MNRAS **000**, 1–16 (2020) Preprint 21 July 2021 Compiled using MNRAS LATEX style file v3.0

Damped Lyman-alpha Absorbers from Sloan Digital Sky Survey DR16Q with Gaussian processes

Ming-Feng Ho[1⋆], Simeon Bird[1]†, Roman Garnett[2]
[1]*Department of Physics and Astronomy, University of California, Riverside, CA*
[2]*Department of Computer Science and Engineering, Washington University in St. Louis, One Brookings Drive, St. Louis, MO*

21 July 2021

ABSTRACT

We present a new catalogue of Damped Lyman-α absorbers from SDSS DR16Q, as well as new estimates of their statistical properties. Our estimates are computed with the Gaussian process models presented in Garnett et al. (2017); Ho et al. (2020) with an improved model for marginalising uncertainty in the mean optical depth of each quasar. We compute the column density distribution function (CDDF) at $2 < z < 5$, the line density ($\mathrm{d}N/\mathrm{d}X$), and the neutral hydrogen density (Ω_{DLA}). Our Gaussian process model provides a posterior probability distribution of the number of DLAs per spectrum, thus allowing unbiased probabilistic predictions of the statistics of DLA populations even with the noisiest data. We measure a non-zero column density distribution function for $N_{\mathrm{HI}} < 3 \times 10^{22}\,\mathrm{cm}^{-2}$ with 95% confidence limits, and $N_{\mathrm{HI}} \lesssim 10^{22}\,\mathrm{cm}^{-2}$ for spectra with signal-to-noise ratios > 4. Our results for DLA line density and total hydrogen density are consistent with previous measurements. Despite a small bias due to the poorly measured blue edges of the spectra, we demonstrate that our new model can measure the DLA population statistics when the DLA is in the Lyman-β forest region. We verify our results are not sensitive to the signal-to-noise ratios and redshifts of the background quasars although a residual correlation remains for detections from $z_{\mathrm{QSO}} < 2.5$, indicating some residual systematics when applying our models on very short spectra, where the SDSS spectral observing window only covers part of the Lyman-α forest.

Key words: quasar: absorption lines - intergalactic medium - galaxies: statistics

1 INTRODUCTION

Damped Lyman-α absorbers (DLAs) are strong Lyman-α absorption features discovered in quasar spectral sightlines. At the densities required to produce neutral hydrogen column densities above the DLA threshold, $N_{\mathrm{HI}} > 10^{20.3}$ cm^{-2} (Wolfe et al. 1986), the gas of DLAs is self-shielded from the ionising effect of the ultra-violet background (UVB) (Cen 2012) but diffuse enough to have a low star formation rate (Fumagalli et al. 2013). DLAs contain a large fraction of the neutral hydrogen budget after reionisation (Gardner et al. 1997; Noterdaeme et al. 2012; Zafar, T. et al. 2013; Crighton et al. 2015), which make them a direct probe of the distribution of neutral gas.

Numerical simulations tell us DLAs are associated with a wide range of halo masses, with a peak value in the range of $10^{10} - 10^{11}\,\mathrm{M}_{\odot}$ (Haehnelt et al. 1998; Prochaska & Wolfe 1997; Pontzen et al. 2008; Rahmati & Schaye 2014). Through cross-correlating the DLAs with the Lyman-α forest, Font-Ribera et al. (2012) measured a DLA bias factor $b_{\mathrm{DLA}} = 2.17 \pm 0.2$. This implies a median host halo mass of $\sim 10^{12}\,\mathrm{M}_{\odot}$, assuming all DLAs arise from halos of the same mass and. However, a model which assumes a power-law distribution function of DLA cross-section as a function of halo mass is only in marginal tension with the data (Bird et al. 2015). Furthermore, a later measurement from SDSS-DR12 (Pérez-Ràfols et al. 2018) found a bias factor $b_{\mathrm{DLA}} = 1.99 \pm 0.11$, and a median host halo mass $\sim 4 \times 10^{11}\,\mathrm{M}_{\odot}$, in good agreement with simulations. Alternative measurements by cross-correlating with CMB lensing data are broadly consistent with both simulated DLAs and Lyman-α clustering (Alonso et al. 2018; Lin et al. 2020).

In the cosmology context, the Lyman-α forest is a successful probe of matter clustering between $2 < z < 6$ (Croft et al. 1998; McDonald et al. 2000; Viel et al. 2004; McDonald et al. 2005b; Iršič et al. 2017; Chabanier et al. 2019). However, high column density absorbers such as DLAs will

⋆ E-mail: mho026@ucr.edu
† E-mail: sbird@ucr.edu

bias cosmological parameter estimates from Lyman-α and thus need to be masked out (McDonald et al. 2005a). Simulations have been performed to study the effect of damped absorbers on the Lyman-α 1-D and 3-D flux power spectrum (Rogers et al. 2018a,b), and a recent Bayesian fitting method has been proposed to better understand how DLA contaminants affect cosmological inference using the BAO peak (Cuceu et al. 2020).

In this work, we present new estimates for the column density distribution function (CDDF), the abundance of DLAs, and the average neutral hydrogen density at $z = 2-5$ for DLAs in the Sloan Digital Sky Survey IV quasar catalogue from Data Release 16 (SDSS-IV/eBOSS DR16) (Dawson et al. 2016; Lyke et al. 2020). We compute DLA population statistics using the Gaussian process (GP) model presented in Ho et al. (2020), a modified version of the machine learning framework from Garnett et al. (2017). We retrain our model on SDSS DR12 (Eisenstein et al. 2011; Dawson et al. 2013; Alam et al. 2015; Pâris, Isabelle et al. 2018) and generate a DLA catalogue from DR16Q (Lyke et al. 2020). We compute DLA population statistics from the DLA catalogue, which update the estimates we made in Bird et al. (2017); Ho et al. (2020).

The pipeline presented in Garnett et al. (2017) provided for the first time probabilistic detections of DLAs in each spectrum, which comes with a posterior distribution on putative DLAs for the column density and the absorber redshift. With the aid of a full posterior probability distribution for the number of DLAs in each quasar spectrum, "soft" detections in noisy data become available. We propagate uncertainties from each individual spectrum into the global population, without setting any hard threshold on the minimum required probability for the presence of DLAs. We are thus able to include even noisy spectra in our sample of DLAs.

Ho et al. (2020) added an alternative model for sub-DLAs, which regularised excessive detections at low column density. We also included absorption from the mean optical depth in the Lyman-α forest in the GP mean function. This helped prevent the pipeline from using DLAs to compensate for Lyman-α forest absorption in the spectrum, essential at high redshift. In this work, we further improve this aspect of our model. We marginalise out uncertainty in the effective optical depth in each spectrum using the measured mean optical depth as a prior when computing the evidence for the null, DLA, and sub-DLA models.

Several other DLA search methods for SDSS spectra have been implemented. These range from visual inspection surveys (Slosar et al. 2011), visually guided Voigt profile fitting (Prochaska et al. 2005; Prochaska & Wolfe 2009), and template fitting (Noterdaeme et al. 2009, 2012), to machine learning based methods such as a convolutional neural network (CNN) approach (Parks et al. 2018) and an unpublished Fisher discriminant analysis (Carithers 2012). The CNN method (Parks et al. 2018) was also run to identify DLAs as part of the SDSS DR16 quasar catalogue (Lyke et al. 2020). We compare the DLAs detected by our GP model and the DLAs in DR16Q in Section 6.

Machine learning methods have also been proposed to classify broad absorption lines (BALs), including a line-finder based convolutional neural network (CNN) (Busca & Balland 2018) and a hybrid of a CNN with a principal component analysis Guo & Martini (2019).

Section 2 will briefly outline our modelling decisions and the changes to the model made in this work. Section 2.1 describes the cuts we applied to SDSS DR16Q. We recap our modelling details in Section 2.2. We present our results in Section 3, including the CDDF in Section 3.1 and the incidence rate of DLAs and total HI density in Section 3.2. In Section 4, we discuss the possible remaining systematics in our method. Section 5 shows population statistics for DLAs in Ly∞ to Lyβ. In Section 6, we briefly compare our DLA catalogue to the DLAs presented in the SDSS DR16Q catalogue, which implemented a CNN model (Parks et al. 2018) to classify DLAs. We conclude in Section 7.

2 METHODS

Here we briefly recap our Gaussian process (GP) based framework for detecting DLAs using *Bayesian model selection*. We summarise the general approach, while more comprehensive mathematical details may be found in Garnett et al. (2017); Ho et al. (2020). A quasar sightline has spectroscopic observations $\mathcal{D} = (\boldsymbol{\lambda}, \boldsymbol{y})$, where $\boldsymbol{\lambda}$ is a vector of rest wavelength bins, and $\boldsymbol{y}$ is a vector of observed flux at these wavelength bins. Suppose we have built likelihood functions for a set of models $\{\mathcal{M}_i\}$. We can evaluate the posterior probability of a model, $\mathcal{M}$, given a quasar observation, $\mathcal{D}$, based on Bayes' rule:

$$\Pr(\mathcal{M} \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \mathcal{M})\Pr(\mathcal{M})}{\sum_i p(\mathcal{D} \mid \mathcal{M}_i)\Pr(\mathcal{M}_i)}, \tag{1}$$

where $p(\mathcal{D} \mid \mathcal{M})$ is the model evidence of the quasar spectrum $\mathcal{D}$ given model $\mathcal{M}$, $\Pr(\mathcal{M})$ is the prior probability of model $\mathcal{M}$, and the denominator on the right-hand-side is the sum of posterior probabilities of all models in consideration.

Concretely, we have the model without DLAs ($\mathcal{M}_{\neg\mathrm{DLA}}$), the model with k DLAs ($\{\mathcal{M}_{\mathrm{DLA}(i)}\}_{i=1}^{k}$), and the model with sub-DLAs ($\mathcal{M}_{\mathrm{sub}}$). We set $k = 3$ here, allowing up to 3 DLAs per spectrum. We consider a posterior probability of a sub-DLA, $\mathcal{M}_{\mathrm{sub}}$, not to be a DLA detection, as in Ho et al. (2020). Section 2.2 describes the details of how we compute the model evidence for each model.

Table 1 lists mathematical notation and definitions of parameters used throughout the paper.

2.1 Data

Our GP model requires a training set *without DLAs* for training the null model, $\mathcal{M}_{\neg\mathrm{DLA}}$. We use the DLAs in SDSS DR12Q detected by Ho et al. (2020) as our true DLA labels. Here we list the subset of DR12 quasars omitted from our training sample:

- Quasars with $z_{\mathrm{QSO}} < 2.15$, which have almost no Lyman-α forest, are removed.
- `BAL`: quasars with a broad absorption line (BAL) probability larger than 0.75 (`BAL_PROB` $\geqslant 0.75$) are removed, as suggested by Lyke et al. (2020). `BAL_PROB` is derived from QuasarNET (Busca & Balland 2018).
- `CLASS_PERSON == 30`: quasars classified as BALs by human visual inspection are removed.

Table 1. Mathematical notations and definitions

Notation	Description
$\mathcal{M}_{\neg\mathrm{DLA}}$	Null model, model without DLAs or subDLAs
$\mathcal{M}_{\mathrm{DLA}}$	Model with DLAs ($20 \leqslant \log_{10} N_{\mathrm{HI}} \leqslant 23$)
$\mathcal{M}_{\mathrm{sub}}$	Model with subDLAs ($19.5 \leqslant \log_{10} N_{\mathrm{HI}} < 20$)
$p(\mathcal{D} \mid \mathcal{M})$	Model evidence, marginalised likelihood
$\Pr(\mathcal{M})$	Model prior
$(\beta_{\mathrm{MF}}, \tau_{0,\mathrm{MF}})$	Parameters of power-law relation of effective optical depth model
$\tau_{\mathrm{eff,HI}}(z; \beta_{\mathrm{MF}}, \tau_{0,\mathrm{MF}})$	Power-law model of effective optical depth
$p(\beta_{\mathrm{MF}})$	Prior of β_{MF}, assumed to be a normal distribution
$p(\tau_{0,\mathrm{MF}})$	Prior of $\tau_{0,\mathrm{MF}}$, assumed to be a normal distribution
$p(z_{\mathrm{DLA}} \mid z_{\mathrm{QSO}}, \mathcal{M}_{\mathrm{DLA}})$	Prior of redshift of DLAs, a uniform distribution
$p(N_{\mathrm{HI}} \mid \mathcal{M}_{\mathrm{DLA}})$	Prior of column density of a DLA, a data-driven distribution
$\boldsymbol{y}$	Vector of normalised observed flux
$\boldsymbol{\lambda}$	Vector of wavelength pixels in restframe
$\boldsymbol{\nu}$	Vector of instrumental noise variance
$\boldsymbol{\mu}$	Vector of GP mean model
$\boldsymbol{\Sigma}$	Matrix of GP covariance
$\mathbf{A}_{\mathrm{F}}$	Matrix of mean flux suppression from the effective optical depth (diagonal matrix)
$\mathbf{K}$	Matrix to describe covariance of quasar emission spectrum (2281×2281 matrix, 20×2281 parameters)
$\boldsymbol{\Omega}$	Matrix of Lyman series absorption noise (diagonal matrix)

• `ZWARNING`: spectra flagged with `ZWARNING` for pipeline redshift estimation are removed, but extremely noisy spectra with `TOO_MANY_OUTLIERS` are kept.

We have in total 89, 408 spectra without DLAs for training the null model.

We also use the same above criteria to select the DR16Q spectra for applying our model. In addition to the above criteria, the DR16Q quasar sample to which our model is applied is a subset of the full DR16Q sample chosen following additional conditions:

• `IS_QSO_FINAL == 1`: We require this flag in the quasar sample, specifying that a spectrum is robustly classified as a quasar.
• `CLASS_PERSON == 3 or 0`: This flag specifies that the spectrum was classified by a human as a quasar (3) or was not visually identified (0).
• `SOURCE_Z`: as suggested in Section 3.2 of (Lyke et al. 2020), spectra with `Z > 5` and `SOURCE_Z == PIPE` have a suspect redshift estimate and should not be used without a careful visual re-inspection. We thus remove these spectra from our analysis.

