The Completed Sloan Digital Sky Survey IV Extended Baryon Oscillation Spectroscopic Survey: The Damped Lyα Systems Catalog

Solène Chabanier, Thomas Etourneau, Jean-Marc Le Goff, James Rich, Julianna Stermer, Bela Abolfathi, Andreu Font-Ribera, Alma X. Gonzalez-Morales, Axel de La Macorra, Ignasi Pérez-Ráfols, et al.

The Astrophysical Journal Supplement Series, 258:18 (10pp), 2022 January https://doi.org/10.3847/1538-4365/ac366e

The Completed Sloan Digital Sky Survey IV Extended Baryon Oscillation Spectroscopic Survey: The Damped Lyα Systems Catalog

Solène Chabanier[1,2], Thomas Etourneau[2], Jean-Marc Le Goff[2], James Rich[2], Julianna Stermer[3], Bela Abolfathi[4], Andreu Font-Ribera[5], Alma X. Gonzalez-Morales[6,7], Axel de la Macorra[8], Ignasi Pérez-Ràfols[3], Patrick Petitjean[9], Matthew M. Pieri[10], Corentin Ravoux[2], Graziano Rossi[11], and Donald P. Schneider[12,13]
[1] Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA; schabanier@lbl.gov
[2] IRFU, CEA, Université Paris-Saclay, F91191 Gif-sur-Yvette, France
[3] Sorbonne Université, Université Paris Diderot, CNRS/IN2P3, Laboratoire de Physique Nucléaire et de Hautes Energies, LPNHE, 4 Place Jussieu, F-75252 Paris, France
[4] Department of Physics and Astronomy, University of California, Irvine, CA 92697, USA
[5] Institut de Física d'Altes Energies, The Barcelona Institute of Science and Technology, Campus UAB, E-08193 Bellaterra (Barcelona), Spain
[6] Consejo Nacional de Ciencia y Tecnología, Av. Insurgentes Sur 1582. Colonia Credito Constructor, Del. Benito Jurez C.P. 03940, México D.F., Mexico
[7] Departamento de Física, División de Ciencias e Ingenierías, Campus Leon, Universidad de Guanajuato, Léon 37150, Mexico
[8] Universidad Nacional Autónoma de México Instituto de Física Apdo. Postal 20-364, Mexico
[9] Institut d'Astrophysique de Paris, Sorbonne Universités and CNRS, 98bis Boulevard Arago, F-75014, Paris, France
[10] Aix Marseille Universités, CNRS, CNES, Laboratoire d'Astrophysique de Marseille, Marseille, France
[11] Department of Physics and Astronomy, Sejong University, Seoul, 143-747, Republic of Korea
[12] Department of Astronomy and Astrophysics, The Pennsylvania State University, University Park, PA 16802, USA
[13] Institute for Gravitation and the Cosmos, The Pennsylvania State University, University Park, PA 16802, USA
Received 2021 July 20; revised 2021 September 27; accepted 2021 October 26; published 2022 January 14

Abstract

We present the characteristics of the damped Lyα (DLA) systems found in data release DR16 of the extended Baryon Oscillation Spectroscopic Survey of the Sloan Digital Sky Survey. The DLAs were identified using the convolutional neural network of Parks et al. (2018). A total of 117,458 absorber candidates were found with $2 \leqslant z_{\rm DLA} \leqslant 5.5$ and $19.7 \leqslant \log(N(H_I)/{\rm cm}^{-2}) \leqslant 22$, including 57,136 DLA candidates with $\log(N(H_I)/{\rm cm}^{-2}) \geqslant 20.3$. Mock quasar spectra were used to estimate the DLA detection efficiency and the purity of the resulting catalog. Restricting the quasar sample to bright forests, i.e., those with mean forest fluxes $\overline{f_\lambda} > 2 \times 10^{-19}$ W m^{-2} nm^{-1}, the efficiency and purity are greater than 90% for DLAs with column densities in the range $20.1 \leqslant \log(N(H_I)/{\rm cm}^{-2}) \leqslant 22$.

Unified Astronomy Thesaurus concepts: Intergalactic medium (813); Quasar absorption line spectroscopy (1317); Catalogs (205)

1. Introduction

Damped Lyα (DLA) are absorption systems with neutral hydrogen column densities, $N(H_I) \geqslant 2 \times 10^{20}$ atoms cm^{-2}, producing broad damping wings in the optical spectra of bright background objects such as quasars (Wolfe et al. 1986).

Such systems are at high enough densities to be self-shielded against ionizing radiation (Vladilo et al. 2001; Cen 2012; Fumagalli et al. 2014), and they are connected to dark matter halos over a large range of masses, from dwarf galaxies to clusters of galaxies (Prochaska & Wolfe 1997; Haehnelt et al. 1998; Pontzen et al. 2008). Observations show that DLAs are the dominant reservoir of neutral hydrogen in the redshift range $0 < z < 5$ and contain 2% of all baryons in the universe (Gardner et al. 1997; Wolfe et al. 2005; Prochaska & Wolfe 2009; Noterdaeme et al. 2012). As such, DLAs are key to understanding galaxy formation and evolution since they are thought to be the reservoir of atomic gas for stellar formation in galaxies. They are thus an important probe of physical conditions in the interstellar medium at high redshifts (Petitjean et al. 2000; Fumagalli et al. 2013; Bird et al. 2014; Ota et al. 2014; Fumagalli et al. 2016; Rudie et al. 2017).

However, DLAs are also contaminants in the measurements of the Lyα forest flux probability distribution function (Lee et al. 2015), its 3D autocorrelation function (Slosar et al. 2011; Bautista et al. 2017; du Mas des Bourboux et al. 2020), or its 1D power spectrum (McDonald et al. 2006; Palanque-Delabrouille et al. 2013; Chabanier et al. 2019). Since DLAs form at high density peaks, they cluster more strongly than diffuse Lyα clouds (Font-Ribera & Miralda-Escudé 2012), thus biasing astrophysical and cosmological parameters if not well accounted for. Therefore, their detection along with the measurements of their physical properties, absorption redshift, and column densities are important in such studies.

With hundreds of thousands detected quasar spectra, the large statistical power of the Sloan Digital Sky Survey (SDSS; York et al. 2000) has fostered the compilation of DLA catalogs (Noterdaeme et al. 2009; Prochaska & Wolfe 2009; Zhu & Ménard 2013; Garnett et al. 2017; Parks et al. 2018; Ho et al. 2020). Given the tremendous number of spectra to analyze, it has also played a critical role in the development of automated detection algorithms over visual inspection, e.g., using Voigt-profile fitting (Prochaska et al. 2005; Noterdaeme et al. 2009, 2012) or machine-learning techniques, such as convolutional neural networks (CNN; Parks et al. 2018), Gaussian processes (Garnett et al. 2017), or random forest classifiers (Fumagalli et al. 2020).

The final SDSS-IV quasar catalog from Data Release 16 (DR16) of the extended Baron Oscillation Spectroscopic

Survey (eBOSS; Dawson et al. 2016; Ahumada et al. 2020), which we will refer to as DR16Q, is the largest quasar spectra sample to date with 920,110 observations of 750,414 quasars (Lyke et al. 2020). In the DR16Q, we used the CNN algorithm from Parks et al. (2018) to include DLA quasar identification for very confident DLAs with $\log(N(H_I)/\mathrm{cm}^{-2}) \geqslant 20.3$ only. Here we present the full sample of absorbing systems detected with the CNN in DR16Q, which includes less confident DLAs and Lyman-limit systems (LLS) with $\log(N(H_I)/\mathrm{cm}^{-2})$ as low as 19.7. The choice of the CNN from Parks et al. (2018) is motivated by the design of the algorithm constructed specifically for low redshift and low signal-to-noise BOSS/eBOSS quasar spectra.

The paper is organized as follows. Section 2 presents the quasar spectra sample, which is scanned for high-column-density absorbing systems. Section 3 introduces the automated algorithm and the CNN architecture from Parks et al. (2018) that we use to detect strong absorbers. We perform efficiency and purity validation of the algorithm with synthetic spectra and a study of biases of DLA parameters, $\log(N(H_I)/\mathrm{cm}^{-2})$ and z_{DLA} in Section 4. Finally, we present the full absorber sample in Section 5 and compare it with existing catalogs. We present concluding remarks in Section 6.

2. Quasar Spectra Sample DR16Q

In this work, we use data measured with BOSS and eBOSS (Dawson et al. 2016) of the SDSS-III and SDSS-IV (Gunn et al. 2006; Smee et al. 2013; Blanton et al. 2017) surveys, respectively. We focus on the Lyα forest regions from the 750,414 quasar spectra available in DR16Q (Lyke et al. 2020), which contains all SDSS spectroscopically observed quasars. The selection of quasars for the BOSS and eBOSS surveys are described in Ross et al. (2012) and Myers et al. (2015).

