The Astrophysical Journal Supplement Series, 259:28 (19pp), 2022 March https://doi.org/10.3847/1538-4365/ac4504

Deep Learning of Dark Energy Spectroscopic Instrument Mock Spectra to Find Damped Lyα Systems

Ben Wang[1], Jiaqi Zou[1], Zheng Cai[1,2], J. Xavier Prochaska[3,4], Zechang Sun[5], Jiani Ding[3], Andreu Font-Ribera[6], Alma Gonzalez[7,8], Hiram K. Herrera-Alcantar[7,8], Vid Irsic[9,10], Xiaojing Lin[1], David Brooks[11], Soléne Chabanier[12], Roger de Belsunce[9,13], Nathalie Palanque-Delabrouille[14], Gregory Tarle[15], and Zhimin Zhou[16]
[1] Department of Astronomy, Tsinghua University, Beijing 100084, People's Republic of China; zcai@mail.tsinghua.edu.cn
[2] Department of Mathematics and Theories, Peng Cheng Laboratory, Nanshan, Shenzhen, People's Republic of China
[3] Department of Astronomy and Astrophysics, UCO, Lick Observatory, University of California, 1156 High Street, Santa Cruz, CA 95064, USA
[4] Kavli IPMU, the University of Tokyo (WPI), Kashiwa 277-8583, Japan
[5] Department of Physics, Tsinghua University, Beijing 100084, People's Republic of China
[6] Institut de Física d'Altes Energies (IFAE), The Barcelona Institute of Science and Technology, E-08193 Bellaterra (Barcelona), Spain
[7] Consejo Nacional de Ciencia y Tecnología, Av. Insurgentes Sur 1582. Colonia Crédito Constructor, Del. Benito Juárez C.P. 03940, México D.F., Mexico
[8] Departamento de Física, División de Ciencias e Ingenierías, Campus Leon, Universidad de Guanajuato, León 37150, Mexico
[9] Kavli Institute for Cosmology, University of Cambridge, Madingley Road, Cambridge CB3 0HA, UK
[10] Cavendish Laboratory, University of Cambridge, 19 JJ Thomson Avenue, Cambridge CB3 0HE, UK
[11] University College London, Dept. of Physics and Astronomy, Gower Street, London, WC1E 6BT, UK
[12] Lawrence Berkeley National Laboratory, Berkeley, CA, 94720, USA
[13] Institute of Astronomy, University of Cambridge, Madingley Road, Cambridge CB3 OHA, UK
[14] IRFU, CEA, Université Paris-Saclay, F-91191 Gif-sur-Yvette, France
[15] Department of Physics, University of Michigan, Ann Arbor, MI 48109, USA
[16] Key Laboratory of Optical Astronomy, National Astronomical Observatories, Chinese Academy of Sciences, Beijing, 100012, People's Republic of China
Received 2021 October 6; revised 2021 December 12; accepted 2021 December 18; published 2022 March 9

Abstract

We have updated and applied a convolutional neural network (CNN) machine-learning model to discover and characterize damped Lyα systems (DLAs) based on Dark Energy Spectroscopic Instrument (DESI) mock spectra. We have optimized the training process and constructed a CNN model that yields a DLA classification accuracy above 99% for spectra that have signal-to-noise ratios (S/N) above 5 per pixel. The classification accuracy is the rate of correct classifications. This accuracy remains above 97% for lower S/N $\approx$1 spectra. This CNN model provides estimations for redshift and H I column density with standard deviations of 0.002 and 0.17 dex for spectra with S/N above 3 pixel^{-1}. Also, this DLA finder is able to identify overlapping DLAs and sub-DLAs. Further, the impact of different DLA catalogs on the measurement of baryon acoustic oscillations (BAO) is investigated. The cosmological fitting parameter result for BAO has less than 0.61% difference compared to analysis of the mock results with perfect knowledge of DLAs. This difference is lower than the statistical error for the first year estimated from the mock spectra: above 1.7%. We also compared the performances of the CNN and Gaussian Process (GP) models. Our improved CNN model has moderately 14% higher purity and 7% higher completeness than an older version of the GP code, for S/N > 3. Both codes provide good DLA redshift estimates, but the GP produces a better column density estimate by 24% less standard deviation. A credible DLA catalog for the DESI main survey can be provided by combining these two algorithms.

Unified Astronomy Thesaurus concepts: Quasar absorption line spectroscopy (1317); Surveys (1671); Astronomy data analysis (1858)

1. Introduction

The absorption systems in the spectra of quasi-stellar objects (QSOs) are widely used to probe the properties of the early universe (e.g., Wolfe et al. 1986; Rauch 1998). Using QSO absorption line systems (e.g., Lyα and metal lines), one can probe a wide range of scales, including the gas properties inside and around galaxies (e.g., Fumagalli et al. 2011). Further, the QSO absorption line systems are used to reconstruct the cosmic web on a few tens of megaparsecs (e.g., McDonald 2003; Lee et al. 2014; Cai et al. 2016, 2017; Li et al. 2021), and test cosmological models on the cosmological scale of hundreds of megaparsecs (e.g., Pérez-Ràfols et al. 2018). Among the absorption systems, damped Lyα systems (DLAs) are a population of strong absorbers with integrated neutral hydrogen (H I) column densities $N_{\rm H\,I} > 2 \times 10^{20}$ cm^{-2} (e.g., Wolfe et al. 2005). DLAs serve as the dominant reservoirs of atomic hydrogen in the universe and offer a unique opportunity to probe the early universe (e.g., Prochaska & Wolfe 1997; Zafar et al. 2013). Their absorption can be described using the Voigt profile, which fits the damping wings driven by the natural broadening of the Lyα transition (e.g., Lee et al. 2020). Recently, DLAs are widely used to probe the circumgalactic medium (CGM) around galaxies, especially high-redshift galaxies at $z > 2$ (e.g., Gardner et al. 1997; Noterdaeme et al. 2019). Simulations show that the majority of gas that gives rise to DLAs is associated with galaxies (e.g., Rahmati et al. 2014; Bird et al. 2014; Grudić et al. 2021). A complete understanding of galaxy evolution is based on the analysis for the properties of neutral gas (e.g., Krogager et al. 2020). Besides, a large sample of QSOs and DLAs can be used for measuring a variety

of cross correlations and autocorrelations, and this could help to fit the baryon acoustic oscillations (BAO) at $z > 2$.

Previously, with the help from visual inspection, Prochaska & Herbert-Fort (2004) and Prochaska et al. (2005) searched for DLA candidates in Sloan Digital Sky Survey (SDSS) spectra by running a window along the spectra to identify the DLA as a region where the signal-to-noise ratio (S/N) is significantly lower than the characteristic S/N in the vicinity. Later, by utilizing a fully automatic procedure based on classical statistics, Noterdaeme et al. (2009, 2012a) identified DLAs in SDSS Data Release (DR) 7 and the SDSS-III Baryon Oscillation Spectroscopic Survey (BOSS). Further, with the rapid increase in the spectral data in the era of Extended Baryon Oscillation Spectroscopic Survey (eBOSS) and future surveys, the efficient and accurate detection of DLAs from low S/N spectra is becoming a technical challenge. An automated technique using Gaussian Process (GP) is applied to detect DLAs along QSO sightlines (Garnett et al. 2017; Ho et al. 2020). Recently, in Parks et al. (2018), a convolutional neural network (CNN) model was designed to detect and characterize DLAs in the QSO spectra of the SDSS and BOSS. This algorithm yields a classification accuracy of 99% on spectra with S/N above 5. The classification accuracy is defined as the proportion of results with correct predictions. The estimation for column densities and redshifts both have median values consistent with the ground truth, with a scattering of standard deviation of column density of $\sigma(\log N_{\rm HI}) = 0.15$ and redshift of $\sigma(z) = 0.002$, respectively. This CNN model is also applied to the SDSS DR16 and a DLA catalog is generated by this algorithm (Chabanier et al. 2022).

The Dark Energy Spectroscopic Instrument (DESI) is a stage IV spectroscopic survey project, and it is a 5 yr survey for galaxies, QSOs, and Milky Way stars, covering 14,000 deg^2 (DESI Collaboration et al. 2016a). The highest redshift coverage of DESI comes from QSOs. At higher redshift, DESI will use QSOs as backlights to measure clustering in the Lyα forest, the series of H I absorption lines in the spectra of distant QSOs. These absorption lines are produced by the Lyα electron transition of the neutral hydrogen (Liske et al. 1998). About 2.4 million QSO spectra are expected to be produced (Yèche et al. 2020), tracing the 3D distribution of the intergalactic gas at $z \gtrsim 2$ with a survey volume of 3 Gpc^3. Comparing with the SDSS survey, it increases the Lyα forest survey volume by 1 order of magnitude. The Lyα forest is now used to provide the BAO measurement at $z \gtrsim 2$. du Mas des Bourboux et al. (2020) showed that the forest with identified DLAs has to be specially treated when BAO analysis is conducted. DLAs will reduce the flux transmission field for the correlation estimate. The spectra pixels where a DLA reduces the transmission by more than 20% should not be used because these pixels will cause a bias in the final BAO measurement. That makes a DLA catalog indispensable for precise and accurate BAO-fitting analysis.

Our paper aims to develop a DLA finder for the DESI survey. The method adopted is based on a CNN model (Parks et al. 2018), and we improve the CNN model using DESI mock spectra. Different DESI mock spectra were chosen to make this algorithm available for a wide range of S/N levels. The minimum S/N of the mock spectra that we used is 0.31. We have updated the framework to TensorFlow2.0. Our neural network was trained on a server with two NVIDIA Tesla V100 GPUs. After developing the CNN model, we used it to get the DLA catalog for the DESI mock spectra. We have also conducted BAO analysis tests to examine the effects of different DLA catalogs with different definitions.

This paper is organized as follows. In Section 2, we introduce the basic physical features of damping wings and the mock spectra for this paper. In Section 3, the description of the training process is presented, including the data set generation, label setting, and training process. The model validation is discussed in Section 4. The comparison of the CNN model and the GP model (Ho et al. 2020) on detecting DLAs is discussed in Section 5. In Section 6, the DLA catalog is generated. Further, the comparison of the BAO-fitting results using our DLA catalog and mock catalog is quantified. All related codes are available at https://github.com/cosmodesi/desi-dlas.

2. DLA Survey

2.1. Basic Terminology of DLA

Among all the Lyα absorbers, DLAs have the highest H I column densities of $N_{\rm H\,I} \geqslant 2 \times 10^{20}$ cm^{-2}. At lower column densities, we designate absorption systems with $10^{17}{\rm cm}^{-2} \leqslant N_{\rm H\,I} \leqslant 2 \times 10^{20}$ cm^{-2} as Lyα limit systems (LLSs) including sub-DLAs with $10^{19}{\rm cm}^{-2} \leqslant N_{\rm H\,I} \leqslant 2 \times 10^{20}$ cm^{-2} or the so-called super Lyman limit systems (SLLSs; e.g., Péroux et al. 2003; Prochaska et al. 2015). At even lower column densities, these systems are called Lyα forest absorbers with $N_{\rm H\,I} \leqslant 10^{17}$ cm^{-2}, corresponding to the intergalactic hydrogen "clouds" along the QSO sightline (e.g., McQuinn 2016). The fundamental difference between DLAs and other Lyα absorbers is that hydrogen is mainly neutral in DLAs, while in all other absorption systems it is ionized (e.g., Wolfe et al. 2005).

DLAs can be fitted by the Voigt profile, which is the convolution of a Lorentz profile and a Gaussian profile (Draine 2011). According to the *Uncertainty Principle*, the energy level of an electron has a finite width. Therefore, when an electron transitions between different energy levels, the corresponding frequency also has a certain range of distribution. This broadening is called natural broadening, and it can be described by a Lorentz profile. Meanwhile, the Doppler effect causes Doppler broadening, which is fitted by a Gaussian profile when the gas satisfies a Maxwellian velocity distribution. The broadening of Lyα absorption lines is mainly caused by natural broadening and Doppler broadening (e.g., Lee et al. 2020).

At higher column densities, absorbers become optically thicker, and the Voigt profile can be characterized by a dark trough and the Lorentz damping wing, making it possible to identify individual absorbers from even a moderate S/N spectrum.

2.2. DESI Mock Spectra

2.2.1. Data Sample

The DESI mock spectra are representative of the data quality (e.g., S/N, resolution) of DESI real data. The location of DLAs and their column density are known in mock spectra, so the mock spectra can be used to train the CNN model and evaluate its performance. We choose four mock spectral database (mocks) from LyaCoLoRe mocks generated by the DESI Lyα Forest Working Group (Farr et al. 2020). To add DLAs in the mock spectra, the location of DLAs can be derived by a Gaussian field, which is used to compute the density and velocities in the spectra (Font-Ribera & Miralda-Escudé 2012).

Table 1
Information of Four DESI Mock Databases

Name[a]	Exposure Time (1000 s)	DLAs	Metals	BALs
desi-0.2-100	100	Y	Y	N
desi-0.2-4	4	Y	Y	N
desi-0.2-1	1	Y	Y	N
desiY1-0.2-DLA	multiple	Y	N	N

Note.
[a] These mocks are named: name-ver-nexp, where name is a short name to differentiate between different sets of runs, ver determines what version of systematics have been used, and nexp determines the number of exposures.

Table 2
Value of Resolution for Three Channels

Channel	Blue Channel	Red Channel	Z Channel
Wavelength (Å)	3570–5950	5625–7740	7435–9833
Δv (km s^{-1})	63.0	44.9	34.7

Then a column density is allocated to each DLA following the observed column density distribution from `pyigm`[17], and the absorption profile is calculated using a Voigt template. After these steps are done in the LyaCoLoRe mock production stage (Farr et al. 2020), DLAs can be inserted into the final synthetic spectra. Table 1 lists the information of the four mock databases we used.

