

Harvesting the Ly α forest with convolutional neural networks

Ting-Yun Cheng ¹,  Ryan J. Cooke ¹ and Gwen Rudie²

¹Centre for Extragalactic Astronomy, Durham University, South Road, Durham DH1 3LE, UK

²The Observatories of the Carnegie Institution for Science, 813 Santa Barbara Street, Pasadena, CA 91101, USA

Accepted 2022 September 9. Received 2022 August 18; in original form 2022 May 25

ABSTRACT

We develop a machine learning based algorithm using a convolutional neural network (CNN) to identify low H I column density Ly α absorption systems ($\log N_{\text{HI}}/\text{cm}^{-2} < 17$) in the Ly α forest, and predict their physical properties, such as their H I column density ($\log N_{\text{HI}}/\text{cm}^{-2}$), redshift (z_{HI}), and Doppler width (b_{HI}). Our CNN models are trained using simulated spectra ($S/N \simeq 10$), and we test their performance on high quality spectra of quasars at redshift $z \sim 2.5$ – 2.9 observed with the High Resolution Echelle Spectrometer on the Keck I telescope. We find that ~ 78 per cent of the systems identified by our algorithm are listed in the manual Voigt profile fitting catalogue. We demonstrate that the performance of our CNN is stable and consistent for all simulated and observed spectra with $S/N \gtrsim 10$. Our model can therefore be consistently used to analyse the enormous number of both low and high S/N data available with current and future facilities. Our CNN provides state-of-the-art predictions within the range $12.5 \leq \log N_{\text{HI}}/\text{cm}^{-2} < 15.5$ with a mean absolute error of $\Delta(\log N_{\text{HI}}/\text{cm}^{-2}) = 0.13$, $\Delta(z_{\text{HI}}) = 2.7 \times 10^{-5}$, and $\Delta(b_{\text{HI}}) = 4.1 \text{ km s}^{-1}$. The CNN prediction costs < 3 min per model per spectrum with a size of 120 000 pixels using a laptop computer. We demonstrate that CNNs can significantly increase the efficiency of analysing Ly α forest spectra, and thereby greatly increase the statistics of Ly α absorbers.

Key words: methods: data analysis – galaxies: high-redshift – intergalactic medium – quasars: absorption lines.

1 INTRODUCTION

The forest of neutral hydrogen (H I) Lyman- α (Ly α) absorption lines imprinted on a quasar spectrum – collectively known as the Ly α forest (Lynds 1971; Sargent et al. 1980) – provides our best understanding of the intergalactic medium (IGM) and circumgalactic medium (CGM), on scales of tens to hundreds of kpc and to Mpc (Cristiani et al. 1995; Fang et al. 1996). The photons emitted by a background quasar are absorbed at the redshifted Ly α transition (rest-frame wavelength = 1215.67 Å) in addition to higher order lines of the H I Lyman series (Sargent et al. 1980).

By number, Ly α absorption systems with low H I column density dominate the Ly α forest and trace the underlying density of the H I clouds (e.g. Schaye 2001). They can be used to probe the distribution and evolution of the baryonic matter, structure formation, and constrain cosmological parameters (e.g. Davé et al. 2010, also see reviews: Rauch 1998; Theuns, Leonard & Efstathiou 1998; Theuns et al. 1999; Tytler et al. 2004; Lehner et al. 2007; Meiksin 2009). Additionally, the thermodynamic properties of these systems are primarily governed by two processes: (1) adiabatic cooling from the expansion of the Universe; and (2) photoheating by the ultraviolet background (UVB) light from quasars and galaxies (Abel & Haehnelt 1999; Theuns et al. 2002; Bolton, Oh & Furlanetto 2009; Puchwein et al. 2015). The competition between these two effects tracks the thermal state of the low-density IGM through a characteristic

temperature–density relation (e.g. Hui & Gnedin 1997; Haehnelt & Steinmetz 1998; Schaye et al. 1999, 2000; Ricotti, Gnedin & Shull 2000; Becker, Rauch & Sargent 2007; Bolton et al. 2008; Rudie, Steidel & Pettini 2012b). Furthermore, the Ly α forest can also be used to probe cosmological models and constrain the properties of dark matter (e.g. Viel et al. 2013; Baur et al. 2016; Garzilli, Boyarsky & Ruchayskiy 2017; Iršič et al. 2017; Boera et al. 2019; Rogers & Peiris 2021).

While the Ly α forest is easily identified in a quasar spectrum, the identification of *individual* Ly α absorption systems within the forest is challenging. Conventionally, these absorption lines in the Ly α forest are fit with Voigt profiles¹ (e.g. Kim et al. 2002, 2013, 2021; Prochaska, Herbert-Fort & Wolfe 2005; Prochaska & Wolfe 2009; Rudie et al. 2012a, hereafter R12); however, a manual fit to the entire Ly α forest is very time-consuming, and requires the aid of visual inspection, and many human hours. To avoid human bias, there are also studies that have developed automated Voigt profile fitting algorithms² (Davé et al. 1997; Carswell & Webb 2014; Bainbridge & Webb 2017; Gaikwad et al. 2017).

With future surveys and facilities such as the WHT Enhanced Area Velocity Explorer (WEAVE; Pieri et al. 2016), and the 4-m

¹For example, the commonly used VPFIT package, which is available from: <https://people.ast.cam.ac.uk/~rfc/vpfit.html>

²We provide a few example codes here, but note that many efforts to generate an automated approach are unpublished. This problem is difficult, and an automated solution is not currently at the same level of accuracy that a human can produce.

* E-mail: ting-yun.cheng@durham.ac.uk

Multi-Object Spectroscopic Telescope (4MOST; de Jong et al. 2019), thousands of high resolution ($R \approx 20000$) quasar spectra are expected in the coming years. It will therefore not be feasible to analyse the enormous number of quasar spectra using conventional analysis methods. To overcome big data problems, such as this, machine learning techniques are essential.

Machine learning techniques, in particular deep learning (LeCun, Bengio & Hinton 2015), have been widely applied to a variety of galaxy studies such as galaxy morphology (Cheng et al. 2020a, 2021; Walmsley et al. 2022), galaxy merger (Bottrell et al. 2019; Ferreira et al. 2020), and strong gravitational lensing (Metcalf et al. 2019; Cheng et al. 2020b; Pearson et al. 2021). Applications to analyse spectroscopic data or time-series data include spectra and their star formation histories (Lovell et al. 2019), gravitational wave analyses (George & Huerta 2018), transient objects (Muthukrishna et al. 2019), and spectral classification (Bailer-Jones, Irwin & von Hippel 1998). Recently, there has been a growing interest in applying machine learning techniques to the Ly α forest, including: (1) a Ly α forest emulator (Bird et al. 2019; Rogers et al. 2019); and (2) the identification and properties of damped Ly α systems (DLAs; Garnett et al. 2017; Parks et al. 2018; Wang et al. 2022). DLAs are defined to have H I column densities that exceed $N_{\text{HI}} \geq 10^{20.3} \text{ cm}^{-2}$, and are easily identified by their strong, damped absorption features superimposed on the Ly α forest. Unlike DLAs, the low H I column density Ly α absorption systems associated with the Ly α forest ($N_{\text{HI}} < 10^{17} \text{ cm}^{-2}$) have a relatively shallow depth and narrow absorption features. Furthermore, Ly α forest absorption features outnumber DLA absorption lines by orders of magnitude, and occupy a wider range of column density. These absorption lines are also often blended and confused with metal lines, making this a challenging and laborious problem. As a result, an efficient and reliable machine learning based solution to harvest the Ly α forest – both line detection and characterization – does not exist. Given the utility of these low column density Ly α systems in studying the physics of the IGM, it is essential to develop a machine-learning-based detection algorithm to identify and characterize these features in preparation for the coming ‘Big Data’ era.

In this paper, for the first time, we apply a convolutional neural network (CNN) to efficiently identify Ly α forest systems ($N_{\text{HI}} < 10^{17} \text{ cm}^{-2}$) and extract their physical properties, including the redshift, Doppler width, and H I column density. While our primary goal is to efficiently extract the properties of the observed Ly α forest, our algorithm can also be used to identify Ly α absorbers in simulated spectra. Since our approach is general, this allows a more direct comparison between spectra extracted from state-of-the-art hydrodynamic cosmological simulations and observations. The paper is arranged as follows. Section 2 describes the generation of our simulated quasar spectra for training and initial testing purposes, and the observed quasar spectra that are used to validate our CNN predictions. Section 3 explains the CNN models and the training strategies, and we describe the evaluation metric in Section 4. In Section 5, we test our pre-trained model with the simulated spectra, while in Section 6, we apply the CNN models to predict the parameters of the Ly α forest from observed spectra, and compare the CNN’s predictions with the results based on Voigt profile fitting and human inspection from R12. Finally, our conclusions are summarized in Section 7.

2 QUASAR SPECTRA

In this section, we describe the simulated and observed quasar Ly α forest data that are used to train and test our network. While the

technique that we employ can be readily applied to quasars at any redshift, the focus of our work is to study Ly α absorption in the optical wavelength range. Since the Ly α forest is blueward of the quasar Ly α emission line, to detect Ly α forest absorption features in the optical range (i.e. ~ 3200 to 7200 \AA), the emission redshift of the quasars is in the range $z = 1.6$ – 5 . To satisfy the observed wavelength range, we generate simulated spectra at $z = 3$ for training our CNN. The details of the spectrum generation are outlined in Section 2.1, while in Section 2.2, we describe the pixel-level labelling of each Ly α absorption system. The observed spectra used to validate our model are described in Section 2.3.

2.1 Mock Spectra

The number of human-analysed quasar spectra that have been fit with Voigt profiles is currently limited by the time effort required to carefully analyse and fit each individual absorption line in every quasar spectrum. The quasar spectra that have been analysed are subject to human choices that may not reflect the true underlying properties of the absorption lines. For this reason, our training data are based on simulated quasar spectra to provide a large quantity of spectra together with ground-truth identifications of Ly α systems and their properties. Our simulated spectra were generated using packages in the PYIGM software.³ The generated spectra represent a typical quasar at redshift $z = 3$ and are convolved with an instrumental full width at half-maximum (FWHM) resolution of $v_{\text{FWHM}} = 7 \text{ km s}^{-1}$. These choices are motivated by the typical properties of high resolution spectra of quasars in current observatory archives. The velocity per pixel of these spectra is set to $2.5 \text{ km s}^{-1} \text{ pixel}^{-1}$. The impact of the model trained with this setup on predicting spectra with different assumptions for the properties of the spectra are discussed in Appendix A.

A catalogue of Ly α forest absorption lines are drawn randomly from the column density distribution function (CDDF), $f(N_{\text{HI},X})$, following the default form implemented in PYIGM (the Hermite spline model of Prochaska et al. 2014), where X is the absorption distance. This provides a distribution of H I absorption systems with $N_{\text{HI}} = 10^{12} - 10^{22} \text{ cm}^{-2}$ that can be imprinted on to a simulated quasar spectrum to generate absorption features with ‘ground-truth’ labels (see Section 2.2). Note that this model was constrained at redshift $z \approx 2.5$. PYIGM uses inverse transform sampling of the $z = 2.4$ CDDF to generate a list of H I column densities; the corresponding Doppler parameters are drawn from the Hui & Rutledge (1999) distribution. The redshifts of the mock lines are generated by inverse transform sampling the redshift-dependent incidence of absorption systems, $l(z)$. Finally, the spectra are generated without noise; additional noise is added later to test the sensitivity of our model to the adopted S/N (Section 3.1). In Fig. 1, we show an example of a simulated spectrum with different choices of the S/N. Our simulated spectra contain only the absorption lines of the H I Lyman series, and do not include metal lines.

Machine learning applications commonly require training samples with a well-defined structure and clear corresponding labels, if possible. Since Ly α absorption features are relatively simple and have a well-defined structure that can be derived by only a few

³Primary Builders include: J. Xavier Prochaska, N. Tejos, and J. Burchett (<https://github.com/pyigm/pyigm>). We also implemented a minor change to this code; when generating Voigt profiles, we constructed a sub-pixelated wavelength array to sample each native pixel by 10 sub-pixels. This accounts for the curvature of the profile within each pixel.

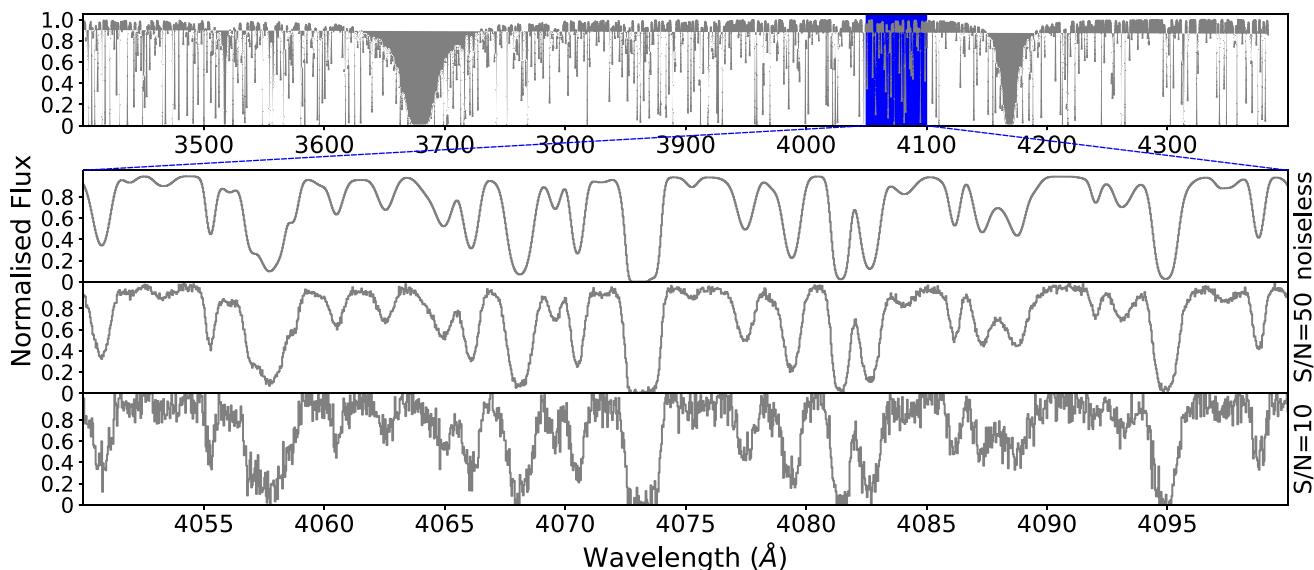


Figure 1. The top panel shows a simulated Ly α forest spectrum of a quasar at redshift $z = 3$ (no noise). The three subsequent panels show a zoom-in of the top panel (see blue box, top panel) with different signal-to-noise ratios ($S/N = \infty, 50$, and 10 per pixel for the second, third, and fourth panels, respectively).

physical properties, i.e. Voigt profiles, having robust labels are more crucial than complexity of data set to avoid confusion in a classification task. Hence, as a first attempt, this training data set defines a clear structure of Ly α absorbers that helps a machine to draw a cleaner decision boundary in a high dimensional parameter space. Additionally, it helps us to analyse the performance of our automated algorithm and identify its limitations. As an alternative, we could generate simulated spectra with cosmological hydrodynamic simulations to account for the clustering and complex structure that exists in a real quasar spectrum. However, since the CGM structures in these simulations are unresolved (Hummels et al. 2019; Rudie et al. 2019; van de Voort et al. 2019), it might be more sensible to train a machine using observed spectra in future works to account for the clustering of absorbers.