We search for DLAs in the 263,201 spectra with $2 \leqslant \mathrm{Z\_PCA} \leqslant 6$, the redshift range over which spectra contain enough pixels to identify DLAs. We use the quasar redshift estimator, Z_PCA, generated by principal component analysis (PCA), using the REDVSBLUE algorithm.[14] The DR16Q catalog is constructed from the SPALL-V5_13_0 (spAll) file containing all SDSS-III/IV observations treated by the version V5_13_0 of the SDSS spectroscopic pipeline.[15] If multiple observations are available for one object in the spAll file, we use the stacked spectrum of all good observations as input to the DLA finder. We identify bad spectra using the ZWARNING parameter. If ZWARNING is SKY, LITTLE COVERAGE, UNPLUGGED, BAD TARGET, or NODATA, we do not use the associated observation in the stack.

Figure 1 shows the redshift distribution of the forest pixels, with a mean of $z = 2.4$. Figure 2 shows the flux and signal-to-noise ratio (S/N) averaged over the forest, with a mean flux of $2.87 \times 10^{-19}\,\mathrm{W\,m^{-2}\,nm^{-1}}$ $(=2.87 \times 10^{-17}\,\mathrm{erg\,s^{-1}\,cm^{-2}\,\AA^{-1}})$ and a mean S/N of 2.90.

We will see in Section 4.2 that the efficiency for finding DLAs is poor for forests with low S/N corresponding generally to forests with low fluxes. Figure 2 therefore also shows the S/N for a "bright sample" of forests with mean forest flux $\overline{f_\lambda} > 2 \times 10^{-19}\,\mathrm{W\,m^{-2}\,nm^{-1}}$. Also shown is the S/N as a function of $\overline{f_\lambda}$ for three redshift ranges. We see that for a given redshift, there is tight correlation between S/N and forest flux. This reflects the relatively uniform sky coverage of SDSS. For reference, forests in the bright sample generally have S/N greater than 2.

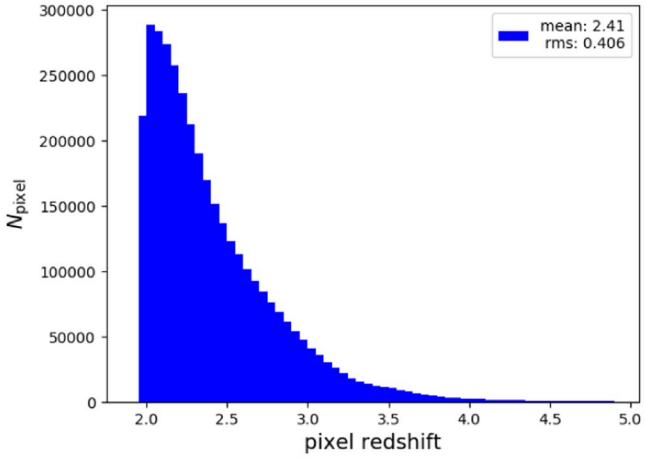

Figure 1. Redshift distribution for pixels in the Lyα forest from the quasar spectra available in DR16Q.

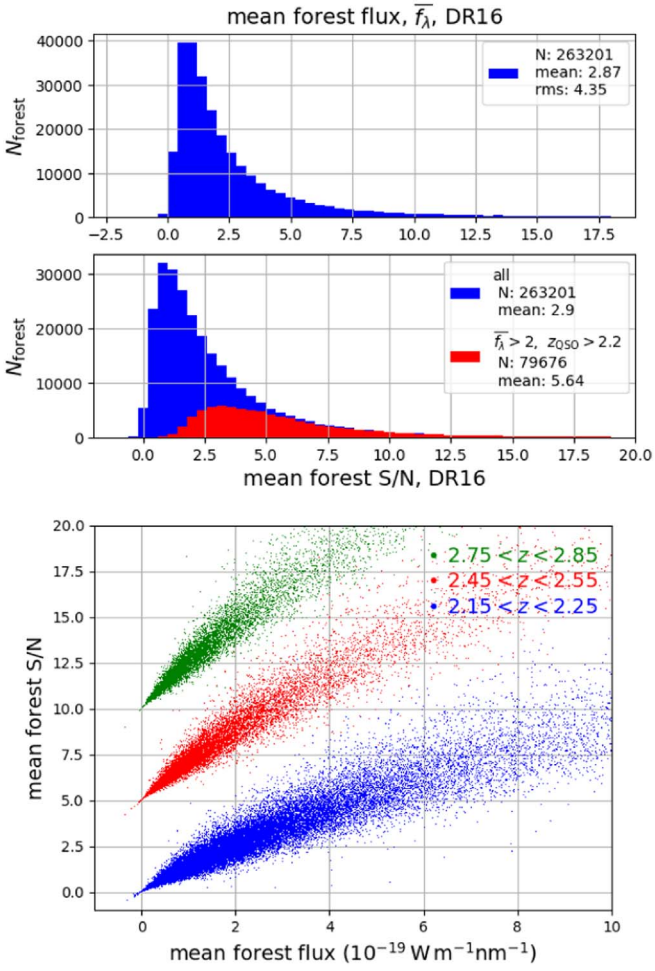

Figure 2. Mean flux and mean signal-to-noise ratio (S/N) for pixels in the Lyα forest. Top panel: the distribution for the complete sample (blue) and for the restricted sample with high DLA detection efficiency; quasar redshift $z_{\mathrm{QSO}} > 2.2$ and mean forest flux $\overline{f_\lambda} > 2.0$ (in units of $10^{-19}\,\mathrm{W\,m^{-2}\,nm^{-1}} = 10^{-17}\,\mathrm{erg\,s^{-1}\,cm^{-2}\,\AA^{-1}}$). Bottom panel: S/N as a function of $\overline{f_\lambda}$ for three ranges of the mean forest redshift, as labeled. For clarity, the S/N is offset by (0,5,10) units.

[14] https://github.com/londumas/redvsblue

[15] https://data.sdss.org/datamodel/files/BOSS_SPECTRO_REDUX/RUN2D/spAll.html

3. DLA Detection Method

We identified DLAs with the algorithm described in Parks et al. (2018), which is based on a multitask learning CNN. We refer the reader to Parks et al. (2018) for a complete description of the detection algorithm, only recalling here the major steps. The CNN architecture and its training aim at constructing an algorithm that works at low redshifts, in noisy regions, and without any input from the user other than raw spectral data. The algorithm therefore does not need quasar continuum or DLA Voigt profile modeling, and it ignores flux errors estimated by the SDSS pipeline. Finally, the model does not include broad absorption lines (BALs), compromising DLA detection. Therefore, we reject lines of sight that the DR16Q pipeline indicates as affected by BALs.

The neural network model uses a standard 2D CNN architecture with four layers. It relies on the Adam (Adaptive Moment Estimation) algorithm to search for the optimal parameters (Kingma & Ba 2014) and is implemented using Google's deep learning framework TensorFlow.[16] It analyzes 1748 pixel long sight lines of $\delta\lambda \simeq 1$ Å in 1748 inference steps with 400 pixel long sliding windows in the 900 Å $\leqslant \lambda \leqslant$ 1346 Å region in order to improve detection of multiple DLAs per sight line. The 400 pixel size is in part imposed by the SDSS resolution. The model produces three outputs for each sliding window: (1) classification of the segment as containing a DLA or not, (2) the DLA absorption redshift Z_DLA, i.e., the pixel center localization, and (3) the HI column density, NHI_DLA, if a DLA is visible. In the case of a detected DLA in a sight line, the authors also define a nonstatistical measure of confidence, the *confidence* parameter over the range (0,1). It is based on how robustly the DLA is localized over the different predictions of the sliding window.

The training sample was constructed using 4113 SDSS sight lines, with quasar redshift $z_{\rm qso} > 2.3$ and S/N > 5, identified as DLA-free from the analysis of Prochaska & Wolfe (2009). The authors generated 200,000 sight lines from the DLA-free sample by inserting DLAs and super LLS (SLLS) with logarithmic column density $19.5 \leqslant \log(N(H_I)/{\rm cm}^{-2}) \leqslant 22.5$ using Voigt profile modeling.

Finally, the algorithm was validated using one catalog with synthetic DLA in real DLA-free spectra and one catalog constituted of visually inspected spectra containing DLAs (Prochaska & Wolfe 2009). The authors found a systematic bias of order ∼0.1 in the predicted $\log(N(H_I)/{\rm cm}^{-2})$ at both low and high ends. They fit this bias with a third degree polynomial (see Figure 9 of Parks et al. 2018) and used this result to correct for the bias in the final automated algorithm.

4. Analysis of DLAs in Mock Spectra

Given that S/N and quasar redshift distributions of the training and validation samples do not exactly match those of the DR16 data, we used synthetic spectra to perform purity and efficiency validation of the algorithm along with an investigation of systematics on the inferred $z_{\rm DLA}$ and $N(H_I)$. The synthetic spectra, hereafter "mocks", were produced for the eBOSS Lyα data analysis (du Mas des Bourboux et al. 2020). In Section 4.1, we briefly describe the construction of mock spectra, and we present our estimates of efficiency and purity in Section 4.2. The accuracy and precision of the estimation of $N(H_I)$ is discussed in Section 4.3.