Integral to our method is a reliable quasar redshift estimate. It is not trivial to reliably estimate quasar redshifts in the large samples provided by DR16Q,[1] and so we are careful to use the redshift estimates suggested by Lyke et al. (2020). To ensure our quasar redshifts are as homogeneous as possible, we use `Z_PCA`, the recommended redshift estimate method for statistical analyses of a large ensemble of quasars. We also remove the spectra where redshift measurements disagree with each other by more than 0.1, which means we remove samples with $|z_i - z_j| > 0.1$ for $z_i, z_j \in \{$`Z_PIPE`, `Z_PCA`, `Z`, `Z_VI`$\}$. If `Z_VI` is not present, we use only the other three redshift estimates. Our final DR16Q sample size contains 159 807 Lyman-α quasar spectra.

[1] Indeed, we have extended our GP framework to provide a quasar redshift estimate (Fauber et al. 2020).

2.2 Gaussian process model

Consider a distant quasar with a known redshift, z_{QSO}. Each spectroscopic observation gives us the observed flux, $\boldsymbol{y}$, on a set of wavelength pixels in observed-frame wavelengths, $\boldsymbol{\lambda}_{\mathrm{obs}}$. Since the quasar redshift is assumed to be known, we shift into the rest frame, $\boldsymbol{\lambda} = \boldsymbol{\lambda}_{\mathrm{obs}}/(1 + z_{\mathrm{QSO}})$. Standard errors are provided with each observed flux pixel, $\sigma(\lambda_i)$, with λ_i the ith pixel in $\boldsymbol{\lambda}$, and we define the noise variance of each observed flux pixel as $\nu_i = \sigma(\lambda_i)^2$. Given the observed flux of a quasar, we normalise all flux measurements by dividing the median flux observed between $[1425\text{Å}, 1475\text{Å}]$ in the rest-frame, a wavelength range redwards of the Lyα emission and avoiding major emission lines.

For each quasar observation, we have data $\mathcal{D} = (\boldsymbol{\lambda}, \boldsymbol{y}, \boldsymbol{\nu}, z_{\mathrm{QSO}})$. We want to build a likelihood function to describe this data:

$$p(\boldsymbol{y}|\boldsymbol{\lambda}, \boldsymbol{\nu}, z_{\mathrm{QSO}}),$$

which is the likelihood of the flux $\boldsymbol{y}$ given all other observed quantities. We model this likelihood as a Gaussian process:

$$p(\boldsymbol{y}|\boldsymbol{\lambda}, \boldsymbol{\nu}, z_{\mathrm{QSO}}) = \mathcal{N}(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

where $\boldsymbol{\mu}$ is the mean vector of the GP, and $\boldsymbol{\Sigma}$ is the covariance matrix of the GP. We will use bold lowercase italics for vectors and bold uppercase letters for matrices.

2.2.1 Learning the GP null model

A GP is fully specified by its first two central moments: the mean function, $\mu(\lambda)$, and the covariance kernel, $K(\lambda, \lambda')$, (Rasmussen & Williams 2005). Our task now is to learn the mean function and the covariance function from the training set. Suppose we have a set of quasar observations without any intervening DLAs, $\{\mathcal{D}_1, \mathcal{D}_2, \cdots, \mathcal{D}_{\mathrm{N_{spec}}}\}$, where $\mathrm{N_{spec}}$ is the number of quasars in the training set. We can then learn

the mean function by taking a precision weighted average:

$$\mu_j = \frac{\sum\limits_i y_{ij} \neq \mathrm{NaN} \left(y_{ij}/\nu_{ij}\right)}{\sum\limits_i y_{ij} \neq \mathrm{NaN} \left(1/\nu_{ij}\right)}, \tag{2}$$

where the summation is over i index. j indicates jth pixel in the observed flux, i represents ith spectrum, and we only average over the non-NaN values. Note this differs from Ho et al. (2020), where we used the mean rather than the precision weighted average. The precision weighted average can be viewed as a result of using an uninformative prior on μ_j and an independent Gaussian likelihood for each y_{ij}. If we have a set of normally disturbed flux pixels with each flux pixel follows $y_{ij} \sim \mathcal{N}(\mu_j, \nu_{ij})$ with known variance ν_{ij} and an unknown μ_j with an uninformative prior, the posterior will be a normal distribution with a new mean equals a precision weighted average.

Instead of training on the raw observed flux $\boldsymbol{y}$ directly, we follow (Ho et al. 2020) to train the mean function and the kernel on the flux after removing the average effect of the Lyman-α forest, the de-forest flux:

$$\begin{aligned} y_{ij} &\leftarrow y_{ij} \cdot \exp(\tau_{\mathrm{eff,HI}}); \\ \nu_{ij} &\leftarrow \nu_{ij} \cdot \exp(2 \cdot \tau_{\mathrm{eff,HI}}), \end{aligned} \tag{3}$$

which means we replace observed flux and its variance with the flux and variance before the suppression of Lyman-α forest. The effective optical depth is parameterised as:

$$\begin{aligned} &\tau_{\mathrm{eff,HI}}(z(\lambda_{\mathrm{obs}}); \beta_{\mathrm{MF}}, \tau_{0,\mathrm{MF}}) = \\ &\qquad \sum_{i=2}^{N} \tau_{0,\mathrm{MF}} \frac{\lambda_{1i} f_{1i}}{\lambda_{12} f_{12}} (1 + z_{1i}(\lambda_{\mathrm{obs}}))^{\beta_{\mathrm{MF}}}, \end{aligned} \tag{4}$$

where λ_{1i} is the transition wavelength from Lyman-α to the ith member in the Lyman series, f_{1i} represents the oscillator strength, z_{1i} is the absorber redshift, and we set $N = 31$. The absorber redshift is written as:

$$\begin{aligned} 1 + z_{1i}(\lambda, z_{\mathrm{QSO}}) &= \frac{\lambda_{\mathrm{obs}}}{\lambda_{1i}} \\ &= \frac{\lambda(1 + z_{\mathrm{QSO}})}{\lambda_{1i}}. \end{aligned} \tag{5}$$

We parameterise the effective optical depth by a power-law relation with $\tau_{0,\mathrm{MF}}$ and β_{MF} parameters. Here we specify a subscript "MF" to annotate the parameters modified by mean flux suppression. Fig 1 shows our new GP mean function, compared to Ho et al. (2020).

Taking this Lyman-α mean flux into account introduces a dependence on quasar redshift into the mean function of the GP for each quasar:

$$\begin{aligned} &\mu(\lambda, z_{\mathrm{QSO}}; \beta_{\mathrm{MF}}, \tau_{0,\mathrm{MF}}) = \\ &\quad \mu(\lambda) \cdot \exp(-\tau_{\mathrm{eff,HI}}(z(\lambda, z_{\mathrm{QSO}}); \beta_{\mathrm{MF}}, \tau_{0,\mathrm{MF}})). \end{aligned} \tag{6}$$

$\mu(\lambda)$ is the mean function we learned from Eq 2. We learn the mean function on a dense grid of wavelengths on a chosen rest-frame wavelength range:

$$\lambda \in [850.75\,\text{Å}, 1420.75\,\text{Å}] \tag{7}$$

with a linearly equal spacing of $\Delta\lambda = 0.25$Å. Ho et al. (2020) only modelled the null model in the Lyman-α region, [911.75Å, 1215.75Å]. We extend the red end of our model to include a part of the metal line region until 1420.75 Å. This empirically improved the column density estimation of DLAs near the Lyman-α emission peak, as otherwise part of the damping wing would go beyond 1215.75 Å when a large DLA is very close to the quasar.

The mean function is thus written as a mean vector $\boldsymbol{\mu}(z_{\mathrm{QSO}}; \beta_{\mathrm{MF}}, \tau_{0,\mathrm{MF}}) = \mu(\boldsymbol{\lambda}, z_{\mathrm{QSO}}; \tau_{0,\mathrm{MF}}, \beta_{\mathrm{MF}})$ and the kernel is written as a matrix $\Sigma(\lambda, \lambda') = \boldsymbol{\Sigma}$. The covariance matrix's optimisation procedure is described in Garnett et al. (2017); Ho et al. (2020). We factorise the covariance matrix as in Ho et al. (2020):

$$\boldsymbol{\Sigma}_i = \mathbf{A}_{\mathrm{F}}^{\top} (\mathbf{K} + \boldsymbol{\Omega}) \mathbf{A}_{\mathrm{F}} + \operatorname{diag} \boldsymbol{\nu}_i. \tag{8}$$

The $\mathbf{K}$ matrix is a positive-definite symmetric matrix corresponding to the covariance between each quasar flux pixel. $\boldsymbol{\Omega}$ is a diagonal matrix describing the absorption noise:

$$\operatorname{diag} \boldsymbol{\Omega} = \boldsymbol{\omega} \circ (1 - \exp(-\tau_{\mathrm{eff,HI}}(\boldsymbol{z}; \beta, \tau_0)) + c_0)^2. \tag{9}$$

$\boldsymbol{\omega}$ is freely optimisable while the Lyman-α flux term, $(1 - \exp(\tau_{\mathrm{eff,HI}}(\boldsymbol{z}; \beta, \tau_0)) + c_0)^2$, includes the redshift dependent noise variance with which we model the Lyman-α forest. The optimised absorption noise parameters used here are:

$$\tau_0 = 0.000119 \qquad \beta = 5.15 \qquad c_0 = 0.146. \tag{10}$$

The $\mathbf{A}_{\mathrm{F}}$ is a diagonal matrix reflecting the mean vector suppression for each spectrum corresponding to the mean flux in the Lyman-α forest:

$$\operatorname{diag} \mathbf{A}_{\mathrm{F}} = \exp\left(-\boldsymbol{\tau}_{\mathrm{eff,HI}}(z_{\mathrm{QSO}}; \beta_{\mathrm{MF}}, \tau_{0,\mathrm{MF}})\right). \tag{11}$$

The parameters of this matrix follow the values given in Kamble et al. (2020), which used a power-law relation to measure the effective optical depth in the Lyman-α forest in SDSS DR12:

$$\tau_{0,\mathrm{MF}} = 0.00554 \qquad \beta_{\mathrm{MF}} = 3.182, \tag{12}$$

with associated uncertainty for each parameter:

$$\sigma_{\tau_{0,\mathrm{MF}}} = 0.00064 \qquad \sigma_{\beta_{\mathrm{MF}}} = 0.074. \tag{13}$$

The instrumental noise is encoded in the diagonal matrix diag $\boldsymbol{\nu}_i$, where i simply denotes the ith quasar observation: The final covariance matrix learned from our data is shown in Fig 2. Comparing the kernel matrix we learned in this work to Ho et al. (2020), the current kernel is less noisy and contains several distinct features of emission lines. The reduction in the noise is due to a larger training set, SDSS DR12Q catalogue, is used for optimising the kernel.

After having learned the GP null model, we can write down the null model likelihood function:

$$\begin{aligned} &p(\boldsymbol{y} \,|\boldsymbol{\lambda}, \boldsymbol{\nu}, z_{\mathrm{QSO}}, \beta_{\mathrm{MF}}, \tau_{0,\mathrm{MF}}, \mathcal{M}_{\neg\mathrm{DLA}}) = \\ &\qquad \mathcal{N}(\boldsymbol{y}; \boldsymbol{\mu}(z_{\mathrm{QSO}}; \beta_{\mathrm{MF}}, \tau_{0,\mathrm{MF}}), \mathbf{A}_{\mathrm{F}}^{\top} (\mathbf{K} + \boldsymbol{\Omega}) \mathbf{A}_{\mathrm{F}} + \operatorname{diag} \boldsymbol{\nu}_i), \end{aligned} \tag{14}$$

where the notation $\mathcal{M}_{\neg\mathrm{DLA}}$ specifies that our null GP model is conditioned on a training set *without DLAs*.

2.2.2 Model evidence for the null model

Once we have trained our GP null model, $\mathcal{M}_{\neg\mathrm{DLA}}$, according to Section 2.2.1, we need to integrate out the nuisance parameters associated with Lyman-α forest absorption to get the model evidence.

In Ho et al. (2020), we only took the mean values of the

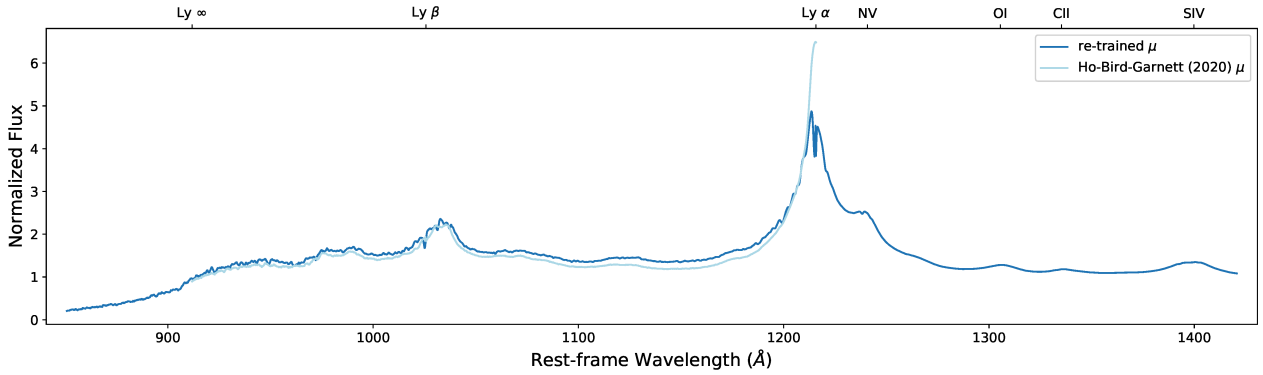

Figure 1. Our GP mean function using a precision weighted average of the rest-frame wavelengths. We extended our model compared Ho et al. (2020) (light blue), both bluewards past the Lyman break at 912Å and redwards past the SiV emission line.