The first three mocks have the same QSO catalog with different exposure times. The redshift of QSOs in the mock catalog is from 1.8 to 3.8. The continuum of these QSOs is generated using the publicly available package *simqso* as implemented in the *desisim* code.[18] The basic procedure is to generate an unabsorbed continuum for each QSO by adding a set of emission lines on top of a broken power-law continuum model. The *simqso* is based on McGreer et al. (2013); however, for these DESI mocks, the emission lines had been tuned to provide a similar mean continuum, in the Lyα and Lyβ forest region, to that observed in eBOSS DR16 (du Mas des Bourboux et al. 2020). A wider description of the *desisim* implementation to generate the continuum, as well as the full synthetic spectra production itself, including the DLA insertion, will be presented in detail in A. Gonzalez-Morales & DESI Lyman α Working Group (2022, in preparation); it is worth noticing that eBOSS DR16 used a quite similar mock set. There are many different versions of mock spectra in DESI; we choose some versions for which the mock spectra are inserted with DLAs to do the training. For the spectra we used, the marked number "0.2"[19] means that these spectra are inserted with DLAs. The last marked number (such as "100", "4", "1") stands for the exposure time these mock spectra have. The mock spectra "desi-0.2-100" have S/N levels equal to the that of the spectra with exposure time 10^5 s. This is the most noise-free sample in our paper. For the "desi-0.2-1" and "desi-0.2-4" mocks, the simulated QSO spectra have a fixed S/N level similar to that of the DESI spectra observed for one or four DESI effective times, 1000 s and 4000 s, respectively.

[17] Publicly available at https://github.com/pyigm/pyigm.
[18] https://github.com/desihub/desisim
[19] 0.0 no extra systems. 0.2 with DLAs.

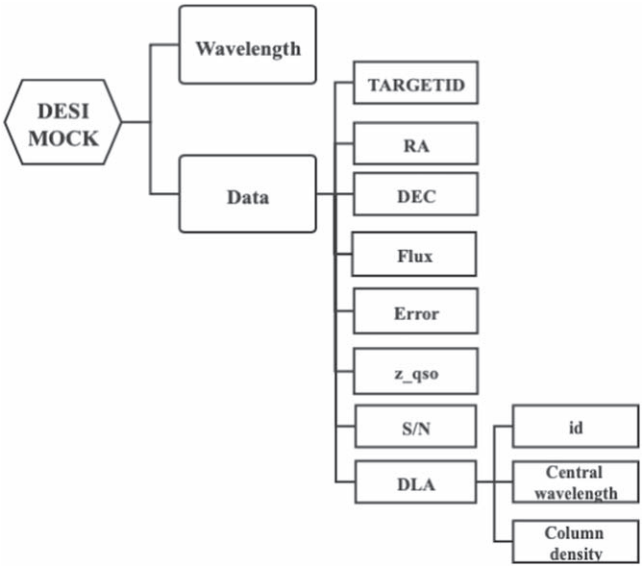


Figure 1. In the data structure of the DESIMOCK class, the wavelength array and the information of every spectrum are stored separately, including TARGETID, RA, DEC, etc.

The last mock, DesiY1-0.2-DLA is more realistic of what we would expect from the DESI first-year observations, as those were constructed using a survey simulation to determine what region of the DESI footprint (Dey et al. 2019) would be covered during DESI's first year, given a random realization of observing conditions. Such survey simulation also includes similar target-selection criteria as the main DESI survey and a simplified fiber assign procedure to reflect that high-redshift QSOs can be observed up to four times (of 1000 s each), as opposed to most targets that are observed only once, depending on what other targets are available to be observed and whether we know the QSO redshift with high significance. This procedure results in a mock spectra sample of low-redshift and high-redshift QSOs, which have a distribution of exposure times ranging from 1000 s to 4000 s, i.e, a more realistic S/N distribution. These simulations were performed using several pieces of desicode[20] and were presented in Herrera-Alcantar (2020).

The broad absorption lines (BALs) have similar profiles to DLAs. We do not simulate the BAL features in all the mock we used. This is to avoid confusion during training and reducing the false-positive predictions of the model.

2.2.2. Data Structure of DESI Spectra

The DESI spectrometer uses three cameras to measure the flux in different wavelength channels. Every spectrum in the same camera shares the same wavelength array. The three channels are given in Table 2.

We used a class named DesiMock to store the various information of every spectrum. As shown in Figure 1, using the spectrum TARGETID as the index, each spectrum contains the flux, error array, celestial coordinates, QSO redshift, S/N, and DLA information, which contains the DLA ID, DLA central wavelength, and neutral hydrogen column density ($N_{\rm H\,I}$). We read the spectral data from this DesiMock class.

[20] https://github.com/desihub/

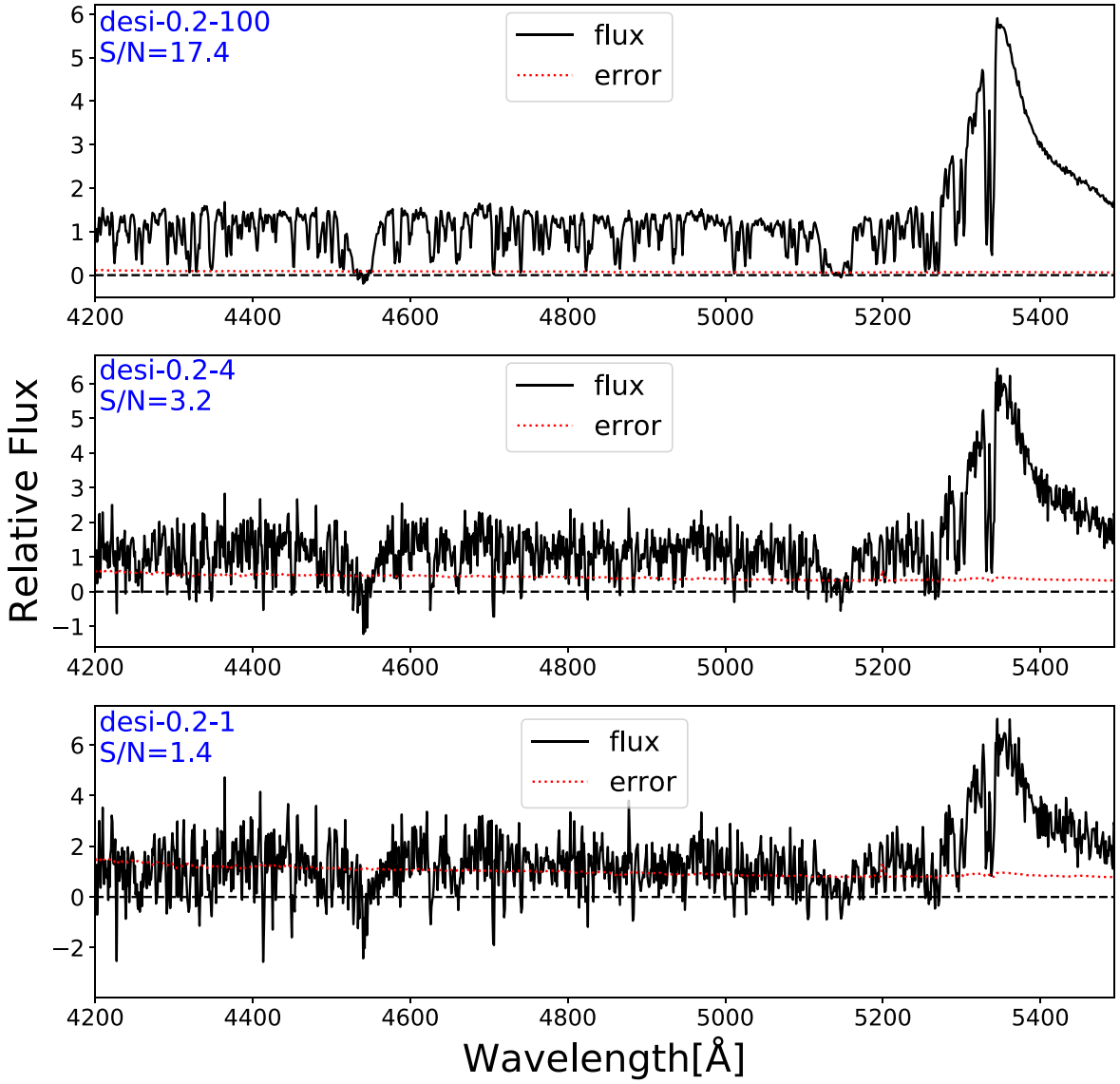


Figure 2. The three mock spectra of a QSO at $z_{\rm em} \approx 3.395$, which exhibits two DLAs at $z_{\rm abs} \approx 2.733$, 3.330 with $N_{\rm H\,I} \approx 10^{20.69}$ cm^{-2}, $10^{20.23}$ cm^{-2} that can be seen on the graph at wavelengths $\lambda_{\rm abs} \approx 4538$ Å, 5142 Å, respectively. The S/N is shown at the upper left of each panel. The black line is the flux, and the red dashed line is the error.

2.2.3. S/N Definition

Figure 2 shows three spectra with different S/N of the same QSO (Mock spectrum ID: 170257611) with emission redshift $z_{\rm em} \approx 3.395$. The S/N of the spectra in Figure 2 is estimated from the median flux of the data to the error array. Note the QSO rest frame 1420–1480 Å does not have strong line features, and thus, we define the S/N of each mock spectrum as follows:

$$\mathrm{S/N_{sightline}} = \mathrm{median}\left(\frac{\mathrm{flux}(\lambda)}{\mathrm{error}(\lambda)}\right),\ \lambda = 1420 - 1480\ \text{Å}. \quad (1)$$

The S/N definition is per pixel. The following S/N values in this paper are all per pixel. Note that the S/N of the Ly α forest is much lower than the S/N defined here. For example, for a spectrum with S/N = 3, the typical S/N in the forest region could be lower than unity.

Figure 3 displays the S/N distribution of the four mock data sets. The spectra used for training and prediction should have similar S/N ratios. Therefore, we used the first three mocks to train models for each mock, respectively, as described in Section 3. The last mock, desiY1-0.2-DLA, is used to validate the three models, as described in Section 4.

3. Training

3.1. Preprocessing

3.1.1. Rebin

Spectral rebinning consists of changing the size of the spectral bins of each spectra depending on the width of the line (Jolly et al. 2020). The dispersion $\Delta\lambda$ of the DESI mock spectra is ≈0.8 Å per pixel, and hence the resolution is not constant along the spectrum $\left(R = \frac{\lambda}{\Delta\lambda}\right)$. The instruments of DESI cover a wavelength range from 360–980 nm with resolution $R = 2000$–5500 depending on the wavelength (DESI Collaboration et al. 2016b). This means that, in the DESI spectra, the number of pixels that span a DLA feature with a

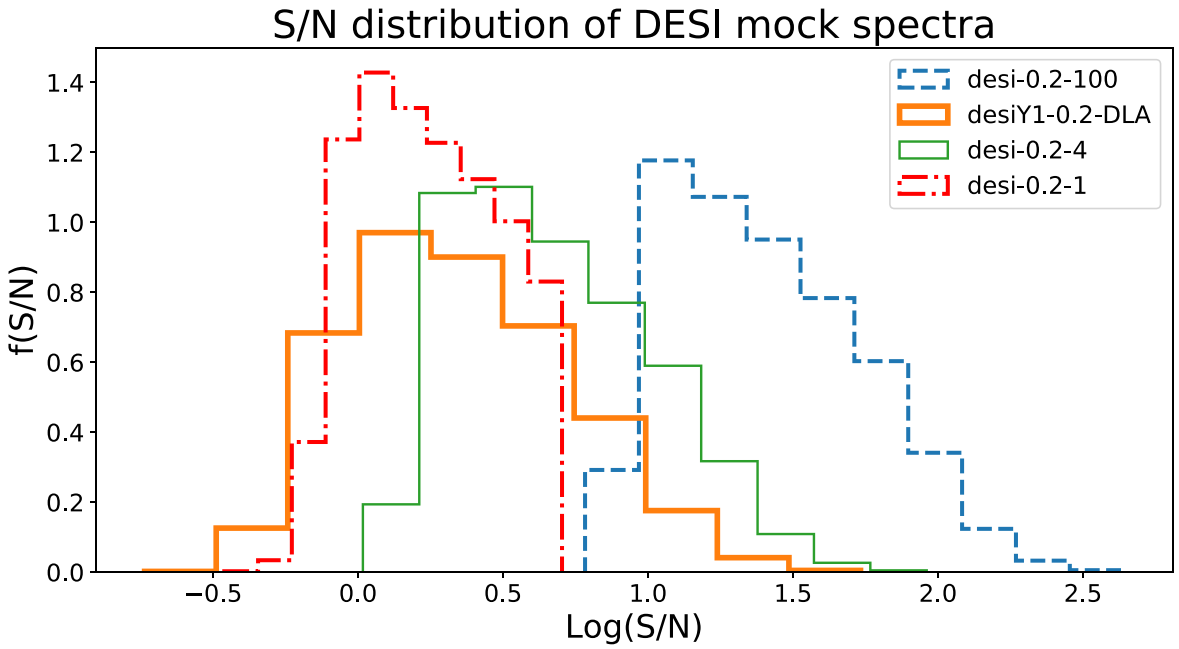


Figure 3. S/N distribution of four mocks. The three different mocks, desi-0.2-100, desi-0.2-4, and desi-0.2-1, are used to train the CNN model at different S/N. It can be seen that the mock desiY1-0.2-DLA has more spectra with S/N < 1.

given $N_{\rm H\,I}$ is proportional to the redshift. A DLA at higher redshift has more pixel numbers than a DLA with the same column density but lower redshift. This will affect the model estimation of the H I column density. To correct this effect, we have to make sure the pixel size is constant in the velocity space. Thus, we set:

$$\frac{\Delta\lambda}{\lambda} = \ln\left(1 + \frac{\Delta v}{c}\right), \tag{2}$$

where $\Delta\lambda$ represents the dispersion per pixel, and Δv represents the median pixel size in velocity. Then, we interpolate the original grid to the rebinned new grid with the pixel size equal to $\Delta v/c$. As seen in Table 2, we set the pixel size as the median velocity value in each channel.