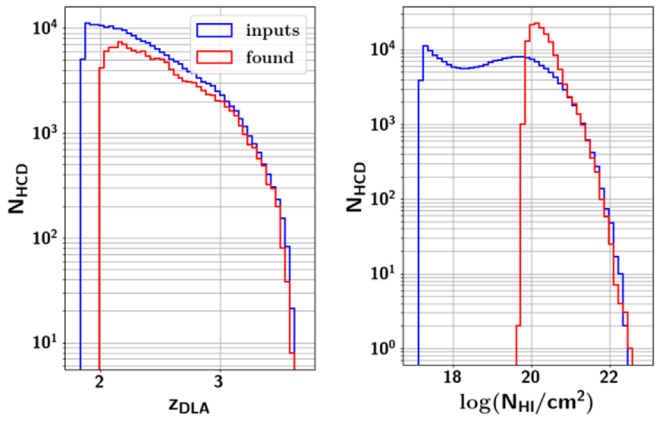

Figure 3. Distributions of $z_{\rm DLA}$ (left) and $\log(N(H_I)/{\rm cm}^{-2})$ (right) for the 218,124 HCD systems placed in mocks (blue) and for the 132,226 DLAs reported by the CNN (red). The excess of detected DLAs at $\log(N(H_I)/{\rm cm}^{-2}) \approx 20$ is due to the poor purity at this $\log N$, as seen in Figure 3.

4.1. Synthetic Quasar Spectra

The production of the mocks is described elsewhere (T. Etourneau et al. 2022, in preparation), and we describe here only the major steps. A low-resolution Gaussian random density field was produced in a box of $2560 \times 2560 \times 1536$ voxels of 2.19 h^{-1} Mpc sides. Quasar positions were drawn proportionally to a lognormal field, with phases in the Fourier space equal to those of the density field. The interpolated values of the density field along the sight lines from the observer position toward the quasars were computed, together with the three components of the velocity and the six components of the velocity gradient tensor. Extra small-scale fluctuations were added to each sight line independently, in order to reproduce the variance in the Lyα forest in the data. Redshift space distortions were implemented by adding the velocity gradient along the sight lines to the density field. We then applied a lognormal transformation to this sum and used the fluctuating Gunn–Peterson approximation to compute the optical depth in each cell.

We define high-column-density (HCD) systems as systems with column density, $N(H_I) > 10^{17.2}\,{\rm cm}^{-2}$, including both DLA and LLSs $10^{17.2}\,{\rm cm}^{-2} < N(H_I) < 2 \times 10^{20}\,{\rm cm}^{-2}$. We selected peaks of the large-scale density field as possible locations of HCD systems and set the threshold to get a constant bias $b_{\rm HCD}(z) = 2$ for the HCD systems, using the formulas in appendix A of Font-Ribera & Miralda-Escudé (2012). We then Poisson sampled the selected peaks to follow the HCD system redshift distribution of the default model from the IGM physics package pyigm.[17] The column density was selected to follow the same model. These distributions are shown as the blue curves in Figure 3. The radial velocities of the HCD systems were obtained from the three velocity component boxes. This information was included in an HCD system catalog.

The redshifts of HCD systems placed in the mocks are in the range of $1.8 \leqslant z_{\rm DLA} \leqslant 3.5$ (blue curve of Figure 3), whereas the CNN has been trained with absorbing redshifts up to $z_{\rm DLA} \sim 4.5$. Therefore, efficiency and purity studies presented

[16] https://www.tensorflow.org/

[17] https://github.com/pyigm/pyigm

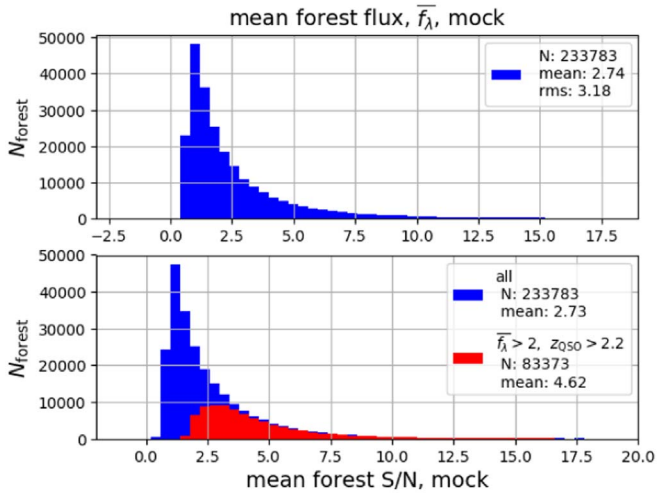

Figure 4. Mean flux and mean S/N for mocks pixels in the Lyα forest for the complete sample (blue) and for the restricted sample with high DLA detection efficiency (red): quasar redshift $z_{QSO} > 2.2$ and mean forest flux $\overline{f_\lambda} > 2.0$.

in Section 4.2 only focus on the 2.0–3.5 range even if the data catalog presents higher absorbing redshifts HCD systems (see Section 5).

As a last step, the quasar spectra are produced by multiplying the transmitted flux fraction by a quasar continuum and adding instrumental noise (A. Gonzales-Morales et al., in preparation). For each HCD system in the catalog, we multiplied the corresponding quasar spectrum by the Voigt profile for the HCD system column density.

Figure 4 shows the mean flux and mean S/N for mock pixels in the Lyα forest. We see that the mock distribution agree qualitatively with the data shown in Figure 2. The data do, however, contain more forests with very low flux and very high flux.

4.2. Efficiency and Purity

The red curves in Figure 3 show the distributions of redshift and of $\log(N(H_I)/\mathrm{cm}^{-2})$ of the 132,226 DLAs found by the CNN. The differences between the blue and red curves are due to many factors affecting the efficiency and purity, as described in the following paragraphs. The most important effect is the insertion of $\log(N(H_I)/\mathrm{cm}^{-2}) < 19.5$ HCD systems into the mocks resulting in a low global efficiency, as seen on the left plot of Figure 3. Also important is the low purity of the found DLAs at $\log(N(H_I)/\mathrm{cm}^{-2}) \approx 20$ (Figure 6) resulting in the excess of found DLAs near $\log(N(H_I)/\mathrm{cm}^{-2}) \approx 20$ (right plot of Figure 3). The low purity is due to the the CNN classifying noise fluctuations as DLAs and to assigning $\log(N(H_I)/\mathrm{cm}^{-2}) > 19.5$ to HCD systems with $\log(N(H_I)/\mathrm{cm}^{-2}) < 19.5$.

The efficiency for DLA detection and the purity of the detected sample were studied by using the mock spectra where the catalog of detected DLAs can be compared with the catalog of generated HCD systems. We define the efficiency as

$$\text{efficiency} = \frac{N_{\mathrm{TP}}}{N_{\mathrm{TP}} + N_{\mathrm{FN}}} \quad (1)$$

and the purity as

$$\text{purity} = \frac{N_{\mathrm{TP}}}{N_{\mathrm{TP}} + N_{\mathrm{FP}}}, \quad (2)$$

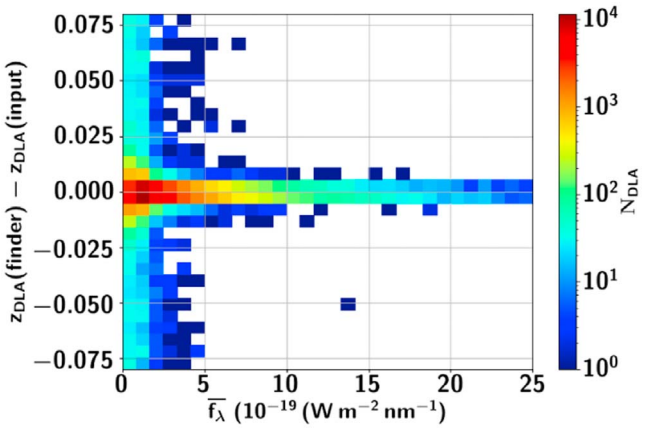

Figure 5. Mean difference between detected and generated DLA redshifts versus the mean forest flux, $\overline{f_\lambda}$.

where N_{TP}, N_{FP}, and N_{TN} are the true positive, false positive, and true negative detected HCD systems, respectively.

Both the efficiency and purity are functions of the characteristics of the forest (S/N and forest mean flux) and of the DLA (redshift and column density). They also depend on the criterion used to define detected DLAs, i.e., the requirements placed on the confidence parameter and on the required agreement between generated and found z_{DLA} and $N(H_I)$.

The criteria for detection and matching are arbitrary to a certain extent. The most important matching criterion concerns the redshift difference between generated and detected DLAs. Figure 5 shows this difference versus the mean forest flux, $\overline{f_\lambda}$, for best-matched DLAs, where the match only requires that the mock and found DLAs are on the same sight line. Here, we adopt a matching criterion requiring that the detected and generated redshifts differ at most by $\Delta z < 0.02$ (about 25 Å). This redshift-matching cut accepts most detected DLAs for $\overline{f_\lambda} > 2 \times 10^{-19}$ W m^{-2} nm^{-1}. However, the redshift resolution degrades substantially for lower $\overline{f_\lambda}$. With the above criterion, the DLA finder recovers 62,847 absorbing systems with $z_{\mathrm{DLA}} > 2$ and $\log(N(H_I)/\mathrm{cm}^{-2}) > 19$ (69% of the absorbing systems put in mocks). Among them 86% (70%) have confidence parameters >0.5 (>0.9). Changing the redshift-matching criterion to $\Delta z < 0.01$ reduces only slightly the number of recovered DLAs from 63,847 to 61,131.