2.2 Ly α absorption systems

The Ly α absorption features in a spectrum can be described with three physical properties: (1) the total H I column density (N_{HI} ; cm^{-2}), (2) the redshift (z_{HI}) of the H I absorbers, and (3) the Doppler width (b_{HI} ; km s^{-1}). The $f(N_{\text{HI}})$ model provides a distribution of H I absorbers that samples the H I column densities of the Ly α forest. With the ‘ground-truth’ information of the three aforementioned properties, we generated four labelling arrays for each pixel in the quasar spectrum (these labels are illustrated in Fig 2):

- (i) **Ly α ID**: set to a value of 1 if a Ly α absorber exists in this pixel, and 0 if not;
- (ii) **$\log N_{\text{HI}}$** : H I column density (in units of cm^{-2}) of the corresponding Ly α absorber on a logarithmic scale;
- (iii) **z_{loc}** : the relative location of the centre of an absorption feature (in units of pixels⁴). A pixel centred on an absorption feature is set to 0, and negative and positive values to pixels at the left and right, respectively. For example, if the centre of a given pixel is 2.4 pixels

⁴Note that z_{loc} is a floating point number, since the centre of the associated absorption line is not coincident with the centre of a pixel.

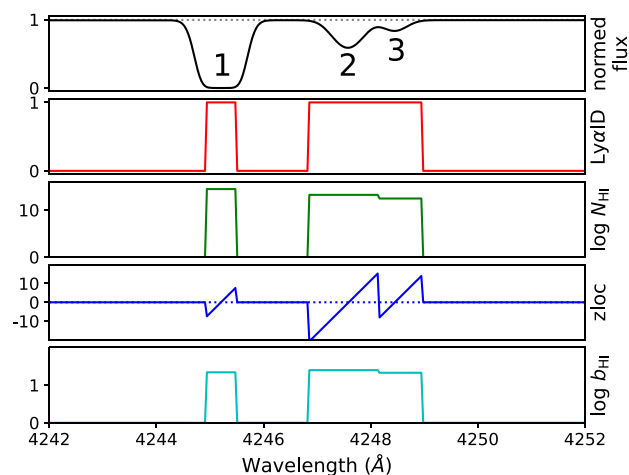


Figure 2. Example of the training labels: Ly α ID, $\log N_{\text{HI}}$, z_{loc} , and $\log b_{\text{HI}}$ in pixel scale from top to bottom. The grey dotted line in the top panel represents the normalized quasar continuum level, and the blue dotted line in the third panel shows $z_{\text{loc}} = 0$. Note that the training labels are only defined when Ly α ID = 1. The absorption features are labelled 1, 2, and 3, ordered by column density.

to the left of the centre of an absorption profile, we assign the label of this pixel to be -2.4 ;

- (iv) **$\log b_{\text{HI}}$** : Doppler width of the corresponding Ly α absorber on a logarithmic scale (km s^{-1}).

First, to ensure that the absorption features used to train our machine are Ly α lines, we applied a cut to exclude the pixels with wavelengths where the Ly β transition of the highest redshift H I absorber appears. The initial pixel values for the four training label arrays were set to 0. The labels were generated for all Ly α systems ordered from the highest H I column density to the lowest H I column density. For each Ly α system, we first check if the optical depth, $\tau = N_{\text{HI}} \sigma_{\alpha}$ (σ_{α} is the absorption cross-section for the Ly α transition), of the pixel is high enough to saturate the absorption line using a criterion

of $\exp(-\tau) < 0.015$, where the threshold is defined by $3/(S/N)$ (where our fiducial $S/N = 200$). If any pixel satisfies this criterion, we store the $\text{Ly}\alpha\text{ID}$, $\log N_{\text{HI}}$, zloc , and $\log b_{\text{HI}}$ of this absorber in the label arrays. Note that ‘zloc’ represents the location of the centre of an absorption feature, where the centre ($\text{zloc} = 0$) is drawn using the redshift of the $\text{Ly}\alpha$ system. If the listed $\text{Ly}\alpha$ system does not saturate a pixel, it is then used to provide values to the corresponding pixels where the flux of the absorption features is < 0.995 . Note, if multiple absorption components contribute to the total optical depth in a pixel, we labelled only the dominant line. This means that in this work we do not consider the impact of a secondary or additional line blends in a single pixel. A more thorough investigation about the effect of blended lines will be carried out in future work. Fig. 2 shows an example of the labelling procedure that we use in this work. In the example shown in Fig. 2, labels are first assigned to the leftmost (strongest) feature, i.e. feature 1. Every pixel associated with this absorption line that has a flux less than 0.015 is assigned a $\text{Ly}\alpha\text{ID} = 1$; the column density and Doppler parameter is the same for all of the associated pixels of this feature, and the zloc label represents the non-integer pixel difference from the centre of the absorption line profile. The next strongest absorption line, feature 2, is then labelled; because the central optical depth is not saturated, we label all pixels that have a flux < 0.995 . The rightmost feature 3, which is partially blended, is labelled using the same approach, however, the labels are only applied to the pixels where the pixel optical depth contributed by this feature is highest.

2.3 Archival quasar observations

To validate our machine’s prediction on real data, we use the 15 quasar spectra observed and reduced by R12. These data were observed with the High Resolution Echelle Spectrometer (HIRES; Vogt et al. 1994) on the Keck I telescope. The redshifts of these quasars are in the range $2.5 \lesssim z \lesssim 2.9$, and the spectra have $R \cong 45\,000$ ($v_{\text{FWHM}} \cong 7\text{ km s}^{-1}$), high signal-to-noise ratio ($S/N \sim 50\text{--}200\text{ pixel}^{-1}$), and cover the wavelength range $3100\text{--}6000\text{ \AA}$. We resampled these spectra to $2.5\text{ km s}^{-1}\text{ pixel}^{-1}$ (while conserving flux) to be consistent with the input of our CNN model (see Section 3.1 and Appendix A). Further details about the observations and data reduction procedure are outlined by R12.⁵

3 DEEP LEARNING MODEL

We employ multi-task learning (Caruana 1998; Ruder 2017) by training with and predicting four outputs (labels): $\text{Ly}\alpha\text{ID}$, $\log N_{\text{HI}}$, zloc , and $\log b_{\text{HI}}$ (see Section 2.2). The network is generalized to approach these four tasks at the same time. The details of the CNN structure for our multi-task learning are described in Section 3.2. The prediction of each variable complements the prediction of the other variables by combining their losses (details in Section 3.3)⁶ as part of the training process.

We employ similar training strategies to that adopted by Parks et al. (2018) to ‘scan’ through a spectrum with a fixed-size window (ws) and a 1-pixel step size. To do this, we used the `fit_generator`

function in KERAS. This method increases the machine’s performance by analysing hundreds of pixels in a segmentation per step rather than tens of thousands of pixels in a whole spectrum in one go. A `fit_generator` has the added benefit that each window is generated at run-time from the full spectrum, and therefore reduces the amount of VRAM required (or, equivalently, allows us to include more training data). The schematic diagram of the scanning process is shown in Fig. 3.

3.1 Data input

In each spectral window used as input (of size ws), there are four training labels, and these labels correspond to the properties of the centre pixel in this window. Our CNN is therefore trained with and only predicts the corresponding values at the central pixel within this window from each labelling array. For example, in Fig. 3, the labels that correspond to the red spectral window are listed in the red labels box, and the ones that correspond to the green spectral window are in the green labels box, etc. The size of the spectral window, ws , is a hyperparameter that is objectively selected using an optimization algorithm (Section 3.2). We scan each training spectrum from left to right during each epoch. Each batch contains one spectral window from each training spectrum. This approach ensures that all training spectra are fully ‘scanned’ and their training losses are taken into account in each epoch (see also Section 3.3).

To ensure that the CNN prediction is primarily sensitive to absorption features that are located at the centre of the window, we define an additional hyperparameter, $cnpix$. This hyperparameter is defined by the absolute value of zloc , $|\text{zloc}| \leq cnpix$, and determines the pixels that are recognized as the ‘centre’ of an absorption feature. For example, in Fig. 3, if $cnpix = 1$, the yellow shaded area is defined as the ‘centre’ region, and the true values outside this range are set to 0 as highlighted by the yellow labels. The CNN is trained with, and predicts the labels associated with, the central pixel of the window. The variable $cnpix$ ensures that the training process only learns from an absorption feature that overlaps with the pixel in the centre of a window.

Additionally, we noticed that training our machine with noiseless spectra results in a significantly worse performance when predicting a noisy spectrum (see Appendix C). To overcome this issue so that our machine can sensibly be applied to predict accurate labels to real data, we included additional noise to each spectrum. The S/N of a given spectrum is drawn from a Gaussian distribution, with a mean of 10 and a standard deviation of 2. Given this S/N value, we randomly perturb every pixel in the perfect normalized spectrum by a Gaussian distribution with a standard deviation of $1/(S/N)$. In previous studies, $\text{Ly}\alpha$ forest analyses have primarily relied on spectra with $S/N > 20$. Hence, we chose a low S/N value as a typical value to allow our machine to produce reliable results when analysing observed quasar spectra that are of somewhat lower S/N . Appendix C outlines the tests we performed to validate this approach, and demonstrate that this stabilizes the predictions for spectra with different S/N .

3.2 CNN architecture

Fig. 4 shows our CNN architecture, which follows the same form as the one used in Parks et al. (2018), including three 1-dimensional convolutional layers (i.e. Conv 1, Conv 2, Conv 3) and each of them is followed by a pooling layer with a kernel size of 2. A dropout is inserted after the third pooling layer (Pool 3), and the array is flattened to connect with a dense layer (Dense 1). Four separate dense layers

⁵Some spectra contain DLA absorption lines. Our CNN model is sensitive to $\text{Ly}\alpha$ systems with low column density, and it then ignores the DLA features. Hence, these features do not impact the results.

⁶The loss quantifies the difference between the expected output (i.e. truth) and the predicted output by a machine learning model, while the loss function is the function used to calculate the loss.

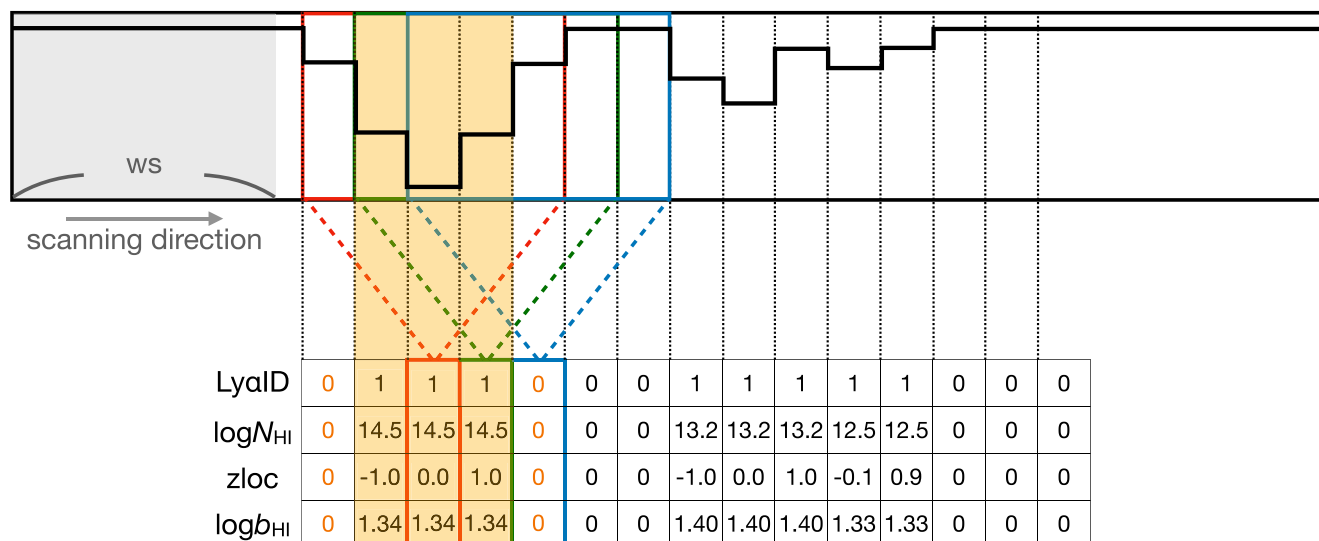


Figure 3. Schematic diagram of the scanning process that we use to train the data. The scanning window size is ws , and the step size is 1 pixel. All of the pixels in a given window are used as input to the CNN, and the corresponding output label of each window is assigned to the pixel located at the centre of the window. The red, green, and blue boxes demonstrate how the corresponding labels change when the window shifts by 1 pixel. Labels are only assigned if the centre pixel of the window is within $cnpix = \pm 1$ pixels of the centre of an absorption feature (this is represented by the yellow box). Outside of this defined area, the corresponding labels are set to 0. Note that the value of $cnpix$ is a hyperparameter of the network.

are then connected with the ‘Dense 1’ layer and dropouts are applied to each dense layer. The dropout rate is consistent throughout the network and is one of the hyperparameters that is selected with an optimization algorithm.

The activation function used before the output layer is consistently ReLU: $f(x) = \max(0, x)$ (Agarap 2018), and the activation functions for the outputs depend on the desired output range of the target variables. Hence, for Ly α ID, we applied the sigmoid function: $f(x) = 1/(1 + e^{-x})$, which outputs a value between 0 and 1 as a probability. For zloc we adopted a linear function: $f(x) = x$, and for both log b_{H1} ⁷ and log N_{H1} we used ReLU, which outputs a value $f(x) = \max(0, x)$. Several crucial hyperparameters in our CNN architecture were objectively selected by a Bayesian optimization process (Snoek et al. 2012, also see appendix B) over a range of possible values. The results are listed in Table 1. In addition to the hyperparameters of the CNN architecture, we include two additional hyperparameters from Section 3.1: (1) the window size (ws) and (2) the number of pixels that are used to define the centre of an absorption feature ($cnpix$). These two hyperparameters are critical in determining the types of Ly α systems that our CNN is sensitive to⁸; the values of these parameters depend on the science question being addressed. We therefore use a Bayesian optimization process to decide their values without human intervention.

Finally, the learning rate was set to 0.0001 and we applied the Adam optimizer (Kingma & Ba 2015). The maximal number of iteration for each training is 20 epochs, but only the model with the minimal validation loss within the 20 epochs is saved.

⁷The minimal value of b_{H1} in this work is 15 km s^{-1} . Hence, the logarithmic value is always > 0 .

⁸One can use a larger size of scanning window to help improve the sensitivity in detecting systems with higher column density. Note that these systems are fewer. To carry out this optimization, one also needs to consider the issues of strongly imbalanced number of different systems.