3.1.2. Generating Data Sets for Training and Validation

Previous DLA surveys (Noterdaeme et al. 2009) first estimate the QSO continua and then normalize the flux. Nevertheless, Parks et al. (2018) claimed that the CNN learned to account for the continuum during the training. The CNN model could potentially detect DLAs without modeling the QSO continuum. Thus, we only use the median flux in the interval of 1420–1480 Å to do the normalization. Then, we construct the appropriate flux data set for training and validation. The sightlines are processed in the following paragraphs. Similar treatment can be found in Parks et al. (2018).

1. We only use a fixed range of the sightline ranging from 900 to 1346 Å in the QSO rest frame. The lower bound ensures that intervening optically thick H I gas below 900 Å does not affect the identification of DLAs, and the upper bound ensures that DLAs on or near the QSO Lyα emission can be recovered. The purpose of choosing 1346 as an upper limit is to avoid missing some high-column-density (HCD)-associated DLAs, which can block the broad-line-region emission from the QSOs (Finley et al. 2013; for example, DLAs with $N_{\rm H\,I} > 10^{22}$ and less than 1500 km s^{-1} from the QSOs redshift). Among more than 40000 DLA candidates detected by CNN in desiY1-0.2-DLA mock spectra, only 79 DLAs are located above the rest frame 1216 Å, and they can be excluded using the redshift cut.
2. Each spectrum contains more than 2000 pixels. Inputting the whole sightline directly into the model leads to difficulties in discriminating multiple DLAs. Thus, we input a sliding window of a fixed pixel region centered on each pixel into the model. The choice of the window size and the hyperparameter selection process are discussed in detail in Section 3.3.
3. As there are far more regions without DLAs than regions with DLAs, our training sets maintain a 50/50 balance between training on positive and negative regions. This means the training data sets have half regions with DLAs and half regions without DLAs. This can help to train the CNN model on both positive and negative samples.
4. Some regions of spectra are not included in the data sets. In the training set, we exclude the fixed pixel regions centered on the DLA boundary and Lyβ absorption regions. 60 pixels on the DLA boundary are avoided in the training set. We mask the 15 Å region around the Lyβ absorption. When we label the data set, the classification value changes abruptly from 1 to 0 on the DLA boundary, which confuses the model. The Lyβ absorption lines corresponding to DLAs may be incorrectly detected as DLAs coming from the lower redshift.
5. Flatten the column density distribution of SLLSs and DLAs. The dashed line in Figure 4 shows the $N_{\rm H\,I}$ distribution of the mixed mock spectra. The mixed spectral database is combined with different DESI mock catalogs, including desi-0.2-100, desi-0.2-4, and desi-0.2-1. There are far more low $N_{\rm H\,I}$ DLAs than high $N_{\rm H\,I}$ DLAs, which would induce bias toward the lower value

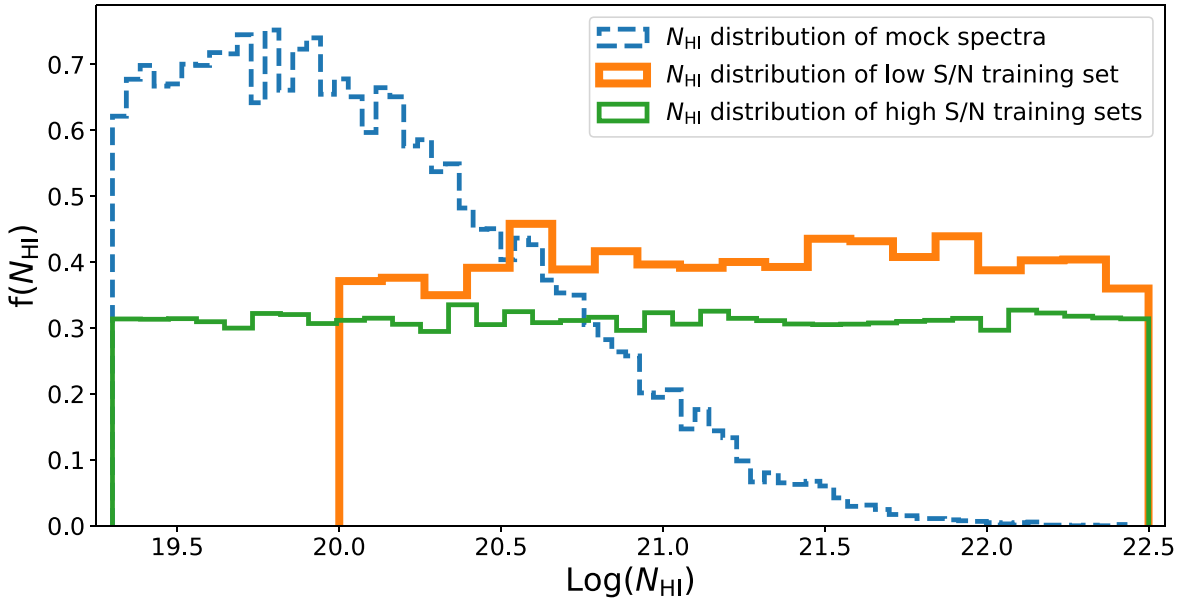


Figure 4. $N_{\rm H\,I}$ distribution of mock spectra and training sets. There are more DLAs with low-column density in the mock data, which follows the empirically measured distribution. The distribution of the column density in the training data is uniformed to reduce the bias of the model to lower $N_{\rm H\,I}$ values.

Table 3
Information of Training Sets

S/N Level	DLAs	High $N_{\rm H\,I}$ DLAs[a]	Total Number of Sightlines	Total Number of Absorbers	Mock Name
S/N 1–3	50,893	34,128	45,748	57,970	desi-0.2-1
S/N 3–6	48,068	17,099	100,000	127,347	desi-0.2-4
S/N >6	58,453	39,453	65,369	85,436	desi-0.2-100

Note.
[a] Here we point out high $N_{\rm H\,I}$ DLAs with $\log N_{\rm H\,I} > 21.0$.

Table 4
Information of Validation Sets

S/N Level	DLAs	High $N_{\rm H\,I}$ DLAs[a]	Total Number of Sightlines	Total Number of Absorbers	Mock Name
S/N 1–3	5366	3657	4938	6040	desi-0.2-1
S/N 3–6	20,365	7267	43,224	55,166	desi-0.2-4
S/N >6	6539	1360	47,424	19,917	desi-0.2-100

Note.
[a] Here we point out high $N_{\rm H\,I}$ DLAs with $\log N_{\rm H\,I} > 21.0$.

when estimating the column density in the algorithm. Thus, we manually inserted DLAs and super-Lyman-limit systems (SLLSs) into sightlines without HCD systems, following the method described in Section 4.2 of Parks et al. (2018). The redshift distribution for the inserted DLAs is uniformed to avoid bias in training. The final $N_{\rm H\,I}$ distribution of our training sets is uniform with log $N_{\rm H\,I}$ ranging from 19.3 to 22.5 for multiple exposure time mocks (desi-0.2-100 and desi-0.2-4) and from 20.0 to 22.5 for a single-exposure-time mock (desi-0.2-1), shown in the solid line of Figure 4. If we do not make the log $N_{\rm H\,I}$ uniformed, the CNN model will perform a bias for the $N_{\rm H\,I}$ estimate. The $N_{\rm H\,I}$ distribution for training is uniformed to avoid the bias.

Following these procedures, we generate DLA training samples.

In Table 3, we list the number of sightlines containing different absorbers for the training sets. With the same distribution as the training sets, we generated the validation sets for Section 4 (see Table 4).

3.1.3. Improvement on Low S/N Spectra

More than 70% of the mock spectra have S/N < 3. The classification accuracy for these low S/N spectra is only 93% using the initial model. To improve the accuracy, we used the median smoothing method to optimize the preprocessing. Smoothing the spectra reduces the resolution but improves the S/N level. Accordingly, for dealing with low S/N spectra, we set the training data as a two-dimensional array (600×4). The first row is the original flux. The other three rows are the median smoothing results for 3, 7, and 15 pixels. The CNN

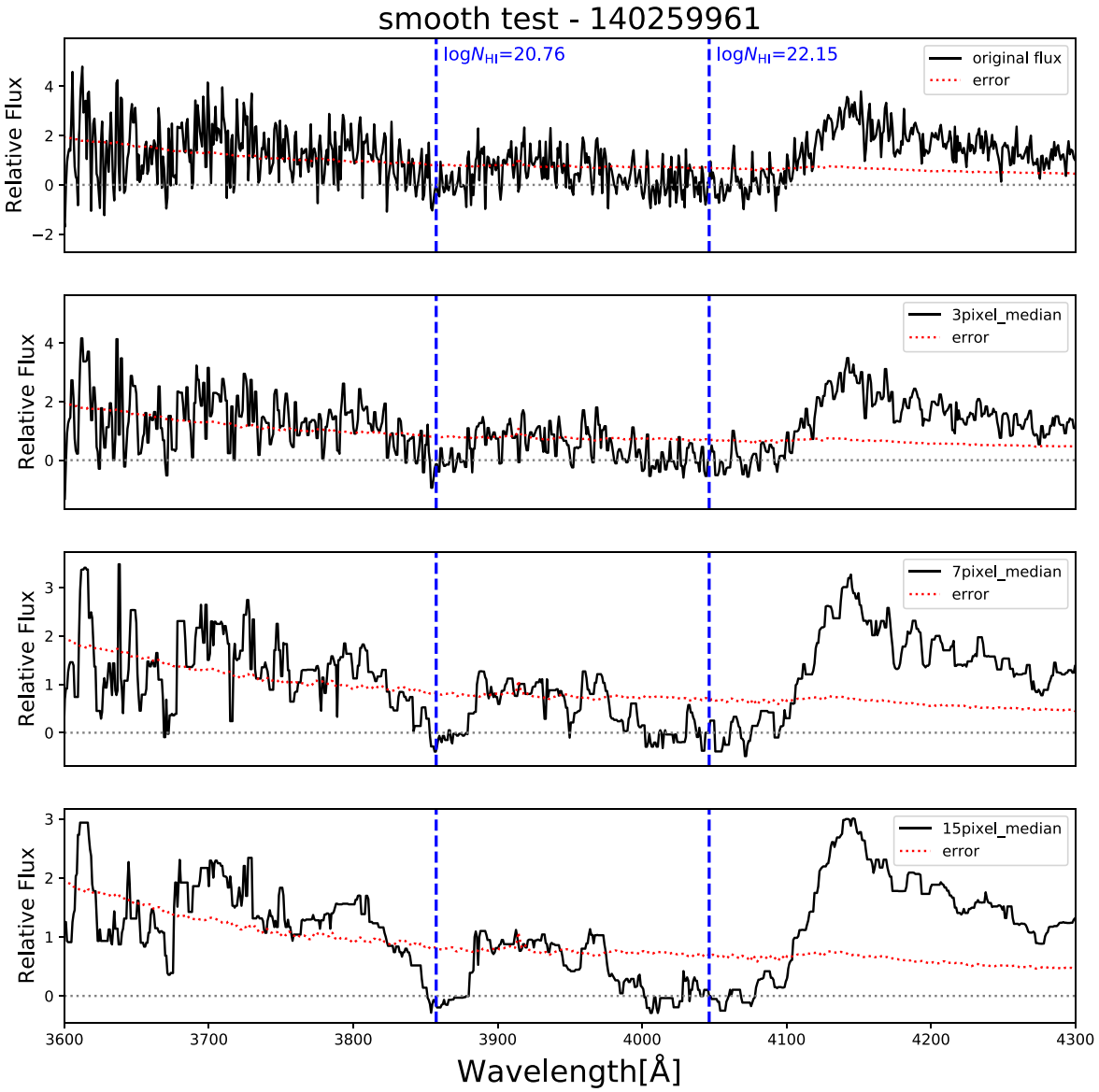


Figure 5. Smoothing result for the spectra. The blue dashed line stands for the center of DLAs. The red dashed line is the error for the spectra. The upper panel is the original flux. The lower three panels show the flux after median smoothing for 3, 7, and 15 pixels.

model is adjusted according to the new training data. The smoothing process and result are shown in Figure 5.

3.2. CNN Structure and Model Training

We followed the standard CNN architecture constructed in Parks et al. (2018). As shown in Figure 6, this model has three convolutional layers, each with a max pooling layer, following a fully connected layer and the last layer containing three separate fully connected sublayers.

We trained our model for 10^6 iterations each time. During the training process, we recorded the training accuracy every 200 steps and testing accuracy every 5000 steps. The classification accuracy improves more than 90% at the first 10^5 steps. However, it achieves the best accuracy of 99% (for $S/N > 5$ spectra; spectra with different S/N levels have different best accuracy) at about 8×10^5 steps. Then, the accuracy improves less than 0.001% for the rest of the steps. Thus, 10^6 iterations are enough for this training. The final classification accuracy for different S/N spectra are shown in Table 5. The training accuracy is the classification accuracy during training, and the definition of testing accuracy is similar. After the smoothing adjustments, the classification accuracy rises from 93% to 97% for spectra with $S/N < 3$.

The definition of the classification accuracy is based on the confusion matrix in Table 6. The label in this confusion matrix is the "pred" label, as described in Section 3.2. The classification accuracy is defined as follows:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}. \quad (3)$$

This model yields four outputs for every window of the spectrum, three labels from the full connected layers, as shown in Figure 6, and the confidence level:

1. "Prediction", labeled as "pred" in the code. This is the classification of the DLA. The value of this label is 0.0 or 1.0. 1.0 stands for that the model detects a DLA in this window, and 0.0 means no detection.