For the adopted matching criterion, $\Delta z_{\mathrm{DLA}} < 0.02$, the efficiency and purity as functions of z_{DLA} and $N(H_I)$ are shown in Figure 6 on the left and right panels, respectively. Figures 7, 8, and 9 show the same measurements but for bright ($\overline{f_\lambda} > 2 \times 10^{-19}$ W m^{-2} nm^{-1}) forests, faint ($\overline{f_\lambda} > 2\times 10^{-19}$ W m^{-2} nm^{-1}) forests, and confident absorbers (confidence >0.9), respectively. Note that we use the HCD system characteristics as returned by the finder to compute the purity and the ones from the mock input for the efficiency, which explains why the right panel does not have data for $\log(N(H_I)/\mathrm{cm}^{-2}) < 19.65$ but the left one does.

For the bright sample, Figure 7 shows that high efficiency (>0.9) *and* purity (>0.9) are obtained for column densities in the range $20.2 < \log(N(H_I)/\mathrm{cm}^{-2}) < 22.0$ and redshifts $z_{\mathrm{DLA}} > 2.2$. For the faint sample, high efficiency and purity are found only for $\log(N(H_I)/\mathrm{cm}^{-2}) > 21.0$ and $z_{\mathrm{DLA}} > 2.2$.

For the efficiency, there is almost no dependence on z_{DLA}. It is degraded for $z_{\mathrm{DLA}} \leqslant 2.2$ but performs quite equally for higher redshifts. The bad performances at low absorbing

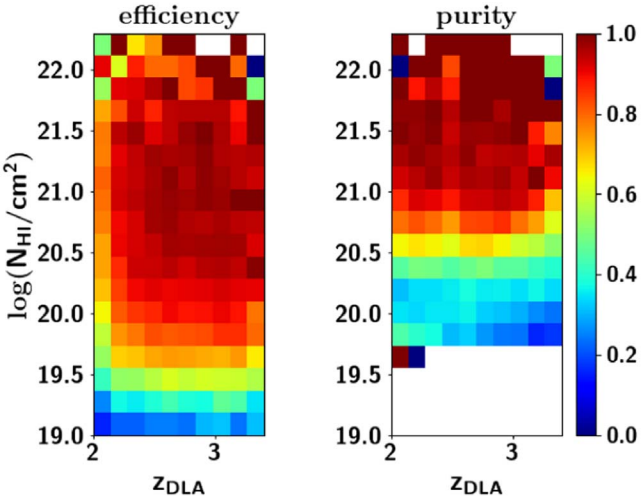

Figure 6. Efficiency (left) and purity (right) versus redshift and column density using matching criterion $\Delta z_{\rm DLA} < 0.02$.

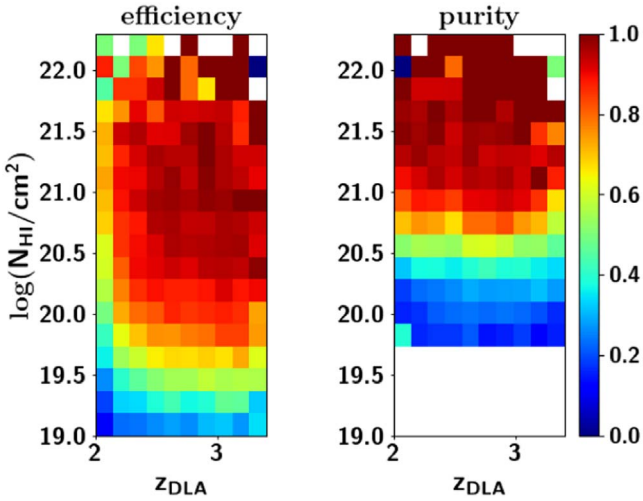

Figure 8. Same as Figure 6 for forests with $\overline{f_\lambda} < 2 \times 10^{-19}\,{\rm W\,m^{-2}\,nm^{-1}}$.

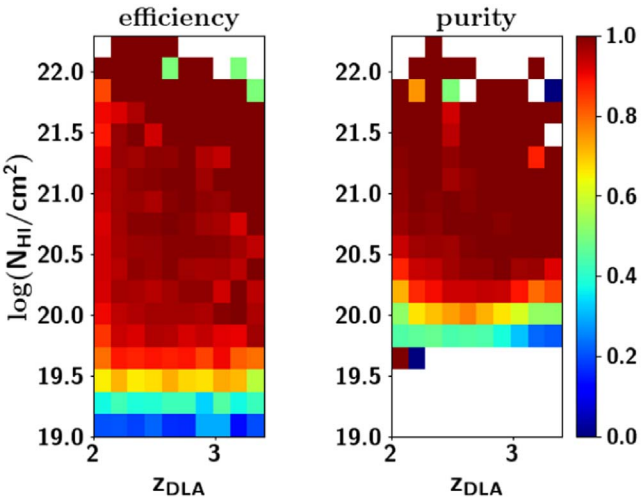

Figure 7. Same as Figure 6 for forests with $\overline{f_\lambda} > 2 \times 10^{-19}\,{\rm W\,m^{-2}\,nm^{-1}}$.

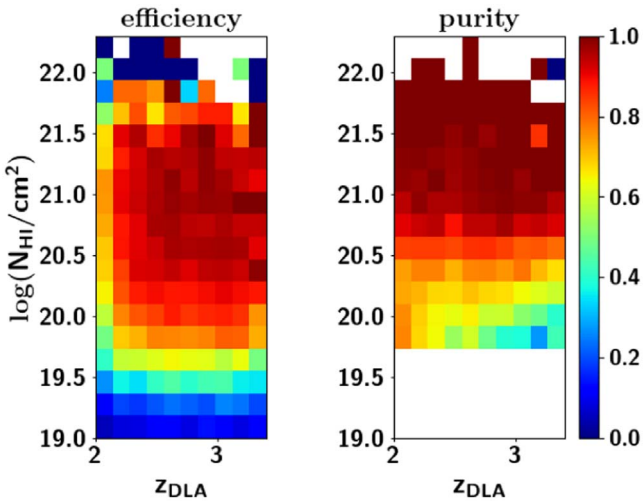

Figure 9. Same as Figure 6 for forests with confidence >0.9.

redshifts occur in the blue and noisy end of the spectra (see the S/N distribution as a function of the quasar redshift in the bottom panel of Figure 2). Indeed, false negatives have a mean forest flux 25% lower than the average. By comparing Figures 6, 7, and 8 we easily deduce that faint forests are driving the bad performances. Also, because the spectra are small in size at low redshifts, i.e., have a low number of pixels, it is harder for the CNN to detect features and to make accurate predictions. The efficiency drops below 0.2 for the low end of $N(H_I)$, for which the CNN has not been specifically designed and trained and for which instrumental noise and resolution make detection difficult. The finder detects HCD systems with $\log(N(H_I)/{\rm cm}^{-2})$ as low as 19 but, as we will see in the next section, overestimates this parameter. This explains the excess of the detected $\log(N(H_I)/{\rm cm}^{-2})$ near the detection threshold compared to the mock distribution on the left panel of Figure 3.

The efficiency also decreases for high $N(H_I)$ where the DLA covers a substantial fraction of the forest. While we observe a trend for a decrease toward the high end of $N(H_I)$, synthetic spectra have a total of 806 HCD systems with $\log(N(H_I)/{\rm cm}^{-2}) > 21.5$. While this makes our results statistically significant for high $N(H_I)$ on average, results can be very noisy when sampled into $z_{\rm DLA}$ bins. Over the 104 missed DLAs with $\log(N(H_I)/{\rm cm}^{-2}) > 21.5$ and $z_{\rm DLA} > 2.2$, 19 are detected by the finder but rejected by the redshift-matching cut criterion ($0.02 < \Delta z < 0.04$) because of low mean forest flux ($\overline{f_\lambda} < 2$). Four have an absorbing redshift extremely close to the Lyα emission line such that the CNN found a $z_{\rm DLA} > z_{\rm QSO}$, and 15 are part of two overlapping DLAs with $\Delta z_{\rm DLA} < 0.03$ detected as one DLA with a higher $N(H_I)$ (as was noted in Parks et al. (2018), the CNN struggles at identifying overlapping DLAs). The 66 remaining DLAs have particularly low mean forest fluxes with an average of $\sim 0.3 \times 10^{-19}\,{\rm W\,m^{-2}\,nm^{-1}}$. When considering bright forests only, with $\overline{f_\lambda} > 2 \times 10^{-19}\,{\rm W\,m^{-2}\,nm^{-1}}$, the efficiency is always >0.9 for $20 < \log(N(H_I)/{\rm cm}^{-2})$. The results are noisy for $\log(N(H_I)/{\rm cm}^{-2}) > 21.5$, especially for high-redshift bins, but the efficiency is close to one on average for $\log(N(H_I)/{\rm cm}^{-2}) > 21.5$ DLAs.