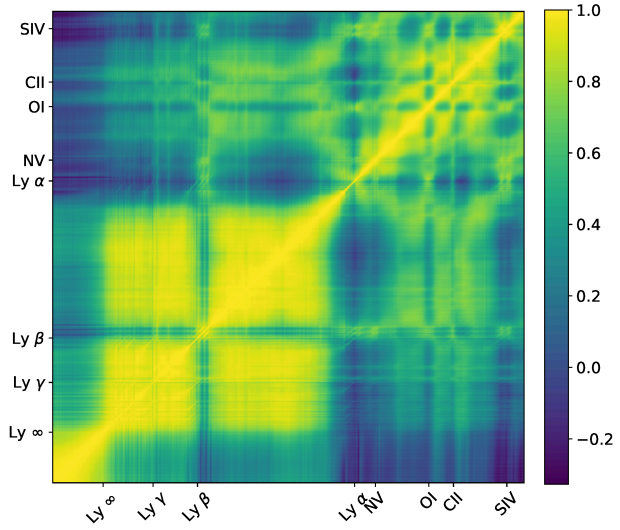

Figure 2. The correlation matrix learned from data, which is the covariance matrix $\mathbf{K}$ normalised by the diagonal elements. Note that the correlation in the plot is pixel-by-pixel, and the matrix dimension is 2281×2281. Different emission lines and the Lyman break are visible in the plot.

meanflux parameters $(\beta_{\mathrm{MF}}, \tau_{0,\mathrm{MF}})$ without their uncertainties, so the model evidence straightforwardly equals to Eq 14 without integration. In this work, we take the uncertainties of meanflux suppression into account and integrate them out, according to Kamble et al. (2020) prior. The model evidence thus will be:

$$p(\mathcal{D} \mid \mathcal{M}_{\neg\mathrm{DLA}}, \boldsymbol{\nu}, z_{\mathrm{QSO}}) \propto p(\boldsymbol{y} \mid \boldsymbol{\lambda}, \boldsymbol{\nu}, z_{\mathrm{QSO}}, \mathcal{M}_{\neg\mathrm{DLA}}), \quad (15)$$

where we integrate out $(\beta_{\mathrm{MF}}, \tau_{0,\mathrm{MF}})$

$$p(\boldsymbol{y} \mid \boldsymbol{\lambda}, \boldsymbol{\nu}, z_{\mathrm{QSO}}, \mathcal{M}_{\neg\mathrm{DLA}}) = \int p(\boldsymbol{y} \mid \boldsymbol{\lambda}, \boldsymbol{\nu}, z_{\mathrm{QSO}}, \beta_{\mathrm{MF}}, \tau_{0,\mathrm{MF}}, \mathcal{M}_{\neg\mathrm{DLA}}) \quad (16)$$
$$p(\beta_{\mathrm{MF}}) p(\tau_{0,\mathrm{MF}}) \mathrm{d}\beta_{\mathrm{MF}} \mathrm{d}\tau_{0,\mathrm{MF}}$$

with

$$\begin{aligned} p(\beta_{\mathrm{MF}}) &= \mathcal{N}(\beta_{\mathrm{MF}} = 3.182, \sigma_{\beta_{\mathrm{MF}}} = 0.074) \\ p(\tau_{0,\mathrm{MF}}) &= \mathcal{N}(\tau_{0,\mathrm{MF}} = 0.00554, \sigma_{\tau_{0,\mathrm{MF}}} = 0.00064)\,. \end{aligned} \quad (17)$$

We then use Quasi-Monte Carlo (QMC) to integrate out the meanflux parameters with 30 000 samples of $(\beta_{\mathrm{MF}}, \tau_{0,\mathrm{MF}})$. QMC takes samples from a so-called low-discrepancy sequence, leading to faster convergence. Here we draw 30 000 samples generated from a scrambled Halton sequence, which gives samples approximately uniformly distributed on a unit square $[0, 1]^2$. We then use inverse transform sampling to transform the Halton sequence to the distribution described in Eq 17.

2.2.3 Model evidence for the DLA model

Suppose we have a trained GP null model in Eq 14, the DLA likelihood function will be the null model likelihood function multiplied by Voigt profiles for each line in the Lyman series of the absorber:

$$\begin{aligned} &p(\boldsymbol{y} \mid \boldsymbol{\lambda}, \boldsymbol{\nu}, z_{\mathrm{QSO}}, \beta_{\mathrm{MF}}, \tau_{0,\mathrm{MF}}, \\ &\qquad \{z_{\mathrm{DLA}\,i}\}_{i=1}^{k}, \{N_{\mathrm{HI}\,i}\}_{i=1}^{k}, \mathcal{M}_{\mathrm{DLA(k)}}) \\ &= \mathcal{N}(\boldsymbol{y}; \boldsymbol{a}_{(k)} \circ \boldsymbol{\mu}(z_{\mathrm{QSO}}; \beta_{\mathrm{MF}}, \tau_{0,\mathrm{MF}}), \\ &\qquad \mathbf{A}_{(k)}(\mathbf{A}_{\mathrm{F}}(\mathbf{K} + \boldsymbol{\Omega})\mathbf{A}_{\mathrm{F}})\mathbf{A}_{(k)} + \mathrm{diag}\,\boldsymbol{\nu}_i)\,. \end{aligned} \quad (18)$$

Here $\mathbf{A}_{(k)} = \mathrm{diag}\,\boldsymbol{a}_{(k)}$ and $\boldsymbol{a}_{(k)}$ is the function with k voigt profiles, which represents k DLAs:

$$\boldsymbol{a}_{(k)} = \prod_{i=1}^{k} a(\boldsymbol{\lambda}; z_{\mathrm{DLA}\,i}, N_{\mathrm{HI}\,i})\,. \quad (19)$$

$a(\boldsymbol{\lambda}; z_{\mathrm{DLA}}, N_{\mathrm{HI}})$ is a Voigt profile parameterised by the DLA's redshift, z_{DLA}, and the column density of the DLA, N_{HI}. The Voigt profile parameterisation used in this work is the same as Garnett et al. (2017). We set the maximum number of DLAs per spectrum at $k = 3$ in this work, as there are rarely more than three DLAs per spectrum. As described in Garnett et al. (2017), the default Voigt profile we use in this work includes Lyα, Lyβ, and Lyγ absorption, which allows us to constrain the DLA column density better when the Lyman-β forest is in the observation window.

To get the model evidence, according to Eq 18, we need to integrate out the prior over the DLA parameters and

the meanflux parameters ($\beta_{\rm MF}, \tau_{0,\rm MF}$). For convenience, we denote the parameters which need to be integrated out by $\theta = \{\{z_{{\rm DLA}\,i}\}_{i=1}^{k}, \{N_{{\rm HI}\,i}\}_{i=1}^{k}, \beta_{\rm MF}, \tau_{0,\rm MF}\}$.

For the model with a single DLA, we have four parameters $\theta = \{z_{\rm DLA}, N_{\rm HI}, \beta_{\rm MF}, \tau_{0,\rm MF}\}$. The model evidence is:

$$p(\boldsymbol{y} \mid \boldsymbol{\lambda}, \boldsymbol{\nu}, z_{\rm QSO}, \mathcal{M}_{\rm DLA}) = \int p(\boldsymbol{y} \mid \boldsymbol{\lambda}, \boldsymbol{\nu}, z_{\rm QSO}, \theta, \mathcal{M}_{\rm DLA}) p(\theta \mid z_{\rm QSO}, \mathcal{M}_{\rm DLA}) {\rm d}\theta. \tag{20}$$

By assuming each parameter is independent of each other, we factorise the parameter prior as:

$$p(\theta \mid z_{\rm QSO}, \mathcal{M}_{\rm DLA}) = p(z_{\rm DLA} \mid z_{\rm QSO}, \mathcal{M}_{\rm DLA}) p(N_{\rm HI} \mid \mathcal{M}_{\rm DLA}) p(\beta_{\rm MF}) p(\tau_{0,\rm MF}), \tag{21}$$

where we assign the Kamble et al. (2020) prior for the meanflux parameters as in Eq 17. We use the same prior for column density, $p(N_{\rm HI} \mid \mathcal{M}_{\rm DLA})$, as Ho et al. (2020). This was trained using kernel density estimation on the $\log_{10} N_{\rm HI}$ distribution from Lee et al. (2013) DR9 DLAs with an addition of a 3% uniform prior.

The $z_{\rm DLA}$ prior is uniform within the search range for DLAs. We set this search range to be from Lyman-β to Lyman-α. Removing DLAs detected in the Lyman-β forest ensures the purity of DLA samples in deriving the statistical properties of the DLA population. However, to generate a complete catalogue, we also consider a search range from the Lyman limit to Lyman-β.

We used the same model and priors for the sub-DLA model as in Ho et al. (2020). The sampling range of the redshifts of sub-DLAs is the same as for the DLA model. Model priors are the same as Ho et al. (2020), based on the DLA catalogue in SDSS DR9 (Carithers 2012).

2.3 Example spectra

In this section, we show some example spectra to demonstrate our proposed model. Figure 3 shows an example with prominent DLA features. As shown in the parameter space (middle plot), the posterior distribution is peaked at the *maximum a posteriori* (MAP) values of those two DLAs. Our GP model estimates the parameters of the DLAs with small uncertainties. As shown in the top plot, our MAP values agree with the column densities measured by the CNN model reported in the DR16Q catalogue.

Figure 4 shows an extremely noisy spectrum, for which our GP model is very uncertain about the effective Lyman-α absorption in the spectrum. The DLA models are degenerate with the absorption from the Lyman-α forest. Without modelling the uncertainty in the mean flux, the GP model does not know that the drop in the spectrum can be explained by Lyman-α forest absorption. It instead fits a big DLA with $N_{\rm HI} = 10^{22.9}\,{\rm cm}^{-2}$ as its preferred explanation for the drop in flux.

2.4 Selection on the strength of Occam's razor

As we use more parameters to compute the DLA or sub-DLA model, the model selection will prefer to fit a Voigt profile to the GP if all candidate models are poorly fit. Thus, the DLA or sub-DLA model's evidence is sometimes too strong compared to the null model.

The most common poor fit situations are quasar spectra with $z_{\rm QSO} < 2.5$ and with low signal-to-noise ratios (SNR). As SDSS optical spectra have a fixed observing window, quasar spectra with $z_{\rm QSO} < 2.5$ have an incomplete Lyman-α forest. The constraining power of the quasar becomes weaker as only part of the data fits into our modelling window, [850.75 Å, 1420.75 Å]. Thus the DLA model and the null model are closer in likelihood space.

To avoid this situation, we introduced an additional Occam's razor in Ho et al. (2020), which is injected in the model selection as:

$$\Pr(\mathcal{M}_{\rm DLA} \mid \mathcal{D}) = \frac{\Pr(\mathcal{M}_{\rm DLA}) p(\mathcal{D} \mid \mathcal{M}_{\rm DLA}) \frac{1}{N}}{\begin{pmatrix}\Pr(\mathcal{M}_{\rm DLA}) p(\mathcal{D} \mid \mathcal{M}_{\rm DLA}) \\ +\Pr(\mathcal{M}_{\rm sub}) p(\mathcal{D} \mid \mathcal{M}_{\rm sub})\end{pmatrix} \frac{1}{N} + \Pr(\mathcal{M}_{\neg \rm DLA} \mid \mathcal{D})}, \tag{22}$$

Here N is the Occam's razor penalty, and we used $N = 10\,000$ in Ho et al. (2020). We previously validated the Occam's razor strength by matching it to the DR9 concordance catalogue (Carithers 2012).

In this work, however, we modify our null model to consider uncertainty from the mean flux measurement, which means it has more parameters. Thus, the null model gains more constraining power, so a weaker Occam's razor may be preferable. To make our model posteriors more consistent with human identifications, we decided to conduct a visual inspection on a small subset of the spectra.

We first train a model without Occam's razor and select at random from this model 239 putative large DLAs with $N_{\rm HI} > 10^{22}\,{\rm cm}^{-2}$ and 243 putative small DLAs with $10^{20} \leqslant N_{\rm HI} < 10^{21}\,{\rm cm}^{-2}$. We visually inspect each spectrum and compute the model posteriors with a range of strengths for Occam's razor, $N = \{1, 10, 100, 1\,000, 30\,000\}$. We then treat each spectrum as a multiple-choice problem: if we think the model posterior of a given Occam's razor describes the given spectrum well, then we record one vote for this value of Occam's razor. Multiple selections are allowed for each spectrum as the model posteriors are often very close. After collecting votes, the winning value of Occam's razor was $N = 1\,000$, a ten times reduction from our earlier value.

For quasar spectra with $z_{\rm QSO} > 2.5$ there are enough data points in the Lyman-α range that the strength of Occam's razor has a small effect. We will discuss the effect of Occam's razor in Section 4. We suggest incorporating variations due to Occam's razor into the uncertainty in population statistics for conservative usage.

2.5 Summary of the modifications

Here we summarise the modifications we made in this work, comparing to the model of Ho et al. (2020):

(i) Our training set is SDSS DR12 quasar spectra with DLAs detected by Ho et al. (2020) removed. We considered a DLA to be detected if the posterior probability of a spectrum containing a DLA is larger than 0.9, $P(\mathcal{M}_{\rm DLA} \mid \mathcal{D}) > 0.9$.

(ii) The wavelength range modelled goes from $\lambda_{\rm rest} =$ 850.75 Å to $\lambda_{\rm rest} = 1\,420.75$ Å.