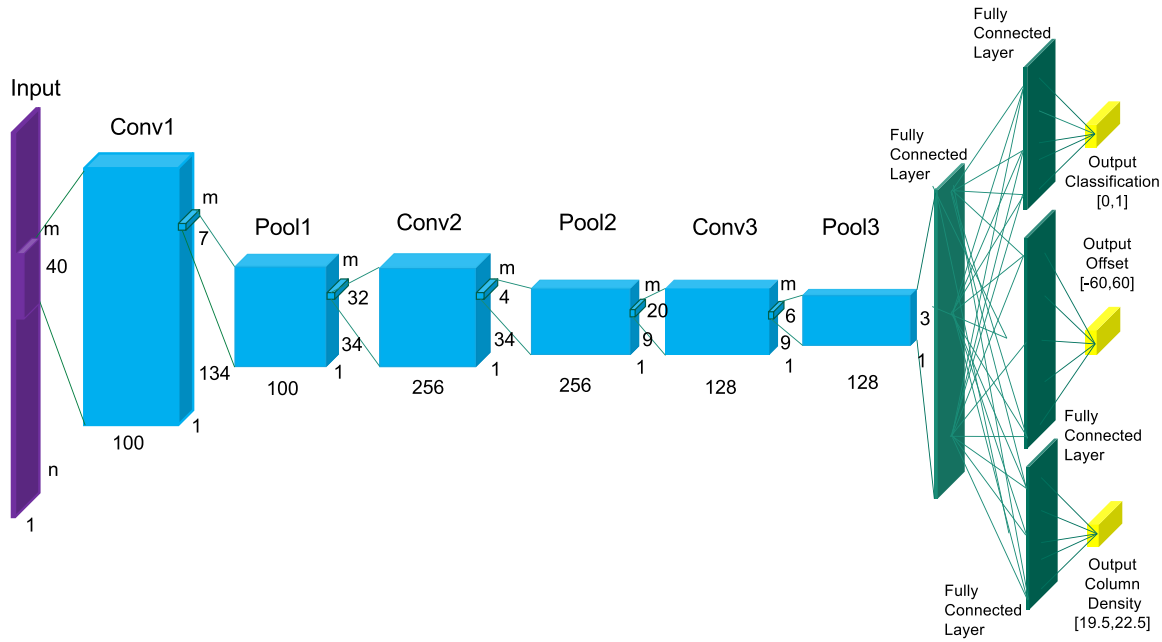


Figure 6. The standard CNN architecture we used, casting DLAs as a 1D image problem. There are three convolutional layers, three pooling layers, one fully connected layer, and three subfully connected layers. We have reset two parameters variable. One is n, it stands for how many pixels one window contains; it could be 400 or 600. The other one is m, and it stands for the dimensions of the input data; the value of it is 1 or 4 (median smoothing for low S/N spectra). The three subfully connected layers are correspond to three labels: classification, offset, and column density.

Table 5
Training and Testing Accuracy

S/N Level	Training Accuracy	Testing Accuracy
S/N > 6	99%	99%
S/N 3–6	98%	97%
S/N 1–3	94%	93%
S/N 1–3 (smooth)	97%	97%

Table 6
Confusion Matrix

Label	GroundTruth 0	GroundTruth 1
Prediction 0	True Negative (TN)	False Negative (FN)
Prediction 1	False Positive (FP)	True Positive (TP)

2. "Offset", labeled as "offset" in the code. This is similar to the generated label. It stands for the distance between the DLA center and the spectra window center. The value of this output is in the range $[-60, +60]$. If there is no DLA in this window, this label will be 0.
3. "Column Density", labeled as "coldensity" in the code. This is the predicted column density of DLAs. This label is 0 if the model predicts that no DLA lies in this window.
4. "Confidence Level", labeled as "conf" in the code. This output is not from the fully connected sublayers. It is not necessary for the prediction but will be helpful for the analysis. It is the confidence level for each prediction. It represents the possibility that one DLA is located in this window (Parks et al. 2018). There are similar concepts in Noterdaeme et al. (2012a, 2012b). The value of label "pred" (0 or 1) is determined by this "conf" label. $C_{\rm min}$ is defined as the minimum of the confidence level of a DLA. In the original set, if label "conf" is above $C_{\rm min}$, then the label "pred" will be 1. The critical value of $C_{\rm min}$ is set as 0.5. Please note that this critical value is adjustable. By changing this critical value, we can optimize the model prediction, especially for low S/N spectra.

These four different output labels are also shown in Figure 7.

3.3. Hyperparameter Search

There are 26 parameters for this CNN model that determine the size of each layer. We have conducted a hyperparameter search for these parameters. As we optimize the hyperparameters for the DESI mock spectra, our best combination of the hyperparameters is different from that of Parks et al. (2018), which is optimized for SDSS spectra.

A normal important parameter in this algorithm is the input size of the spectral window. In Parks et al. (2018), each QSO spectrum was cut into windows with the size of 400 pixels, and the classification accuracy depends on the size of the window. We have measured the classification accuracy by changing the window size from 300 to 700 pixels, with a step of 100 pixels. We find that a window size of 600 pixels, combined with the hyperparameters determined above, gives the best accuracy.

4. Validation

4.1. Purity and Completeness

Beside the classification accuracy, the purity and completeness are also important results for the CNN model. Purity and completeness are both for DLAs because TN samples are excluded. The confusion matrix definition is similar to Table 6 but without TN samples. GroundTruth stands for the label of DLAs in the mock spectra, and the prediction is the DLA label from our CNN. A DLA in mock spectra will have a

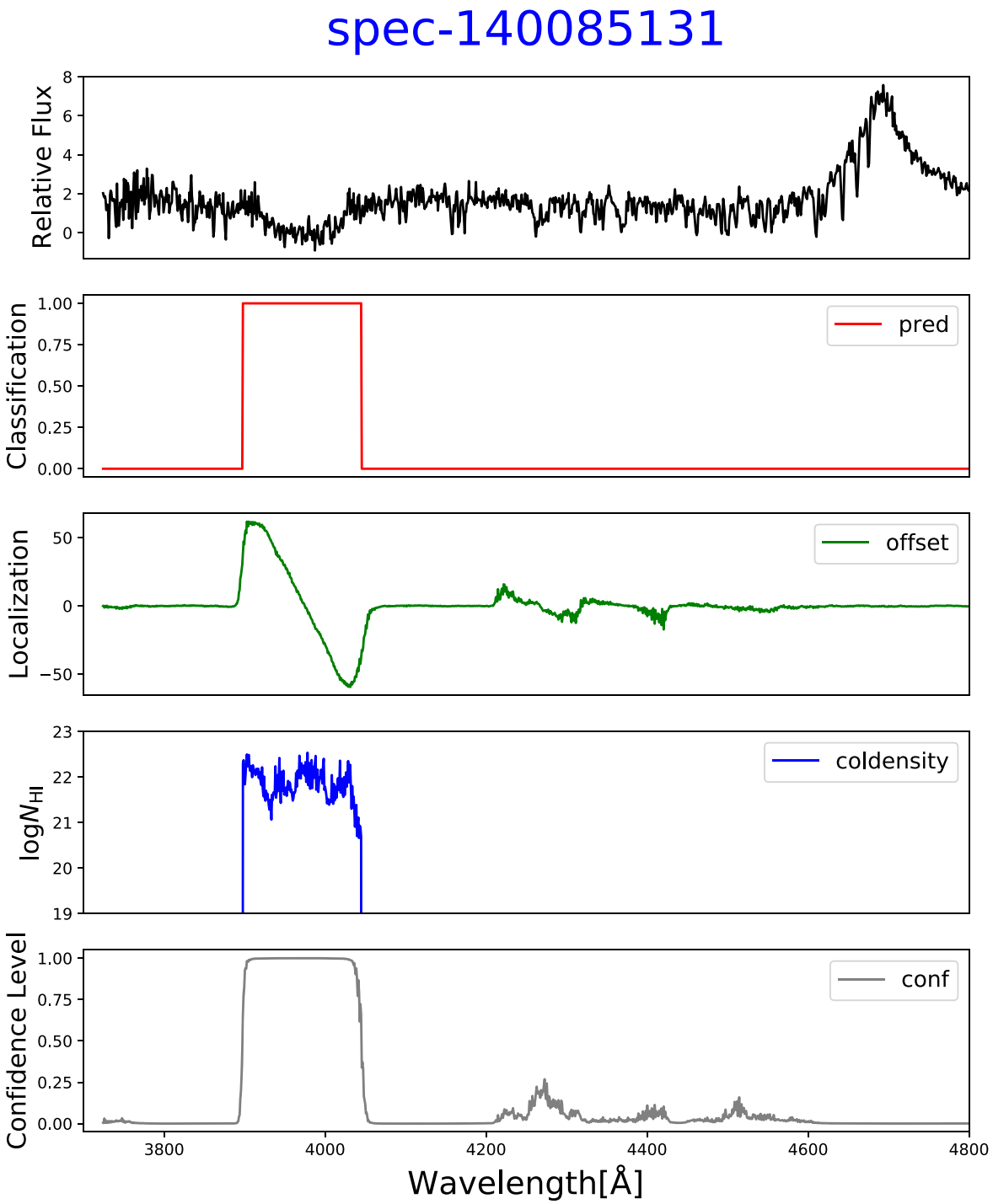


Figure 7. Four outputs for every window: "pred", "offset", "coldensity", and "conf". The red line is "pred"; the value for 120 pixels near the center of a DLA is 1 and for other pixels is 0. The green line is the "offset"; it shows the distance between every pixel and the center pixel of a DLA. The value of the offset is close to 0 for the pixels without a DLA. The blue line is the column density. The last gray line is the confidence level. This is an example with high confidence level; the confidence value for pixels with a DLA is above 0.9.

GroundTruth value of 1. Similarly, a prediction of 1 means that a DLA is detected by the CNN model. If our CNN misses one DLA, then an FN sample is produced (GroundTruth is 1 but prediction is 0). After the prediction, two DLA catalogs are generated. One is produced by the CNN model, and another one is the mock DLA catalog. For each DLA in the mock catalog, we search DLAs in the same sightline among the predicted DLA catalog and then compare the distance between the center of these two DLAs. If the distance is less than 10 Å, this is a true-positive prediction, and these two DLAs will be both marked as TP. We choose 10 Å as the critical value to make the redshift estimate for TP DLAs less than 0.008. This difference in the redshift estimate is acceptable. This critical value can make the CNN model provide an accurate redshift estimate ($<$0.008) and high classification accuracy ($>$97%) when comparing the result to the mock spectra. If the distance is above 10 Å, this is a false-negative prediction, and the DLA in the mock catalog will be marked as an FN. After that, DLAs

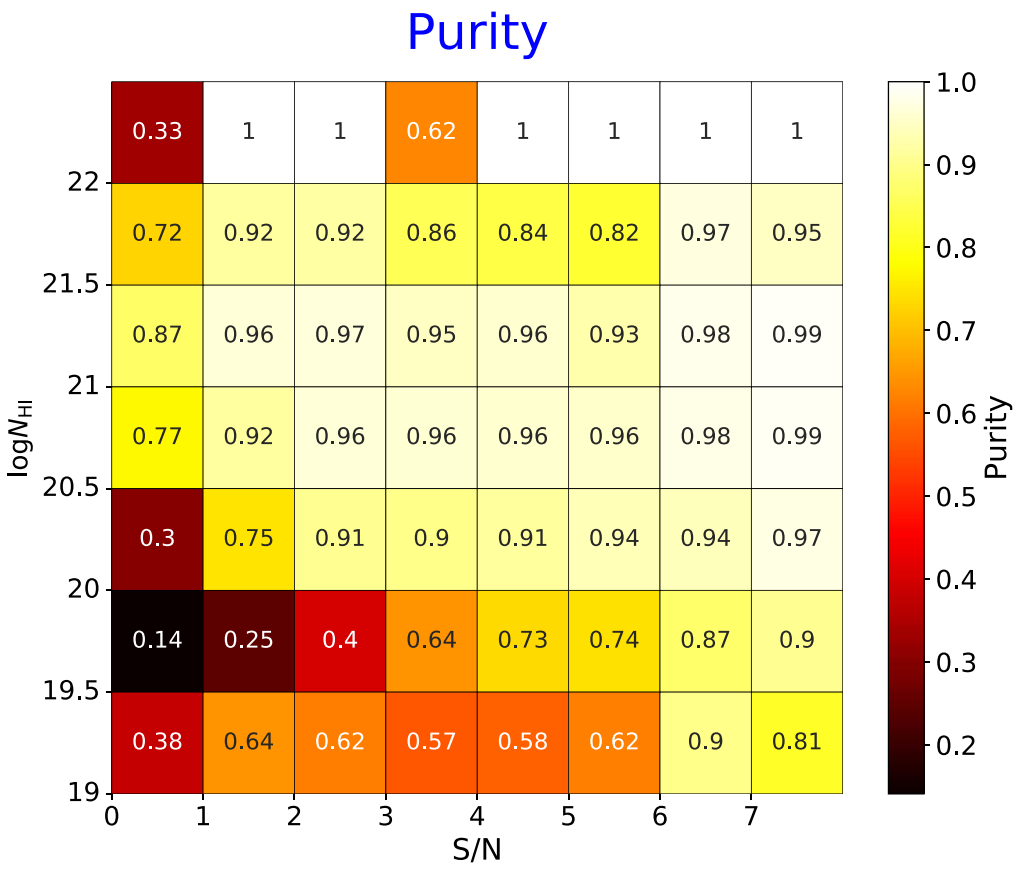


Figure 8. Purity results for different S/N levels and column densities using the desi-Y1 mock spectra. This is the result choosing the critical value of $C_{\rm min}$ as 0.5. Some bins on the top have a value of 1; this is because there are few DLAs with $\log N_{\rm H\ I}$ above 22 from the Y1 mock. For example, there are just two DLAs in the bin with S/N 4 to 5 and $\log N_{\rm H\ I}$ above 22. The DLA finder detects them both and no FP samples, so the purity goes to 1. There are just five DLAs in the bin with S/N 3 to 4 and $\log N_{\rm H\ I}$ above 22. The DLA finder detects three of them and produces two FP samples, so the purity goes down to 0.6. That is why the purity for HCD DLAs changes a lot in different bins.

in the predicted catalog without a TP label will be considered as a FP sample. The purity and completeness are defined as

$$\text{Purity} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{4}$$

$$\text{Completeness} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{5}$$

By changing the critical value of $C_{\rm min}$, we obtain different DLA catalogs. The default critical value is 0.5, as introduced in Section 3.2. Changing this critical value changes the FP rate and FN rate, and yields different DLA catalogs. The higher critical values improve the purity but reduce the completeness. To balance the purity and completeness, this value is still set as 0.5 in our model.