3.3 Loss function

With our multi-task learning model, four outputs were produced for a given input: Ly α ID, log N_{H1} , zloc, log b_{H1} . With the `fit_generator` function, the final loss per epoch for each output is an average value of the losses of all steps. For a binary classification task, the loss of ‘Ly α ID’ uses a binary cross-entropy loss function

$$L_{ID} = -\frac{1}{N} \sum_{i=1}^N y_{c,i} \log(p_{c,i}) + (1 - y_{c,i}) \log(1 - p_{c,i}), \quad (1)$$

where N is the total number of training windows per epoch, i.e. the number of input training spectra (= number of data per step) times the number of pixels in each spectrum (= number of steps),⁹ y_c represents the true classification label (i.e. Ly α ID = $y_c = 1$ for a Ly α absorption system), and p_c is the probability of being a Ly α system predicted by the CNN. The loss functions of the remaining outputs use a masked mean square error (MSE)

$$L_j = \frac{1}{N'} \sum_{i=1}^N y_{c,i} (y_{j,i} - \hat{y}_{j,i})^2, \quad (2)$$

where N' is the total number of training windows per epoch where y_c equals to 1, and $j = \{N_{H1}, z, b_{H1}\}$ represents the loss functions of log N_{H1} , zloc, log b_{H1} , respectively. The $y_{j,i}$ are the true values of $j = \{N_{H1}, z, b_{H1}\}$, while the $\hat{y}_{j,i}$ are the predicted values from the CNN. With this ‘masked’ form of the loss function for log N_{H1} , z , and log b_{H1} , losses are only contributed to the final loss per epoch when $y_c = 1$.¹⁰ The final loss function of the CNN training process per epoch is the sum of the above-mentioned losses

$$L = L_{ID} + L_{N_{H1}} + L_z + L_{b_{H1}} \quad (3)$$

⁹Recall that the model is trained by scanning through all spectra.

¹⁰We note that this masked loss function ensures that our machine is not biased by the log N_{H1} , z , and log b_{H1} labels in pixels where there is no absorption.

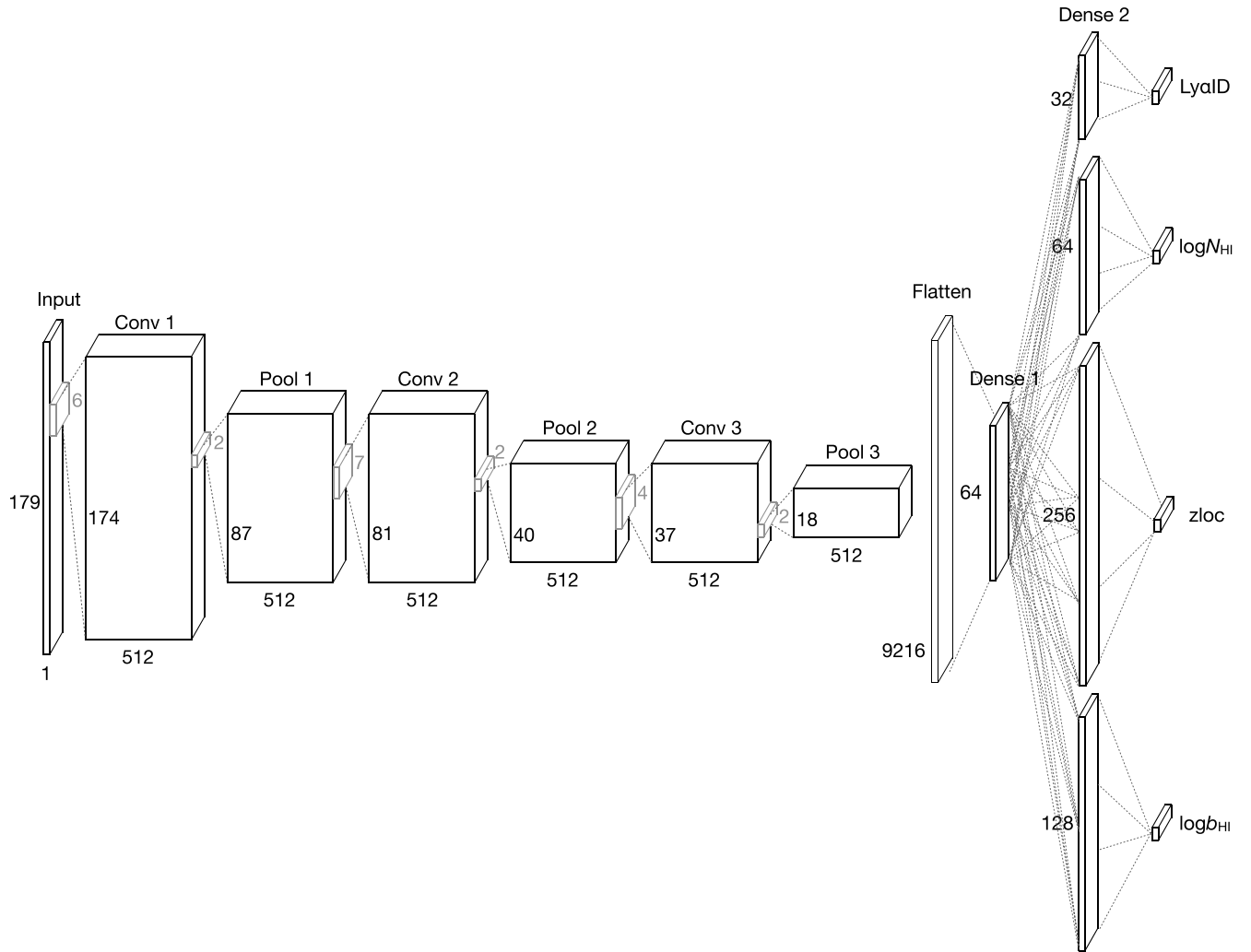


Figure 4. Schematic diagram of the CNN architecture used in this work. It is composed of three 1-dimensional convolutional layers with pooling layers following each, one dense layer to connect each component, and four dense layers for four target outputs. The values of relevant hyperparameters are listed in Table 1.

Table 1. Hyperparameters used in our CNN architecture. These values are selected using a Bayesian optimization algorithm (Snoek, Larochelle & Adams 2012).

	Hyperparameters	Optimized value
Data input	window size (ws)	179
	central pixels ($cnpix$)	1
CNN Architecture	L2	0.0
	dropout	0.1
	conv_filter_1	512
	conv_filter_2	512
	conv_filter_3	512
	conv_kernel_1	6
	conv_kernel_2	7
	conv_kernel_3	4
	dense_1	64
	dense_2_ID	32
	dense_2_N	64
	dense_2_z	256
	dense_2_b	128

Note that the scale of each loss needs to be comparable in order to prevent a biased weighting due to a single label that contributes most of the loss. For example, in our preliminary test, we found that a large uncertainty in predicting linear b_{HI} values (range of 15–75 km s⁻¹) contributes a significant loss which therefore decreases the CNN’s capability of precisely predicting the other labels. Hence, we opted to predict the logarithmic b_{HI} values in this work.

4 EVALUATION METRICS

Before showing the results of our CNN models, we first introduce the metrics that were used to evaluate the CNN performance. For the classification of Ly α absorbers, we use recall and precision, as defined below, to evaluate the CNN performance.

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}; \quad \text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad (4)$$

where ‘TP’ is a true positive (i.e. a correct classification), ‘FP’ is a false positive (i.e. a mis-classified system), and ‘FN’ is false negative corresponding to true systems that are missed by our CNN. Recall is a measure of completeness: the fraction of true absorbers identified by the CNN. Precision is a measure of the fraction of identified systems

that are real. We have designed the CNN to have high precision at the expense of recall, so that we are confident that a CNN-classified Ly α system is a real Ly α system. This choice may need to be different, depending on the scientific question being addressed.

On the other hand, when estimating the physical properties of a Ly α absorber such as redshift, H I column density, and Doppler width, we consider two metrics: (1) the root mean square error (RMSE) and (2) mean absolute error (MAE), to assess the ‘accuracy’ of the CNN predictions. The RMSE is defined as

$$\text{RMSE} = \sqrt{\frac{\sum_{k=1}^{N_{\text{Ly}\alpha}} (y_k - \hat{y}_k)^2}{N_{\text{Ly}\alpha}}}, \quad (5)$$

where $N_{\text{Ly}\alpha}$ is the number of matched Ly α systems, and y_k and \hat{y}_k represent the ‘true’ and ‘predicted’ values of each Ly α system, respectively. The RMSE is strongly impacted by the outliers due to the square of the residual. Hence, we also introduce MAE (equation 6) which is more resilient to outliers than the RMSE.

$$\text{MAE} = \frac{\sum_{k=1}^{N_{\text{Ly}\alpha}} |y_k - \hat{y}_k|}{N_{\text{Ly}\alpha}}, \quad (6)$$

where the definition of each variable is the same as equation (5). The MAE is more useful, since we do not expect the CNN to be absolutely correct. For example, in many cases our CNN predicts that a single Ly α absorber is required to recover an absorption feature, while there are in fact many neighbouring lines that contribute to this absorption feature (see discussion in Section 5.3). For this example, we will have poor estimates of the physical properties when comparing with the ‘true’ values, and this yields strong outliers. Thus, the MAE is a more robust indicator of the CNN performance than the RMSE in the context of this study. In later sections, we will list both quantities, but the discussion will be based on the MAE.

5 PREDICTION TO SIMULATED SPECTRA

With the aforementioned setups in Section 3, we trained a CNN model with 900 simulated spectra¹¹ with an S/N randomly drawn from a Gaussian distribution with a mean of 10 and a standard deviation of 2. An independent set of 25 simulated spectra (a noiseless set and a S/N = 10 set) were used to examine our pre-trained CNN model.

5.1 CNN-classified Ly α forest systems

With a CNN prediction for each pixel in all spectra, we used the following two criteria to identify Ly α systems: (1) Ly α ID > 0.3, and (2) |zloc| ≤ *cnpix*, where *cnpix* = 1. The former criterion judges if a pixel contains a Ly α system by the binary classification probability. The initial probability threshold > 0.3 used for Ly α ID is considered to have the maximum number of identified systems without decreasing the precision by selecting pixels with low predicted probabilities. The second criterion is applied in order to identify the centre of an absorber.

To compare the CNN-classified Ly α systems with the ground-truth labels, we match the input and predicted catalogue of systems; our matching criteria require that the velocity difference between the input and prediction is smaller than half of the minimum FWHM that

can be detected by a machine. This FWHM threshold is estimated by the minimum $b_{\text{H I}}$ value our CNN predictor can detect, i.e. $b_{\text{min}} = 15 \text{ km s}^{-1}$, using the relation: $\text{FWHM} = 2\sqrt{\log(2)}b$. Hence, the threshold applied is $\sqrt{\log(2)}b_{\text{min}} \sim 12.5 \text{ km s}^{-1}$.

The comparisons between the true and predicted systems with Ly α ID > 0.3 using noiseless spectra and S/N = 10 spectra are shown in Fig. 5. This figure provides an indication of the upper and lower ranges of CNN predictions for mock spectra of different noise levels. With low S/N, the total number of matched systems decreases when using the same probability threshold. However, the overall CNN performance remains consistent, with only a minor increase of the MAE. We summarize the evaluation metrics of different data sets in Table 2.

When applying a higher probability threshold to Ly α ID for spectra of the same noise, fewer systems with a high accuracy are matched. For example, in Table 2, when applying Ly α ID > 0.7 to predict noiseless spectra, the recall drops while our CNN predictions show an improvement.

For either noiseless or S/N = 10 simulated spectra, the overall precision of our CNN is over 0.98. In the following sections, we investigate the causes of the FP and FN classifications.

5.2 False positive

A false positive (FP) is an absorption system identified by our CNN classifier that cannot be matched to a Ly α system in the simulated true label catalogue. When predicting the labels of a simulated spectrum that only contains Ly α absorbers, our CNN reaches a precision of over 0.98 for spectra with S/N = 10 (over 0.99 for noiseless spectra). The FP in this case is exclusively a simple mismatch due to the velocity threshold ($\sim 12.5 \text{ km s}^{-1}$) used in matching systems between true and predicted catalogues. An example is shown in Fig. 6. The velocity difference between our CNN-classified system and the closest neighbour is 17.5 km s^{-1} in this example. FPs occur when an absorption feature comprises multiple nearby lines, while our CNN tends to use one line to describe the absorption feature. This results in a shift of the defined centre and the mismatch of the true and predicted Ly α systems. Through visual inspection, we noticed that the parameters of the FPs predicted by the CNN classifier fit the absorption feature as well as the true labels, especially given that our classifier is trained on data of S/N $\simeq 10$. This type of failure can also happen when using conventional methods such as Voigt profile fitting (e.g. human bias, or indistinguishable absorption profiles). This reflects a potential underestimation of the number of Ly α systems that are indistinguishable due to confusion or insufficient S/N, or because they are due to sub-structures of H I gas within a larger H I gas cloud. This may be improved by including higher order Lyman series lines as part of the CNN training process in future work.

5.3 False negative

False negatives (FN) occur when a system is listed in the true label catalogue, but it is not identified by our CNN. Examples are shown in Fig. 7. Since we train our CNN with noisy spectra, some detailed structures are buried in the noise, resulting in either a non-detection or low predicted probability (Ly α ID) to these pixels. This is a compromise between the accuracy and the feasibility of a CNN technique to spectra with low S/N. One can train a CNN with a higher S/N to increase the accuracy of a CNN detection and therefore reduce potential FN that are impacted by noise. In our test, when training

¹¹This number of training set is sufficient since each spectrum includes over 20 000 segmentation windows for the training process. Using additional spectra did not improve our result.

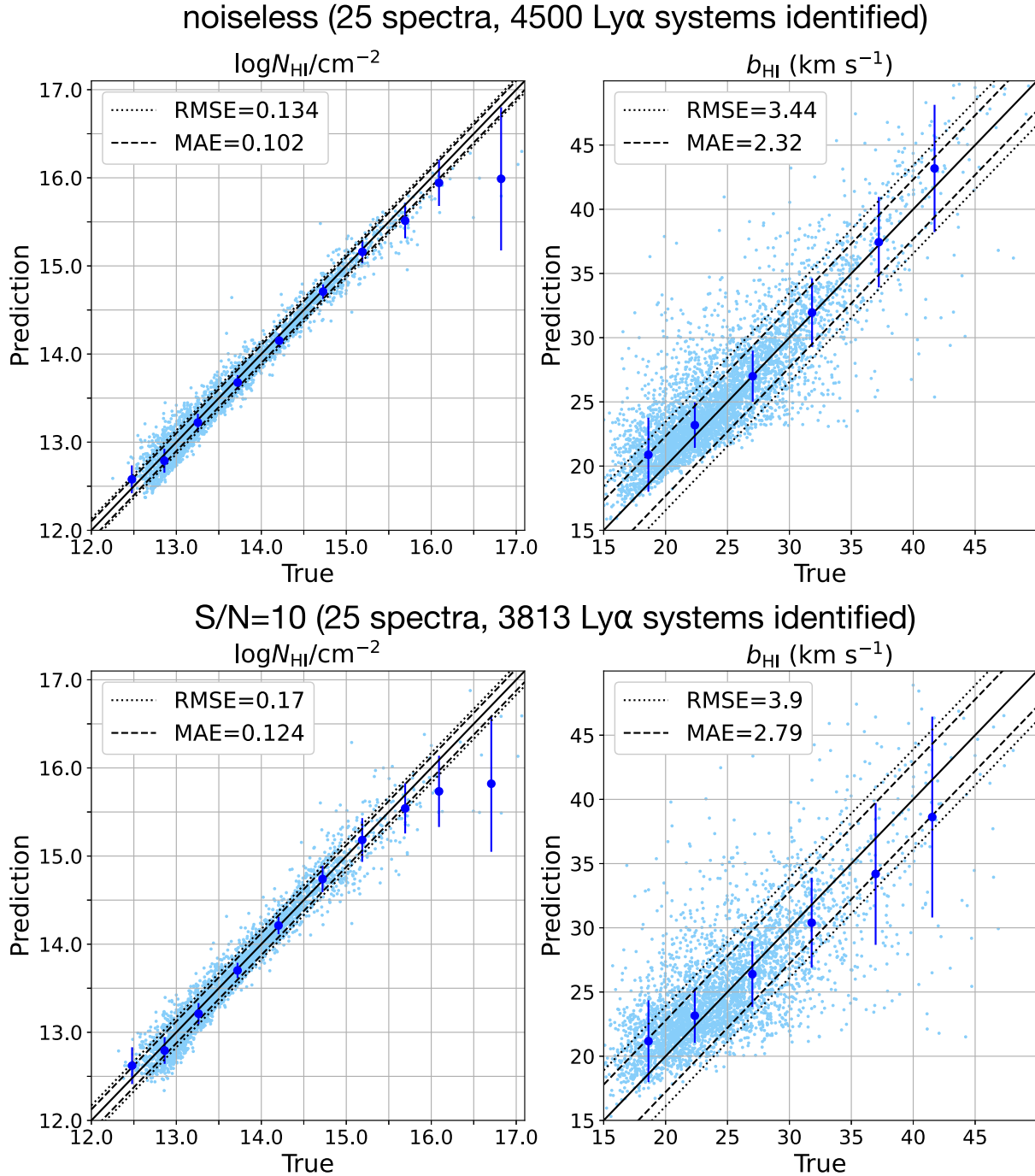


Figure 5. Comparisons between the true and predicted values of $\log N_{\text{HI}}/\text{cm}^{-2}$ and b_{HI} with Ly α ID > 0.3 using noiseless spectra (top) and spectra with S/N = 10 (bottom). The black solid line shows $f(x) = x$, the dashed lines indicate the scatter range defined by the MAE of each plot, and the dotted lines are the range defined by the RMSE of each plot. Dark blue data points show the median values within different bins of the true values. The y-axis error bar presents the MAE of each bin.

our CNN with noiseless spectra,¹² the machine reaches a high recall (~ 0.88) and a high precision (~ 0.90) when testing on a noiseless spectrum. However, the ability to predict physical properties such as b_{HI} and N_{HI} drops significantly for a CNN that is trained on noiseless data, but applied to noisy spectra (for further details, see

¹²Note that this result used a CNN architecture with hyperparameters that were specifically tuned to noiseless data.