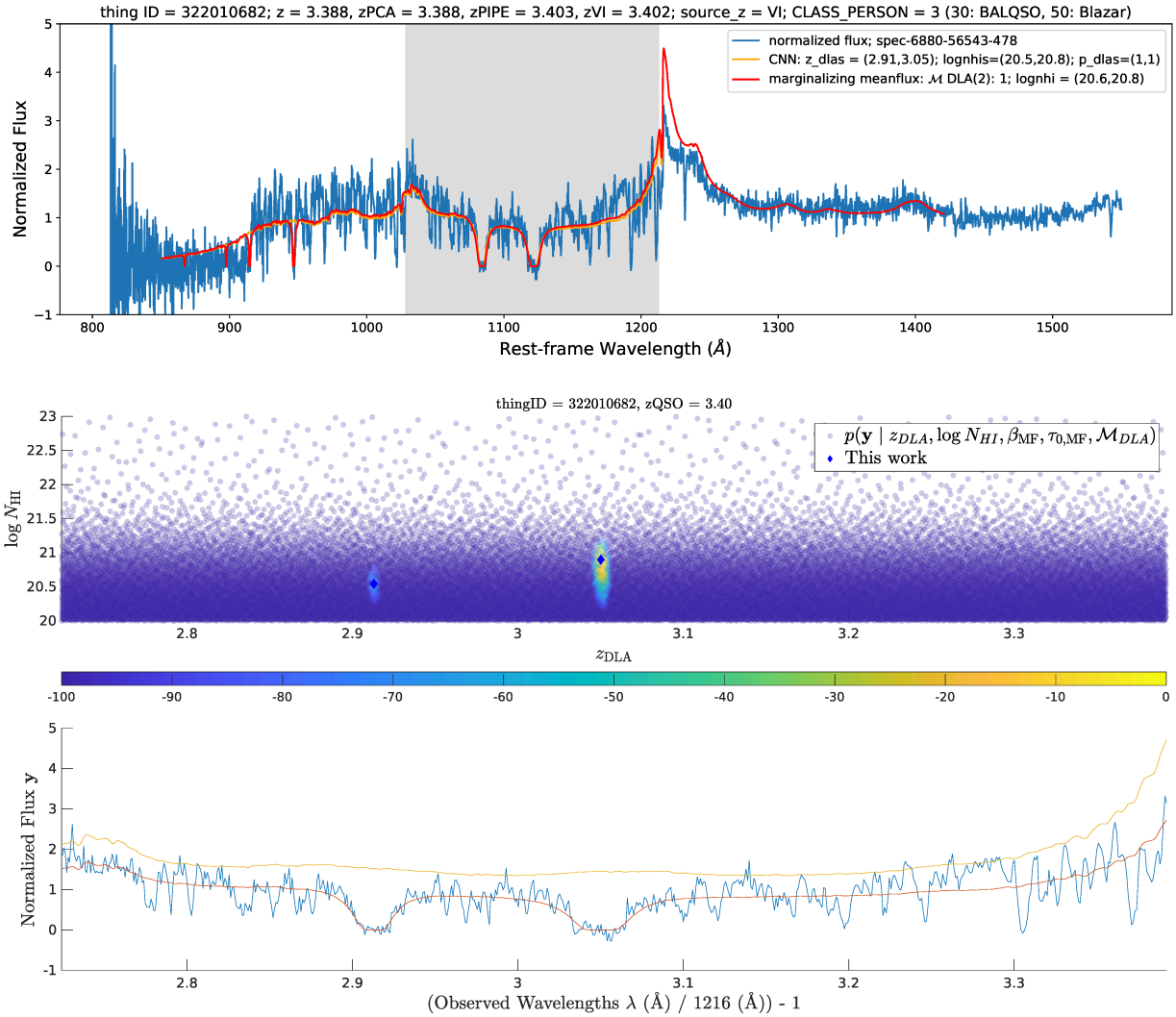

Figure 3. An example of a spectrum with distinct DLA features. **(Top):** The normalised observed spectrum in rest-frame wavelengths (blue) with the GP model (red) and the detection from the CNN model reported in DR16Q (orange). The title shows a series of column values in SDSS DR16Q catalogue, including SDSS identifier, best available redshift, PCA redshift, SDSS pipeline redshift, redshift from visual inspection, source for the best available redshift, and object classification from visual inspection (0: not inspected; 1: star; 3: quasar; 4: galaxy; 30: BAL quasar; 50: Blazar(?)). Shaded area (grey) shows the sampling range of z_{DLA}, which is from Ly$\beta + 3\,000\,\mathrm{km\,s^{-1}}$ to $z_{\mathrm{QSO}} - 3\,000\,\mathrm{km\,s^{-1}}$. The legend shows the spectrum is from spec-6880-56543-478 (spec-`plate-mjd-fiber_id`). The CNN model (orange) detected two DLAs, with redshifts of $z_{\mathrm{DLA}} = 2.91, 3.05$ and column densities of $\log_{10} N_{\mathrm{HI}} = 20.5, 20.8$, at DLA confidence = 1 for each DLA. Our GP model (red) also detected two DLAs with the model posterior $p(\mathcal{M}_{\mathrm{DLA\,(2)}} \mid \mathcal{D}) = 1$ and column densities $\log_{10} N_{\mathrm{HI}} = 20.6, 20.8$. **(Middle):** The sample likelihoods of detecting DLAs in the parameter space, $\theta \in (z_{\mathrm{DLA}}, \log_{10} N_{\mathrm{HI}})$. Colour bar shows the normalised log likelihoods, $\log p(\boldsymbol{y} \mid z_{\mathrm{DLA}}, \log_{10} N_{\mathrm{HI}}, \tau_{0,\mathrm{MF}}, \beta_{\mathrm{MF}}, z_{\mathrm{QSO}}, \mathcal{M}_{\mathrm{DLA}})$, with the maximum log likelihood to be zero. We also show the maximum a posteriori estimates of DLAs in the blue squares. The posterior distribution sharply peaks at the parameter space, indicating the detection of these DLAs in high confidence. **(Bottom):** The observed flux (blue) as a function of absorber redshifts with the GP model (red) and the GP model before the meanflux suppression (yellow). The position on the x-axis directly corresponds to the x-axis in the middle plot.

(iii) The effective optical depth prior, (τ_0, β), is updated from Kim et al. (2007) to Kamble et al. (2020).

(iv) The uncertainty in the mean flux suppression parameters, τ_0 and β, is marginalised while computing the model evidence.

The first modification gives us a training set size containing 89,408 spectra without DLAs. The larger training set better learns the covariance structure of quasar emission lines, which allows the second modification: expanding the model to cover the Lyman break and SiIV line. The expansion enables the model to use the metal lines to constrain the correlations of the emission lines in the Lyman-α forest. When using the previous modelling range, [911.75Å, 1215.75Å], we found that we often detected DLAs with high N_{HI} in the red end of the spectrum, where the code inserts a DLA at the quasar redshift to compensate for an oddly shaped Lyman-α emission line. This was possible because when we cut the spectrum at a rest frame wavelength 1215.75Å, half of the damping wings were removed, allowing for more model freedom and dubious N_{HI} estimation.

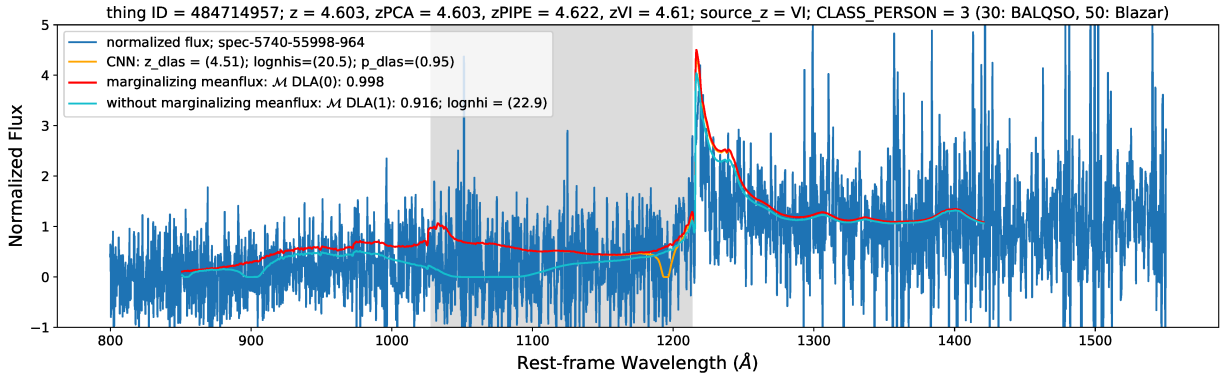

Figure 4. An example of a noisy spectrum with an uncertain meanflux. The normalised observed spectrum in rest-frame wavelengths (blue) with the GP model (red) and the detection from the CNN model reported in DR16Q (orange). We also plot the result without marginalising the uncertainty of meanflux prior (cyan). Shaded area (grey) shows the sampling range of $z_{\rm DLA}$, which is from Lyβ + $3\,000\,{\rm km\,s^{-1}}$ to $z_{\rm QSO} - 3\,000\,{\rm km\,s^{-1}}$. Our proposed model (red) indicates no DLA in the spectrum, with the null model posterior $p(\mathcal{M}_{\neg{\rm DLA}} \mid \mathcal{D}) = 0.998$. On the other hand, if our model ignores the uncertainty of $(\beta_{\rm MF}, \tau_{0,{\rm MF}})$, it would falsely detect a DLA with $p(\mathcal{M}_{\rm DLA} \mid \mathcal{D}) = 0.916$ with $\log_{10} N_{\rm HI} = 22.9$ (cyan). When marginalising over the uncertainty in effective optical depth, our proposed model (red) avoids detecting a false-positive large DLA.

Third, to make the mean flux suppression prior for (τ_0, β) consistent with the DR12Q training set, we switched to the mean flux measurement based on BOSS DR12Q (Kamble et al. 2020). Our last modification is marginalising the uncertainty of Kamble et al. (2020)'s parameters while marginalising the DLA parameters.

To compute the statistical properties of DLAs, we need to convert the posterior distribution of a DLA in each spectrum into the expected number of DLAs per redshift or column density bin, for which we use the method described in Bird et al. (2017). We briefly summarise the modelling decisions we made to produce the DLA samples in the result section:

- Search range: from Lyman-β to Lyman-α.[2].
- Maximum number of DLAs: three.
- Maximum $z_{\rm QSO}$: quasar redshifts < 7.
- DLA redshift $2 < z_{\rm DLA} < 5$.

3 RESULTS

3.1 Column density distribution function

Figure 5 shows the CDDF we estimate from DR16Q spectra. In the following sections, the CDDF is computed for $N_{\rm HI} \in [10^{20}, 10^{23}]$, while the DLA incidence rate ${\rm d}N/{\rm d}X$ and the total HI density in DLAs $\Omega_{\rm DLA}$ are computed for $N_{\rm HI} \in [10^{20.3}, 10^{23}]$. Ho20 refers to Ho et al. (2020), a DR12 DLA catalogue that used a modified GP model from Garnett et al. (2017).

The CDDF is a histogram of column densities normalised by the effective spectral path that could contain DLAs. We count all spectral path with an absorber with $z_{\rm DLA} < 5$. Error bars denote the 68% confidence limits, and the grey band represents the 95% confidence limits. Note that the uncertainties here are the statistical uncertainties associated with the GP model. They do not include uncertainty due to potential systematics. Section 4 will describe how possible systematics would affect the CDDF.

As shown in Figure 5, we observe non-zero column density until $3 \times 10^{22}\,{\rm cm^{-2}}$. Our DR16 measurement is mostly consistent with our previous DR12 measurement until $N_{\rm HI} \leqslant 9 \times 10^{21}$. For $N_{\rm HI} \geqslant 3 \times 10^{22}$, both our DR12 and DR16 measurement are consistent with zero at 95% confidence level, though there is one bin from DR16 not consistent with zero (see Table A3).

We also measure no turn over for the CDDF at the high column end, $N_{\rm HI} \sim 10^{21.5}\,{\rm cm^{-2}}$. It was suggested in Schaye (2001) that molecular hydrogen sets a maximum $N_{\rm HI}$ so that steepen the CDDF at the high end. The latest simulated CDDF from SIMBA (Hassan et al. 2020), which included molecular hydrogen formation in their star formation recipe, predicts no turn over at the high end, consistent with our measurements.

In Figure 6, we plot the CDDF with different Occam's razor strengths. When the Occam's razor strength is weak ($N = 30$), model selection will find DLAs even though the SNR is low, so we get more absorbers at both high and low column density ends. On the other hand, if the razor strength is strong ($N = 30\,000$), model selection will prefer to avoid finding DLAs at low SNR spectra, which results in a decrease.

However, in general, in Figure 6, we observe the razor strength only marginally affects the CDDF. Thus the small tension at the low end, $N_{\rm HI} \in [10^{20}, 10^{20.3}]$, between our CDDF and N12 is more likely due to other reasons than Occam's razor.

We show the redshift evolution of the CDDF in Figure 7. The downward pointing symbols indicate the 68% upper confidence limit when the data is consistent with zero at 68% confidence limits. As we can anticipate, for high-

[2] For the CDDF in Ho et al. (2020), we used a sampling range from Lyβ + $3\,000\,{\rm km\,s^{-1}}$ to Lyα − $30\,000\,{\rm km\,s^{-1}}$ to avoid finding DLAs in the proximity zone. Here, we instead use Lyβ + $3\,000\,{\rm km\,s^{-1}}$ to Lyα − $3\,000\,{\rm km\,s^{-1}}$. This has a very moderate effect on our results, however, we provide a check of systematics due to removing DLAs near to the quasar redshift in Section 4.2.

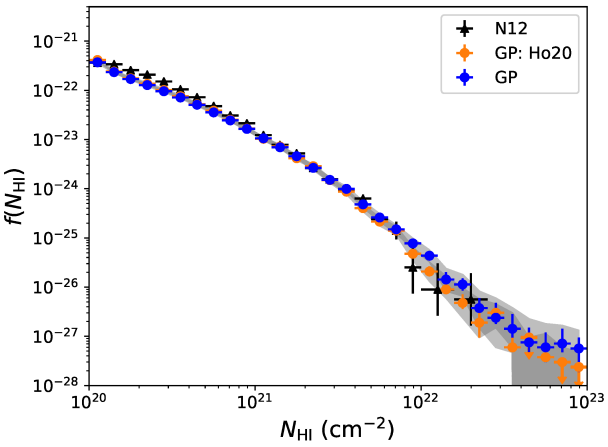

Figure 5. The CDDF, integrated over all $z < 5$ spectral path, derived from SDSS DR16Q spectra with our proposed Gaussian process models (GP; blue). The CDDF measurements from Ho et al. (2020) (GP: Ho20; orange) are plotted as a comparison. Error bars show the 68% confidence limits, while grey areas show the 95% confidence limits. Black dots are from Noterdaeme et al. (2012) (N12).

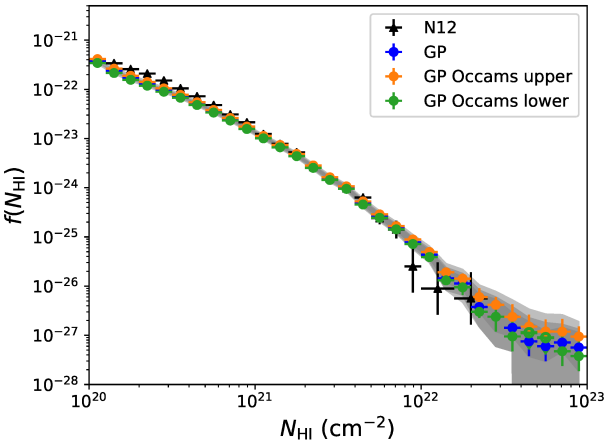

Figure 6. The CDDFs with different Occam's razor strengths, which discussed in Section 2.4. Occam's upper (orange) represents $N = 30\,000$ while Occam's lower (green) represents $N = 30$. We present our main result (GP; blue) with an optimal strength $N = 1\,000$, which we selected from visually inspecting a subset of the dataset. Note that the difference between different Occam's strengths is well within 95% confidence limits. Black dots are from Noterdaeme et al. (2012) (N12).

redshift quasars with $z_{\rm QSO} > 4$, since the flux is highly absorbed, we detect DLAs with larger uncertainties, and the number of large DLAs is consistent with zero.

