100.0	0.00
<b>Unmasked BALs</b>	98.7	1.3
<b>Masked BALs</b>	99.99	0.01

**Table A2.** ZWARNING flags reported by REDROCK when comparing quasars without BALs, Unmasked BALs and Masked BALs. We split the entire sample of 116750 quasars in different exposure times: 64 per cent with a single exposure, 14.5 per cent with 2000 s, 9.7 per cent with 3000 s, and 11.7 per cent with 4000 s.

ZWARNING = 0 (per cent)			ZWARNING ≠ 0 (per cent)		
Exposure time = 1000 s					
<b>No BALs</b>	97.25		2.75		
<b>Unmasked BALs</b>	96.68		3.32		
<b>Masked BALs</b>	97.03		2.97		
Exposure time = 2000 s					
<b>No BALs</b>	99.99		0.01		
<b>Unmasked BALs</b>	99.68		0.32		
<b>Masked BALs</b>	99.97		0.03		
Exposure time = 3000 s					
<b>No BALs</b>	99.54		0.46		
<b>Unmasked BALs</b>	99.20		0.80		
<b>Masked BALs</b>	99.38		0.62		
Exposure time = 4000 s					
<b>No BALs</b>	100.00		0.0		
<b>Unmasked BALs</b>	99.36		0.64		
<b>Masked BALs</b>	99.53		0.47		



**Figure A1.** Distribution of failed fits, missed and lost opportunities as a function of the estimated redshift by REDROCK for different exposure times, which is a proxy for SNR. We compare synthetic spectra with unmasked BAL features and masked BAL features in the left and right columns, respectively. Conversely, the exposure times are presented in rows, from top to bottom: 1000, 2000, 3000, and 4000 s. Note that in the masked BAL case lost opportunities barely appear, which demonstrates that masking is improving REDROCK’s performance.

**Table A3.** Success rates achieved by REDROCK for different exposure times.

	Good (per cent)	Failed (per cent)	Missed (per cent)	Lost (per cent)
Exposure time = 1000 s				
$z_{rr,BAL} - z_{rr,noBAL}$	96.24	0.44	3.30	0.02
$z_{rr,mas} - z_{rr,noBAL}$	96.94	0.08	2.97	0.01
Exposure time = 2000 s				
$z_{rr,BAL} - z_{rr,noBAL}$	98.06	0.44	1.49	0.01
$z_{rr,mas} - z_{rr,noBAL}$	98.75	0.11	1.13	0.01
Exposure time = 3000 s				
$z_{rr,BAL} - z_{rr,noBAL}$	98.77	0.44	0.78	0.01
$z_{rr,mas} - z_{rr,noBAL}$	99.26	0.12	0.62	0.00
Exposure time = 4000 s				
$z_{rr,BAL} - z_{rr,noBAL}$	98.93	0.43	0.63	0.01
$z_{rr,mas} - z_{rr,noBAL}$	99.41	0.12	0.47	0.00

times result is nearly perfect classifications in both the non-BAL and the masked-BAL cases, with only some residual misclassifications in the case of unmasked BALs.

Table A2 shows the percentage of ZWARNING flags when the entire sample of synthetic quasars are split by exposure time. As before, the most significant number of errors occur for a single exposure (1000 s), and the warning flags drop off significantly for more significant exposure times in the observations.

We also re-calculate the fits, considering the different exposure times in the sample. Table A3 and Fig. A1 show the results on this analysis while taking the difference between  $z_{rr}$  and  $z_{rr,noBAL}$ .

## APPENDIX B:

In addition to the test presented in Appendix A, we assess the accuracy of our results by comparing the goodness of REDROCK fits with the same number of spectra in each exposure time (11546) in each realization in Table B1.

Interestingly, Table B1 reveals that good fits are always above 90 per cent regardless of the exposure time considered. However, there is a higher success rate when BAL-QSOs are only 16 per cent of the total quasars population, as explored in Table A3 compared with the results in this Appendix, in Table B1.

**Table B1.** Success rates in fits achieved by REDROCK, when considering the possible exposure times.

	Good (per cent)	Failed (per cent)	Missed (per cent)	Lost (per cent)
Exposure time = 1000 s				
$z_{rr,BAL} - z_{rr,noBAL}$	91.82	0.99	5.09	2.10
$z_{rr,mas} - z_{rr,noBAL}$	95.15	0.11	4.14	0.60
Exposure time = 2000 s				
$z_{rr,BAL} - z_{rr,noBAL}$	95.89	1.11	2.16	0.84
$z_{rr,mas} - z_{rr,noBAL}$	98.69	0.06	1.19	0.06
Exposure time = 3000 s				
$z_{rr,BAL} - z_{rr,noBAL}$	96.97	1.24	1.29	0.50
$z_{rr,mas} - z_{rr,noBAL}$	99.29	0.06	0.64	0.01
Exposure time = 4000 s				
$z_{rr,BAL} - z_{rr,noBAL}$	97.34	1.30	0.94	0.42
$z_{rr,mas} - z_{rr,noBAL}$	99.49	0.07	0.44	0.0

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