Appendix C). This severely limits the utility of a CNN model, since most spectroscopic data are of low S/N.

We summarize four main cases that contribute to FNs:

- (i) Weak absorption features that are not significantly detected in the noisy data (S/N ~ 10) used for the training process. Our CNN then has difficulty distinguishing these features from the noise, and provides either a non-detection or a low predicted probability, i.e. Ly α ID < 0.3 , even for data with much higher S/N.

Table 2. Evaluation metrics (precision, recall, MAE) of CNN-classified systems on different test data sets. All of the results tabulated here are based on the same model, which is trained on noisy spectra (where the S/N is drawn from a Gaussian distribution with a mean = 10 and a standard deviation = 2).

Test sets	Precision	Recall	$\Delta \log N_{\text{HI}}/\text{cm}^{-2}$	Δz_{HI}	Δb_{HI} (km s $^{-1}$)
Noiseless mock spectra (Ly α ID > 0.7)	0.992	0.127	0.08	1.7×10^{-5}	1.4
Noiseless mock spectra (Ly α ID > 0.3)	0.994	0.322	0.10	2.4×10^{-5}	2.3
S/N = 10 mock spectra (Ly α ID > 0.3)	0.987	0.273	0.12	3.0×10^{-5}	2.8
R12 spectra (Ly α ID > 0.3):					
All predictions	0.782	0.260	0.14	2.7×10^{-5}	4.2
$12.5 \leq \log N_{\text{HI}}/\text{cm}^{-2} < 15.5$	0.792	0.258	0.13	2.7×10^{-5}	4.1

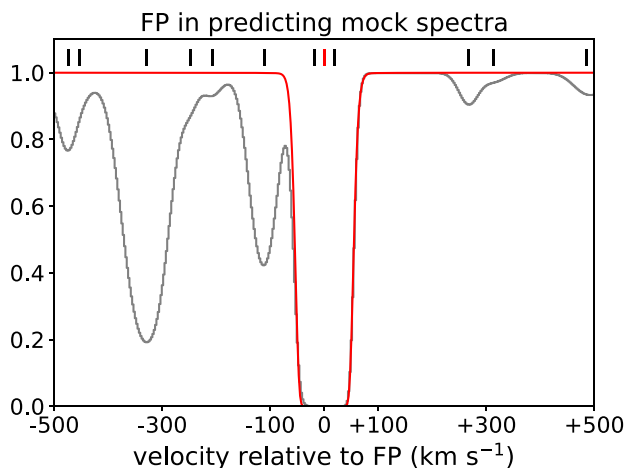


Figure 6. An example of a FP (red curve). The grey curve shows the contribution of all absorption lines. Short tick marks above the spectra indicate the centre of a Ly α system (black for catalogue lines, red for FP). The velocity between the red prediction and the closest neighbour system from the true catalogue of primary absorption lines is $\Delta V = 17.5$ km s $^{-1}$.

(ii) A strong absorption feature composed of multiple neighbouring lines (e.g. see Fig. 6). This type of FN occurs when our CNN uses one line to fit an absorption feature while this feature is in fact composed of several Ly α systems (Section 5.2). This mismatch therefore contributes several FNs, and one FP.

(iii) A strong, broad absorption feature with a size that is larger than the scanning window, i.e. $\Delta V > 447.5$ km s $^{-1}$. Due to the fixed size of our scanning window, our CNN is restricted to features that are well-defined within the window size.

(iv) Complex absorption features contributed by many nearby lines. Similar to case (ii) above, our CNN only fits a dominant feature from this complex structure and misses other overlapped absorption features formed by nearby, usually weaker, Ly α systems.

We find that the dominant FN contribution comes from weak absorption features that are within 3σ of the continuum; our 25 test spectra indicate this type of FN contributes ~ 86 per cent of the total number of FNs. These weak absorption features have an average value of $\log N_{\text{HI}}/\text{cm}^{-2} = 12.40 \pm 0.25$ and $b = 28.9 \pm 9.8$ km s $^{-1}$. If we exclude this kind of FN, the recall improves from ~ 0.32 to ~ 0.77 for noiseless spectra.

In Fig. 8, we present the change of recall and the MAE values of $\log N_{\text{HI}}$, z_{HI} , and b_{HI} grouped by column density using noiseless spectra (solid line) and S/N = 10 spectra (dashed line). This demonstrates that our CNN model has better recall to Ly α systems with a column density range of $13 \leq \log N_{\text{HI}}/\text{cm}^{-2} < 16$. By visual

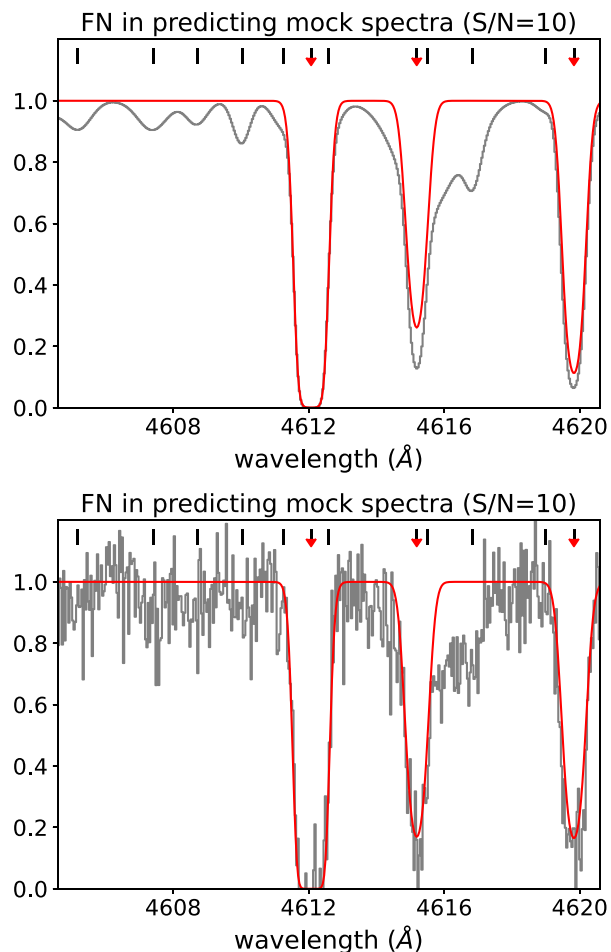


Figure 7. Several examples of FN. The centres of each true Ly α system are labelled by black short tick marks above the spectra, while the red arrow indicates a match between the true and predicted systems. Hence, a black tick mark without a matching red arrow represents a FN. The top panel shows a noiseless spectrum and the bottom panel is a spectrum with S/N = 10. The grey histogram represents the data and the red curve is a reconstruction based on all of the predicted lines by our CNN.

inspection, we found that the FNs for systems with column density in this range are only the cases (ii)–(iv) listed above. Compared to other bins, low H I column density systems (i.e. $\log N_{\text{HI}}/\text{cm}^{-2} < 13$) contribute weak absorption features which can be hidden in the noise, and result in much lower recall value (≤ 0.1), i.e. a higher fraction of FNs.

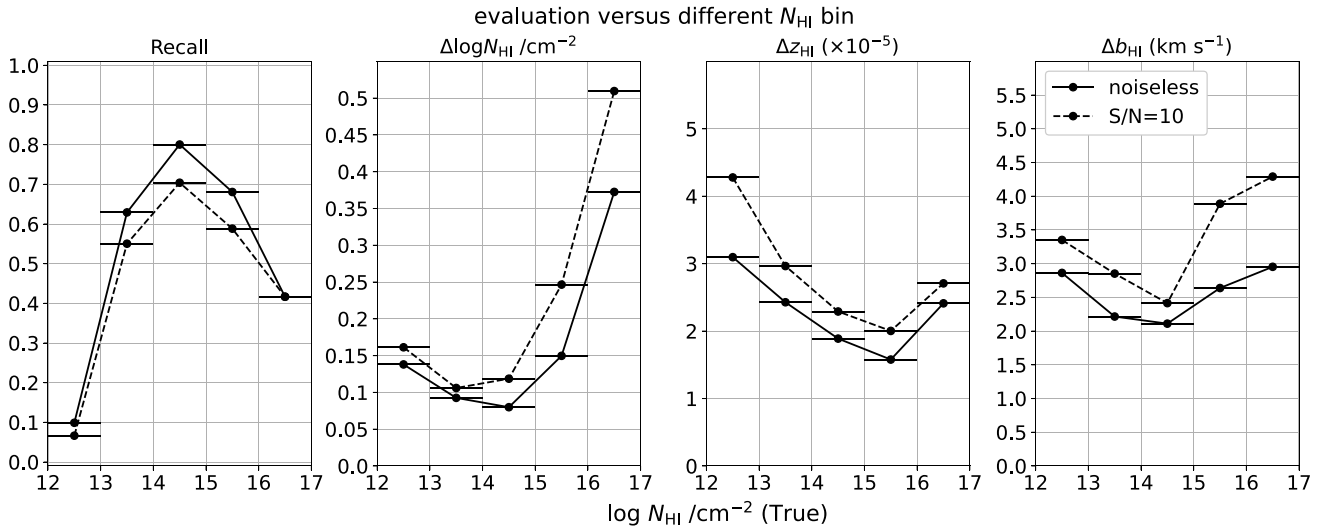


Figure 8. Evaluation graphs of recall and the MAE of $\log N_{\text{HI}}/\text{cm}^{-2}$, redshift (z_{HI}), and b_{HI} values, within different $\log N_{\text{HI}}/\text{cm}^{-2}$ bins. The horizontal lines of each data point represent the range of each bin.

6 APPLICATION TO OBSERVATIONAL DATA

Following the setup described in Section 3, we train five individual CNN models, and each model is trained with a set of 900 noiseless spectra that we perturb with Gaussian noise (Section 2.1). We then use these CNN models to predict the Ly α forest parameters of the 15 HIRES quasar spectra from R12 (Section 2.3) and build a catalogue of Ly α absorption systems with minimum Ly α ID > 0.3 for each quasar spectrum. The final determinations of the physical properties (i.e. $\log N_{\text{HI}}$, z_{HI} , and $\log b_{\text{HI}}$) are a weighted-average of the predictions for a given absorption line, using Ly α ID as the weights. The mean value of Ly α ID is used as the final probability of a CNN prediction. In order to evaluate the performance of our CNN models, we compare the CNN-classified Ly α absorbers with the catalogue built by R12. These authors identified Ly α systems and estimated the HI column density, redshift, and Doppler width by Voigt profile fitting. To avoid the proximity effect, Ly α systems are excluded if they are within 3000 km s⁻¹ of the quasar. Additionally, each Ly α system identified by R12 was validated by confirming the existence of at least one other higher order Lyman series transition; when higher order Lyman series lines were available, they were jointly fit. Our network only uses the Ly α absorption line.

6.1 Predicting 15 HIRES observed spectra

Although our CNN reaches high precision (equation 4) when predicting fake spectra that only contain Ly α absorbers, observed quasar spectra are much more complex and challenging due to the existence of other Lyman series absorption lines and metal lines. Hence, we carry out a post-processing procedure to exclude CNN-classified absorption line systems that: (1) have a redshift greater than the quasar Ly α emission redshift; (2) are within a region including higher order Lyman series lines such as Ly β lines; or (3) do not exhibit a Ly β absorption line.

In detail, we first remove the systems with a CNN-estimated redshift larger than or equal to the quasar redshift. To avoid regions including other higher order Lyman series lines, we focus on the region where only Ly α absorbers of the Lyman series exist. This is carried out by removing systems located at the wavelengths blueward

from the potential highest-redshift Ly β absorber estimated by the quasar emission redshift. Additionally, as in R12, we remove systems that are within 3000 km s⁻¹ of the quasar to avoid the proximity effect. Finally, we examine the corresponding Ly β absorption lines for each CNN-classified Ly α system using the CNN-predicted redshift, column density, and Doppler width. A CNN-classified Ly α system is removed if the following criteria are satisfied: (1) the estimated Ly β flux is much lower than the observed flux, i.e. the difference of fluxes (CNN Ly β flux – observed flux) is negative and its absolute value > 1σ , where σ is the median value of the noise spectrum near the centre of the absorption line, defined by the FWHM (i.e. pixels within $\sim\pm 12.5$ km s⁻¹, see Section 5.1); and (2) the Ly β absorption line is not saturated, i.e. Ly β observed flux > 3σ (following the definition of saturation in Section 2.2).

Additionally, we add two additional flags to our CNN catalogue – *Ly β _inspec_flag* and *sml_Ly β _flag*. The former decides if the wavelength of a corresponding Ly β absorber of a CNN-classified system is within the observed wavelength range; thus, 1 if yes and 0 if no. The latter flag assesses if the estimated flux of a corresponding Ly β absorber can be hidden within the noise level, i.e. Ly β flux > $(1 - \sigma)$. If the Ly β absorption feature can be hidden within 1σ , this flag *sml_Ly β _flag* is set to 1, and the opposite case has *sml_Ly β _flag* = 0.

6.2 Comparison with R12 catalogue

To assess the confidence of the CNN results by the R12 catalogue, we focus on the spectral region that only contains Ly α absorption lines (Section 6.1). We match the CNN-classified Ly α absorbers with the R12 catalogue using the same criteria for simulated spectra described in Section 5.1, i.e. the velocity difference between the identified systems of our CNN and R12 is smaller than half of the minimum FWHM that can be detected by our machine: ~ 12.5 km s⁻¹.