The purity is >0.5 for $z_{\rm DLA} < 3.2$ and $20.3 < \log(N(H_I)/{\rm cm}^{-2}) < 21.5$, and >0.9 for $z_{\rm DLA} < 3.2$ and $20.8 < \log(N(H_I)/{\rm cm}^{-2}) < 21.5$.

Our matching criterion does not use the confidence parameter, and using it could increase the purity at the cost

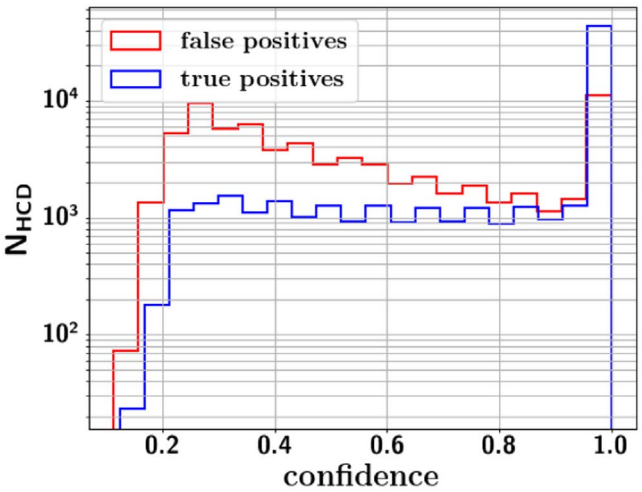

Figure 10. Confidence parameter distributions for the 132,226 absorbing systems found by the CNN: 69,380 false positives (red) and 62,846 true positives (blue). False positives are defined as HCD systems found by the CNN but that do not match the HCD system input. To match an input HCD system, it must be on the same sight line with $\Delta z_{\rm DLA} < 0.02$.

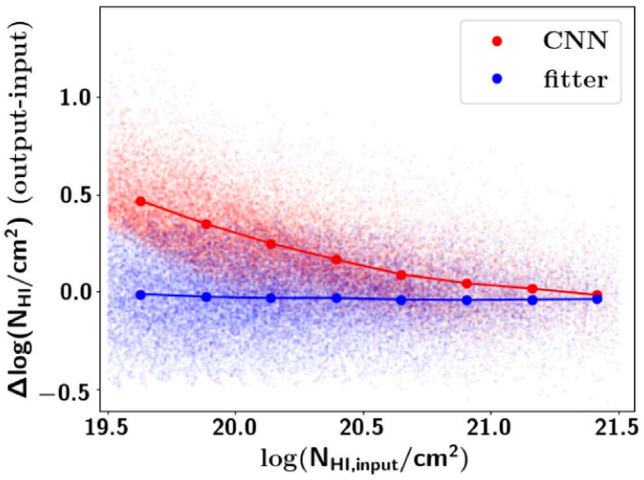

Figure 11. Difference in output and input values of $N(H_I)$ (in red for the CNN and blue for the fitter) vs. input for the 25,696 found DLAs in the Lyα forest, i.e., in the rest-frame range 1040 Å $\leqslant \lambda_{\rm RF} \leqslant$1216 Å.

of decreasing efficiency. Figure 10 shows the distribution of the confidence parameter for all found HCD systems, true positives and false positives. False positives are all HCD systems found by the CNN but that have not been matched to an HCD systems input (same sight line with $\Delta z_{\rm DLA} < 0.02$). Only 18% (44%) of false positives are *confident* HCD systems with confidence parameters of >0.9 (>0.5). Taking only HCD systems with confidence parameters of >0.9 results in the purity always being >0.9 for $\log(N(H_I)/{\rm cm}^{-2}) > 20.3$. As stated in Parks et al. (2018), the confidence parameter is nonstatistical measure based on how tightly the model predicted the location of the DLA over the sliding (redshift) window (see their Section 5.1.3). Figure 10 shows that, as expected, false positives have a lower mean value of confidence than true positives. However, even false positives exhibit a peak at confidence ≈1. This shows that a significant fraction of false positives yield stable values of $z_{\rm DLA}$. Because of the peak at confidence ≈1 for false positives, requiring confidence of >0.9 to increase purity is not very efficient. A better procedure would be to use the mock data to set a confidence cut that depends on $\log(N(H_I)/{\rm cm}^{-2})$, $z_{\rm DLA}$, and the mean flux or S/N in a way that best fits the purity–efficiency requirements of the user.

The net decrease in the purity toward low $\log(N(H_I)/{\rm cm}^{-2})$ seen in the right panel of Figure 6 occurs because it gets more and more difficult for the CNN to distinguish between real but relatively small absorptions and noise. Indeed, when considering bright forests only (see right panel of Figure 7), with $\overline{f_\lambda} > 2 \times 10^{-19}$ W m^{-2} nm^{-1}, the purity is always >0.9 for $20.1 < \log(N(H_I)/{\rm cm}^{-2})$. We observe a decrease in purity at high redshifts for both bright and faint samples (see Figures 6, 7, 8) because the mean flux decreases for high-redshift quasars, making it harder for the CNN to distinguish between Lyα absorptions and DLAs.

To summarize, the main parameter for maximal efficiency and purity of the absorber catalog is the mean flux of the forest. Taking $\overline{f_\lambda} > 2$ ensures that the efficiency and purity are >0.9 for $\log(N(H_I)/{\rm cm}^{-2}) > 20.1$. However, degrades the size of the sample. If taking all bright and faint forests, the efficiency is >0.9 for $z_{\rm DLA} > 2.2$ and $20 < \log(N(H_I)/{\rm cm}^{-2}) < 21.5$, the purity is >0.9 for $z_{\rm DLA} < 3.2$ and $20.5 < \log(N(H_I)/{\rm cm}^{-2}) < 21.5$, and the confidence is >0.9.

4.3. Parameter Estimation

The CNN cannot be expected to give an unbiased estimate of $\log(N(H_I)/{\rm cm}^{-2})$ because the DLA sample was selected by the CNN. The mocks contain a large number of low-column-density HCD systems (Figure 3) and some, through noise, may appear as detectable DLAs with $\log(N(H_I)/{\rm cm}^{-2}) > 20$. As such, we expect the estimated $\log(N(H_I)/{\rm cm}^{-2})$ to be on average greater than the true $\log(N(H_I)/{\rm cm}^{-2})$. This expectation is confirmed by Figure 11, which compares the values of $N(H_I)$ returned by the finder with the input value from the mocks.

We investigate the dependence of this systematic bias on the confidence parameter in Figure 12 showing the difference between input and CNN values of $\log(N(H_I)/{\rm cm}^{-2})$ for four ranges of CNN values of $\log(N(H_I)/{\rm cm}^{-2})$. First, as already shown in Figure 11, the bias is worse for low $\log(N(H_I)/{\rm cm}^{-2})$ as the mean increases toward 0 with increasing $\log(N(H_I)/{\rm cm}^{-2})$. More importantly, the confidence parameter is a good indicator of biased $N(H_I)$ as the blue curves always tend toward more negative values than the red curves. The $\Delta\log(N(H_I)/{\rm cm}^{-2})$ tail of nonconfident HCD systems is particularly long on the top left panel. This is because even if these low-$N(H_I)$ candidates are matched to input HCD systems, we are close to the $N(H_I)$ detection threshold, so that many detected HCD systems are in fact noise fluctuations close to a low-$N(H_I)$ HCD system. As such, the confidence parameter is also very useful for increasing the purity in the low-$N(H_I)$ regime.

To provide a more unbiased estimate of $\log(N(H_I)/{\rm cm}^{-2})$, we developed a DLA fitter and applied it to the 25,696 DLA candidates in the rest-frame range 1040 Å $\leqslant \lambda_{\rm RF} \leqslant$1216 Å. Figure 11 shows the difference between the input $N(H_I)$ and Voigt profile fitted $N(H_I)$, which are more accurate that the CNN ones.