In both Bird et al. (2017) and Ho et al. (2020), we found that the CDDF is getting shallower at $z > 4$. However, given our detection for $N_{\rm HI} > 4 \times 10^{21}$ at $z > 4$ is highly uncertain and consistent with zero detection, this trend is not significant in our current dataset. Instead, the detection of DLAs with $N_{\rm HI} < 4 \times 10^{21}$ at $z > 4$ is consistent with the measurements at $z \in [2.5, 4]$. We find no strong evidence for an evolution of the slope of the CDDF at $z > 4$.

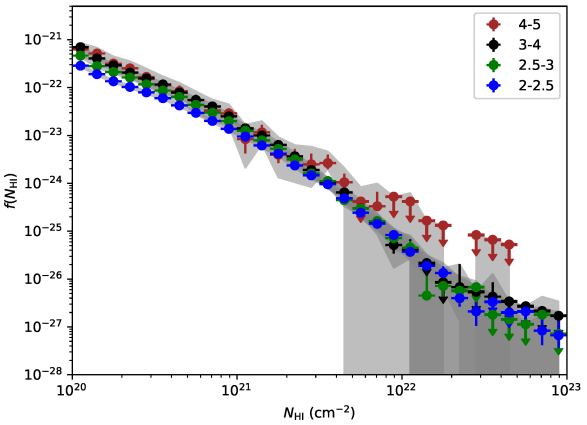

Figure 7. The CDDF derived from DLAs in a variety of redshift bins. Labels show the redshift bins in used. We show 68% confidence limits in error bars and 95% confidence limits in grey areas. If the bin is consistent with no detection at 68% limits, we show a down-pointing arrow indicating the 68% confidence upper limit.

One possible reason why we found the CDDF was shallower at $z > 4$ in Bird et al. (2017) and Ho et al. (2020) is absorption due to the Lyman-α forest. When the spectrum is highly absorbed, there is a degeneracy between a large DLA and the forest's absorption. In Bird et al. (2017), we did not model the GP mean as a function of effective optical depth, so it is possible the model was trying to use DLAs to compensate the excess absorption due to the forest, which results in a shallower CDDF at $z > 4$. In Ho et al. (2020), the slope of the CDDF is less shallow at $z > 4$, as we modelled the effective optical depth into our GP mean. In this work, we integrated out the measurement uncertainty of the mean flux, and the slope is almost indistinguishable from the CDDF at $z \in [2.5, 4]$. This may indicate that, to understand the DLAs at $z > 4$ better, we need a better measurement for the effective optical depth at $z > 4$.

One interesting feature in Figure 7 is the drop in the amplitude of the CDDF at $z \in [2, 2.5]$. As we will discuss in Section 3.2, the drop of CDDF at the low redshifts also shows in the DLA incident rate, $\mathrm{d}N/\mathrm{d}X$. We will discuss this in more detail in the next section.

3.2 Redshift evolution of DLAs

Figure 8 shows the incident rate of DLAs, $\mathrm{d}N/\mathrm{d}X$, as a function of absorber redshift. Our results are consistent with Prochaska & Wolfe (2009) and Ho et al. (2020) and are slightly lower than Noterdaeme et al. (2012). $\mathrm{d}N/\mathrm{d}X$ is sensitive to the weaker DLAs, so the difference between Noterdaeme et al. (2012) and Prochaska & Wolfe (2009) is likely due to the false positive rate.

Prochaska & Wolfe (2009) performed a visually-guided Voigt profile fitting on SDSS-DR5. Though their sample size is smaller, with the help of the human eye, their method is likely less prone to false positives than the automated template fitting used in Noterdaeme et al. (2012). This difference may also explain the drop in amplitude of the CDDF at $z \in [2, 2.5]$. Prochaska & Wolfe (2009) and Noterdaeme et al.

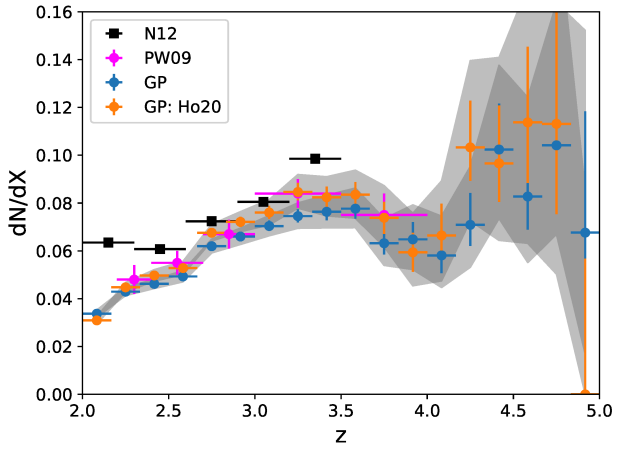

Figure 8. The incident rate of DLAs as a function of redshift, integrated over $\log_{10} N_{\rm HI} > 20.3$ spectra from our catalogue (GP; blue). We also plot the line densities from Noterdaeme et al. (2012) (N12; black) and Prochaska & Wolfe (2009) (PW09; pink), and Ho et al. (2020) (GP: Ho20; orange) as comparisons.

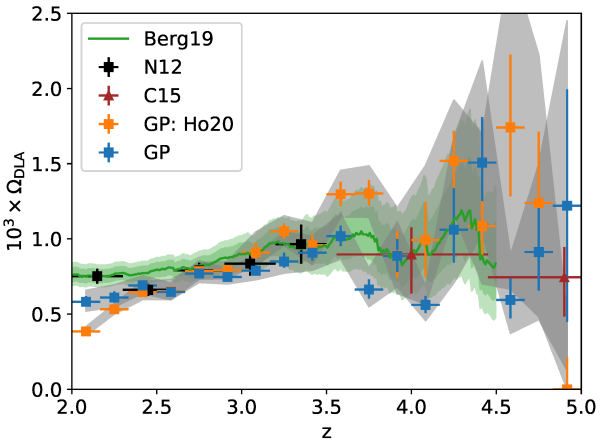

Figure 9. The total HI density in DLAs, integrated over DLAs with $\log_{10} N_{\rm HI} > 20.3$ in our catalogue (GP; blue). For comparison, we plot the measurements from Berg et al. (2019) (Berg19; green line and shaded area), Noterdaeme et al. (2012) (N12; black), Crighton et al. (2015) (C15; red), and Ho et al. (2020) (GP: Ho20; orange).

(2012) have a larger discrepancy at $z \in [2, 2.5]$, and Noterdaeme et al. (2012) may overestimate the weak absorbers at this redshift range where the spectra are short. Our measurement is consistent with Prochaska & Wolfe (2009), which implies that we detect fewer weak absorbers in low redshift bins and explains the small tension in the CDDF at low-$N_{\rm HI}$.

One noticeable feature in $\mathrm{d}N/\mathrm{d}X$, which we have not discussed before, is the decrease of the line density from $z = 3.5$ to $z = 4.0$ and another increase at $z > 4.0$. This feature is also shown in our Ho20 measurement. The drop of $\mathrm{d}N/\mathrm{d}X$ at $z \in [3.5, 4]$ is consistent with Prochaska & Wolfe (2009) at 95% confidence limits. One interesting question is whether the increase from $z \in [4.0, 4.5]$ is real. The measurements at $z > 4$ still have large error bars, so it is hard to say whether $\mathrm{d}N/\mathrm{d}X$ at $z > 4$ is an increase or a flat line. More data, especially with high SNR, are needed to determine the trend of line density at $z > 4$.

In Figure 9, we show the total HI density in DLAs in terms of cosmic density. Our results are mostly consistent with Noterdaeme et al. (2012) at $z \in [2.5, 3.5]$. At higher redshift bins, $z > 3.5$, our measurements are consistent with Prochaska & Wolfe (2009) and Crighton et al. (2015). Crighton et al. (2015) used high signal-to-noise spectra from a smaller survey, so they have larger error bars. Comparing to Ho20, our current $\Omega_{\rm DLA}$ has more mass at low-redshifts ($z_{\rm DLA} \sim 2$) and less mass at $z_{\rm DLA} \in [3.5, 4]$. The trend of $\Omega_{\rm DLA}$ in DR16 is shallower than Ho20.

We also plot the $\Omega_{\rm DLA}$ measured by Berg et al. (2019) in Figure 9. We see our DR16 measurement is consistent with Berg et al. (2019) even at $z > 3.5$. There is a slight tension at $z < 2.5$, which may be because some low-redshift spectra are too short and noisy to measure column density confidently using our model. SDSS spectra with $z_{\rm QSO} < 2.5$ only covers a region from the Lyα to Lyβ or shorter. When the signal-to-noise is low, it is difficult to identify DLAs even using human eyes. As shown in Fig 10, a different selection of Occam's razor could moderately affect the two bins with $z < 2.33$. The strength of Occam's penalty corresponds to a prior belief in detecting a DLA in a short and noisy spectrum, as discussed in Section 2.4.

Note that, from Figure 7 we see there are no solid detections for $N_{\rm HI} > 3 \times 10^{21}$ DLAs at $z > 4$. In Bird et al. (2017), $\Omega_{\rm DLA}$ was skewed towards high values at $z > 4$ even without real detections of large DLAs. Our result in Figure 9 does not have this issue. This may indicate our proposed method of integrating out the uncertainty on meanflux measurement helps us avoid the forest biasing the posterior density of $N_{\rm HI}$ towards the high end.

In general, we observe an increase of $\Omega_{\rm DLA}$ from $z = 2$ to $z = 3.5$, and a slight decrease from $z = 3.5$ to $z = 4$. For $z_{\rm QSO} > 4$, the measurement error is larger and less correlated between redshift bins, as in Ho20. This is reasonable given the lower quasar number density at high redshift.

Figure 10 shows the line density and $\Omega_{\rm DLA}$ with various Occam's razor strengths. As expected, the razor strengths only moderately affect the statistics of low redshift spectra. For $\mathrm{d}N/\mathrm{d}X$, our results are consistent with Prochaska & Wolfe (2009), even with the weakest razor. N12 still detects somewhat more weak DLAs than we do, even though we only apply a small penalty for these short and noisy spectra.

4 CHECKS FOR SYSTEMATICS

4.1 Effect of signal to noise ratios

Figure 11 and Figure 12 show the abundance of DLAs from subsets of our catalogue with various signal-to-noise cuts (SNR), SNR > 2 and > 4. We define our SNR as the median of the ratio between the flux and the instrumental noise within the quasar spectrum redwards of the Lyman-α emission peak. This specific choice is to avoid introducing correlations between the detected DLAs and the SNR. With this definition, 80% of the quasar spectra have SNR > 2, and 46% of the spectra have SNR > 4.

We verify that, in Figure 11, the CDDF is not sensitive

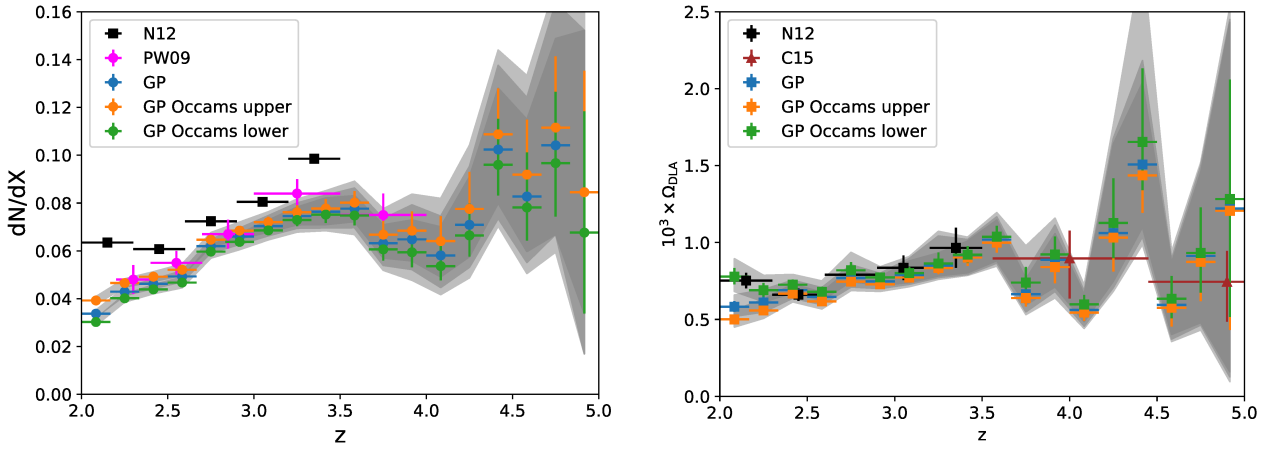

Figure 10. The line density **(left)** and $\Omega_{\rm DLA}$ **(right)** in DLAs as a function of redshifts with different Occam's razor strengths. Occam's upper (orange) represents $N = 30\,000$ while Occam's lower (green) represents $N = 30$. The main result (GP; blue) is computed with $N = 1\,000$.

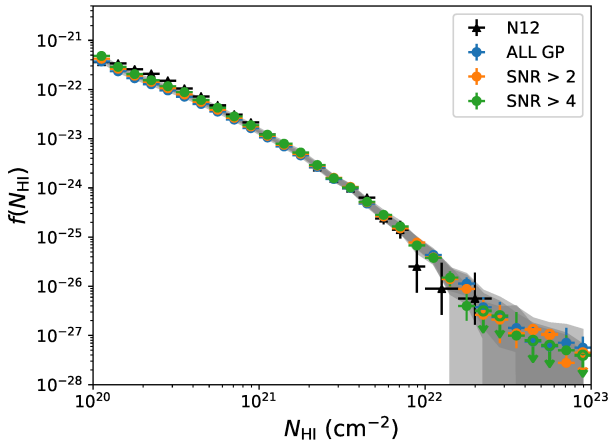

Figure 11. The CDDF of DLAs for a subset of samples with different minimal SNRs. SNR > 2 (orange) excludes 20% of the noisiest spectra, and SNR > 4 (green) excludes 54% of the spectra. 68% confidence limits are drawn as error bars, while 95% confidence limits are shown as a grey filled band.

to the SNR when $N_{\rm HI} < 10^{22}\,{\rm cm}^{-2}$. However, we note that the highest non-zero column density at 95% confident limits changed from $N_{\rm HI} < 3 \times 10^{22}\,{\rm cm}^{-2}$ to $N_{\rm HI} \lesssim 10^{22}\,{\rm cm}^{-2}$ for samples with SNR > 4. This is likely because there are too few high column density absorbers to constrain the CDDF sufficiently at the high end in the smaller high SNR sample.