Since the R12 catalogue is based on a consistent fit of all available Lyman series lines (i.e. Ly α and at least one other Lyman series transition), a mismatch between R12 and our CNN could happen if: (1) the corresponding Ly β absorber of a CNN-classified system is out of the wavelength range of the spectrum, or (2) weak Ly β absorption lines that are buried within the noise level of a broad absorption

feature. By applying two additional flags¹³: (1) $Ly\beta_inspec_flag = 1$ and (2) $sml_Ly\beta_flag = 0$, ~ 78 percent of the ML-classified Ly α systems are matched with R12, i.e. precision = 0.78.¹⁴ The comparison of column density ($\log N_{HI}/\text{cm}^{-2}$) and Doppler width (b_{HI}) between our CNN and R12 are shown in Fig. 9 (also check Table 2). Dark blue data points are the median values of each bin of R12. The bin interval is 0.5 for $\log N_{HI}/\text{cm}^{-2}$ and 5 km s⁻¹ for b_{HI} . There is a statistical uncertainty associated with each quantity in the R12 catalogue based on Voigt profile fitting; the x-axis error bar uses the median value of the deviations to represent the typical error of each quantity in R12 within different bins. On the other hand, the y-axis error bar presents the MAE of the data points in each bin.

Compared to the simulated spectra results in Fig. 5, the CNN performance decreases when predicting real spectra. This is due to the more complex blending of features that are seen in observational data. As discussed in Section 5.3, our CNN tends to use only one line to recover a broad absorption feature while it is generally composed of multiple neighbouring lines. In this case, even though there is a matched Ly α system between the two catalogues, the CNN predictions of $\log N_{HI}$ and b_{HI} will not be consistent with the values listed in the R12 catalogue.

Nevertheless, our CNN models do a good job in predicting HI column density $\log N_{HI}/\text{cm}^{-2}$ with MAE = 0.139. In particular, the range between $12.5 \leq \log N_{HI}/\text{cm}^{-2} < 15.5$ shows a tight one-to-one relation with MAE = 0.135 (also see Table 2). Outside this column density range, the number of CNN-classified Ly α systems are much fewer (42 out of 1930 CNN-classified Ly α systems) which results in a larger scatter within this range. This indicates that our CNN has difficulty in correctly classifying these absorption lines and leads to a relatively poor estimate of the HI column density for Ly α systems with $\log N_{HI}/\text{cm}^{-2} < 12.5$ or $\log N_{HI}/\text{cm}^{-2} > 15.5$. Note that the R12 data are of considerably higher S/N compared to the simulated data that were used to train our CNN model. As described in Section 5.3, weak low HI column density systems are often buried in noise near the continuum level for data of S/N $\simeq 10$. To improve the prediction of lower HI column density systems, one may train a model with higher S/N input spectra, and apply this model to observational data of comparably high S/N (see Appendix D).

The catalogue comparison of b_{HI} (bottom panel of Fig. 9) shows a similar trend to the results of simulated spectra in Fig. 5 with a larger scatter. Within the range of $b_{HI} < 20$ km s⁻¹, our CNN tends to overestimate the b_{HI} value, because the CNN models use a single Ly α line to recover a feature that is composed of multiple lines. On the other hand, for larger b_{HI} values, our CNN has difficulty in predicting systems with broader absorption features due to the restriction of the scanning window size [see point (iii) of the FN summary in Section 5.3]. The window size is a hyperparameter tuned to optimize the predictions for the majority of Ly α absorbers (Section 3.2). There are ~ 93 per cent of Ly α absorbers with $b_{HI} < 35$ km s⁻¹ from R12 (~ 97 per cent of them with $b_{HI} < 40$ km s⁻¹), and the predictions for the systems with a larger b_{HI} value are worse (this also occurred for the simulated spectra).

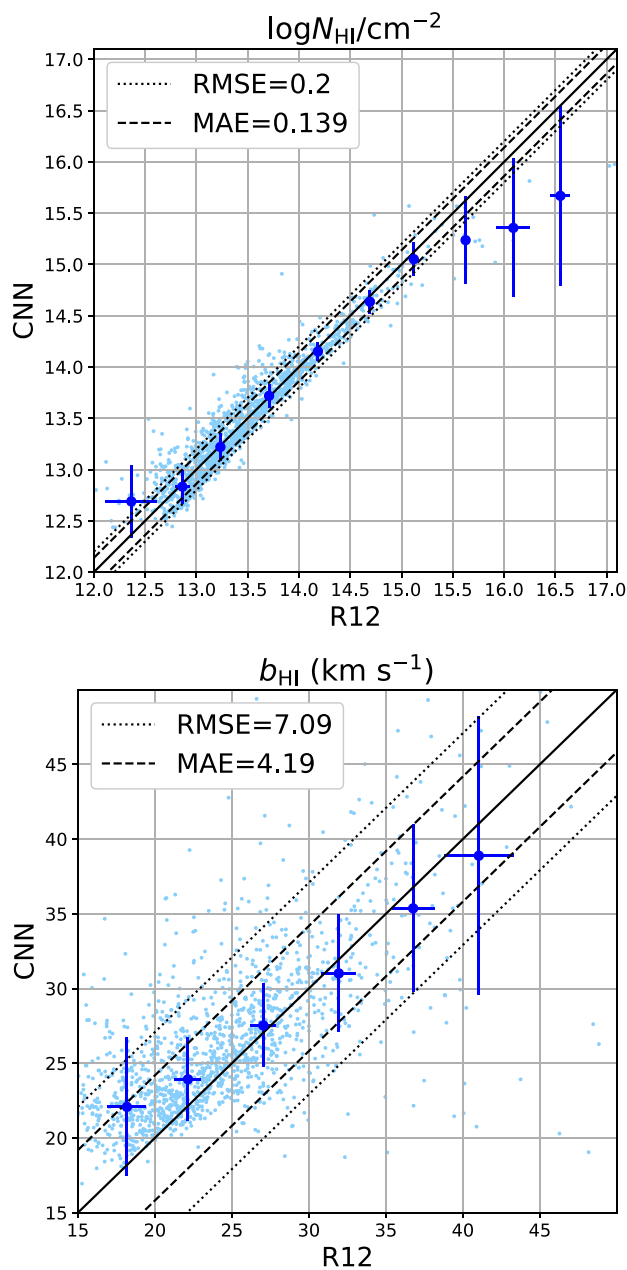


Figure 9. Comparisons between the CNN and R12 values of $\log N_{HI}/\text{cm}^{-2}$ (top) and b_{HI} (bottom). The black solid line shows a one-to-one relation, the dashed lines indicate the scatter defined by the MAE of each plot, and the dotted lines are defined by the RMSE of each plot. Dark blue data points show the median values of CNN and R12 within different bins of R12. The x-axis error bar is defined by the median value of the estimation errors of the data points provided in R12 within different bins of R12, while the y-axis error bar presents the MAE of each bin.

Finally, for different column density bins (defined using the R12 catalogue), we present the recall and the MAE of $\log N_{HI}/\text{cm}^{-2}$, redshift (z_{HI}), and b_{HI} in Fig. 10. Based on the column density comparison shown in Fig. 9, we separate matched samples into four bins: $\log N_{HI}/\text{cm}^{-2} = 12.0-12.5$, $12.5-14.0$, $14.0-15.5$, and $15.5-17.0$. Fig. 10 demonstrates that the CNN predictions of different physical properties are most consistent with R12 within the HI column density range $12.5 \leq \log N_{HI}/\text{cm}^{-2} < 15.5 \text{ cm}^{-2}$ (Table 2).

¹³Without applying the two flags, one can just compare the systems that are within the same redshift ranges as the ones that were fit in R12. The precision, recall, and MAE of $\Delta \log N_{HI}/\text{cm}^{-2}$, Δz_{HI} , and Δb_{HI} are 0.778, 0.284, 0.14 dex, 2.7×10^{-5} , and 4.3 km s⁻¹, respectively.

¹⁴It may be possible to include Ly β lines in the training process. This may improve the precision of the predictions of the Ly α lines. We leave this as an exercise for future work.

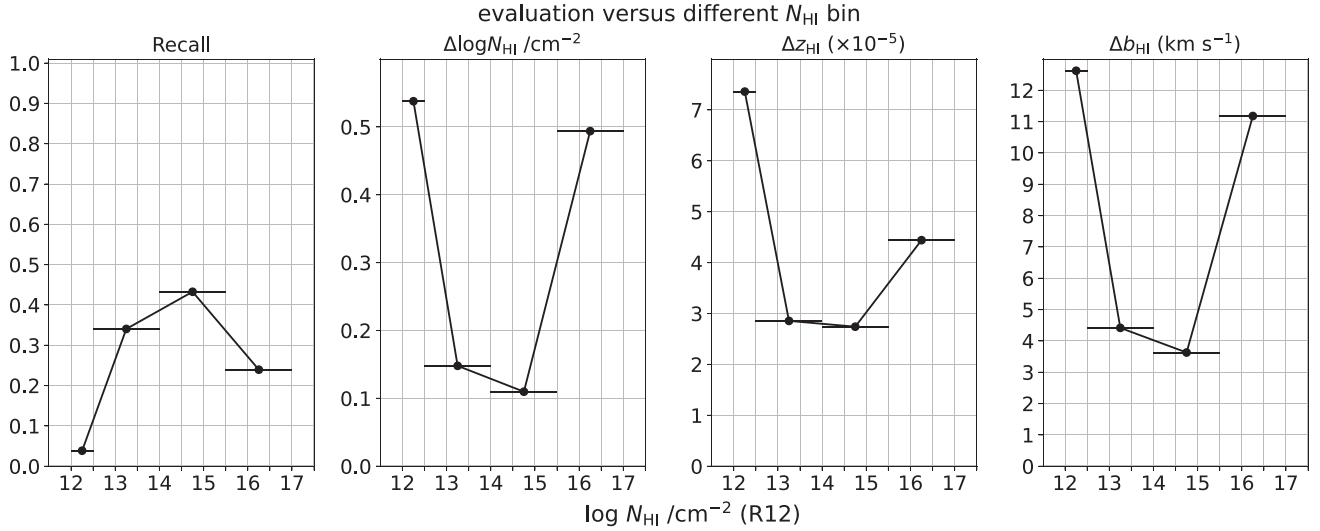


Figure 10. Evaluation graphs of recall and the MAE of $\log N_{\text{HI}}/\text{cm}^{-2}$, redshift (z_{HI}), and b_{HI} values, within different $\log N_{\text{HI}}/\text{cm}^{-2}$ bins. The horizontal lines of each data point represent the range of each bin; the intervals are 0.5, 1.5, 1.5, and 1.5, from low to high values of $\log N_{\text{HI}}/\text{cm}^{-2}$.

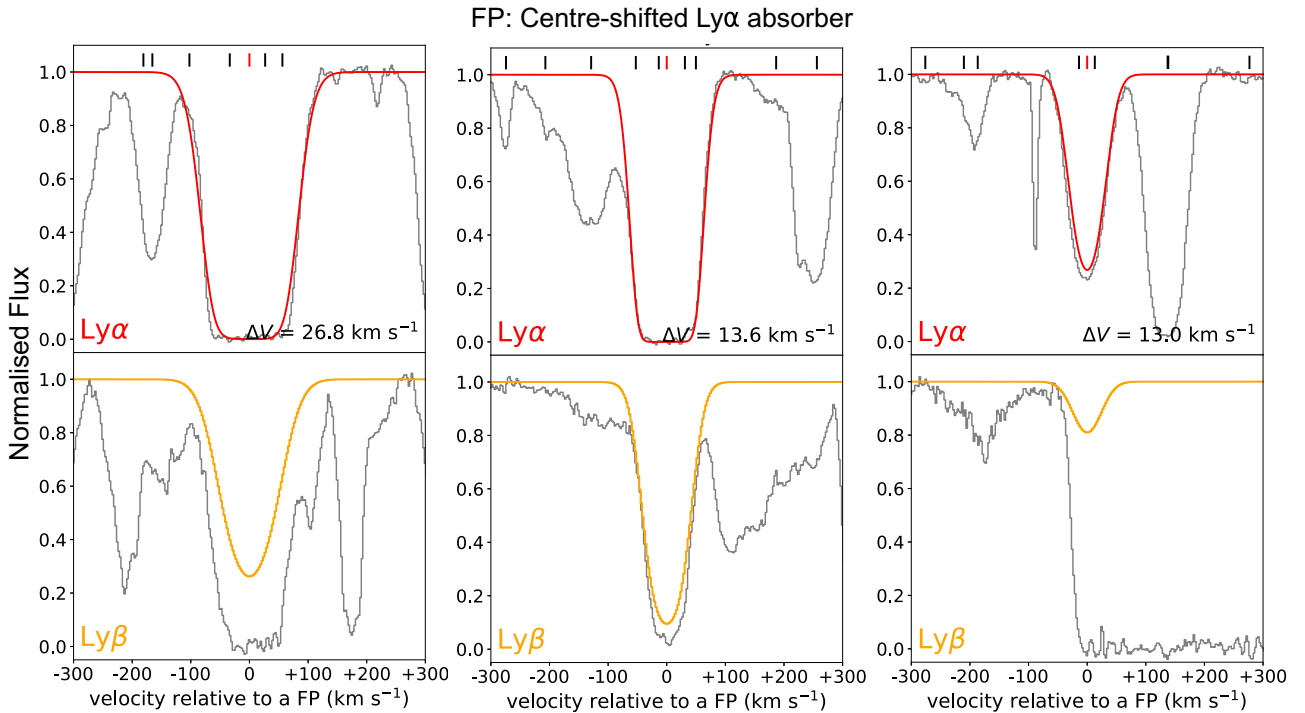


Figure 11. Examples of potential $\text{Ly}\alpha$ absorbers (red curves; top panels) that are not matched with an absorber in **R12** due to multiple components in **R12** being fit with a single absorber by the CNN. Due to velocity differences between the two catalogues, they are flagged as FPs although the absorber is identified in both catalogues. The centre of the FP is labelled by a red short tick mark. These examples of FPs are due to the matching criterion (see Section 5.2). The orange curves in the bottom panels show the corresponding $\text{Ly}\beta$ absorber. The grey histograms show the observational data. Black short tick marks above the spectra indicate the centres of the $\text{Ly}\alpha$ systems from **R12**. The ΔV values shown in the top panels provide the velocity difference between the prediction and the nearest absorption line from the catalogue of **R12**.

Note that the CNN-classified $\text{Ly}\alpha$ absorbers for the above results are identified by at least one CNN model out of five models. To further improve the CNN predictions, one can impose a selection criterion to the number of the CNN models that identify a $\text{Ly}\alpha$ system. For example, by requiring that a $\text{Ly}\alpha$ absorber must be identified by at least two CNN models, the precision increases from 0.78 to 0.85, and the overall MAE for $\log N_{\text{HI}}/\text{cm}^{-2}$ improves slightly, and the MAE

for b_{HI} drops to 3.8 km s^{-1} , respectively (see the results of other test data sets in Table 2).