5. The DR16 DLA Catalog

We applied the automated algorithm to the 263,201 DR16 quasar spectra sample described in Section 2. A total of

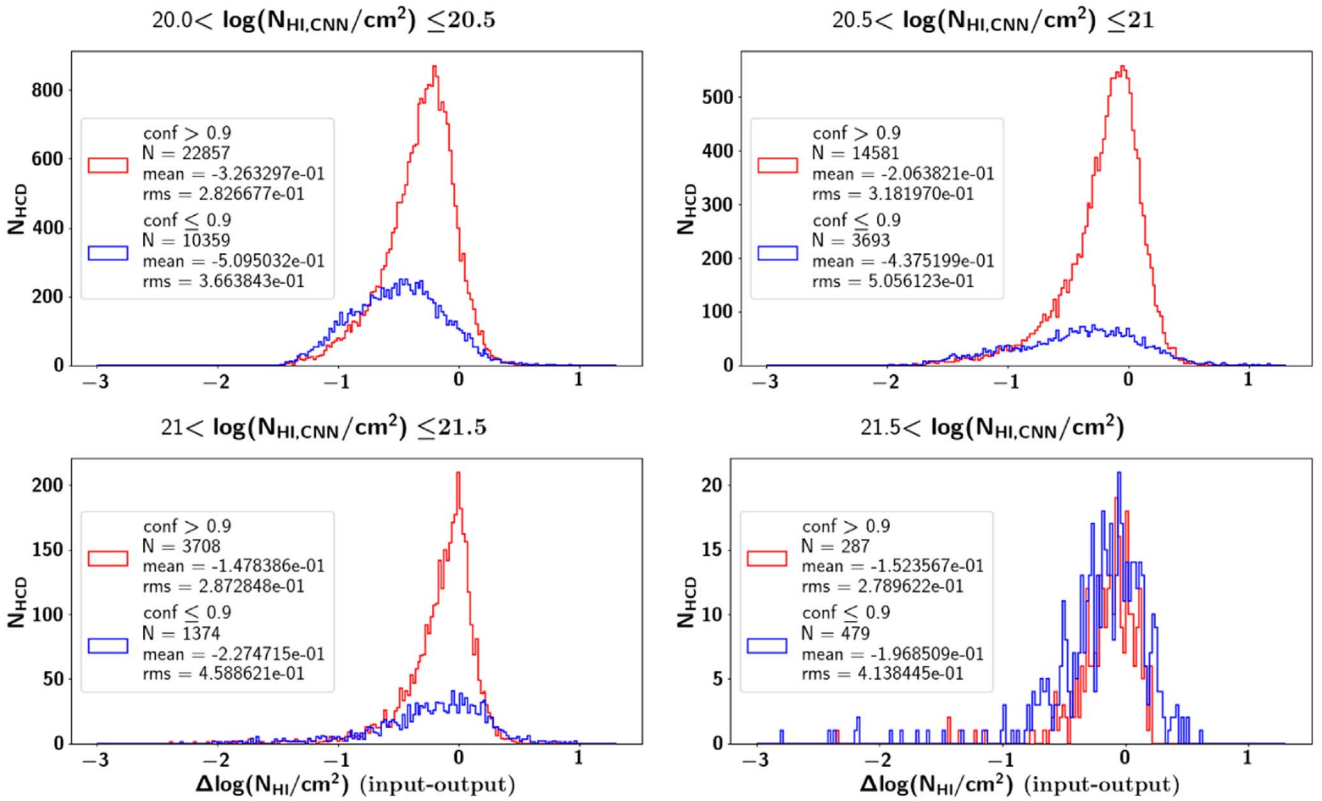

Figure 12. Distributions of the difference in $\log(N(H_I)/\mathrm{cm}^{-2})$ between the CNN predictions and the mock inputs for DLA candidates with confidence >0.9 (red) and confidence ⩽0.9 (blue). The distributions are shown for four ranges of $\log(N(H_I)/\mathrm{cm}^{-2})$ (CNN value), as labeled.

176,807 HCD systems were found with $z_{\mathrm{QSO}} > z_{\mathrm{DLA}}$ and $z_{\mathrm{DLA}} \geqslant 2$ in 112,155 sight lines. These numbers are reduced to 117,458 absorbers in 78,018 sight lines when we reject BAL quasars with BAL_PROB > 0; among them, 39,067 (33%) are classified as confident with confidence >0.9. Figure 13 shows the z_{DLA} and $\log(N(H_I)/\mathrm{cm}^{-2})$ distributions for the 20,375 bright forests and the remaining 97,083 faint forests of the 117,458 total sample. The sample was further reduced to 57,136 absorbers with $\log(N(H_I)/\mathrm{cm}^{-2}) \geqslant 20.3$ in 20,016 sight lines, yielding a purity of ∼0.3 given that the number of DLAs per los is roughly <1 (Noterdaeme et al. 2012; Bird et al. 2014). Only considering bright forests raises the purity to >0.9 for DLAs with $\log(N(H_I)/\mathrm{cm}^{-2}) \geqslant 20.3$ since the CNN found 6996 such absorbers in 6293 lines of sight.

The DLA sample we presented in DR16Q (Lyke et al. 2020) only includes DLAs absorbers with $\log(N(H_I)/\mathrm{cm}^{-2}) \geqslant 20.3$ and confidence >0.9, and we did not reject BAL quasars. As presented in Section 4.2, the confidence cut highly degrades the efficiency toward high- and low-$N(H_I)$ absorbing redshifts. The DLA sample presented here is consequently more complete and less pure. As discussed in Section 4.2, users of this catalog can construct their own selection criteria, depending on their specific needs.

We compared our sample of 117,458 absorbers with two other catalogs based on BOSS and eBOSS data. The first is the 12,081 absorber sample of Noterdaeme et al. (2012), hereafter N12, based on the DR9 SDSS data release that uses the Voigt profile fitting procedure. The second, based on DR16 SDSS data, was provided by Ho et al. (2021), hearafter H21, that extends the Gaussian processes method presented in Garnett et al. (2017). We reject BAL quasars with BAL_PROB > 0 and consider only DLA with a high probability, $p_{\mathrm{DLA}} > 0.9$. With these criteria, their sample contains a total of 25,160 absorbers.

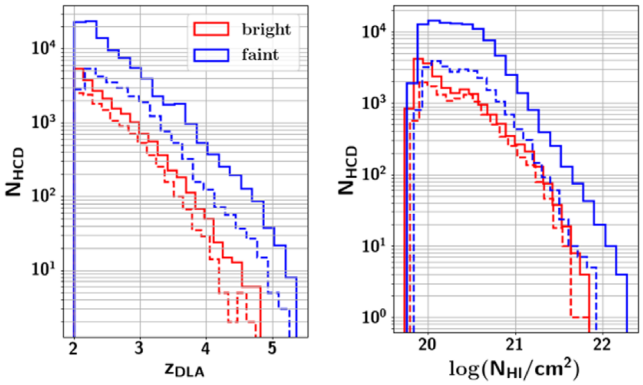

Figure 13. Distribution of z_{DLA} and $\log(N(H_I)/\mathrm{cm}^{-2})$ of the 117,458 absorbers detected by the CNN in DR16Q for the 20,375 bright forests with $\overline{f_\lambda} > 2 \times 10^{-19}$ W m^{-2} nm^{-1} (red) and the remaining 97,083 faint forests with $\overline{f_\lambda} \leqslant 2 \times 10^{-19}$ W m^{-2} nm^{-1} (blue). The dashed lines show the same samples reduced to confident absorbers only with confidence ⩾0.9.

Given that the efficiency and purity of the catalogs are functions of cuts on S/N, mean forest flux, and the DLA parameters $\log(N(H_I)/\mathrm{cm}^{-2})$ and z_{DLA}, we do not expect perfect overlap between the catalogs. This fact is illustrated in Figure 14 for N12. As for the mock study in Section 4.2, DLAs are matched if they are in the same sight line and have absorbing redshifts such that $\Delta z_{\mathrm{DLA}} < 0.02$ The figure shows their distribution of $\log(N(H_I)/\mathrm{cm}^{-2})$ for candidate DLAs that are found and not found in our catalog. The distribution is shown for the "statistical" and "nonstatistical" samples of N12.

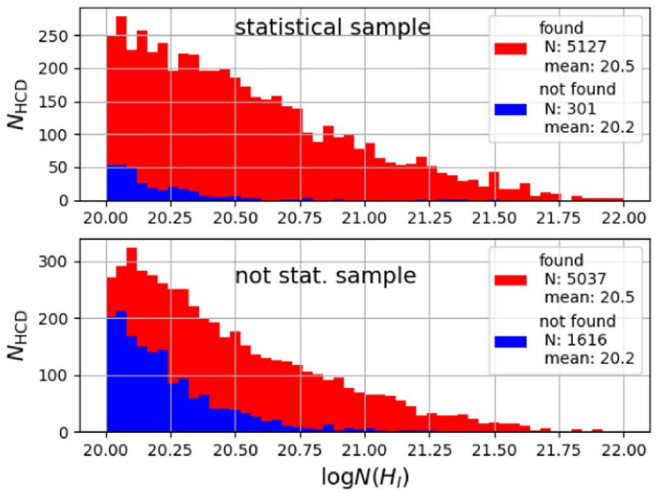

Figure 14. Distribution of $N(H_I)$ for the 12,081 DLAs of N12 found and not found by the CNN in red and blue, respectively. The statistical sample consists of confident DLA candidates with high S/N (> 3 with DR9 measurements), which are used in N12 to measure the $\log(N(H_I)/\mathrm{cm}^{-2})$ distribution and the cosmological mass density of neutral gas.

The statistical sample consists of confident DLA candidates with sufficiently high S/N to be used in N12 to measure the $\log(N(H_I)/\mathrm{cm}^{-2})$ distribution and the cosmological mass density of neutral gas. We see that the overlap is very good for the statistical sample with $\log(N(H_I)/\mathrm{cm}^{-2}) > 20.2$. On the other hand, the overlap for the nonstatistical sample is good only for $\log(N(H_I)/\mathrm{cm}^{-2}) > 20.6$.

Figure 15 shows the same distribution for the catalog of H21 where a similar behavior is seen.