We find that our $\Omega_{\rm DLA}$ measurement exhibits no systematic correlation with the SNR cuts. We notice a dependence of SNR on $\mathrm{d}N/\mathrm{d}X$ at $z \in [2.0, 2.5]$, which is due to the difficulty of finding DLAs in short and noisy spectra. As discussed in Section 2.4, it is hard to find features in these spectra, and the observing window cannot fully cover a high-$N_{\rm HI}$ DLA profile with damping wings. To secure our samples' purity, we use an Occam's razor penalty which may also introduce this SNR dependence at $z \in [2, 2.5]$.

As mentioned in Krogager et al. (2019), the colour and magnitude criteria used in SDSS for quasar target selection is biased against dusty DLAs, which harbour a certain amount of cold neutral gas. Krogager et al. (2019) showed that in SDSS DR7 this caused $\Omega_{\rm DLA}$ to be underestimated by $10-50\%$ at $z \sim 3$. Also, redder quasars containing metal rich dusty DLAs will have lower SNR in the blue part of the spectrum and thus may be excluded from the sample of Noterdaeme et al. (2009), who enforced CNR > 3. This effect is likely to be substantially reduced in our sample, if present at all, as we use all quasars irrespective of SNR. We also define SNR using the region redwards of the Lyα emission peak specifically to avoid this kind of selection effect, and we are using DR16, which has a different and more complex selection function. More quantitatively, the XQ-100 targets in Berg et al. (2019) use only radio-selected quasars, or quasars previously found by other techniques, and so avoids any SDSS colour selection bias. Figure 9 shows that our $\Omega_{\rm DLA}$ mostly agrees with Berg et al. (2019), implying that colour effects in our sample are smaller than those in DR7.

4.2 Effect of quasar redshifts

In Figure 13, we test our measured $\mathrm{d}N/\mathrm{d}X$ with different quasar redshift bins. In a perfect scenario without systematics, we expect that the absorber properties be uncorrelated with the background quasars, as they are widely separated in physical space. However, Figure 13, shows some residual correlation between absorber properties and the redshifts of the background quasars for DLAs in spectra with $z_{\rm QSO} < 3$.

In Figure 14, we have investigated removing the sampling range near the quasar redshift, $|z_{\rm QSO} - z| < 30\,000\,{\rm km\,s}^{-1}$. We found removing the putative absorbers near the Lyman-α emission is sufficient to remove the correlation between quasar redshifts and DLA properties at $z < 3$. A small tension still exists for the $z = 2$ bin within $2.5 > z_{\rm QSO} > 2.0$ for $\mathrm{d}N/\mathrm{d}X$, which may be due to the effect discussed above for SNR, as these very short spectra are often also noisy.

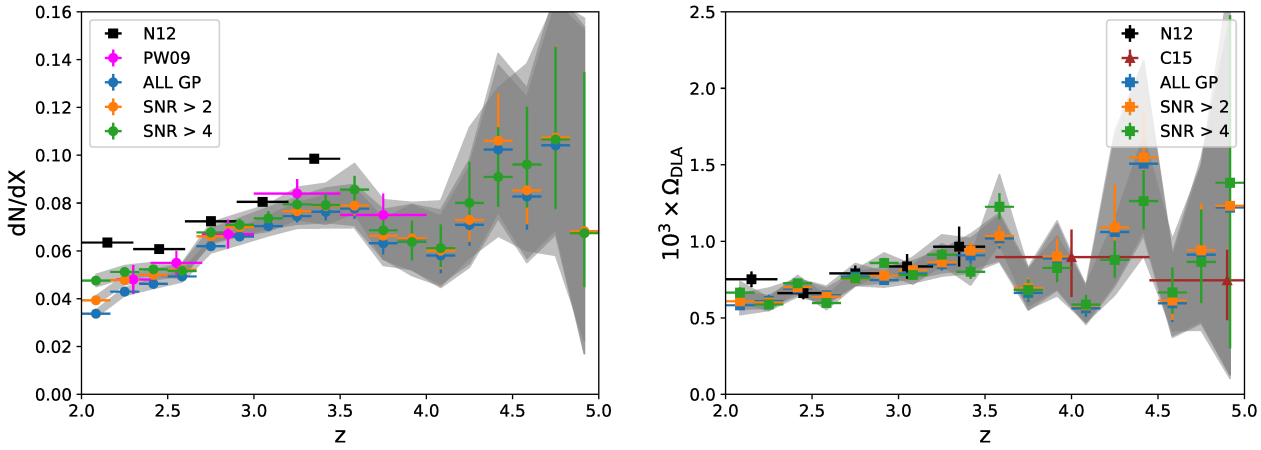

Figure 12. The line density **(left)** and total N_{HI} mass **(right)** in DLAs as a function of absorber redshift from subsets of samples with different minimal SNRs. SNR > 2 (orange) excludes 20% of the noisiest spectra, and SNR > 4 (green) excludes 54% of the spectra.

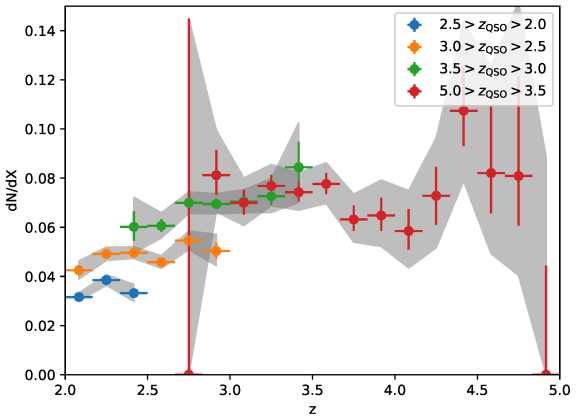

Figure 13. The redshift evolution of the incident rate of DLAs, cutting with different quasar redshift intervals. Any correlation between the absorber properties and the background quasars redshifts might indicate systematics.

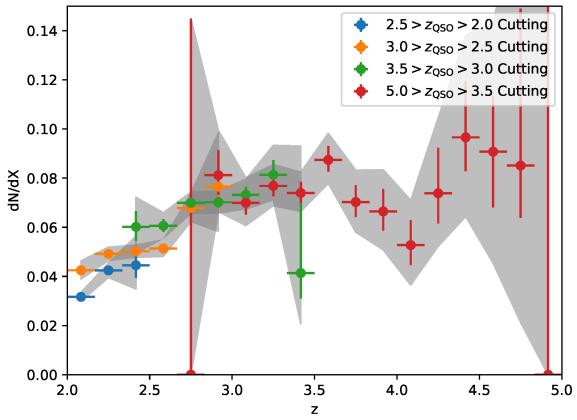

Figure 14. The redshift evolution of the incident rate of DLAs , cutting with different quasar redshift intervals. Unlike Figure 13, we remove the putative absorbers near the Lyman-α emission line with $|z_{QSO} - z_{DLA}| < 30\,000\,\mathrm{km\,s^{-1}}$.

4.3 Additional noise test

To understand the implication of applying Occam's razor to the model posteriors, we conduct a test based on adding noise to a DLA spectrum. We choose a quasar spectrum that we are very confident contains a DLA and add additional Gaussian noise with zero mean and standard deviation σ to the flux and noise variance.

We then examine changes in the DLA model posterior $p(\{\mathcal{M}_{DLA}\} \mid \mathcal{D})$. This test will mimic the effect of SNR on the model's ability to detect the underlying DLAs. For Occam's razor $N = 30\,000$, the model posterior is $p(\{\mathcal{M}_{DLA}\} \mid \mathcal{D}) \simeq 0.9$ for $\sigma \leqslant 1.5$, which corresponds to SNR $\simeq 0.9$. On the other hand, for a model without Occam's razor, the model posterior is $p(\{\mathcal{M}_{DLA}\} \mid \mathcal{D}) \simeq 0.9$ for $\sigma \leqslant 3$, which means SNR $\simeq 0.5$. A strong Occam's razor thus introduces false negatives in very noisy spectra. However, by visually inspecting the flux with $\sigma = 3$ we determined that it is almost impossible for humans to identify the underlying DLA. Therefore, we choose to follow the value ($N = 1\,000$) we determined in Section 2.4.

We were unable to quantify the number of false positives, as our simple assumption of Gaussian noise rarely produces correlated structures that resemble DLAs. In practice, false positives are likely caused by oscillatory structure embedded in the noise, present when the SNR is extremely low.

5 RESULTS WITH DLAS IN THE LYMAN β REGION

We have shown the CDDF, $\mathrm{d}N/\mathrm{d}X$, and Ω_{DLA} of our GP model in Section 3. In this section, instead of using a sampling range from Lyβ to Lyα, we only compute the population statistics of DLAs detected *within the Lyβ forest region.* We set the sampling range to be Lyman limit $+30\,000\,\mathrm{km\,s^{-1}}$ to Lyman-β. We cut off a wider velocity width at the blue

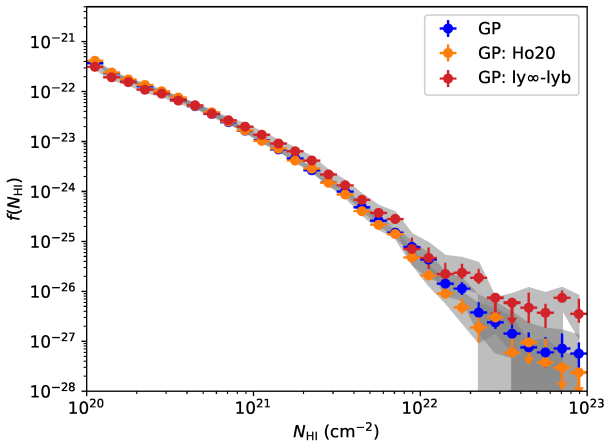

Figure 15. Comparing the CDDFs between the sampling range from Lyβ-Lyα (Blue; GP) and **Ly∞-Lyβ** (Red; GP: ly∞-lyb). The error bars are 68% confidence limits, and the shaded areas are 95% confidence limits. Ho et al. (2020) (GP: Ho20; Orange) also used a sampling range from Lyβ-Lyα.

end to avoid counting DLAs detected right on the edge of the Lyman break.

Figure 15 shows the CDDF for DLAs in the Lyman β region. As we can see from the figure, it is mostly consistent with the CDDF from Lyβ-Lyα for $N_{\rm HI} < 10^{21}$, and it starts to diverge for $N_{\rm HI} > 20^{22}$. We visually inspected those spectra and found that they are mostly due to fitting large DLAs on the spectra's noisy left edges. This may indicate that additional regularisation is still needed to avoid spurious detections at the blue end of high redshift spectra. In particular, if the redshift measurement is slightly inaccurate, parts of the Lyman break move into our modelling window.

We also show the dN/dX and $\Omega_{\rm DLA}$ for DLAs in the Lyman-β forest region in Figure 16. dN/dX in the Lyman-β region is broadly consistent with other measurements, with the detection consistent with zero at $z_{\rm DLA} > 4$.

For $\Omega_{\rm DLA}$, in the right panel of Figure 16, we observe our measurement is biased high and highly uncertain for $z_{\rm DLA} > 3.5$. This may be because our current model can only poorly estimate the column density from the Lyman-β region from high-redshift quasar spectra, perhaps due to the high level of absorption from the Lyman-β and Lyman-α forests at these redshifts. Alternatively, it could again reflect that the mean flux measure is not certain at these redshifts, so the degeneracy between large DLAs and the effective Lyman-α/Lyman-β absorption is not fully broken by sampling ($\tau_{0,\rm MF}, \beta_{\rm MF}$).

6 COMPARISON TO THE CNN MODEL

SDSS DR16Q includes DLA measurements using the convolutional neural network (CNN) model of Parks et al. (2018). The DLAs from the CNN model are recorded as `CONF_DLA`, `Z_DLA`, and `NHI_DLA` columns in the DR16Q catalogue.[3]

[3] This column is the log column density of the given DLA.

To compare our model and the CNN model, we restrict the $z_{\rm DLA}$ sampling range of the CNN DLAs to be the same as our GP DLAs. Table 2 shows the confusion matrix. On the existence of DLAs, which means the binary classification of having at least one DLA or no DLA, the GP model is $\sim$ 94.8% in agreement with the CNN model. If we only consider only spectra with SNR > 6, the rate of agreement climbs to $\sim$ 96.5%.

We have also checked the CDDF of the CNN DLAs, as shown in Figure 17. The sampling range is restricted to be the same as ours, and we only count the DLAs with `CONF_DLA` larger than 0.98. The CDDF of the CNN model under-detects DLAs with $N_{\rm HI} > 7 \times 10^{20}$, compared to N12. We have discussed this issue in Figure 19 of Ho et al. (2020). The CDDF of the CNN model in the DR16Q catalogue shows improvements in detecting more high column density systems comparing to Parks et al. (2018), but it is still an order of magnitude lower than N12 for $N_{\rm HI} > 2 \times 10^{21}$. Thus the lack of high column density systems in the CNN DLAs, as identified in Ho et al. (2020), is still present in the latest catalogue.

The dN/dX of the CNN model, in contrast, mostly agree with our GP measurements. Bins with $z > 4.5$ are even consistent at the 1-σ level. Since dN/dX is sensitive to low column density systems, it shows these two codes find consistent small DLAs, but differ in their column density estimates.

We compare DLAs detected by the CNN and GP codes on a spectrum-by-spectrum basis in Figure 18. As anticipated, the CNN and the GP code have a perfect agreement in $z_{\rm DLA}$, but the CNN predicts slightly lower $\log_{10} N_{\rm HI}$ than the GP code, consistent with the CDDF plot in Figure 17.

We visually inspected 319 quasar spectra, where the CNN code strongly disagrees with the GP code's detections. As expected, most cases are spectra with low SNRs, where even human experts will have difficulty identifying DLAs. Besides those low-SNR cases, in general, the CNN code has false negatives on DLAs overlapping with sub-DLAs or DLAs very close to each other. There are 24 out of 319 cases which show a clear pattern where the CNN missed the DLAs when multiple absorption systems are overlapping or nearby.[4] Some of these are ambiguous detections, but 9 out of 24 have apparent damping wings on the absorber.

We show two examples in Figure 19. The first one shows a sub-DLA intervening on the right of the DLA damping wings. Though the damping wings are disturbed by the sub-DLA, the pattern of a DLA is still visible. The second example shows two DLAs close to each other, but not close enough to overlap. We suspect these non-detections for the CNN code are due to the lack of training data for multiple absorption systems (sub-DLAs or DLAs) close to each other. Since these overlapping cases are rare in the real dataset, we think one might need to implement simulated DLAs/sub-DLAs to augment the CNN training set.