6.2.1 False positive and false negative

Except for misclassification, which is dominated by contaminating metal lines, one of the primary causes of FPs in the observational data

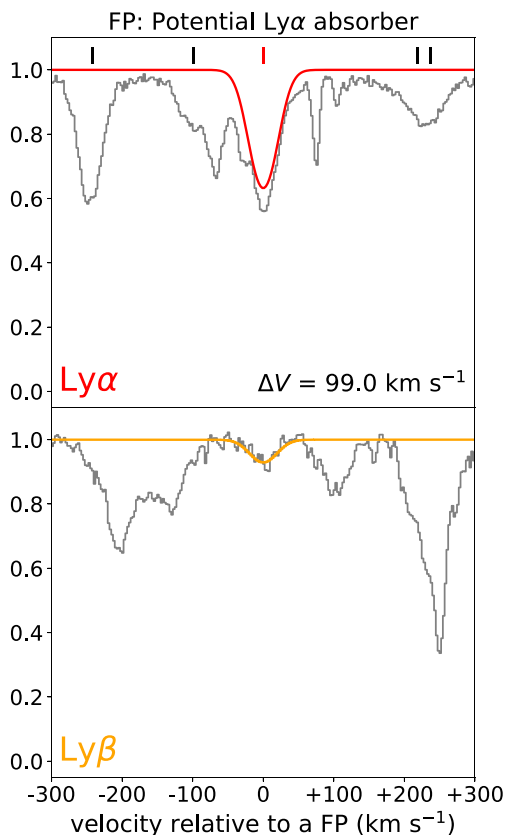


Figure 12. Same as Fig. 11, but shows an example of a potential Ly α absorber that is not listed in the R12 catalogue.

is due to the velocity threshold ($\sim 12.5 \text{ km s}^{-1}$) used to match systems between the CNN and R12 catalogues (as discussed in Section 5.2). Fig. 11 shows three examples of this FP case. The CNN tends to fit a broad absorption feature with one line, while it is composed of multiple neighbouring lines in the R12 catalogue. We notice that the broader an absorption feature is, the worse CNN predictions are obtained, e.g. the leftmost panel in Fig. 11. Additionally, since the classifications in R12 might have missed some Ly α systems from manual Voigt profile fitting, some FPs by our CNN could be a potential Ly α absorber. Examples are shown in Fig. 12. We compared the CNN identified systems with the robust Ly α absorbers from R12 which were validated with higher order lines, e.g. Ly β , Ly γ , etc. Hence, there may be mismatch because the corresponding Ly β or Ly γ lines of a potential Ly α absorber are difficult to detect. For example, in Fig. 12, we show an example with possible Ly α and Ly β absorption consistent with a true absorption system. However, this example may instead be due to a metal line absorber, given that there are several neighbouring metal line absorbers nearby.

The FNs identified with the observational data are contributed by the same sources as the ones discussed in Section 5.3 (i.e. low column density absorption features that are buried in the noise). As mentioned in Section 5.3, our choice to train a model on low S/N data is a compromise to allow a CNN technique to be applied to spectra with both low and high S/N. For completeness, we have also trained a model with S/N closer to the quasar spectra from R12 and we test this model using observed spectra. This comparison is discussed in Appendix D.

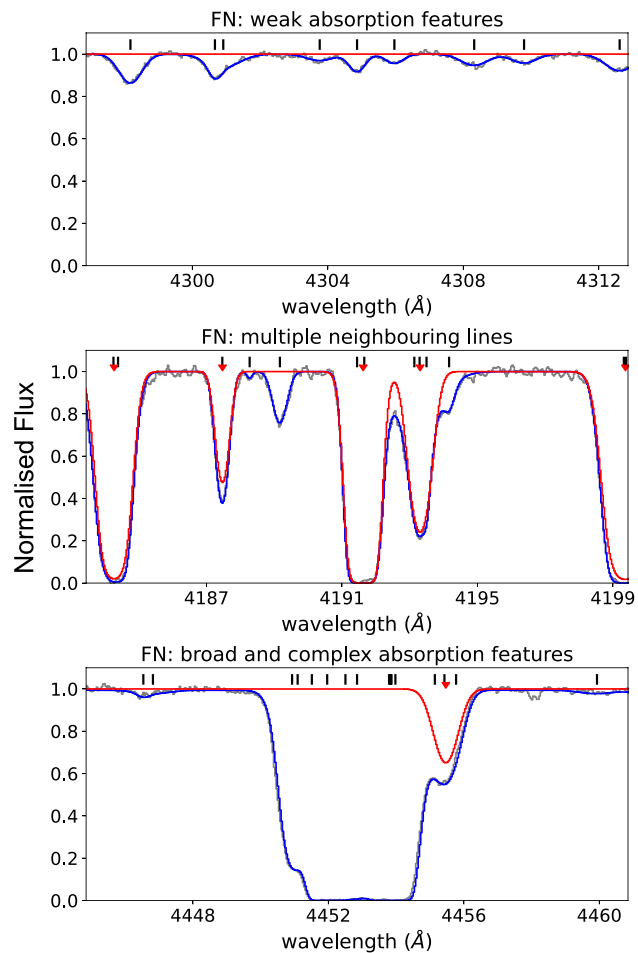


Figure 13. Several examples of FNs. The centres of each Ly α system from R12 are labelled by black short tick marks above the spectra, while the red arrow indicates a match between the CNN and R12 results. From the top to bottom panel, we showcase different reasons responsible for FNs. The grey curve represents the data, the blue curves represent a reconstruction of the data based on the Ly α absorbers listed in the R12 catalogue, and the red curve is a reconstruction based on all of the predicted lines by our CNN.

In Fig. 13, we showcase examples of the different cases of FNs. The top panel presents the case of FNs having weak absorption features that are missed by our CNN, which is trained with noisy spectra. In the middle panel, our CNN uses one Ly α line to describe the absorption feature, while there are multiple nearby lines listed in R12 responsible for this feature. This specific case also contributes a FP depending on the distance between a CNN-classified system and its closest system from R12. Finally, we showcase a broad and complex absorption feature containing many Ly α absorbers in the bottom panel. Our CNN has difficulty analysing a broader feature such as the one showcased here, since our CNN is trained with only primary lines. When an absorption feature is broader than the structure that our CNN can reconstruct with one Ly α line, the CNN fails to classify.

Since the behaviour of FNs using our CNN can be determined empirically, a correction factor can be calculated to convert the predicted distribution of Ly α forest absorbers to the intrinsic (i.e. input) distribution of Ly α absorbers. We will consider this approach in a future paper.

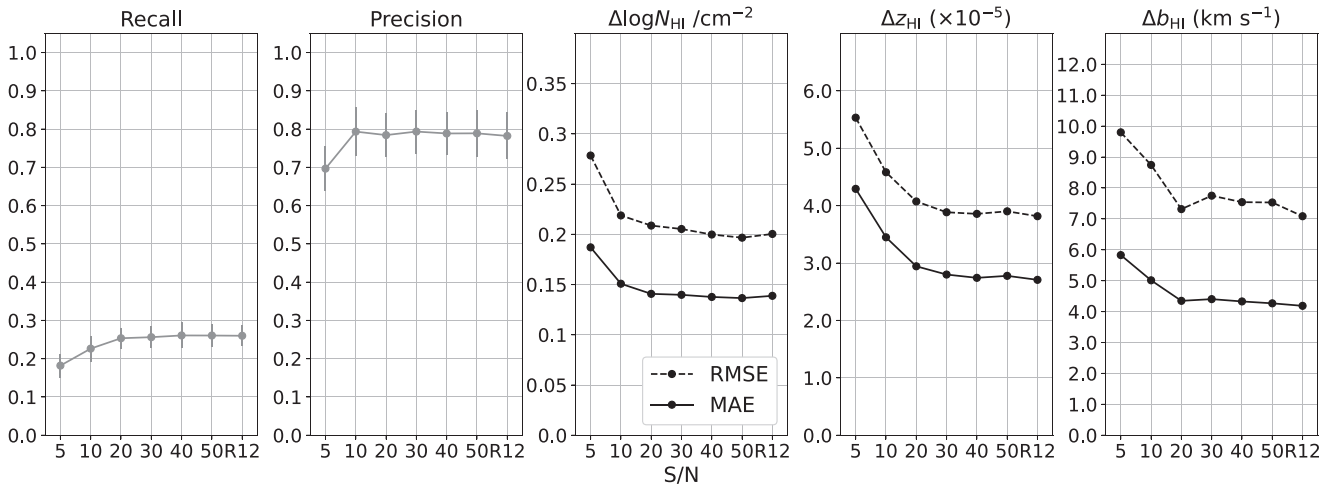


Figure 14. Our CNN predictions of the R12 HIRES spectra are stable when we artificially degrade the R12 data. Note that the ‘R12’ data points represent the predictions to the high quality HIRES spectra from R12 ($S/N \gtrsim 50$). From left to right, we show the recall, precision, the RMSE (black dashed line) and MAE (black solid line) of H I column density ($\Delta \log N_{\text{HI}}/\text{cm}^{-2}$), redshift (Δz_{HI}), and the Doppler width (Δb_{HI}).

6.2.2 Predicting HIRES spectra with different S/N

In this section, we test if our CNN is capable of predicting observed Keck/HIRES spectra of different S/N. Additional noise is added to the high quality HIRES spectra from R12 to degrade the S/N. We test different cases from $S/N = 5$ to 50 (i.e. the latter case represents the lower S/N end of the R12 spectra). This test is to ensure that in future works, we can further apply our trained CNN models to predict spectra in the HIRES archives such as the Keck Observatory Database of Ionized Absorption toward Quasars (KODIAQ) survey (O’Meara et al. 2017, 2021). Fig. 14 demonstrates that the performance of our CNN is consistent with the predictions of simulated spectra (see Appendix C). This again confirms that training the CNN with noisy spectra is of great importance to stabilize the predictions of spectra with different noise levels. Although there is a drop in the CNN performance at $S/N < 20$, the changes are still within an acceptable range for further scientific analyses. By training and testing a CNN applied to high redshift quasar spectra, we have opened up the possibility to efficiently and effectively harvest the information buried in the Ly α forest. This is an important step towards understanding and analysing the significant amount of data that will be acquired with future facilities.

7 SUMMARY

We have developed a machine learning based detection algorithm using CNNs to derive the physical parameters of Ly α absorbers within the forest of high-resolution QSO absorption line spectra. In particular, we focus on the low HI column density systems ($N_{\text{HI}} < 10^{17} \text{ cm}^{-2}$) and predict their physical properties such as HI column density ($\log N_{\text{HI}}/\text{cm}^{-2}$), redshift (z_{HI}), and Doppler width (b_{HI}). The low column density Ly α absorbers serve as a great tracer to the thermal history of the low-density IGM and can be used to probe the baryonic matter distribution. However, since they can be easily contaminated by other Lyman series and metal lines, previous applications of machine learning to the Ly α forest have focused on identifying DLAs ($N_{\text{HI}} \geq 10^{20.3} \text{ cm}^{-2}$) which show strong, damped absorption features.

Our CNN model is trained with 900 noisy simulated spectra with a S/N drawn from a Gaussian distribution of mean = 10 and

standard deviation = 2. This training strategy stabilizes the CNN performance when predicting spectra of different S/N (Appendix C and D) and allows us to apply our CNN models to the current archives of spectroscopic data, as well as future surveys. The simulated spectra that we use for training our model represent quasars at redshift $z = 3$ and are convolved with an instrumental resolution of $v_{\text{FWHM}} = 7 \text{ km s}^{-1}$. These values are typical of the data in current observatory archives. Different FWHM values have no impact on the performance of the CNN model (i.e. the Ly α forest absorption lines are fully resolved), while at higher redshifts there is increased blending due to neighbouring absorption features, which negatively impacts the accuracy of the CNN predictions (see Appendix A). The pixel size of the simulated spectra is set to 2.5 km s^{-1} .

We first examine the CNN performance with simulated spectra, and match the CNN prediction and true systems using a velocity threshold defined by half of the minimum FWHM $\sim 12.5 \text{ km s}^{-1}$ (estimated by $b_{\text{HI}} = 15 \text{ km s}^{-1}$). By matching the predicted systems with the systems listed in the true catalogue, over 99 per cent of the CNN-classified Ly α systems are true. However, the completeness is low (~ 32 per cent), i.e. only a small fraction of the Ly α systems are identified by our CNN. We summarize three types of FN: (1) weak absorption features that might be neglected by our CNN due to the limitation of the noisy training spectra; (2) a strong absorption feature composed of multiple neighbouring lines, contributing one FP and many FNs; (3) broad and complex absorption features that cannot be represented by one Ly α absorber. Case (1) dominates the FN; the completeness increases to 77 per cent when excluding this case of FN.

We then train five individual CNN models to predict 15 HIRES spectra and compare the CNN predictions with the results of manual Voigt profile fitting by R12. While the manual method costs 1–2 yr for the 15 spectra in R12, the prediction process by our CNN costs less than 3 min per quasar spectrum with a size of $\sim 120\,000$ pixels using a MacBook Pro with a 2.3-GHz Intel Core i7 processor and Intel Iris Plus Graphics 1536 MB.

Since an observed spectrum contains complex structures and contamination such as metal lines, a post-processing procedure is carried out to exclude unreliable ML-classified Ly α systems. Around 78 per cent of ML-classified Ly α systems are matched with R12.

There are three sources responsible for FPs: (1) a simple mismatch due to the chosen velocity threshold ($\sim 12.5 \text{ km s}^{-1}$); (2) a potential Ly α absorber that is not listed in R12 due to weak or hidden Ly β absorption; and (3) misclassification due to broad absorption features formed by multiple blended metal lines. We further conclude that the CNN models provide the most reliable predictions within the range of $12.5 \leq \log N_{\text{HI}}/\text{cm}^{-2} < 15.5$. Within this range, the MAE of $\log N_{\text{HI}}/\text{cm}^{-2}$, z_{HI} , and b_{HI} are 0.13 dex, 2.7×10^{-5} , and 4.1 km s^{-1} , respectively, demonstrating the accuracy of our CNN predictions. We conclude that a general-purpose CNN applied to the Ly α forest may not be as effective as one that is trained for a specific science goal, and it is important to better understand the parameter space where a model succeeds or under performs. We found that the FNs occur under the same conditions for both simulated and observed spectra.

Although we train the CNN models with noisy ($S/N \simeq 10$) simulated spectra, they provide consistent performance when predicting much higher quality ($S/N \gtrsim 70$) observational spectra. This gives us confidence that our model can be applied to both cosmological simulations and observations of the Ly α forest, and help to provide an insight into some of the missing ingredients in simulations.

Finally, we examine the CNN performance when predicting observed Keck/HIRES spectra of different S/N, and draw the same conclusions as the analysis of the simulated spectra. An investigation can be further carried out to quantify the impact of different S/N on the ‘accuracy’ of the conventional analyses to observed spectra. More importantly, this result validates the possibility to apply a CNN model with our approach to analyse the enormous quantity of data that will be obtained with future facilities.

ACKNOWLEDGEMENTS

We thank an anonymous referee for a timely and thorough report that helped to clarify various aspects of the paper. T.-Y. Cheng acknowledges the support of STFC grant ST/T000244/1 and Royal Society grant RF/ERE/210326, hosted at Durham University, and the support by Towards Turing 2.0 under the EPSRC Grant EP/W037211/1 and The Alan Turing Institute. During this work, RJC was supported by a Royal Society University Research Fellowship. RJC acknowledges support from STFC (ST/T000244/1).

DATA AVAILABILITY

The observed Keck/HIRES spectra are publicly available on the Keck Observatory Archive. The machine learning code is not published, but may be shared upon request.