We have calculated the purity and completeness for different S/N and column densities. The results are shown in Figures 8 and 9. For DLAs in spectra with S/N $>$ 3, our model can achieve both purity and completeness more than 90% for almost all column density levels. Although efforts have been made to optimize the DLA identification accuracy, FN and FP cases are inevitable. The majority of these occur in spectra with low S/N ($<$3). As shown in Figure 10, these are examples that are difficult to classify even for an expert.

We also test our model on the desiY1-0.14 mock spectra. This mock spectra contains both DLAs and BALs. We show the purity and completeness in Appendix B. The completeness is still as good as the results for desiY1-0.2-DLA mock spectra. The purity drops about 10% to 20% in different bins. This is because some BALs are identified as DLAs. Nevertheless, the DESI has a formal BAL catalog, which will get rid of more than 98.6% BALs (Guo & Martini 2019) from the catalog. Then, we can run the DLA finder on the BAL-removed spectra. Therefore, we think that the purity result shown in Figure 8 is still valid.

4.2. Redshift Estimation

According to the offset label predicted by the CNN model, we locate the central wavelength of DLAs. This can be used to estimate the redshift of DLAs. The direct result for our model is the DLA location in every window. We need to transfer this result to the location in the sightline. This procedure is similar in Parks et al. (2018). We make the histogram of all the offset values, and a cluster of values at the center of a true DLA is expected. A confidence parameter for the whole spectra is further defined as the sum of the histogram over the nearest five pixels. After normalizing by the 9 pixel median filter, the maximum limit is set to one. Every detection with this confidence parameter above the critical value is considered as a DLA. Then we can calculate the redshift of DLAs according to the central wavelength.

The difference in the redshift estimation compared to the true value (value in the mock spectra) is shown in Figure 11. This result is for the "desi-0.2-100" mock spectra. The mean value of the difference is -0.00012, and the standard deviation $\sigma(z)$ is 0.002.

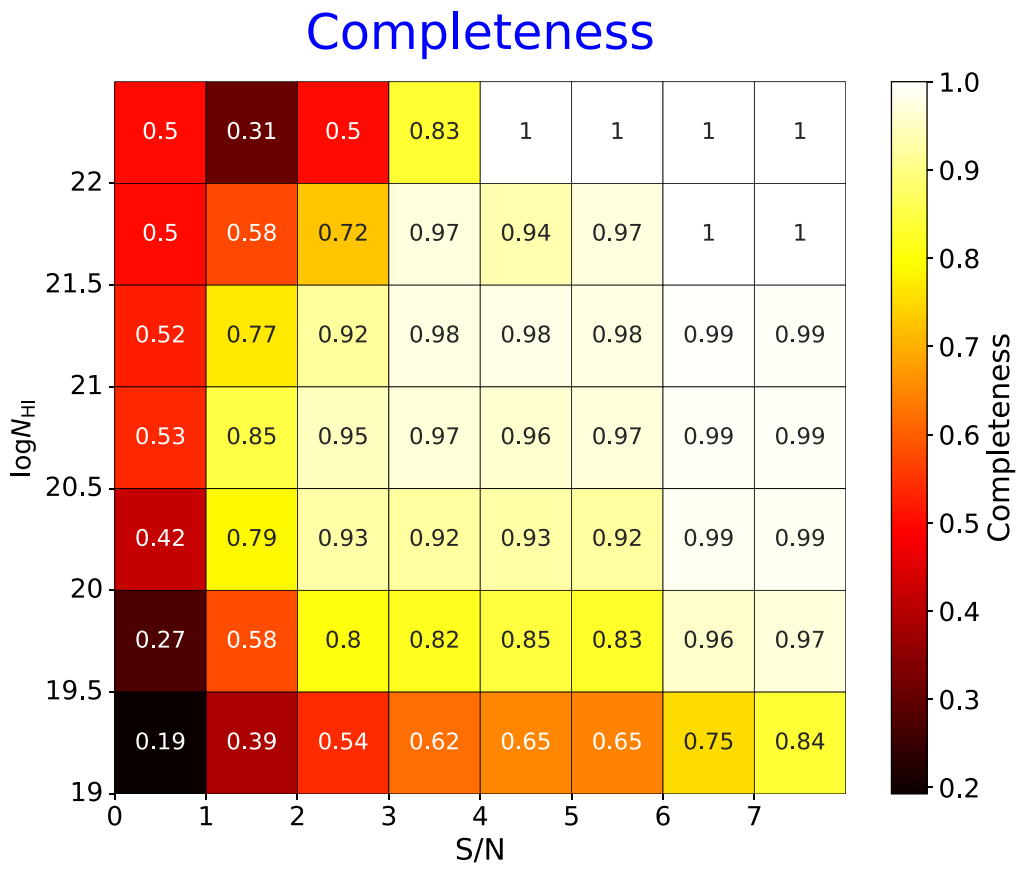


Figure 9. Completeness results for different S/N levels and column densities using the desi-Y1 mock spectra. This is the result choosing the critical value of $C_{\rm min}$ as 0.5. Some bins on the top have a value of 1; this is because there are few DLAs with $\log N_{\rm H\,I}$ above 22 from the Y1 mock. For example, there are just two DLAs in the bin with S/N 4 to 5 and $\log N_{\rm H\,I}$ above 22. The DLA finder detects them both, so the purity goes to 1. There are just five DLAs in the bin with S/N 3 to 4 and $\log N_{\rm H\,I}$ above 22. The DLA finder detects four of them and misses one DLA, so the purity goes down to 0.8. That is why the completeness for HCD DLAs changes a lot in different bins.

4.3. Column-density Estimation

Our model can also give the estimation for the column density of DLAs. For every window of the spectra, we can get an estimated value of the column density. After locating the central wavelength of a DLA, we can get the $N_{\rm H\,I}$ estimation results for the 40 pixels near the center and take the average value of these 40 $N_{\rm H\,I}$ estimates as the final estimate. The difference in the column density estimation compared to the true value is shown in Figure 12. This result is for the "desi-0.2-100" mock spectra. The mean value of the difference is -0.007, and the standard deviation $\sigma(\log N_{\rm H\,I})$ is 0.17.

5. Comparison with the GP Model

Recent advances in the GP model facilitated the investigation of detecting DLAs from the SDSS (Garnett et al. 2017; Ho et al. 2020, 2021). Ho et al. (2021) presented a DLA catalog from SDSS DR16Q, with an improved GP model. Currently, it might be the most solid approach to compare the performance between the CNN model and the GP model using DESI mock spectra with a given DLA catalog.

Here we briefly introduce the GP model based on the Bayesian model selection in Ho et al. (2020). A set of models $\mathcal{M}_i$ are developed, including the model without DLAs ($\mathcal{M}_{\rm DLA}$), the models with four DLAs ($\mathcal{M}_{\rm DLA(i)}{}_{i=1}^{4}$), and the model with sub-DLAs ($\mathcal{M}_{\rm sub}$). These models use the GP to describe the QSO emission function, a QSO's true emission spectrum $f(\lambda)$. Then they add the instrumental noise and absorption due to the intervening intergalactic medium (IGM) to obtain the observed flux as a function $y(\lambda)$. With a given spectroscopic sightline $\mathcal{D}$, they can evaluate the posterior probability of these model based on Bayes's rule:

$$\Pr(\mathcal{M}|\mathcal{D}) = \frac{p(\mathcal{D}|\mathcal{M})\Pr(\mathcal{M})}{\sum_i p(\mathcal{D}|\mathcal{M}_i)\Pr(\mathcal{M}_i)}, \quad (6)$$

where $p(\mathcal{D}|\mathcal{M})$ is the model evidence of the QSO 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.

Following the pipeline described in Ho et al. (2020), we first retrained the null model $M_{\rm DLA}$ using 70,255 spectra without DLAs in the "desiY1-0.2-DLA" mock. Then we extended the null model $M_{\rm DLA}$ to a model with k intervening DLAs, $\mathcal{M}_{\rm DLA(k)}$(k up to 4). The model prior and model evidence for these models are approximated by using the "desiY1-0.2-DLA" mock DLA catalog. Applying the new GP model, we obtain a DLA catalog of 248,512 sightlines in the "desiY1-0.2-DLA" mock. Note that the default Voigt profile used in Ho et al. (2020) includes Lyα, Lyβ, and Lyγ absorption, but we set the number of absorption lines *num_lines* to one as this mock only contains Lyα absorption lines.[21] Also we

[21] Note that the modification of this parameter may limit the performance of the GP model; Ho et al. (2021) claimed that the GP model performs better in the Lyβ forest than the CNN model, but this is not discussed in this article.

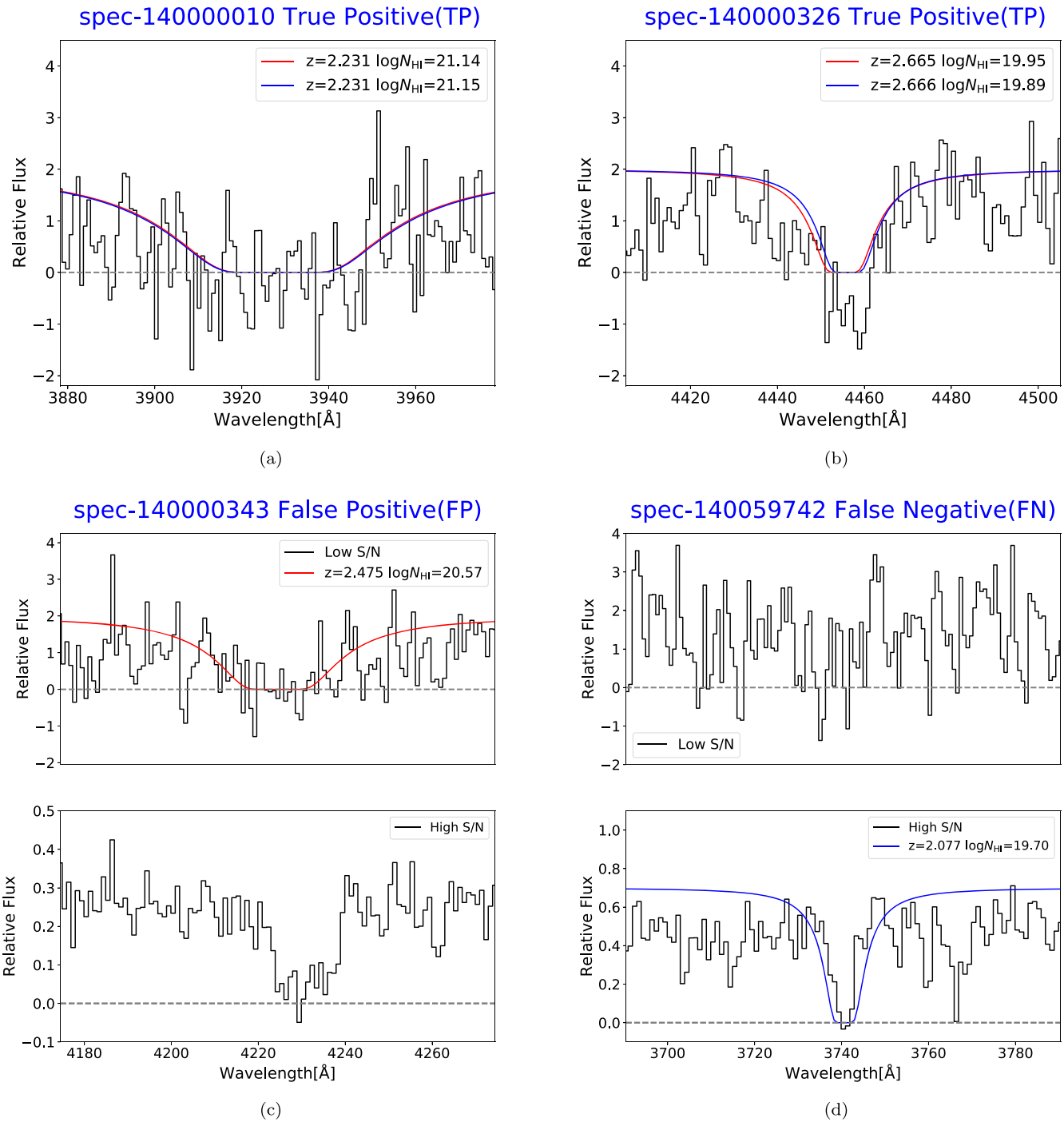


Figure 10. Validation samples: red lines are the DLAs detected by our CNN model, and blues lines are the DLAs in mock spectra. (a),(b) TP case: our model can detect DLAs with different column density levels. Even sub-DLAs can be characterized. (c) FP case: the DLA finder identifies a DLA in this window, but there is no such DLA in the mock catalog. In the lower panel of (c) is the same spectra as in the upper panel but with a higher S/N. It is clear that there is no DLA if we check the high S/N spectra. However, this is very difficult to identify even for the human being in the upper panel of panel (c) because of the low S/N. (d) FN sample: the DLA finder misses a sub-DLA in this wavelength range. In the lower panel of (d) is the same spectra as in the upper panel but with higher S/N. This shows a missing sub-DLA with low S/N(<2). The flux range for the lower and upper panels of (c) and (d) is quite different because of the random array as the noise is inserted to get a lower S/N.

modify the parameters about the minimum distance between DLAs min_*z_separation* to zero since we want to identify very close overlapping DLAs. All codes related to the GP are available at https://github.com/zoujiaqi99/GP_DLA_DESI.