[4] We put the figures for these 24 spectra in here `http://tiny.cc/overlapping_dlas` for future investigators.

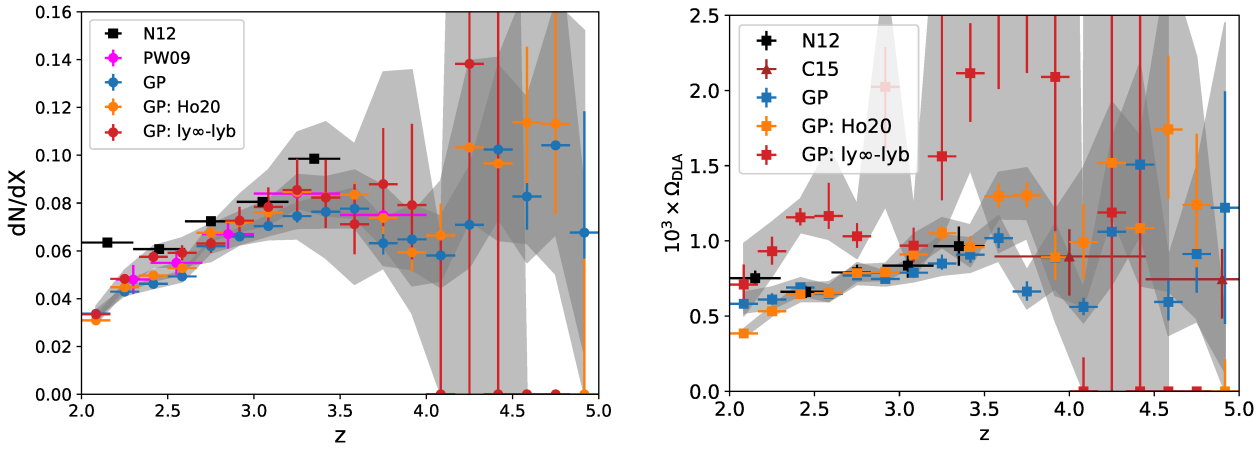

Figure 16. **(Left)** The comparison of $\mathrm{d}N/\mathrm{d}X$ with different sampling ranges, Lyβ-Lyα (blue) and Ly∞-Lyβ (red). Other plot settings are the same as Figure 8. **(Right)** The comparison of $\Omega_{\rm DLA}$ with difference sampling ranges, Lyβ-Lyα (blue) and Ly∞-Lyβ (red). Other plot settings are the same as Figure 9.

Table 2. The confusion matrix for multi-DLAs detections between the GP and the CNN model (Parks et al. 2018). Note we require both the model posteriors of our GP model and DLA confidence in Parks to be larger than 0.98. We also require $\log_{10} N_{\rm HI} > 20.3$. The maximum number of DLAs is fixed to three, and everything larger than three is considered three.

CNN GP with Multi-DLAs	0 DLA	1 DLA	2 DLAs	3 DLAs
0 DLA	**142759**	5686	93	2
1 DLA	2397	**8007**	208	1
2 DLAs	117	234	**333**	5
3 DLAs	8	6	11	**4**

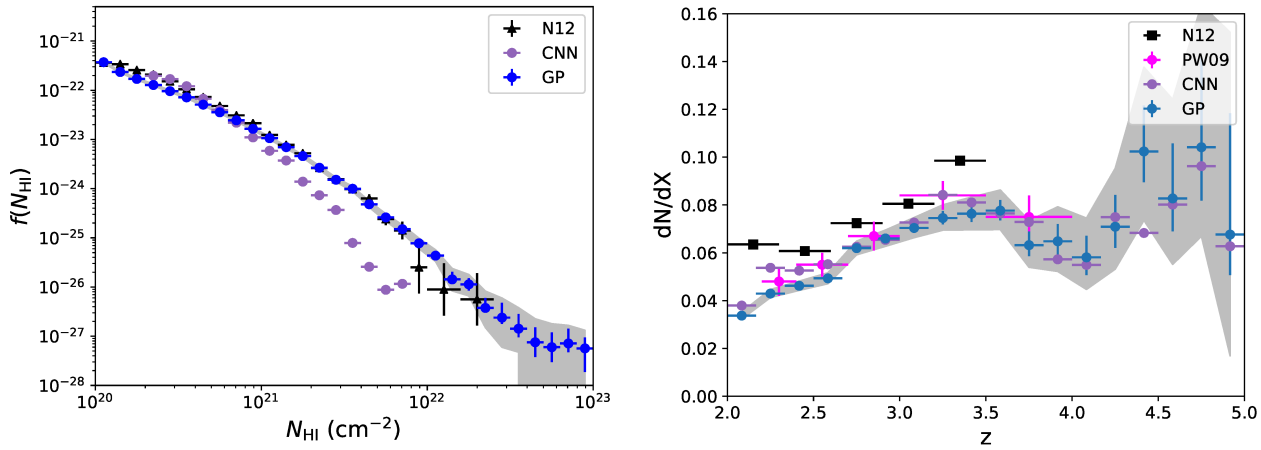

Figure 17. **(Left)** The CDDF of the DLAs detected by the CNN model presented in Parks et al. (2018). The $z_{\rm DLA}$ and $\log_{10} N_{\rm HI}$ values are taken from the SDSS DR16Q catalogue in column `Z_DLA` and `NHI_DLA`. We require the confidence of DLAs to be larger than 0.98 and set the search range of the CNN DLAs to be the same as our search range, which is Lyman-β $+3\,000\,\mathrm{km\,s^{-1}}$ to $z_{\rm QSO} - 3\,000\,\mathrm{km\,s^{-1}}$. **(Right)** The line density of the DLAs detected by the CNN model. All three measurements, GP, PW09, and CNN are consistent on the line density.

7 CONCLUSION

We have presented a new estimate of the abundance of DLAs from $z = 2$ to $z = 5$ and a DLA catalogue built from SDSS DR16Q spectra (Lyke et al. 2020) using our Gaussian process model Garnett et al. (2017); Ho et al. (2020). We verify our results are in good agreement with previous measurements from Noterdaeme et al. (2012), Prochaska & Wolfe (2009), and Crighton et al. (2015). We newly integrate out the uncertainty in the measured mean flux, which improves our modelling of DLA detection uncertainties for $z_{\rm QSO} > 4$ without biasing towards high $N_{\rm HI}$ detections.

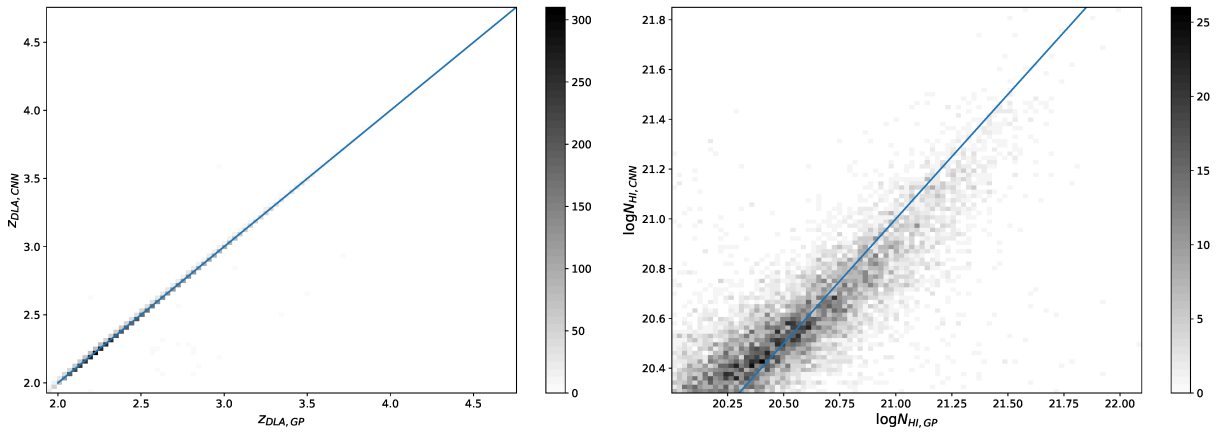

Figure 18. The 2D histograms for $z_{\rm DLA}$ **(left)** and $\log_{10} N_{\rm HI}$ **(right)** estimated by the GP code and the CNN. We use the maximum a posteriori (MAP) estimate for parameter estimation for the GP code. The colourbars indicate the number of DLAs within the bin. The blue line is a straight line that shows the diagonal line of the 2D histogram.

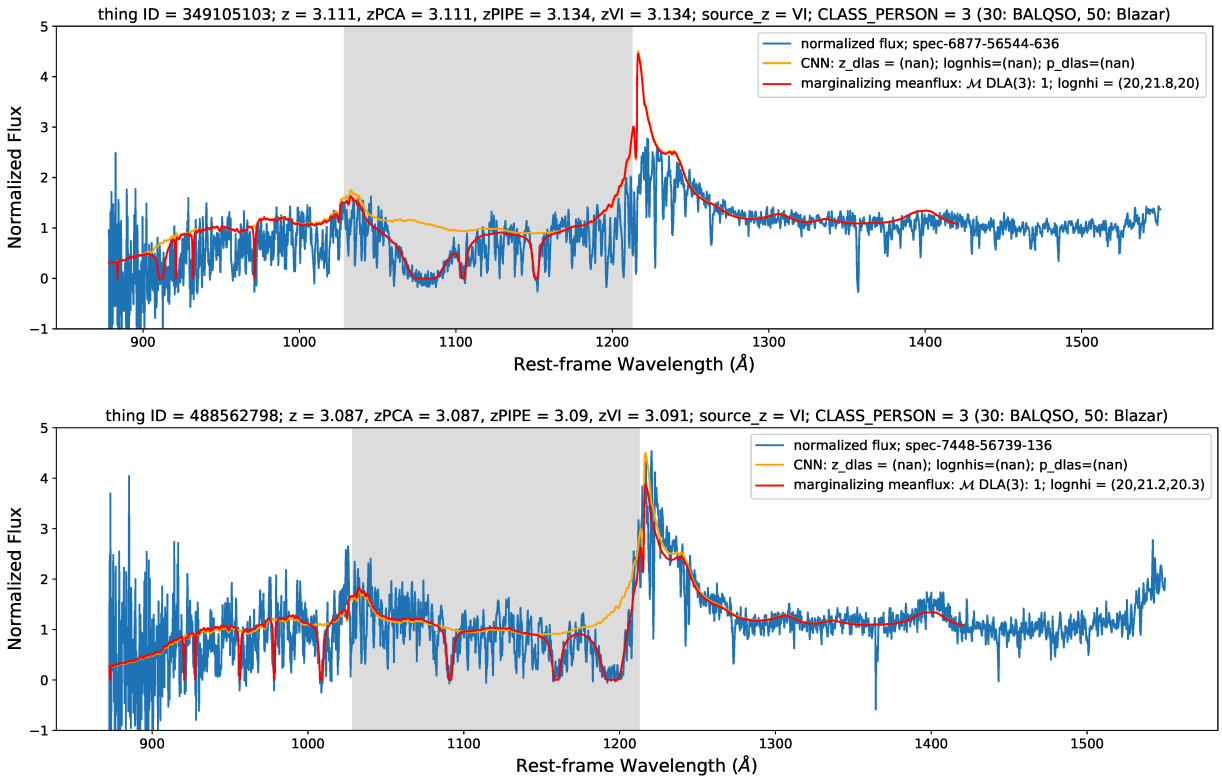

Figure 19. Examples showing **(top)** the case of a sub-DLA overlapping a DLA and **(bottom)** the case of a DLA near to another DLA. The red line indicates the GP code predictions, and we describe the $\log_{10} N_{\rm HI}$ in the legend. We intervene the DLAs from the CNN model in the DR16Q catalogue onto our null model in the orange line. Both spectra have high enough SNR: the upper one has SNR = 3.45 while the bottom one has SNR = 7.52. The damping wings and the Lyman-β absorption lines of the DLAs are visible in the plots.

We note, nevertheless, that there is a residual dependence on low-redshift spectra with $z_{\rm QSO} < 2.5$. This could be due to unmodelled systematics or simply because the low-redshift optical spectra are incomplete in the Lyman series range, so we can not securely detect DLAs in low $z_{\rm QSO}$. Incorporating spectra with shorter observed wavelengths could potentially verify these detections at $z_{\rm QSO} < 2.5$.

Our measurement shows the abundance of DLAs and neural hydrogen increases moderately over $2 < z < 4$, while the trend beyond $z = 4$ is unclear due to statistical uncertainties. Larger datasets and better mean flux measurements are needed to give more robust constraints for DLA detections at $z > 4$.

DATA AVAILABILITY

Our DLA catalogue is publicly available at `http://tiny.cc/gp_dla_dr16q`, including both MATLAB catalogue and JSON catalogue. A sub-DLA candidate catalogue is available in JSON format. README files are included to describe the data formats of both catalogues. The data files for DLA population statistics are also included, including CDDF, $\mathrm{d}N/\mathrm{d}X$, and Ω_{DLA} with or without SNR cuts. A tutorial for manipulating the MATLAB catalogue is publicly available at `https://github.com/jibanCat/gp_dla_detection_dr16q_public/tree/master/notebooks` as a notebook file. Our GP code is also publicly available at `https://github.com/jibanCat/gp_dla_detection_dr16q_public/`.

ACKNOWLEDGEMENTS

We thank the anonymous referee for the constructive comments and suggestions. We thank Reza Monadi and Bryan Scott for useful discussions and comments. SB was supported by NSF AST-1817256. RG was supported by the NSF under award numbers IIS–1939677, OAC–1940224, and IIS–1845434. SB and RG were supported by an Amazon.com Machine Learning Research Award, which also provided computing time. We hope the world can recover from the COVID pandemic soon.