REFERENCES

Abel T., Haehnelt M. G., 1999, *ApJ*, 520, L13
 Agarap A. F., 2018, preprint (arXiv:1803.08375)
 Bailer-Jones C. A. L., Irwin M., von Hippel T., 1998, *MNRAS*, 298, 361
 Bainbridge M. B., Webb J. K., 2017, *MNRAS*, 468, 1639
 Baur J., Palanque-Delabrouille N., Yèche C., Magneville C., Viel M., 2016, *J. Cosmol. Astropart. Phys.*, 2016, 012
 Becker G. D., Rauch M., Sargent W. L. W., 2007, *ApJ*, 662, 72
 Bird S., Rogers K. K., Peiris H. V., Verde L., Font-Ribera A., Pontzen A., 2019, *J. Cosmol. Astropart. Phys.*, 2019, 050
 Boera E., Becker G. D., Bolton J. S., Nasir F., 2019, *ApJ*, 872, 101
 Bolton J. S., Viel M., Kim T. S., Haehnelt M. G., Carswell R. F., 2008, *MNRAS*, 386, 1131
 Bolton J. S., Oh S. P., Furlanetto S. R., 2009, *MNRAS*, 395, 736
 Bottrell C. et al., 2019, *MNRAS*, 490, 5390

Carswell R. F., Webb J. K., 2014, Astrophysics Source Code Library, record ascl:1408.015
 Caruana R., 1998, *Multitask Learning*. Kluwer Academic Publishers, USA, p. 95
 Cheng T.-Y. et al., 2020a, *MNRAS*, 493, 4209
 Cheng T.-Y., Li N., Conselice C. J., Aragón-Salamanca A., Dye S., Metcalf R. B., 2020b, *MNRAS*, 494, 3750
 Cheng T.-Y. et al., 2021, *MNRAS*, 507, 4425
 Cristiani S., D’Odorico S., Fontana A., Giallongo E., Savaglio S., 1995, *MNRAS*, 273, 1016
 Davé R., Hernquist L., Weinberg D. H., Katz N., 1997, *ApJ*, 477, 21
 Davé R., Oppenheimer B. D., Katz N., Kollmeier J. A., Weinberg D. H., 2010, *MNRAS*, 408, 2051
 de Jong R. S. et al., 2019, *The Messenger*, 175, 3
 Dekker H., D’Odorico S., Kaufer A., Delabre B., Kotzłowski H., 2000, in Iye M., Moorwood A. F., eds, Proc. SPIE Conf. Ser. Vol. 4008, Optical and IR Telescope Instrumentation and Detectors. SPIE, Bellingham, p. 534
 Fang Y., Duncan R. C., Crofts A. P. S., Bechtold J., 1996, *ApJ*, 462, 77
 Ferreira L., Conselice C. J., Duncan K., Cheng T.-Y., Griffiths A., Whitney A., 2020, *ApJ*, 895, 115
 Frazier P. I., 2018, preprint (arXiv:1807.02811)
 Gaikwad P., Srikanand R., Choudhury T. R., Khaire V., 2017, *MNRAS*, 467, 3172
 Garnett R., Ho S., Bird S., Schneider J., 2017, *MNRAS*, 472, 1850
 Garzilli A., Boyarsky A., Ruchayskiy O., 2017, *Phys. Lett. B*, 773, 258
 George D., Huerta E. A., 2018, *Phys. Rev. D*, 97, 044039
 GPpyOpt, 2016, GPpyOpt: A Bayesian Optimization framework in python. Available at: <http://github.com/SheffieldML/GPpyOpt>
 Haehnelt M. G., Steinmetz M., 1998, *MNRAS*, 298, L21
 Hui L., Gnedin N. Y., 1997, *MNRAS*, 292, 27
 Hui L., Rutledge R. E., 1999, *ApJ*, 517, 541
 Hummels C. B. et al., 2019, *ApJ*, 882, 156
 Iršič V., Viel M., Haehnelt M. G., Bolton J. S., Becker G. D., 2017, *Phys. Rev. Lett.*, 119, 031302
 Jones D. R., Schonlau M., Welch W. J., 1998, *J. Glob. Optimiz.*, 13, 455
 Kim T. S., Carswell R. F., Cristiani S., D’Odorico S., Giallongo E., 2002, *MNRAS*, 335, 555
 Kim T. S., Partl A. M., Carswell R. F., Müller V., 2013, *A&A*, 552, A77
 Kim T. S. et al., 2021, *MNRAS*, 501, 5811
 Kingma D. P., Ba J., 2015, in Bengio Y., LeCun Y., eds, 3rd Int. Conf. Learning Representations Vol. 1, Adam: A Method for Stochastic Optimization. ICLR, San Diego, CA
 LeCun Y., Bengio Y., Hinton G., 2015, *Nature*, 521, 436
 Lehner N., Savage B. D., Richter P., Sembach K. R., Tripp T. M., Wakker B. P., 2007, *ApJ*, 658, 680
 Lovell C. C., Acquaviva V., Thomas P. A., Iyer K. G., Gawiser E., Wilkins S. M., 2019, *MNRAS*, 490, 5503
 Lynds R., 1971, *ApJ*, 164, L73
 Meiksin A. A., 2009, *Rev. Mod. Phys.*, 81, 1405
 Metcalf R. B. et al., 2019, *A&A*, 625, A119
 Muthukrishna D., Narayan G., Mandel K. S., Biswas R., Hložek R., 2019, *PASP*, 131, 118002
 O’Meara J. M., Lehner N., Howk J. C., Prochaska J. X., Fox A. J., Peeples M. S., Tumlinson J., O’Shea B. W., 2017, *AJ*, 154, 114
 O’Meara J. M., Lehner N., Howk J. C., Prochaska J. X., 2021, *AJ*, 161, 45
 Parks D., Prochaska J. X., Dong S., Cai Z., 2018, *MNRAS*, 476, 1151
 Pearson J., Maresca J., Li N., Dye S., 2021, *MNRAS*, 505, 4362
 Pieri M. M. et al., 2016, in Reylé C., Richard J., Cambrésy L., Deleuil M., Pécontal E., Tresse L., Vaugin I., eds, SF2A-2016: Proceedings of the Annual Meeting of the French Society of Astronomy and Astrophysics. Société Française d’Astronomie et d’Astrophysique, Lyon, p. 259
 Prochaska J. X., Wolfe A. M., 2009, *ApJ*, 696, 1543
 Prochaska J. X., Herbert-Fort S., Wolfe A. M., 2005, *ApJ*, 635, 123
 Prochaska J. X., Madau P., O’Meara J. M., Fumagalli M., 2014, *MNRAS*, 438, 476
 Puchwein E., Bolton J. S., Haehnelt M. G., Madau P., Becker G. D., Haardt F., 2015, *MNRAS*, 450, 4081

- Rasmussen C. E., Williams C. K. I., 2006, *Gaussian Processes for Machine Learning: Adaptive Computation and Machine Learning*. MIT Press, Cambridge, MA
- Rauch M., 1998, *ARA&A*, 36, 267
- Ricotti M., Gnedin N. Y., Shull J. M., 2000, *ApJ*, 534, 41
- Rogers K. K., Peiris H. V., 2021, *Phys. Rev. Lett.*, 126, 071302
- Rogers K. K., Peiris H. V., Pontzen A., Bird S., Verde L., Font-Ribera A., 2019, *J. Cosmol. Astropart. Phys.*, 2019, 031
- Ruder S., 2017, preprint ([arXiv:1706.05098](https://arxiv.org/abs/1706.05098))
- Rudie G. C. et al., 2012a, *ApJ*, 750, 67 (R12)
- Rudie G. C., Steidel C. C., Pettini M., 2012b, *ApJ*, 757, L30
- Rudie G. C., Steidel C. C., Pettini M., Trainor R. F., Strom A. L., Hummels C. B., Reddy N. A., Shapley A. E., 2019, *ApJ*, 885, 61
- Sargent W. L. W., Young P. J., Boksenberg A., Tytler D., 1980, *ApJS*, 42, 41
- Schaye J., 2001, *ApJ*, 559, 507
- Schaye J., Theuns T., Leonard A., Efstathiou G., 1999, *MNRAS*, 310, 57
- Schaye J., Theuns T., Rauch M., Efstathiou G., Sargent W. L. W., 2000, *MNRAS*, 318, 817
- Snoek J., Larochelle H., Adams R. P., 2012, *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 2. NIPS'12*. Curran Associates Inc., Red Hook, NY, p. 2951
- Theuns T., Leonard A., Efstathiou G., 1998, *MNRAS*, 297, L49
- Theuns T., Leonard A., Schaye J., Efstathiou G., 1999, *MNRAS*, 303, L58
- Theuns T., Bernardi M., Frieman J., Hewett P., Schaye J., Sheth R. K., Subbarao M., 2002, *ApJ*, 574, L111
- Tytler D. et al., 2004, *ApJ*, 617, 1
- van de Voort F., Springel V., Mandelker N., van den Bosch F. C., Pakmor R., 2019, *MNRAS*, 482, L85
- Viel M., Becker G. D., Bolton J. S., Haehnelt M. G., 2013, *Phys. Rev. D*, 88, 043502
- Vogt S. S. et al., 1994, in Crawford D. L., Craine E. R., eds, *Proc. SPIE Conf. Ser. Vol. 2198, Instrumentation in Astronomy VIII*. SPIE, Bellingham, p.362
- Walmsley M. et al., 2022, *MNRAS*, 509, 3966
- Wang B. et al., 2022, *ApJS*, 259, 28

APPENDIX A: THE IMPACT ON PREDICTING SPECTRA WITH DIFFERENT INITIAL SETUPS

We test our pre-trained CNN model on 25 newly generated spectra with: (1) quasars at different redshifts; (2) different instrument

resolution (FWHM); and (3) data that are sampled with different pixel size (vpix). This test is to validate the feasibility of our CNN model to predict observational spectra that were acquired with different setups.

First, for different quasar redshifts (top row in Fig. A1), the recall and precision remain consistent and the $\log N_{\text{H I}}/\text{cm}^{-2}$ prediction does not show significant deviation. However, the RMSE and MAE of $z_{\text{H I}}$ and $b_{\text{H I}}$ estimations increase as a quasar emission redshift increases. This means that the prediction accuracy decreases as a quasar redshift increases. This result is caused by the increased blending due to Ly α forest absorption lines at higher redshift. During the early Universe, H I gas clouds are more abundant and their absorption features overlap in velocity space. This overlap introduces additional uncertainty in the predicted physical properties; this is true for both a machine-learning based algorithm or conventional Voigt profile fitting. Even though the RMSE and MAE of $z_{\text{H I}}$ and $b_{\text{H I}}$ are within a factor of ~ 1.5 of the RMSE and MAE at $z \simeq 3$, we conclude that a CNN tailored to a specific redshift may further improve the results, depending on the science application.

In the middle row of Fig. A1, we generated spectra with different instrument FWHM resolution over a narrow range $5 < \text{FWHM}/(\text{km s}^{-1}) < 9$, which samples the relevant resolutions of current high dispersion spectrographs like Keck/HIRES and ESO/UVES (European Southern Observatory Ultraviolet and Visual Echelle Spectrograph; Dekker et al. 2000). Overall, different FWHMs in this range show no impact to both the detection and physical property estimates. This is because the widths of the Ly α forest absorption lines ($\text{FWHM} \gtrsim 15 \text{ km s}^{-1}$) are usually fully resolved at the instrument resolution of typical spectrographs, such as Keck/HIRES and VLT/UVES ($\text{FWHM} \simeq 7 \text{ km s}^{-1}$). Finally, in the bottom row of Fig. A1, we show that the choice of pixel size introduces a serious impact to the results, in particular the Doppler width, $b_{\text{H I}}$. However, this issue can be circumvented by resampling. If an input spectrum is not sampled with $\text{vpix} = 2.5 \text{ km s}^{-1}$, we resample the input data to ensure that our CNN model produces reliable results. This resampling process does not impact the trained network, nor the results, since the Ly α forest absorption lines are fully resolved.

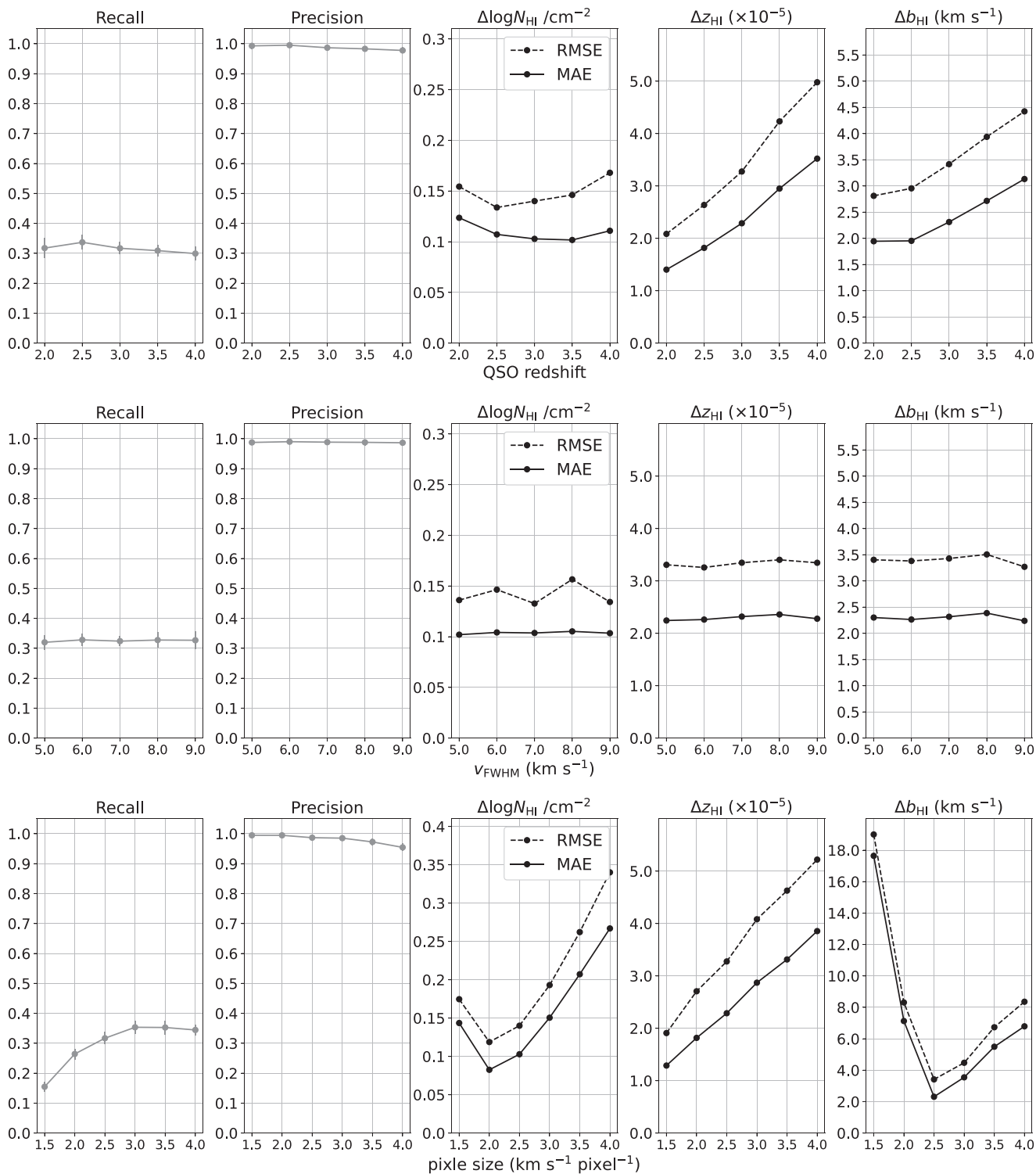


Figure A1. The overall performance of our CNN on noiseless spectra with different observing setups. We illustrate the sensitivity to quasar redshift (top row), instrumental FWHM resolution (middle row), and pixel size (bottom row). Note that the CNN model is trained using simulated noisy quasar spectra at redshift $z = 3$ and are observed using an instrumental FWHM resolution of $v_{\text{FWHM}} = 7 \text{ km s}^{-1}$ and sampled with $v_{\text{pix}} = 2.5 \text{ km s}^{-1} \text{ pixel}^{-1}$ (Section 2.1). From left to right-hand, the panels show the recall, precision, RMSE (black dashed line), and MAE (black solid line) of H I column density ($\Delta \log N_{\text{HI}} / \text{cm}^{-2}$), redshift (Δz_{HI}), and the Doppler width (Δb_{HI}).