With $S/N > 3$ and $\log(N_{H\,I}) > 20.0$, there are 18,613 real DLAs in the mock DLA catalog. 17,571 DLAs are predicted by the CNN model while 23,212 DLAs are predicted by the GP model. As discussed in Section 4, we also present the purity and completeness for the S/N and column density level, as shown in Figures 13 and 14. For the GP model, completeness and purity are both greater than 88% for $S/N > 3$. Note that our CNN model can measure the redshift and column density with $\log(N_{H\,I}) > 19.3$. The GP model we used provides the model posterior probability of whether the sightline containing

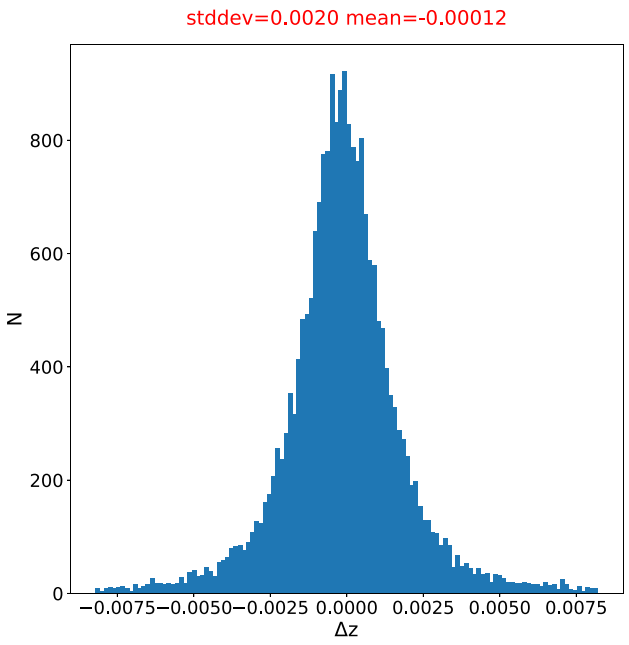


Figure 11. Redshift estimation for the DLAs matched between the CNN model and the true value in mock spectra with S/N > 3 and log($N_{\rm H\,I}$) > 20.0.

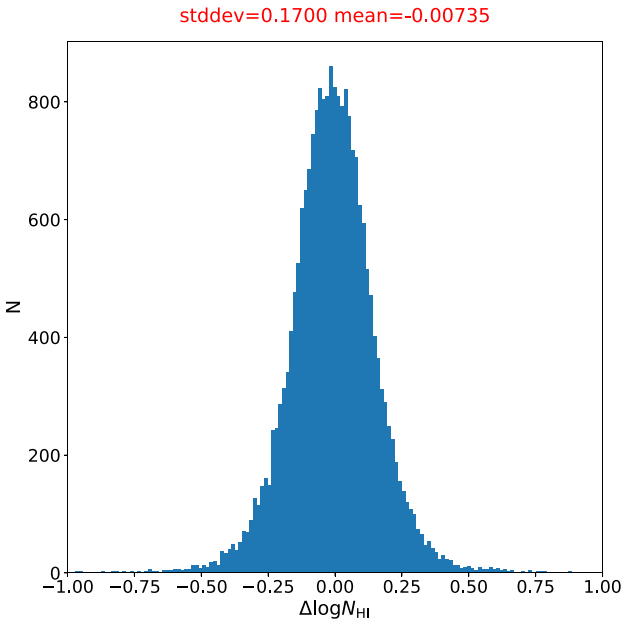


Figure 12. The column density estimation for the DLAs matched between the CNN model and the true value in mock spectra with S/N > 3 and log ($N_{\rm H\,I}$) > 20.0.

absorbers with log(NHI) < 20.0 but does not save the exact redshifts or column densities. This is due to the difference in the training sets between the CNN and GP methods. We fairly compare the two models under the same conditions in this article. However, the DESI mock spectra allow us to build data sets with low $N_{\rm H\,I}$ absorbers to retrain the GP model in order to decrease its minimum to $\log(N_{\rm H\,I}) = 19.3$.

Figures 15 and 16 show histograms of the offsets in redshift and $N_{\rm H\,I}$ between the GP model's predictions and real values in the mock catalog. The mean redshift offset is 0.00001 with a standard deviation of 0.0016. The mean log column density

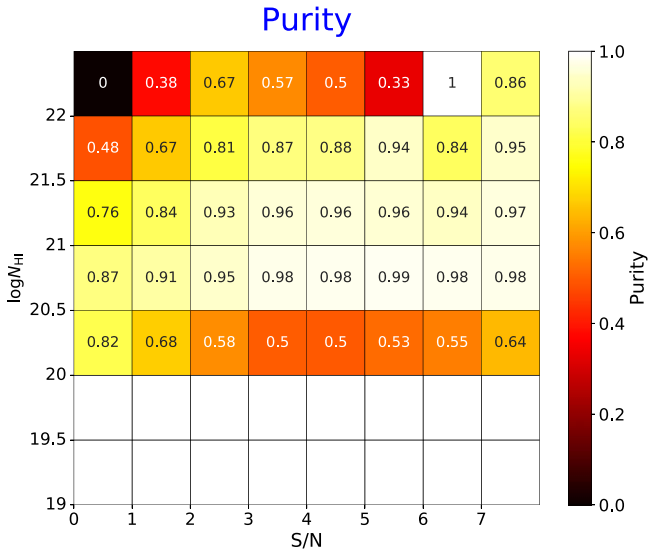


Figure 13. GP purity results for different S/N levels and column densities using the desi-Y1 mock spectra. Here we only select absorbers with $N_{\rm H\,I} > 20.0$ to provide a simple comparison because the current GP model is well developed only using $N_{\rm H\,I} > 20.0$ absorbers.

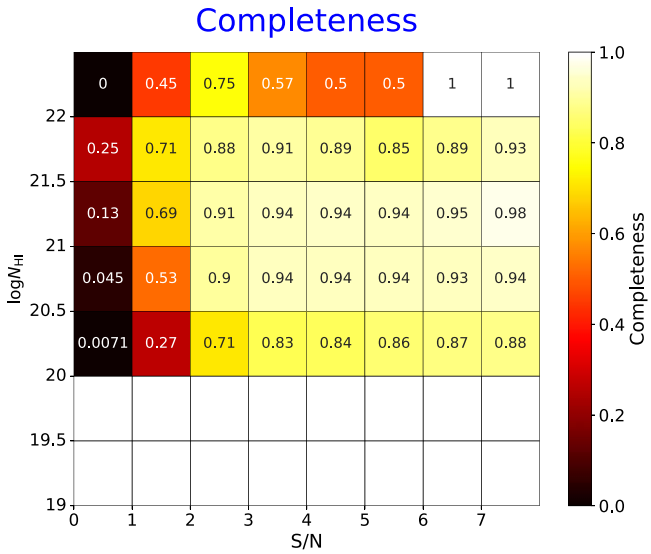


Figure 14. GP completeness results for different S/N levels and column densities using the desi-Y1 mock spectra. Here we only select absorbers with $N_{\rm H\,I} > 20.0$ to provide a simple comparison because the current GP model is well developed only using $N_{\rm H\,I} > 20.0$ absorbers.

offset is $\Delta\log(N_{\rm H\,I}) = 0.005$ with a standard deviation of 0.13 dex.

We also present the column density distribution function (CDDF) in Figure 17. This figure contains both the CNN DLA catalog and GP DLA catalog in comparison to the real mock DLA catalog with $z < 3.8$. The distribution of both models does not significantly differ from that of the real catalog. Due to the more accurate $N_{\rm H\,I}$ estimation, the GP model is in better agreement with the real catalog except for the Monte Carlo sampling boundary ($\log(N_{\rm H\,I}) = 20.0$), which induces the overdetection of absorbers with $\log(N_{\rm H\,I}) < 20.0$. Ho et al. (2020) showed the CDDF and concluded that the previous CNN model (Parks et al. 2018) fails to detect >60% of DLAs with $\log(N_{\rm H\,I}) > 21$. This problem does not exist in our results. The lack of HCD absorbers in the Parks catalog may be due to a lack of HCD systems in their training set.

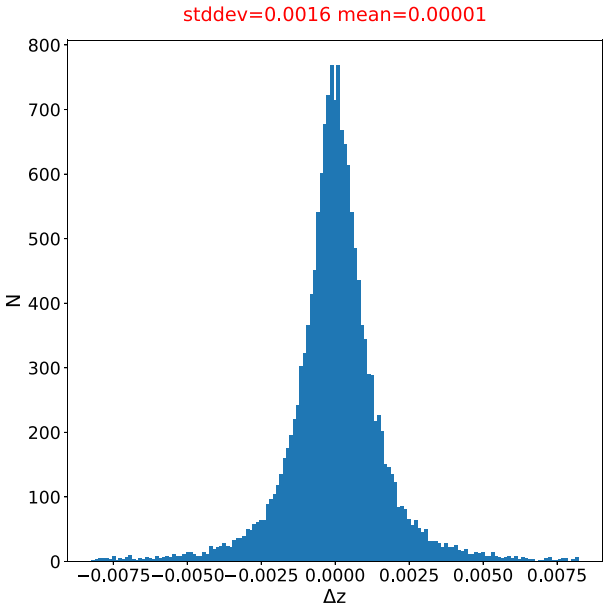


Figure 15. The redshift estimation for the DLAs matched between the GP model and the true value in mock spectra with S/N > 3 and log$N_{\rm H\,I}$ > 20.0

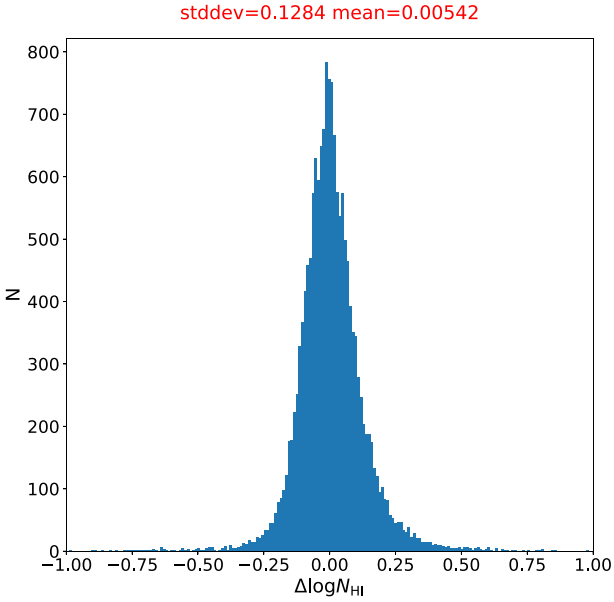


Figure 16. The column density estimation for the DLAs matched between the GP model and the true value in mock spectra with S/N > 3 and log$N_{\rm H\,I}$ > 20.0

Comparing the performance of the same mock, we find that the CNN model performs better in purity and completeness while the GP model has a more accurate estimation of the column density. In terms of the BAO measurement using the DESI Y1 mock, the CNN model is more effective, as described in Section 6. Besides, the GP model takes 8 to 10 times longer than the CNN model to predict the same data set. Consequently, a combined DLA catalog that takes the best of both models might be a better choice for the DESI real data. CNNs can be mainly used to detect DLAs and estimate the redshift because of the higher completeness and purity. The GP can be applied to further improve the column density estimate. The DLAs detected by the GP are an important part in completing

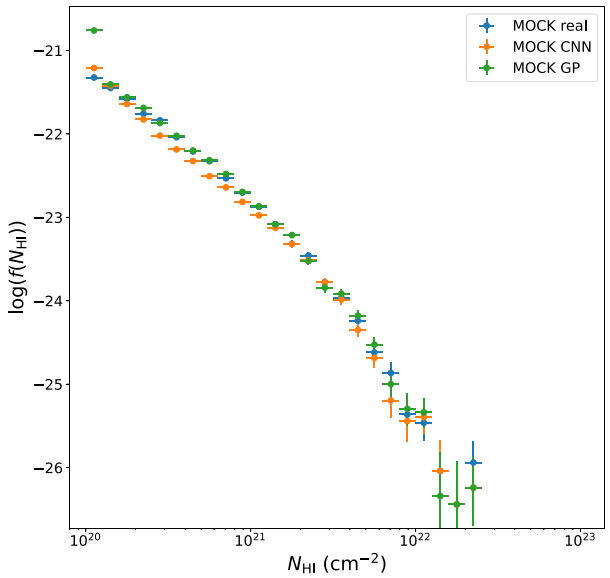


Figure 17. The CDDF from both the CNN DLA catalog and GP DLA catalog in comparison to the real mock DLA catalog with $z < 3.8$. The error bars in the y-axis represent the 68% confidence limits.

the DLA catalog. Besides, we can also provide a DLA catalog that only contains DLAs detected by both CNN and GP. The DLAs detected by both algorithms may be a smaller sample but with high confidence.