REFERENCES

Alam S., et al., 2015, ApJS, 219, 12 (arXiv:1501.00963)
Alonso D., Colosimo J., Font-Ribera A., Slosar A., 2018, J. Cosmology Astropart. Phys., 2018, 053 (arXiv:1712.02738)
Berg T. A. M., et al., 2019, MNRAS, 488, 4356 (arXiv:1907.07703)
Bird S., Haehnelt M., Neeleman M., Genel S., Vogelsberger M., Hernquist L., 2015, MNRAS, 447, 1834
Bird S., Garnett R., Ho S., 2017, MNRAS, 466, 2111
Busca N., Balland C., 2018, arXiv e-prints, p. arXiv:1808.09955 (arXiv:1808.09955)
Carithers W., 2012, Published internally to SDSS
Cen R., 2012, ApJ, 748, 121
Chabanier S., et al., 2019, J. Cosmology Astropart. Phys., 2019, 017 (arXiv:1812.03554)
Crighton N. H. M., et al., 2015, MNRAS, 452, 217
Croft R. A. C., Weinberg D. H., Katz N., Hernquist L., 1998, ApJ, 495, 44 (arXiv:astro-ph/9708018)
Cuceu A., Font-Ribera A., Joachimi B., 2020, J. Cosmology Astropart. Phys., 2020, 035 (arXiv:2004.02761)
Dawson K. S., et al., 2013, AJ, 145, 10 (arXiv:1208.0022)
Dawson K. S., et al., 2016, AJ, 151, 44 (arXiv:1508.04473)
Eisenstein D. J., et al., 2011, AJ, 142, 72 (arXiv:1101.1529)
Fauber L., Ho M.-F., Bird S., Shelton C. R., Garnett R., Korde I., 2020, MNRAS, 498, 5227 (arXiv:2006.07343)
Font-Ribera A., et al., 2012, J. Cosmology Astropart. Phys., 2012, 059 (arXiv:1209.4596)
Fumagalli M., O'Meara J. M., Prochaska J. X., Worseck G., 2013, ApJ, 775, 78 (arXiv:1308.1101)
Gardner J. P., Katz N., Weinberg D. H., Hernquist L., 1997, ApJ, 486, 42
Garnett R., Ho S., Bird S., Schneider J., 2017, MNRAS, 472, 1850
Guo Z., Martini P., 2019, ApJ, 879, 72 (arXiv:1901.04506)
Haehnelt M. G., Steinmetz M., Rauch M., 1998, ApJ, 495, 647
Hassan S., Finlator K., Davé R., Churchill C. W., Prochaska J. X., 2020, MNRAS, 492, 2835 (arXiv:1910.07541)
Ho M.-F., Bird S., Garnett R., 2020, MNRAS, 496, 5436 (arXiv:2003.11036)
Iršič V., et al., 2017, MNRAS, 466, 4332 (arXiv:1702.01761)
Kamble V., Dawson K., du Mas des Bourboux H., Bautista J., Scheinder D. P., 2020, ApJ, 892, 70 (arXiv:1904.01110)
Kim T.-S., Bolton J. S., Viel M., Haehnelt M. G., Carswell R. F., 2007, MNRAS, 382, 1657
Krogager J.-K., Fynbo J. P. U., Møller P., Noterdaeme P., Heintz K. E., Pettini M., 2019, MNRAS, 486, 4377 (arXiv:1904.06966)
Lee K.-G., et al., 2013, AJ, 145, 69
Lin X., Cai Z., Li Y., Krolewski A., Ferraro S., 2020, arXiv e-prints, p. arXiv:2011.01234 (arXiv:2011.01234)
Lyke B. W., et al., 2020, ApJS, 250, 8 (arXiv:2007.09001)
McDonald P., Miralda-Escudé J., Rauch M., Sargent W. L. W., Barlow T. A., Cen R., Ostriker J. P., 2000, ApJ, 543, 1 (arXiv:astro-ph/9911196)
McDonald P., Seljak U., Cen R., Bode P., Ostriker J. P., 2005a, MNRAS, 360, 1471 (arXiv:astro-ph/0407378)
McDonald P., et al., 2005b, ApJ, 635, 761 (arXiv:astro-ph/0407377)
Noterdaeme P., Petitjean P., Ledoux C., Srianand R., 2009, A&A, 505, 1087 (arXiv:0908.1574)
Noterdaeme et al., 2012, A&A, 547, L1
Pâris, Isabelle et al., 2018, A&A, 613, A51
Parks D., Prochaska J. X., Dong S., Cai Z., 2018, MNRAS, 476, 1151
Pérez-Ràfols I., et al., 2018, MNRAS, 473, 3019 (arXiv:1709.00889)
Pontzen A., et al., 2008, MNRAS, 390, 1349
Prochaska J. X., Wolfe A. M., 1997, ApJ, 487, 73
Prochaska J. X., Wolfe A. M., 2009, ApJ, 696, 1543
Prochaska J. X., Herbert-Fort S., Wolfe A. M., 2005, ApJ, 635, 123
Rahmati A., Schaye J., 2014, MNRAS, 438, 529 (arXiv:1310.3317)
Rasmussen C. E., Williams C. K. I., 2005, Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning). The MIT Press
Rogers K. K., Bird S., Peiris H. V., Pontzen A., Font-Ribera A., Leistedt B., 2018a, MNRAS, 474, 3032 (arXiv:1706.08532)
Rogers K. K., Bird S., Peiris H. V., Pontzen A., Font-Ribera A., Leistedt B., 2018b, MNRAS, 476, 3716 (arXiv:1711.06275)
Schaye J., 2001, ApJ, 562, L95 (arXiv:astro-ph/0109280)
Slosar A., et al., 2011, J. Cosmology Astropart. Phys., 2011, 001
Viel M., Haehnelt M. G., Carswell R. F., Kim T. S., 2004, MNRAS, 349, L33 (arXiv:astro-ph/0308078)
Wolfe A. M., Turnshek D. A., Smith H. E., Cohen R. D., 1986, ApJS, 61, 249
Zafar, T. Péroux, C. Popping, A. Milliard, B. Deharveng, J.-M. Frank, S. 2013, A&A, 556, A141

APPENDIX A: TABLES OF THE MEASUREMENTS

Table A1. Table of $\mathrm{d}N/\mathrm{d}X$ values, integrated over all putative absorbers with $N_{\mathrm{HI}} > 10^{20.3}$ in our catalogue.

z	dN/dX	68% limits	95% limits
$2.00 - 2.17$	0.0337	$0.0330 - 0.0345$	$0.0323 - 0.0352$
$2.17 - 2.33$	0.0429	$0.0421 - 0.0438$	$0.0413 - 0.0446$
$2.33 - 2.50$	0.0462	$0.0452 - 0.0472$	$0.0443 - 0.0481$
$2.50 - 2.67$	0.0493	$0.0482 - 0.0505$	$0.0471 - 0.0516$
$2.67 - 2.83$	0.0620	$0.0606 - 0.0634$	$0.0592 - 0.0649$
$2.83 - 3.00$	0.0660	$0.0643 - 0.0678$	$0.0627 - 0.0695$
$3.00 - 3.17$	0.0704	$0.0683 - 0.0726$	$0.0663 - 0.0747$
$3.17 - 3.33$	0.0745	$0.0719 - 0.0774$	$0.0695 - 0.0800$
$3.33 - 3.50$	0.0763	$0.0729 - 0.0800$	$0.0696 - 0.0833$
$3.50 - 3.67$	0.0777	$0.0735 - 0.0821$	$0.0697 - 0.0862$
$3.67 - 3.83$	0.0632	$0.0586 - 0.0688$	$0.0539 - 0.0735$
$3.83 - 4.00$	0.0648	$0.0585 - 0.0720$	$0.0522 - 0.0792$
$4.00 - 4.17$	0.0581	$0.0507 - 0.0670$	$0.0447 - 0.0745$
$4.17 - 4.33$	0.0709	$0.0620 - 0.0842$	$0.0532 - 0.0953$
$4.33 - 4.50$	0.1024	$0.0896 - 0.1216$	$0.0736 - 0.1376$
$4.50 - 4.67$	0.0827	$0.0689 - 0.1057$	$0.0552 - 0.1241$
$4.67 - 4.83$	0.1041	$0.0818 - 0.1413$	$0.0669 - 0.1636$
$4.83 - 5.00$	0.0676	$0.0507 - 0.1184$	$0.0169 - 0.1522$

Table A2. Ω_{DLA} values, integrated over all putative absorbers with $N_{\mathrm{HI}} > 10^{20.3}$ in our catalogue.

z	$\Omega_{\mathrm{DLA}}(10^{-3})$	68% limits	95% limits
$2.00 - 2.17$	0.582	$0.550 - 0.619$	$0.520 - 0.659$
$2.17 - 2.33$	0.610	$0.576 - 0.651$	$0.548 - 0.694$
$2.33 - 2.50$	0.691	$0.664 - 0.722$	$0.638 - 0.755$
$2.50 - 2.67$	0.647	$0.621 - 0.676$	$0.596 - 0.706$
$2.67 - 2.83$	0.770	$0.738 - 0.809$	$0.711 - 0.855$
$2.83 - 3.00$	0.747	$0.723 - 0.773$	$0.701 - 0.799$
$3.00 - 3.17$	0.789	$0.758 - 0.829$	$0.729 - 0.896$
$3.17 - 3.33$	0.850	$0.810 - 0.909$	$0.773 - 1.042$
$3.33 - 3.50$	0.908	$0.855 - 0.962$	$0.792 - 1.019$
$3.50 - 3.67$	1.019	$0.953 - 1.087$	$0.866 - 1.166$
$3.67 - 3.83$	0.664	$0.604 - 0.731$	$0.550 - 0.806$
$3.83 - 4.00$	0.887	$0.781 - 1.000$	$0.683 - 1.112$
$4.00 - 4.17$	0.562	$0.508 - 0.622$	$0.457 - 0.684$
$4.17 - 4.33$	1.061	$0.843 - 1.337$	$0.708 - 1.675$
$4.33 - 4.50$	1.507	$1.252 - 1.810$	$1.038 - 2.182$
$4.50 - 4.67$	0.595	$0.473 - 0.737$	$0.373 - 0.892$
$4.67 - 4.83$	0.913	$0.657 - 1.208$	$0.465 - 1.498$
$4.83 - 5.00$	1.221	$0.449 - 1.995$	$0.127 - 2.449$

Table A3. The column density distribution function integrated over all spectral lengths within $2 < z < 5$.

$\log_{10} N_{HI}$	$f(N_{HI})$ (10^{-21})	68% limits (10^{-21})	95% limits (10^{-21})
$20.0 - 20.1$	0.371	$0.365 - 0.378$	$0.358 - 0.385$
$20.1 - 20.2$	0.235	$0.230 - 0.240$	$0.225 - 0.244$
$20.2 - 20.3$	0.170	$0.166 - 0.173$	$0.162 - 0.177$
$20.3 - 20.4$	0.128	$0.125 - 0.131$	$0.122 - 0.134$
$20.4 - 20.5$	9.58×10^{-2}	$[9.36 - 9.80] \times 10^{-2}$	$[9.15 - 10.02] \times 10^{-2}$
$20.5 - 20.6$	7.16×10^{-2}	$[6.99 - 7.33] \times 10^{-2}$	$[6.83 - 7.50] \times 10^{-2}$
$20.6 - 20.7$	5.09×10^{-2}	$[4.97 - 5.23] \times 10^{-2}$	$[4.85 - 5.35] \times 10^{-2}$
$20.7 - 20.8$	3.56×10^{-2}	$[3.47 - 3.66] \times 10^{-2}$	$[3.38 - 3.75] \times 10^{-2}$
$20.8 - 20.9$	2.45×10^{-2}	$[2.38 - 2.52] \times 10^{-2}$	$[2.31 - 2.59] \times 10^{-2}$
$20.9 - 21.0$	1.64×10^{-2}	$[1.59 - 1.69] \times 10^{-2}$	$[1.55 - 1.74] \times 10^{-2}$
$21.0 - 21.1$	1.06×10^{-2}	$[1.02 - 1.09] \times 10^{-2}$	$[9.92 - 11.29] \times 10^{-3}$
$21.1 - 21.2$	6.96×10^{-3}	$[6.72 - 7.22] \times 10^{-3}$	$[6.48 - 7.47] \times 10^{-3}$
$21.2 - 21.3$	4.58×10^{-3}	$[4.41 - 4.77] \times 10^{-3}$	$[4.25 - 4.94] \times 10^{-3}$
$21.3 - 21.4$	2.66×10^{-3}	$[2.55 - 2.79] \times 10^{-3}$	$[2.43 - 2.91] \times 10^{-3}$
$21.4 - 21.5$	1.51×10^{-3}	$[1.44 - 1.60] \times 10^{-3}$	$[1.36 - 1.68] \times 10^{-3}$
$21.5 - 21.6$	9.95×10^{-4}	$[9.43 - 10.56] \times 10^{-4}$	$[8.91 - 11.08] \times 10^{-4}$
$21.6 - 21.7$	4.82×10^{-4}	$[4.52 - 5.23] \times 10^{-4}$	$[4.18 - 5.57] \times 10^{-4}$
$21.7 - 21.8$	2.60×10^{-4}	$[2.39 - 2.84] \times 10^{-4}$	$[2.18 - 3.08] \times 10^{-4}$
$21.8 - 21.9$	1.50×10^{-4}	$[1.35 - 1.69] \times 10^{-4}$	$[1.23 - 1.83] \times 10^{-4}$
$21.9 - 22.0$	7.73×10^{-5}	$[6.98 - 8.86] \times 10^{-5}$	$[6.03 - 9.81] \times 10^{-5}$
$22.0 - 22.1$	4.34×10^{-5}	$[3.74 - 5.09] \times 10^{-5}$	$[3.30 - 5.69] \times 10^{-5}$
$22.1 - 22.2$	1.43×10^{-5}	$[1.19 - 2.02] \times 10^{-5}$	$[8.33 - 23.80] \times 10^{-6}$
$22.2 - 22.3$	1.13×10^{-5}	$[8.51 - 15.12] \times 10^{-6}$	$[6.62 - 17.96] \times 10^{-6}$
$22.3 - 22.4$	3.75×10^{-6}	$[3.00 - 6.01] \times 10^{-6}$	$[1.50 - 8.26] \times 10^{-6}$
$22.4 - 22.5$	2.39×10^{-6}	$[1.79 - 4.77] \times 10^{-6}$	$[5.96 - 59.63] \times 10^{-7}$
$22.5 - 22.6$	1.42×10^{-6}	$[9.47 - 28.42] \times 10^{-7}$	$[4.74 - 37.90] \times 10^{-7}$
$22.6 - 22.7$	7.53×10^{-7}	$[3.76 - 15.05] \times 10^{-7}$	$0 - 2.26 \times 10^{-6}$
$22.7 - 22.8$	5.98×10^{-7}	$[2.99 - 11.96] \times 10^{-7}$	$0 - 1.79 \times 10^{-6}$
$22.8 - 22.9$	7.12×10^{-7}	$[4.75 - 14.24] \times 10^{-7}$	$[2.37 - 16.62] \times 10^{-7}$
$22.9 - 23.0$	5.66×10^{-7}	$[1.89 - 9.43] \times 10^{-7}$	$0 - 1.32 \times 10^{-6}$