APPENDIX B: BAYESIAN OPTIMIZATION

The predictions by a network strongly depend on its hyperparameters such as the number of neurons, dropout rate, the kernel sizes, etc. A failed prediction of a network might be simply due to an unoptimized architecture used for training. Hence, it is of great importance to select a set of hyperparameters that provide the most optimal combination for a specific goal. This selection is often done by a brute-force method – grid searches – such that all possible combinations of each hyperparameter are evaluated. This method is therefore extremely time-consuming, and the tested sets of hyperparameters are limited due to computational allowance.

Unlike grid searches, Bayesian optimization (Snoek et al. 2012) provides a ‘smart guess’ to approach an optimal combination of hyperparameters: x_1, x_2, \dots, x_n , where n represents the number of hyperparameters. This process is much faster than the grid searches to find a set of hyperparameters that performs well. The concept is to model the network’s function $f(x_1, x_2, \dots, x_n)$ to a surrogate analytical function. In this work, we use a Gaussian process (Rasmussen & Williams 2006) which forms the prior distribution as multivariate normal distributions. As providing data, the posterior probability distribution f given $f(x_1, x_2, \dots, x_n)$ is computed, and approaches to the prior using a chosen acquisition function which we use the default function – Expected Improvement (Jones, Schonlau & Welch 1998). A detailed tutorial is described in Frazier (2018).

Our optimization process uses GPYOPT (GPpyOpt 2016)¹⁵ by running 60 iterations to search for the optimal set. A final set of hyperparameters for the model trained with noisy quasar spectra (where the S/N is drawn from a Gaussian distribution of a mean = 10 and a standard deviation = 2) is listed in Table 1.

¹⁵<http://sheffieldml.github.io/GPyOpt/>

APPENDIX C: THE IMPACT OF DIFFERENT S/N ON MOCK SPECTRA

We tested the impact of training a CNN model using spectra of different noise levels. In Fig. C1, we show the results of training a CNN model with noiseless spectra (top; hyperparameters used in the CNN architecture are shown in Table C1) and noisy spectra (bottom; Table 1) with S/N drawn from a Gaussian distribution with a mean of 10 and a standard deviation of 2 (Section 3.1). This figure clearly shows that a model trained with noiseless spectra cannot be used to predict new spectra with even a modest amount of noise. Although this model can reach a higher overall recall for noiseless data, the precision and the parameter determinations of $\log N_{\text{H I}}/\text{cm}^{-2}$, $z_{\text{H I}}$, and $b_{\text{H I}}$ are poorly known when noise is added to the testing spectra.

Compared to this, a CNN model trained with spectra involving a distribution of S/N shows stable performance when predicting spectra with different noise levels. A drop in performance occurs to spectra with S/N < 20. Testing on spectra with S/N = 10, there is a drop in recall which does not decrease the precision. This indicates that many true Ly α absorbers might be hidden in the noise, and our CNN has difficulty to identify them. However, ~98 per cent of the CNN-classified Ly α systems are classified correctly compared with the list of true systems.

The precision then drops to ~0.94 when analysing spectra of S/N = 5, and there are significant changes to the RMSE and MAE for the estimates of the physical properties. We did not expect our CNN model to perform well when analysing spectra with S/N = 5, since this noise level is beyond the range included in our training spectra. However, the changes to the predictions are minor compared to the top panels of Fig. C1 using the model trained with noiseless spectra. Additionally, they are still within an acceptable range for scientific analyses. Hence, with caution, this CNN model can be used to analyse spectra with S/N > 5.

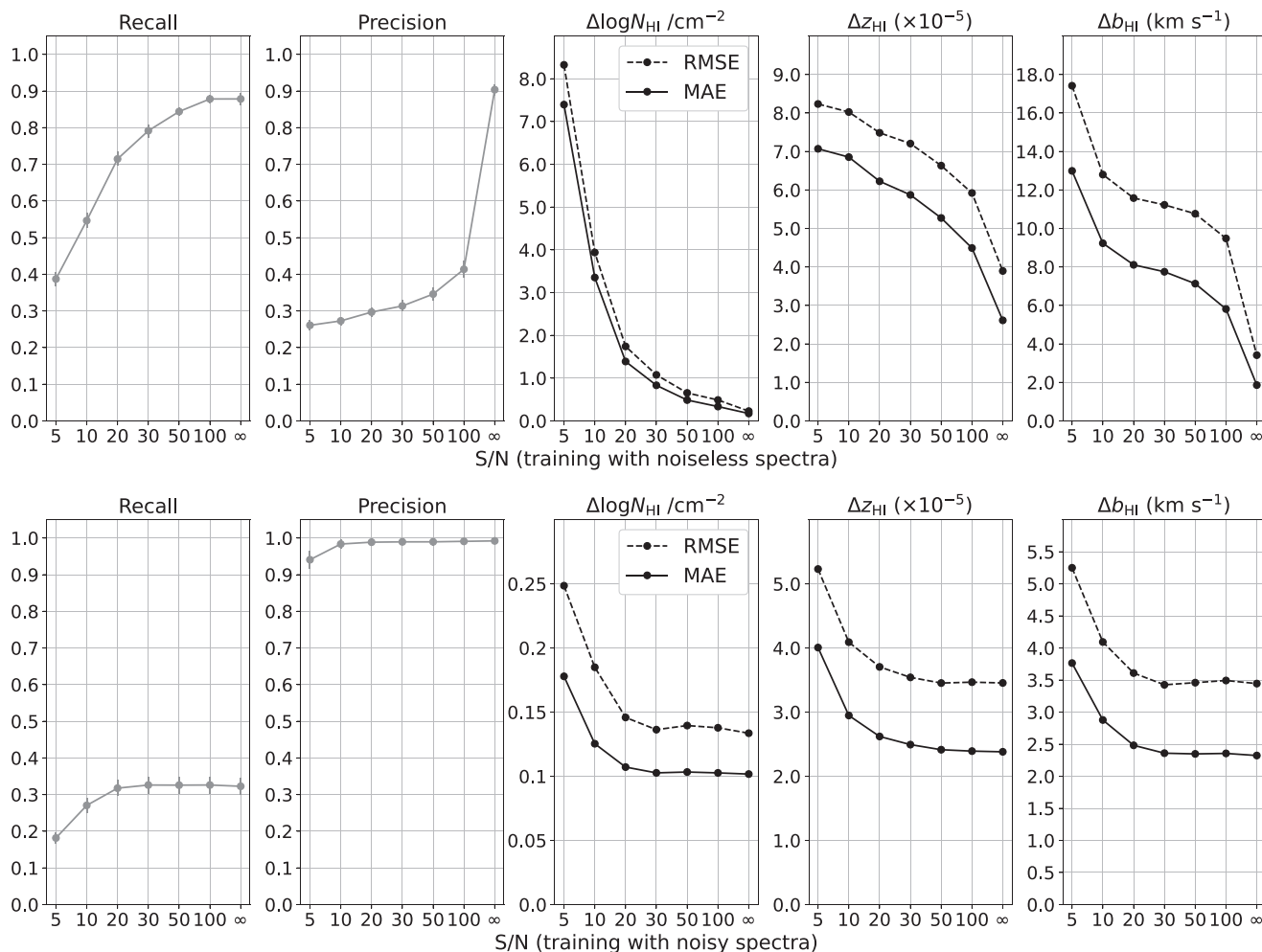


Figure C1. CNN analysis of spectra with different S/N using two CNN models trained with noiseless spectra (top row) and a Gaussian distribution of S/N with a mean = 10 and a standard deviation = 2 (bottom row), respectively. From left- to right-hand panels, we present the recall, precision, the RMSE (black dashed line) and MAE (black solid line) of HI column density ($\Delta \log N_{\text{HI}}/\text{cm}^{-2}$), redshift (Δz_{HI}), and the Doppler width (Δb_{HI}). Note that all metrics in the bottom row show a high level of stability for S/N > 20, demonstrating the more general success of training a CNN model with somewhat low S/N data.

Table C1. Hyperparameters used in a CNN architecture trained with noiseless spectra. These values are selected using a Bayesian optimization algorithm.

	Hyperparameters	Optimized value
Data Input	window size (w_s)	259
	$cnpix$	5
CNN Architecture	L2	0.0
	dropout	0.0
	conv_filter_1	128
	conv_filter_2	128
	conv_filter_3	128
	conv_kernel_1	8
	conv_kernel_2	7
	conv_kernel_3	8
	dense_1	128
	dense_2_ID	256
dense_2_N	512	
dense_2_z	256	
dense_2_b	128	

APPENDIX D: THE IMPACT OF DIFFERENT S/N ON R12 SPECTRA

Extending the discussion in Section C, we have trained a CNN model with spectra of a higher S/N than the one used in the main work, and tested this model with the observed spectra from R12. The S/N of the training data is drawn from a Gaussian distribution of S/N with a mean = 90 and a standard deviation = 30; these values are chosen to be close to the S/N of the R12 data. The optimized hyperparameters of this CNN architecture are listed in Table D1. Comparing Fig. D1 with Fig. 14, the recall increases slightly but the performance drops when S/N < 50. However, Fig. D2 shows that by training a model with high S/N spectra, it helps to improve the predictions of systems within the range of lower ($\log N_{\text{HI}}/\text{cm}^{-2} < 12.5$) and higher ($\log N_{\text{HI}}/\text{cm}^{-2} \geq 15.5$) column density.

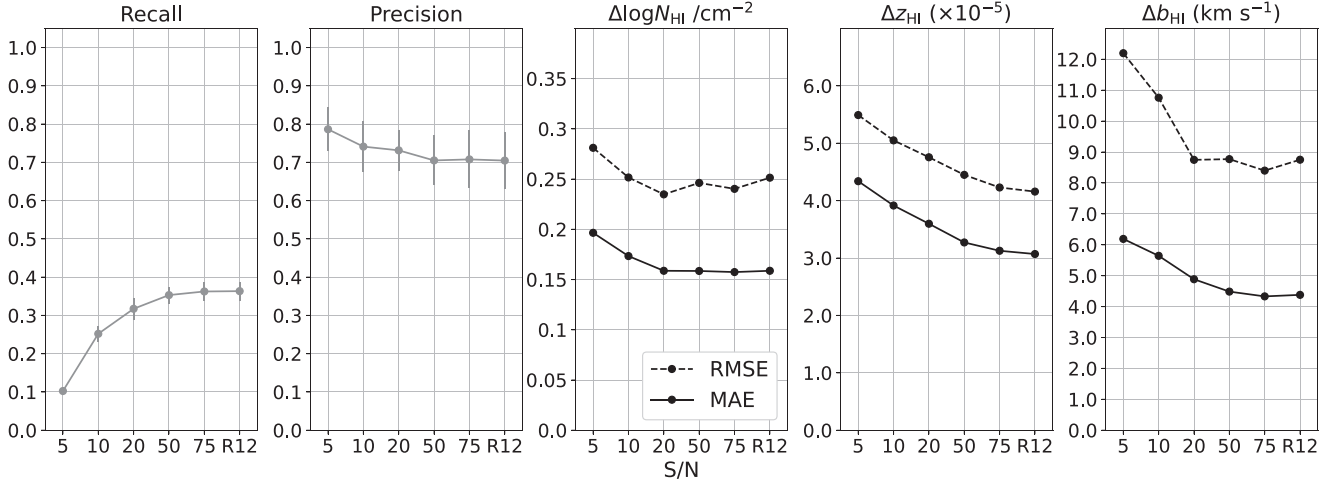


Figure D1. CNN analysis of R12 spectra using a Gaussian distribution of S/N with a mean = 90 and a standard deviation = 30 (closer to the distribution of R12 spectra). From left- to right-hand panel, we present the recall, precision, the RMSE (black dashed line), and MAE (black solid line) of H I column density ($\Delta \log N_{\text{HI}} / \text{cm}^{-2}$), redshift (Δz_{HI}), and the Doppler width (Δb_{HI}).

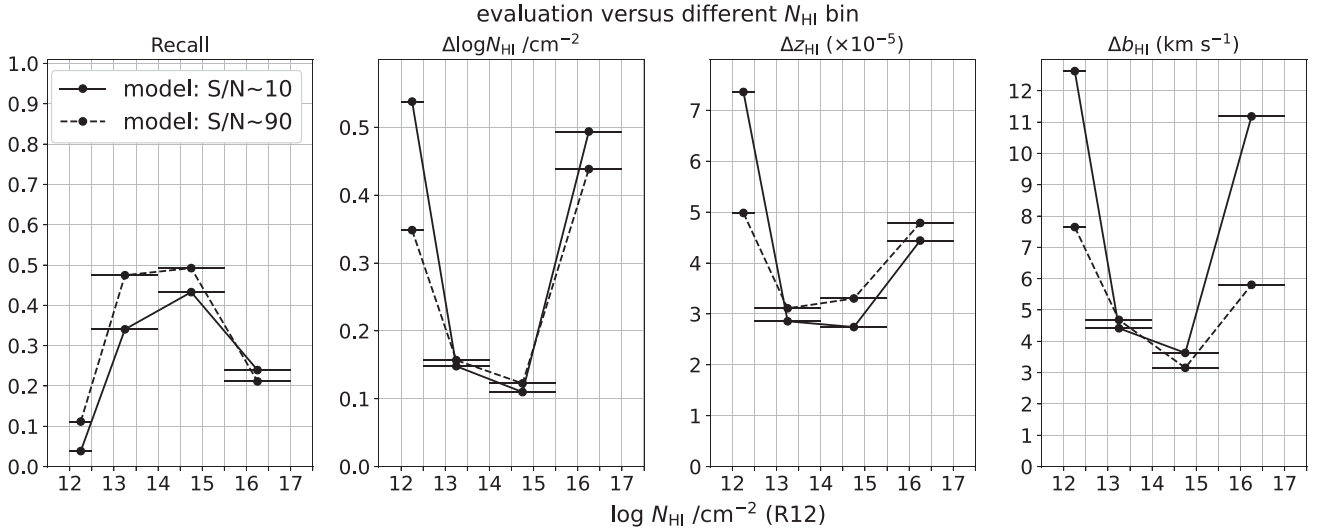


Figure D2. Evaluation graphs of recall and the MAE of $\log N_{\text{HI}} / \text{cm}^{-2}$, redshift (z_{HI}), and b_{HI} values, within different $\log N_{\text{HI}} / \text{cm}^{-2}$ bins. The horizontal lines of each data point represent the range of each bin; the intervals are 0.5, 1.5, 1.5, and 1.5, from low to high values of $\log N_{\text{HI}} / \text{cm}^{-2}$. The solid line shows the results using models trained by lower S/N (mean = 10) while the dashed line shows the ones using models trained by higher S/N (mean = 90).

Table D1. Hyperparameters used in a CNN architecture trained using spectra with a higher S/N. These values are selected using a Bayesian optimization algorithm.

	Hyperparameters	Optimized value
Data Input	window size (<i>ws</i>)	285
	<i>cpix</i>	1
CNN	L2	0.0
Architecture	dropout	0.6
	conv_filter_1	256
	conv_filter_2	256
	conv_filter_3	512
	conv_kernel_1	10
	conv_kernel_2	7
	conv_kernel_3	6
	dense_1	64
	dense_2_ID	512
	dense_2_N	512
	dense_2_z	256
dense_2_b	128	

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