Figures 16 and 17 compare the values of $\log(N(H_I)/\mathrm{cm}^{-2})$ from N12 and H21 with our values as determined by the CNN and by our fitter. The displayed samples are restricted to absorbers in the Lyα forest in order to have values of $\log(N(H_I)/\mathrm{cm}^{-2})$ for the fitter as well, i.e., 3070 for N12 and 11,438 for H21. However, the trend of the difference with $\log(N(H_I)/\mathrm{cm}^{-2})$ as predicted by the CNN (the blue curves) is similar when using the full sample of matched DLAs for both N12 and H21. For both N12 and H21, the bias with the CNN is $\log(N_{HI})$-dependent. The CNN values are typically slightly greater than N12 and H21 for $\log(N(H_I)/\mathrm{cm}^{-2}) \leqslant 20.5$ and slightly lower for $\log(N(H_I)/\mathrm{cm}^{-2}) \geqslant 20.5$, whereas the bias with our fitter is almost flat, ∼0.12 larger when compared to N12 and ∼0.05 larger when compared to H21.

H21 correctly pointed out that the DLA catalog from DR16Q (Lyke et al. 2020) does not include some obvious absorbers (see for instance Figure 19 of H21). This is due to the extremely conservative cuts on confidence >0.9 and $\log(N(H_I)/\mathrm{cm}^{-2}) > 20.3$ in the previously published DR16Q catalog. We visually inspected 24 of such spectra identified by H21 [18] and found that the majority of those absorbers are actually detected by the CNN. We also note that a few of them are in BAL sight lines so they are actually removed from our final 117,458 absorber catalog.

We make the catalog available as a FITS file.[19] There is a line for each detected absorber with $z_{QSO} > z_{DLA}$ and $z_{DLA} \geqslant 2$ in the sight line with BAL_PROB = 0. Each of the 117,458 line contains the following information:

[18] They can be found at http://tiny.cc/overlapping_dlas.

[19] https://drive.google.com/drive/folders/1UaFHVwSNPpqkxTbcbR8mVRJ5BUR9KHzA?usp=sharing

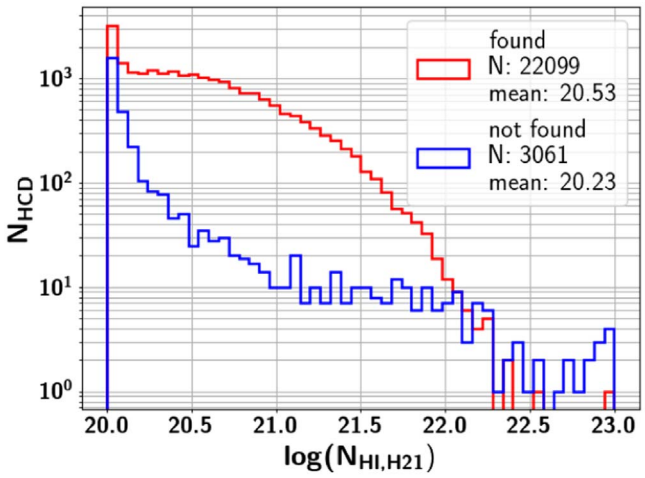

Figure 15. Distribution of $N(H_I)$ for the 25,160 confident DLAs ($p_{DLA} > 0.9$) of H21 with BAL_PROB = 0 found and not found by the CNN in red and blue respectively.

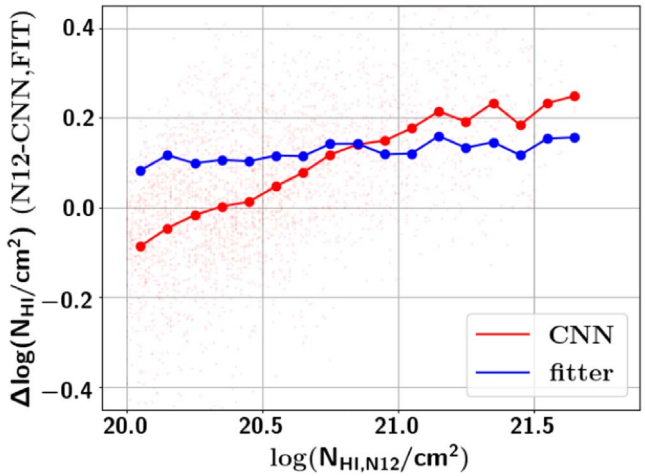

Figure 16. Comparison of the 3070 $N(H_I)$ of the DR9 absorbers in the forest, i.e., in the rest-frame range 1040 Å $\leqslant \lambda_{RF} \leqslant$ 1216 Å, found in both N12 and this study, using either the CNN (red) or the fitter (blue) result.

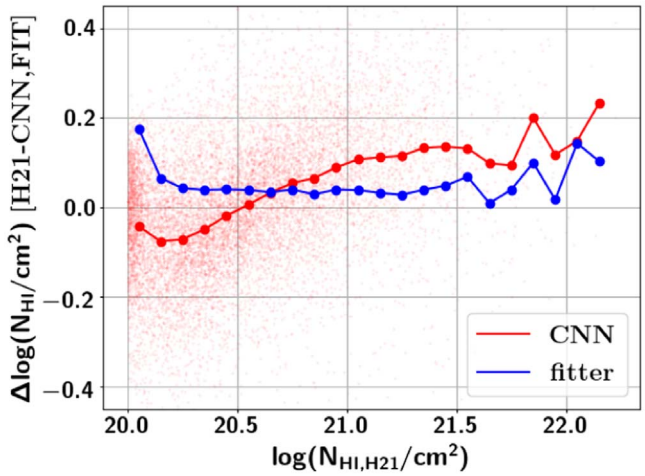

Figure 17. Same as Figure 16 for H21. The sample contains 12,449 absorbers in the forest.

1. THING_ID: the SDSS identifier as found in DR16Q
2. Z_QSO: the quasar redshift of the sight line using the Z_PCA estimator of DR16Q
3. PLATE: SDSS spectroscopic plate of the sight line as found in DR16Q
4. MJD: SDSS modified Julian date of observation of the sight line as found in DR16Q
5. FIBERID: SDSS spectroscopic fiber identification of the sight line as found in DR16Q
6. RA: R.A. of the sight line as found in DR16Q, in degrees
7. DECL.: decl. of the sight line as found in DR16Q, in degrees
8. SNR: mean S/N of the sight line
9. MEAN_FLUX: mean forest flux in $10^{-19}\,\mathrm{W\,m^{-2}\,nm^{-1}}$. The efficiency and purity of sight lines with MEAN_FLUX $> 2\times10^{-19}\,\mathrm{W\,m^{-2}\,nm^{-1}}$ are greater than 90% for absorbers with $20.1 \leqslant \log(N(H_I)/\mathrm{cm^{-2}}) \leqslant 22$.
10. Z_CNN: absorber redshift as found by the CNN
11. NHI_CNN: logarithm of the absorber column density as found by the CNN
12. CONF_CNN: confidence parameter of the CNN over the range (0,1). Absorbers with confidence >0.5 are considered as highly confident absorbers.
13. NHI_FIT: logarithm of the absorber column density as found by the Voigt profile fitter for absorbers in the rest-frame range $1040\,\mathring{A} \leqslant \lambda_{\mathrm{RF}} \leqslant 1216\,\mathring{A}$. This parameter is set to -1 for absorbers that do not meet the criteria or if the fitter could not converge on one value.

6. Conclusions

We presented here the production of the strong-absorber catalog in the 263,201 Lyα quasar spectra of the final SDSS-IV quasar catalog from DR16 (Lyke et al. 2020). We used the CNN pipeline from Parks et al. (2018) to identify absorbers and estimate their properties, z_{DLA} and $N(H_I)$. This choice was motivated by the fact that the algorithm has been constructed for low redshift and low S/N BOSS/eBOSS quasar spectra.

We performed efficiency and purity studies of the algorithm with synthetic spectra (T. Etourneau et al. 2022, in preparation) produced for the eBOSS Lyα data analysis (du Mas des Bourboux et al. 2020) that reproduce the characteristics of the data sample, in terms of the redshift and S/N distribution. The comparison between finder outputs and mock inputs showed that the algorithm performs well for confident DLAs with $2.2 \leqslant z_{\mathrm{DLA}} \leqslant 3.5$, $20.5 \leqslant \log(N(H_I)/\mathrm{cm^{-2}}) \leqslant 21.5$ and confidence parameter >0.9 with both purity and efficiency >0.9 Figure (9)). Taking only the sample of bright forests with $\overline{f_\lambda} > 2\times10^{-19}\,\mathrm{W\,m^{-2}\,nm^{-1}}$ (and no cut in the confidence parameter) increases the efficiency and purity to >0.9 values for a wider parameter range, for absorbers with $\log(N(H_I)/\mathrm{cm^{-2}}) \geqslant 20.1$ (Figure 7).