6. BAO-fitting Analysis

6.1. BAO-fitting Procedure

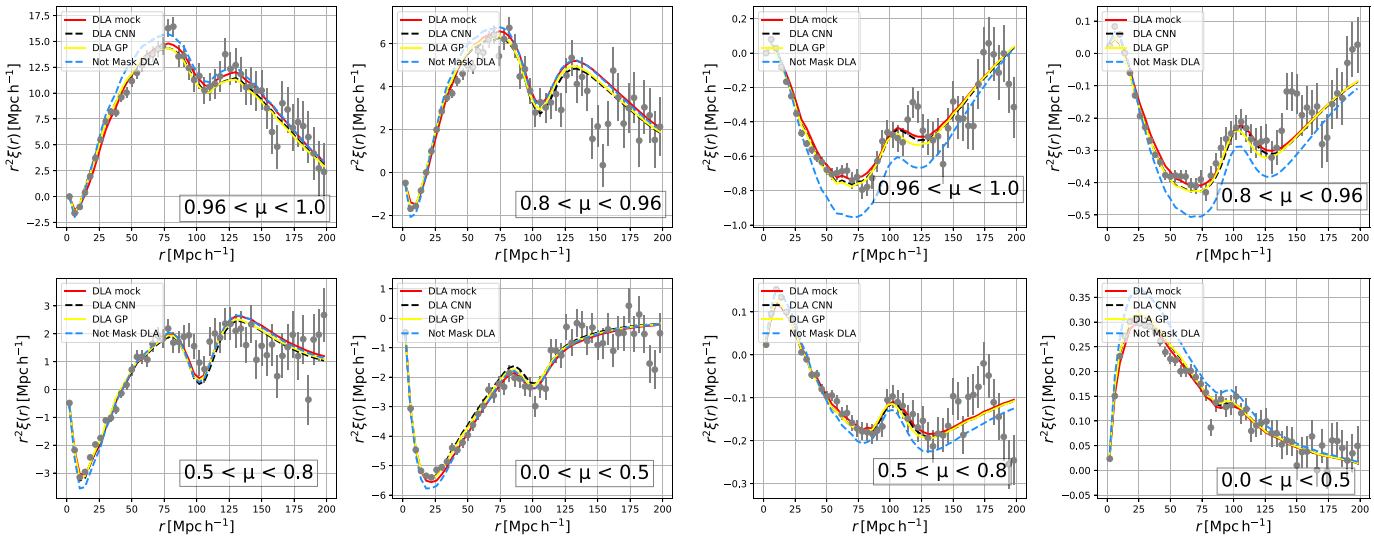
After the DLA catalogs are generated, we tested the influence of different DLA catalogs on the measurement of BAO fitting. Two different DLAs catalogs are generated. One catalog contains all the DLAs in the mock spectra; we will take this catalog as the real DLA catalog. Another catalog only contains the DLAs detected by our CNN model. The BAO fitting is conducted using the Lyα-QSO cross correlation and Lyα–Lyα autocorrelation. According to the definition of du Mas des Bourboux et al. (2020), the flux-transmission field is

$$\delta_q(\lambda) = \frac{f_q(\lambda)}{F(\lambda)C_q(\lambda)} - 1, \tag{7}$$

where $f_q(\lambda)$ is the observed flux, and $F(\lambda)C_q(\lambda)$ is the mean expected flux. The BAO-fitting procedure is followed by the pipeline in du Mas des Bourboux et al. (2020). For the Lyα–QSO cross correlation and Lyα–Lyα autocorrelation, the first step is to perform the continuum fitting. In this step, we obtain the flux-transmission field from the observed flux and the mean expected flux. DLAs should be masked in this step.

The next steps and the meaning of parameters are described in detail in du Mas des Bourboux et al. (2020). The pipeline we use for these steps is called PICCA,[22] which was developed by the eBOSS Lyα working group (du Mas des Bourboux et al. 2020). This pipeline can mask DLAs for the BAO fitting if we provide a DLA catalog as an input.

[22] https://github.com/igmhub/picca



(a) (b)

Figure 18. BAO-fitting result: (a) QSO–Lya cross correlation and (b) Lya–Lya autocorrelation. The red line is the best-fit model by masking the DLAs in the mock catalog. The black dashed line is the best-fit model by masking DLAs detected by our CNN model. The blue dashed line is the best-fit model if we do not mask any DLA. The yellow dashed line is the best-fit model by masking DLAs detected by the GP. The gray points are the data points from the analyze by masking the DLAs in the mock spectra.

Table 7
Best Fitting Parameters

Parameters	DLA Mock	DLA CNN	DLA GP	No Mask
$\alpha_{\parallel}$	0.981	0.983	0.977	0.992
σ	0.0173	0.0174	0.0195	0.0212
$d_{\alpha_{\parallel}}$		0.16%	0.41%	1.12%
ratio		0.116	0.209	0.636
$\alpha_{\perp}$	1.019	1.025	1.025	1.028
σ	0.0172	0.0177	0.0194	0.0227
difference		0.61%	0.61%	0.94%
ratio		0.349	0.304	0.523
χ^2/DOF	1.081	1.091	1.098	1.136

In this fitting, we have set the following parameters in PICCA as free parameters to fit. HCD systems are also considered in the fitting:

1. $\alpha_{\parallel}$, $\alpha_{\perp}$: BAO-peak position parameters.
2. $b_{\mathrm{Ly}\alpha}$, $\beta_{\mathrm{Ly}\alpha}$: bias parameters for Ly$\alpha$ absorption.
3. b_{HCD}, β_{HCD}: bias parameters for HCD systems.

Other parameters in PICCA has been set as fixed parameter.

6.2. BAO-fitting Results

We have used the "desiY1-0.2-DLA" mock spectra to do the BAO-fitting analysis. These mock spectra contain DLAs and HCDs but do not have metal components.

These mock spectra contain 212,238 sightlines with 36,212 DLAs and 43,909 sub-DLAs. Our CNN model has detected 43,530 DLA candidates and 19,687 sub-DLAs. 38,410 DLA candidates are detected by GP. The sub-DLAs in the mock catalog are added to the GP catalog when we do the BAO fitting. FP and FN samples are inevitable for the lower S/N spectra (<3). We have conducted the Lyα–QSO cross correlation and Lyα–Lyα autocorrelation by masking these two different DLA catalogs. We also plot the fitting result if we do not mask any DLAs and mask the DLA catalog generated by the GP. which is described in Section 5. The results are shown in Figure 18.

After masking DLAs, the reduced χ^2 of the fitting decreases from 1.136 to 1.081. The impact of DLAs on BAO fitting is obvious in the autocorrelation fitting. The BAO-fitting result from the mock DLAs catalog was set as the ground truth. The two parameters describing the position of the BAO peak (Busca et al. 2013), $\alpha_{\parallel}$ and $\alpha_{\perp}$ are estimated to quantify the fitting results. The best-fit parameters are shown in Table 7. This table also contains the standard deviation for $\alpha_{\parallel}$, $\alpha_{\perp}$, and the reduced χ^2 in two different fittings. The difference of the results can be quantified by the following equations:

$$d_{\alpha_{\parallel}} = \frac{|\alpha_{\parallel}(\mathrm{mock}) - \alpha_{\parallel}(\mathrm{pred})|}{\alpha_{\parallel}(\mathrm{mock})}, \tag{8}$$

$$d_{\alpha_{\perp}} = \frac{|\alpha_{\perp}(\mathrm{mock}) - \alpha_{\perp}(\mathrm{pred})|}{\alpha_{\perp}(\mathrm{mock})}, \tag{9}$$

where $a_{\parallel}$(mock) is the fitting result of the $\alpha_{\parallel}$ parameter by using the mock DLA catalog, and $a_{\parallel}$(pred) stands for the result using our DLA catalog. Similarly, $a_{\perp}$(mock) and $a_{\perp}$(pred) are the fitting values of the $\alpha_{\perp}$ parameter using the mock DLA catalog and our DLA catalog. We also define the ratio of difference and error as follows:

$$\mathrm{ratio}_{\alpha_{\parallel}} = \frac{|\alpha_{\parallel}(\mathrm{mock}) - \alpha_{\parallel}(\mathrm{pred})|}{\sigma\alpha_{\parallel}(\mathrm{mock})}, \tag{10}$$

$$\mathrm{ratio}_{\alpha_{\perp}} = \frac{|\alpha_{\perp}(\mathrm{mock}) - \alpha_{\perp}(\mathrm{pred})|}{\sigma\alpha_{\perp}(\mathrm{mock})}. \tag{11}$$

The difference between these two fitting results is less than 0.61%. This difference is lower than the statistical error using the DESI first-year mock spectra (above 1.7%; shown in Table 7). The statistical error will be reduced in the next 4 yr DESI survey. With increasing S/N in the next few years, the

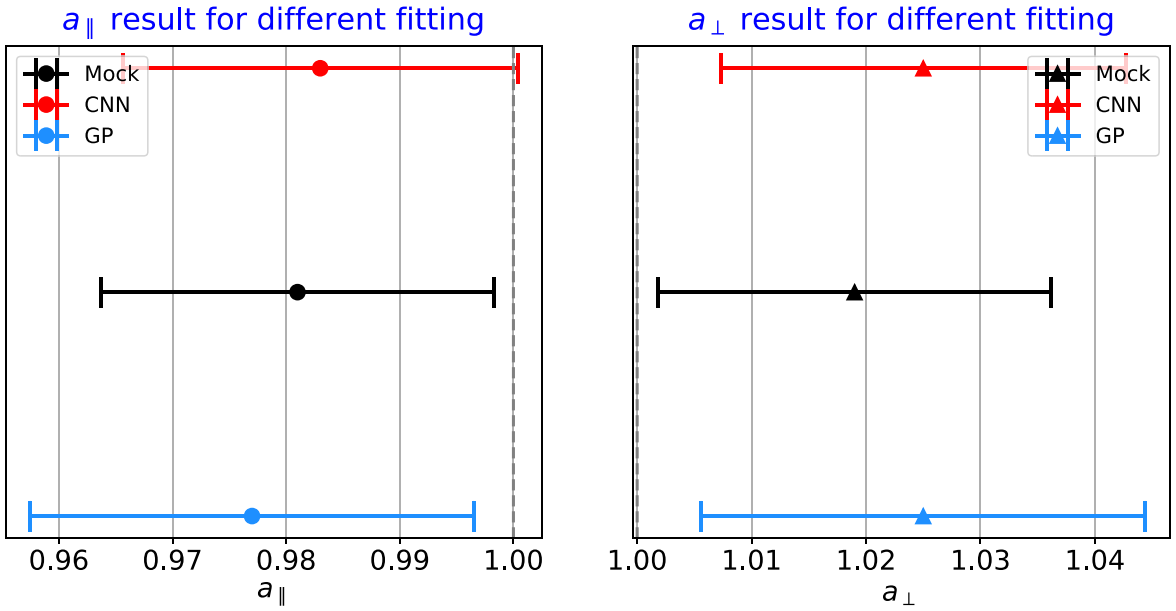


Figure 19. The value and error bar for three different fittings: masking DLAs in the mock catalog, masking DLAs detected by the CNN, and masking DLAs detected by the GP. It is clear that the result from masking the GP DLA catalog has larger difference from that of the mock DLA catalog. The gray line is the theoretical value for these two parameters.

Table 8
Best-fit Parameters

Parameters	DLA Mock	DLA CNN	DLA GP	No Mask
$\alpha_\parallel$	0.981	0.983	0.977	0.992
σ	0.0173	0.0174	0.0195	0.0212
$\alpha_\perp$	1.019	1.025	1.025	1.028
σ	0.0172	0.0177	0.0194	0.0227
$\mathrm{beta_{LYA}}$	1.5603	1.5515	1.6684	1.4066
σ	0.0259	0.0265	0.0584	0.0214
$\mathrm{bias_{LYA}}$	−0.1359	−0.1391	−0.0755	−0.1581
σ	0.0013	0.0014	0.0018	0.0015
$\mathrm{bias_{eta}(LYA)}$	−0.2186	−0.2224	−0.2077	−0.2292
σ	0.0021	0.0022	0.0028	0.0021
$\mathrm{beta_{HCD}}$	0.8237	0.8703	0.8710	0.8891
σ	0.0689	0.0685	0.0850	0.0657
$\mathrm{bias_{HCD}}$	−0.0591	−0.0621	−0.0701	−0.0795
σ	0.0022	0.0022	0.0022	0.0028
χ^2/DOF	1.081	1.091	1.098	1.136

DLA finder can give a better detection than that in the first year, and this difference can be further reduced. The DLA catalog generated by our CNN model can be used for the BAO-fitting analysis.

We also compare the performance on correlation fitting using the DLA catalog generated by the GP. The difference in the best-fit parameters between the GP and the mock shown in Table 7 is about 0.25% larger than the difference between the CNN and the mock, but this difference is still less than the statistical error. So the BAO-fitting result for the mock spectra is not affected much in any of the analyses presented. The CNN DLA catalog can give a better BAO analysis support because of the higher completeness. The difference in the fitting is clear when checking the difference in the best-fit parameter values. We have also plotted the result in Figure 19. Details about more parameters are shown in Table 8. This BAO analysis result is based on the mock spectra for DESI's first-year survey. We plan to use different mock spectra to do the fitting comparison in the future.

7. Conclusion

In this article, we have applied deep-learning techniques on QSO spectra to classify and characterize DLAs. We improved the CNN model created by Parks et al. (2018) using DESI mock spectra to make it work successfully on the DESI spectra. We optimized the preprocess, training procedure, parameter selection, and performance on low S/N and HCD DLAs. Our model can give effective and accurate estimation of the redshift and column density. We also improved the performance of the CNN on low S/N spectra by smoothing the input flux; this method may also be used to other algorithms for low S/N signals. This CNN model can detect DLAs even when the S/N of DLAs is only about 1. We believe that there is still room to further improve this algorithm.

Besides, our DLA catalog can also help perform BAO analysis. When we want to use correlation to perform BAO fitting, a DLA catalog is necessary. The results produced by our DLA catalog are very close to the real mock results, within a difference of about 0.61% for the best-fit parameters.

Finally, we compare our CNN DLA finder with the GP model from Ho et al. (2020) based on DESI mock spectra. Note that it may not be possible to show all the advantages and disadvantages of the two models due to the limitation of the mock. However, it is sufficient to see the differences between the two models by comparing their predictions with the mock truths. The CNN model has a higher completeness and purity on detecting DLAs. The fitting differences using the CNN and GP DLA catalogs are less than the statistical error on BAO fitting using the Y1 mock spectra. The CNN results yield a systematic uncertainty of 0.25%, less than that of GP. The BAO-fitting result is not affected much by using either DLA finders. Both algorithms can estimate the redshift of DLAs well while the GP model has more accurate column density estimation. Combining the catalogs given by the two models,

we can obtain a credible DLA catalog that can be widely used for real DESI spectra release.