We found a bias for $N(H_I)$ toward the lowest end because the finder detects absorbers with $\log(N(H_I)/\mathrm{cm^{-2}})$ as low as 19 but overestimates this parameter just above the threshold it has been trained with. To alleviate this issue, we fit detected strong absorptions in the rest-frame range $1040\,\mathring{A} \leqslant \lambda_{\mathrm{RF}} \leqslant 1216\,\mathring{A}$ with Voigt profiles, which returns more accurate value of $N(H_I)$ than the CNN (Figure 11).

The algorithm detect 117,458 strong absorbers with $\log(N(H_I)/\mathrm{cm^{-2}}) > 19.7$ and 57,136 DLAs with $\log(N(H_I)/\mathrm{cm^{-2}}) > 20.3$, which is the largest DLA sample to date. We provided the complete results of the finder for absorbers with $z_{\mathrm{QSO}} > z_{\mathrm{DLA}}$, $z_{\mathrm{DLA}} \geqslant 2$ and in sight lines without BALs detected in the DR16Q such that BAL_PROB = 0. We also provided $N(H_I)$ information on the Voigt profile fitting for confident absorbers in the rest-frame range $1040\,\mathring{A} \leqslant \lambda_{\mathrm{RF}} \leqslant 1216\,\mathring{A}$. We compared our results to previously published catalogs from N12 and Ho et al. (2020) showing consistent findings but the CNN still appears to miss very-high-column-density absorbers, with $\log(N(H_I)/\mathrm{cm^{-2}}) > 22$, as noted by H21.

This comprehensive analysis will enable users of this catalog to construct their own selection criteria matching the needs of their study. In addition, it highlights the regimes where DLA finders need to be improved, in particular the low-S/N regime.

S.C. thanks Xavier Prochaska for his precious help and the many discussions for using the Convolutional Neural Network. S.C. thanks Simeon Bird and Ming-Feng Ho for useful discussions that greatly helped to improve the paper. The authors thank the DESI Lyα working group for providing the software to simulate the mock spectra. S.C. was partially supported by DOE's Office of Advanced Scientific Computing Research and Office of High Energy Physics through the Scientific Discovery through Advanced Computing (SciDAC) program.

This work was supported by the A*MIDEX project (ANR-11-IDEX-0001-02) funded by the "Investissements d'Avenir" French Government program, managed by the French National Research Agency (ANR), and by ANR under contracts ANR-14-ACHN-0021 and ANR-16-CE31–0021.

Funding for the SDSS IV has been provided by the Alfred P. Sloan Foundation, the U.S. Department of Energy Office of Science, and the Participating Institutions. SDSS acknowledges support and resources from the Center for High-Performance Computing at the University of Utah. The SDSS website is www.sdss.org. SDSS is managed by the Astrophysical Research Consortium for the Participating Institutions of the SDSS Collaboration including the Brazilian Participation Group, the Carnegie Institution for Science, Carnegie Mellon University, the Chilean Participation Group, the French Participation Group, Harvard-Smithsonian Center for Astrophysics, Instituto de Astrofísica de Canarias, The Johns Hopkins University, Kavli Institute for the Physics and Mathematics of the Universe (IPMU)/University of Tokyo, the Korean Participation Group, Lawrence Berkeley National Laboratory, Leibniz Institut für Astrophysik Potsdam (AIP), Max-Planck-Institut für Astronomie (MPIA Heidelberg), Max-Planck-Institut für Astrophysik (MPA Garching), Max-Planck-Institut für Extraterrestrische Physik (MPE), National Astronomical Observatories of China, New Mexico State University, New York University, University of Notre Dame, Observatório Nacional/MCTI, The Ohio State University, Pennsylvania State University, Shanghai Astronomical Observatory, United Kingdom Participation Group, Universidad Nacional Autónoma de México, University of Arizona, University of Colorado Boulder, University of Oxford, University of Portsmouth, University of Utah, University of Virginia, University of Washington, University of Wisconsin, Vanderbilt University, and Yale University.

In addition, this research relied on resources provided to the eBOSS Collaboration by the National Energy Research Scientific Computing Center (NERSC). NERSC is a U.S.

Department of Energy Office of Science User Facility operated under Contract No. DE-AC02-05CH11231.

Our analysis makes use of the following algorithms:

Software: pyigm (Prochaska et al. 2017), redvsblue https://github.com/londumas/redvsblue.

ORCID iDs

Solène Chabanier https://orcid.org/0000-0002-5692-5243
Donald P. Schneider https://orcid.org/0000-0001-7240-7449

References

Ahumada, R., Prieto, C. A., Almeida, A., et al. 2020, ApJS, 249, 3
Bautista, J. E., Busca, N. G., Guy, J., et al. 2017, A&A, 603, A12
Bird, S., Vogelsberger, M., Haehnelt, M., et al. 2014, MNRAS, 445, 2313
Blanton, M. R., Bershady, M. A., Abolfathi, B., et al. 2017, AJ, 154, 28
Cen, R. 2012, ApJ, 748, 121
Chabanier, S., Palanque-Delabrouille, N., Yèche, C., et al. 2019, JCAP, 2019, 017
Dawson, K. S., Kneib, J.-P., Percival, W. J., et al. 2016, AJ, 151, 44
du Mas des Bourboux, H., Rich, J., Font-Ribera, A., et al. 2020, ApJ, 901, 153
Font-Ribera, A., & Miralda-Escudé, J. 2012, Journal of Cosmology and Astro-Particle Physics, 2012, 028
Fumagalli, M., Fotopoulou, S., & Thomson, L. 2020, MNRAS, 498, 1951
Fumagalli, M., O'Meara, J. M., & Prochaska, J. X. 2016, MNRAS, 455, 4100
Fumagalli, M., O'Meara, J. M., Prochaska, J. X., Kanekar, N., & Wolfe, A. M. 2014, MNRAS, 444, 1282
Fumagalli, M., O'Meara, J. M., Prochaska, J. X., & Worseck, G. 2013, ApJ, 775, 78
Gardner, J. P., Katz, N., Hernquist, L., & Weinberg, D. H. 1997, ApJ, 484, 31
Garnett, R., Ho, S., Bird, S., & Schneider, J. 2017, MNRAS, 472, 1850
Gunn, J. E., Siegmund, W. A., Mannery, E. J., et al. 2006, AJ, 131, 2332
Haehnelt, M. G., Steinmetz, M., & Rauch, M. 1998, ApJ, 495, 647
Ho, M.-F., Bird, S., & Garnett, R. 2020, MNRAS, 496, 5436
Ho, M.-F., Bird, S., & Garnett, R. 2021, MNRAS, 507, 704
Kingma, D. P., & Ba, J. 2014, arXiv:1412.6980
Lee, K.-G., Hennawi, J. F., Spergel, D. N., et al. 2015, ApJ, 799, 196
Lyke, B. W., Higley, A. N., McLane, J. N., et al. 2020, ApJS, 250, 8
McDonald, P., Seljak, U., Burles, S., et al. 2006, ApJS, 163, 80
Myers, A. D., Palanque-Delabrouille, N., Prakash, A., et al. 2015, ApJS, 221, 27
Noterdaeme, P., Petitjean, P., Carithers, W. C., et al. 2012, A&A, 547, L1
Noterdaeme, P., Petitjean, P., Ledoux, C., & Srianand, R. 2009, A&A, 505, 1087
Ota, K., Walter, F., Ohta, K., et al. 2014, ApJ, 792, 34
Palanque-Delabrouille, N., Yèche, C., Borde, A., et al. 2013, A&A, 559, A85
Parks, D., Prochaska, J. X., Dong, S., & Cai, Z. 2018, MNRAS, 476, 1151
Petitjean, P., Srianand, R., & Ledoux, C. 2000, A&A, 364, L26
Pontzen, A., Governato, F., Pettini, M., et al. 2008, MNRAS, 390, 1349
Prochaska, J. X., Herbert-Fort, S., & Wolfe, A. M. 2005, ApJ, 635, 123
Prochaska, J. X., Tejos, N., Cwotta, et al. 2017, Pyigm/pyigm: Initial Release for Publications, v1.0, Zenodo, doi:10.5281/zenodo.1045480
Prochaska, J. X., & Wolfe, A. M. 1997, ApJ, 487, 73
Prochaska, J. X., & Wolfe, A. M. 2009, ApJ, 696, 1543
Ross, N. P., Myers, A. D., Sheldon, E. S., et al. 2012, ApJS, 199, 3
Rudie, G. C., Newman, A. B., & Murphy, M. T. 2017, ApJ, 843, 98
Slosar, A., Font-Ribera, A., Pieri, M. M., et al. 2011, JCAP, 2011, 001
Smee, S. A., Gunn, J. E., Uomoto, A., et al. 2013, AJ, 146, 32
Vladilo, G., Centurión, M., Bonifacio, P., & Howk, J. C. 2001, ApJ, 557, 1007
Wolfe, A. M., Gawiser, E., & Prochaska, J. X. 2005, ARA&A, 43, 861
Wolfe, A. M., Turnshek, D. A., Smith, H. E., & Cohen, R. D. 1986, ApJS, 61, 249
York, D. G., Adelman, J., Anderson, J. E. J., et al. 2000, AJ, 120, 1579
Zhu, G., & Ménard, B. 2013, ApJ, 770, 130