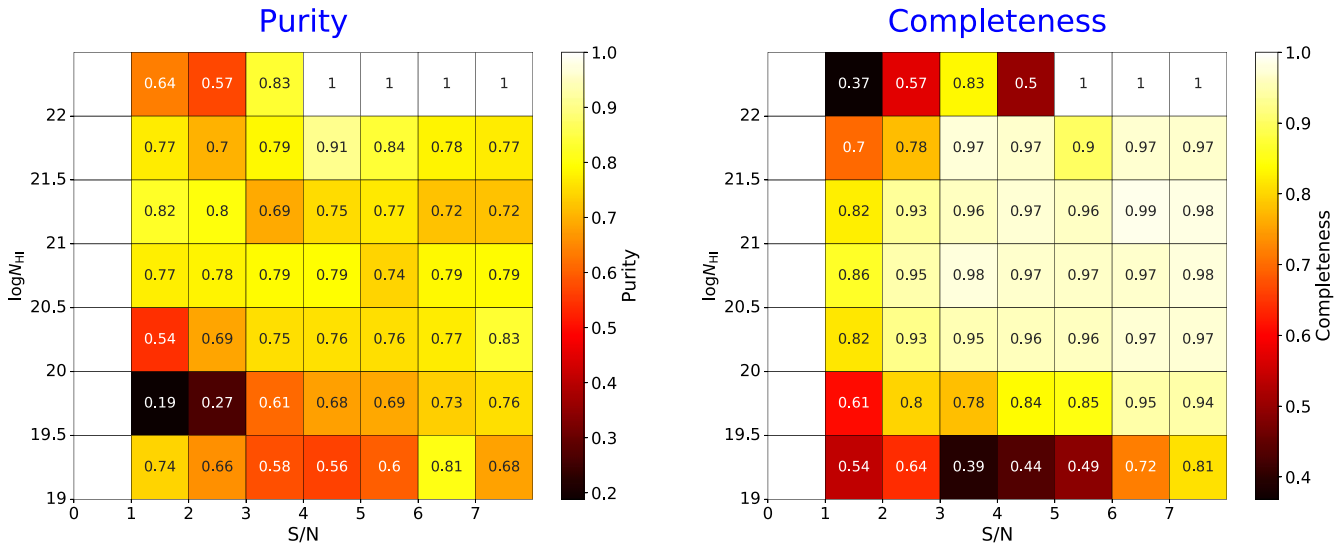
This research is supported by the Director, Office of Science, Office of High-Energy Physics of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231, and by the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility under the same contract; additional support for DESI is provided by the U.S. National Science Foundation, Division of Astronomical Sciences under Contract No. AST-0950945 to the NSF's National Optical-Infrared Astronomy Research Laboratory; the Science and Technologies Facilities Council of the United Kingdom; the Gordon and Betty Moore Foundation; the Heising-Simons Foundation; the French Alternative Energies and Atomic Energy Commission (CEA); the National Council of Science and Technology of Mexico; the Ministry of Economy of Spain, and by the DESI Member Institutions. The authors are honored to be permitted to conduct scientific research on Iolkam Du'ag (Kitt Peak), a mountain with particular significance to the Tohono O'odham Nation. B.W., J.Z., Z.C. are supported by the National Key R & D Program of China (grant No. 2018YFA0404503), the National Science Foundation of China (grant No. 12073014). A.F.R. acknowledges support by the FSE funds through the program Ramon y Cajal (RYC-2018-025210) of the Spanish Ministry of Science and Innovation. V.I. is supported by the Kavli foundation. B.W., J.Z., Z.C. acknowledge the fruitful discussion with Dr. Tao Qin at Microsoft Research Asia. B.W. and J.Z. appreciate the great help from Huaizhe Xu of the International Digital Economy Academy (IDEA). The authors also appreciate Mr. Ming-Feng Ho from the University of California Riverside for the great help in running the GP algorithm.

Appendix A
BAO Combined Fitting Parameters

The details of BAO-fitting parameters are shown in Table 8.

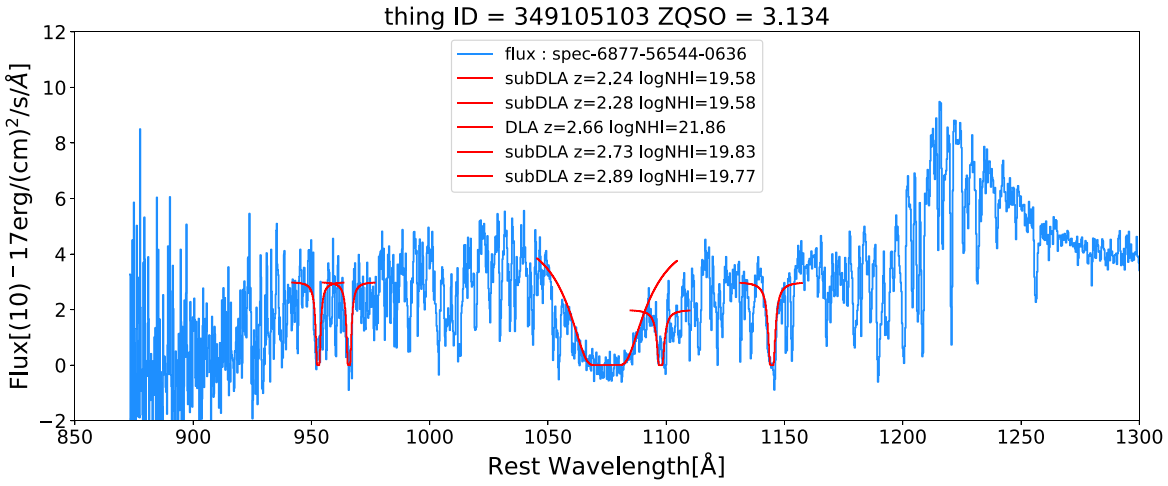
Appendix B
Purity and Completeness for the desiY1-0.14 Mock Spectra

The purity and completeness for the mock spectra including both DLAs and BALs are shown in Figure 20.

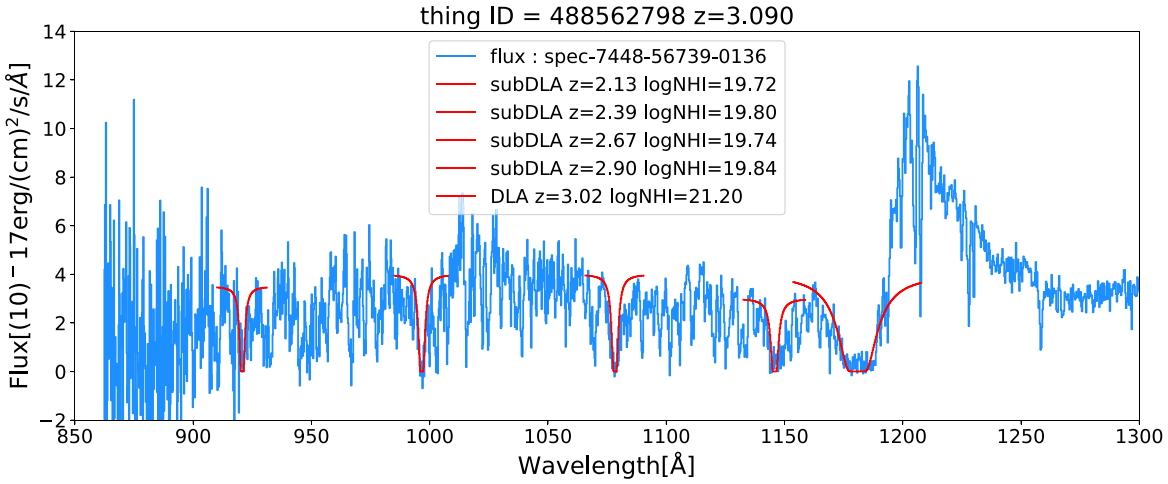


(a) (b)

Figure 20. We have applied our CNN model on the desiY1-0.14 mock spectra. These mock spectra contain both DLAs and BALs. The purity and completeness are shown in (a) and (b). The minimum S/N for this mock is 1.0, and thus the left panel (S/N < 0) is blank. The completeness is still above 96% for S/N > 3 spectra. The purity drops about 10%–20% in different bins. Nevertheless, the DESI has a formal BAL catalog, which will get rid of more than 98.6% BALs from the catalog. Then, we can run the DLA finder on the BAL-removed spectra. Therefore, we think that the purity result shown in Figure 8 is still valid.



(a)



(b)

Figure 21. We have applied our CNN model on these two spectra mentioned by Ho et al. (2021). It is clear that the DLA and sub-DLA candidates are detected.

Appendix C
DLA Detection in SDSS Spectra

In Ho et al. (2021), the author mentioned that the previous version of the CNN DLA finder (Parks et al. 2018) missed HCD DLAs in two SDSS spectra. We test our model on these two spectra; both the DLAs and sub-DLAs can be detected, as shown in Figure 21.

ORCID iDs

Ben Wang https://orcid.org/0000-0003-4877-1659
Jiaqi Zou https://orcid.org/0000-0001-9189-0368
Zheng Cai https://orcid.org/0000-0001-8467-6478
J. Xavier Prochaska https://orcid.org/0000-0002-7738-6875
Zechang Sun https://orcid.org/0000-0002-8246-7792
Jiani Ding https://orcid.org/0000-0003-4651-8510
Hiram K. Herrera-Alcantar https://orcid.org/0000-0002-9136-9609
Vid Irsic https://orcid.org/0000-0002-5445-461X
Xiaojing Lin https://orcid.org/0000-0001-6052-4234
David Brooks https://orcid.org/0000-0002-8458-5047
Soléne Chabanier https://orcid.org/0000-0002-5692-5243
Nathalie Palanque-Delabrouille https://orcid.org/0000-0003-3188-784X
Gregory Tarle https://orcid.org/0000-0003-1704-0781

References

Bird, S., Vogelsberger, M., Haehnelt, M., et al. 2014, MNRAS, 445, 2313
Busca, N. G., Delubac, T., Rich, J., et al. 2013, A&A, 552, A96
Cai, Z., Fan, X., Peirani, S., et al. 2016, ApJ, 833, 135
Cai, Z., Fan, X., Bian, F., et al. 2017, ApJ, 839, 131
Chabanier, S., Etourneau, T., Le Goff, J.-M., et al. 2022, ApJS, 258, 18
DESI Collaboration, Aghamousa, A., Aguilar, J., et al. 2016a, arXiv:1611.00036
DESI Collaboration, Aghamousa, A., Aguilar, J., et al. 2016b, arXiv:1611.00037
Dey, A., Schlegel, D. J., Lang, D., et al. 2019, AJ, 157, 168
Draine, B. T. 2011, Physics of the Interstellar and Intergalactic Medium (Princeton, NJ: Princeton Univ. Press)
du Mas des Bourboux, H., Rich, J., Font-Ribera, A., et al. 2020, ApJ, 901, 153
Farr, J., Font-Ribera, A., du Mas des Bourboux, H., et al. 2020, JCAP, 2020, 068
Finley, H., Petitjean, P., Pâris, I., et al. 2013, A&A, 558, A111
Font-Ribera, A., & Miralda-Escudé, J. 2012, JCAP, 2012, 028
Fumagalli, M., Prochaska, J. X., Kasen, D., et al. 2011, MNRAS, 418, 1796
Gardner, J. P., Sharples, R. M., Frenk, C. S., & Carrasco, B. E. 1997, ApJL, 480, L99

Garnett, R., Ho, S., Bird, S., & Schneider, J. 2017, MNRAS, 472, 1850
Grudić, M. Y., Guszejnov, D., Hopkins, P. F., Offner, S. S. R., & Faucher-Giguère, C.-A. 2021, MNRAS, 506, 2199
Guo, Z., & Martini, P. 2019, ApJ, 879, 72
Herrera-Alcantar, H. K. 2020, Master's thesis, Universidad de Guanajuato
Ho, M.-F., Bird, S., & Garnett, R. 2020, MNRAS, 496, 5436
Ho, M.-F., Bird, S., & Garnett, R. 2021, MNRAS, 507, 704
Jolly, J.-B., Knudsen, K. K., & Stanley, F. 2020, MNRAS, 499, 3992
Krogager, J.-K., Møller, P., Christensen, L. B., et al. 2020, MNRAS, 495, 3014
Lee, C. C., Webb, J. K., & Carswell, R. F. 2020, MNRAS, 491, 5555
Lee, K.-G., Hennawi, J. F., Stark, C., et al. 2014, ApJL, 795, L12
Li, Z., Horowitz, B., & Cai, Z. 2021, ApJ, 916, 20
Liske, J., Webb, J. K., & Carswell, R. F. 1998, MNRAS, 301, 787
McDonald, P. 2003, ApJ, 585, 34
McGreer, I. D., Jiang, L., Fan, X., et al. 2013, ApJ, 768, 105
McQuinn, M. 2016, ARA&A, 54, 313
Noterdaeme, P., Balashev, S., Krogager, J. K., et al. 2019, A&A, 627, A32
Noterdaeme, P., Petitjean, P., Ledoux, C., & Srianand, R. 2009, A&A, 505, 1087
Noterdaeme, P., Petitjean, P., Carithers, W. C., et al. 2012a, A&A, 547, L1
Noterdaeme, P., Laursen, P., Petitjean, P., et al. 2012b, A&A, 540, A63
Parks, D., Prochaska, J. X., Dong, S., & Cai, Z. 2018, MNRAS, 476, 1151
Pérez-Ràfols, I., Font-Ribera, A., Miralda-Escudé, J., et al. 2018, MNRAS, 473, 3019
Péroux, C., McMahon, R. G., Storrie-Lombardi, L. J., & Irwin, M. J. 2003, MNRAS, 346, 1103
Prochaska, J. X., & Herbert-Fort, S. 2004, PASP, 116, 622
Prochaska, J. X., Herbert-Fort, S., & Wolfe, A. M. 2005, ApJ, 635, 123
Prochaska, J. X., O'Meara, J. M., Fumagalli, M., Bernstein, R. A., & Burles, S. M. 2015, ApJS, 221, 2
Prochaska, J. X., & Wolfe, A. M. 1997, ApJ, 487, 73
Rahmati, A., Cravens, T., Larson, D. E., et al. 2014, AGUFM, 2014, P51B-3932
Rauch, M. 1998, ARA&A, 36, 267
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
Yèche, C., Palanque-Delabrouille, N., Claveau, C.-A., et al. 2020, RNAAS, 4, 179
Zafar, T., Péroux, C., Popping, A., et al. 2013, A&A, 556, A141