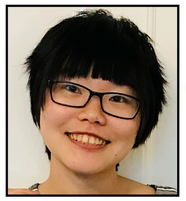


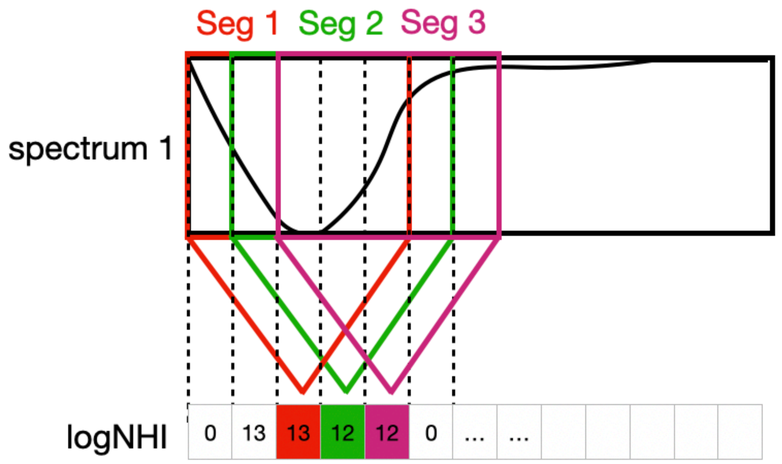
Predicting physical properties of Lyman alpha forest with deep learning

Ting-Yun Cheng (Durham), Ryan Cooke (Durham)

Abstract

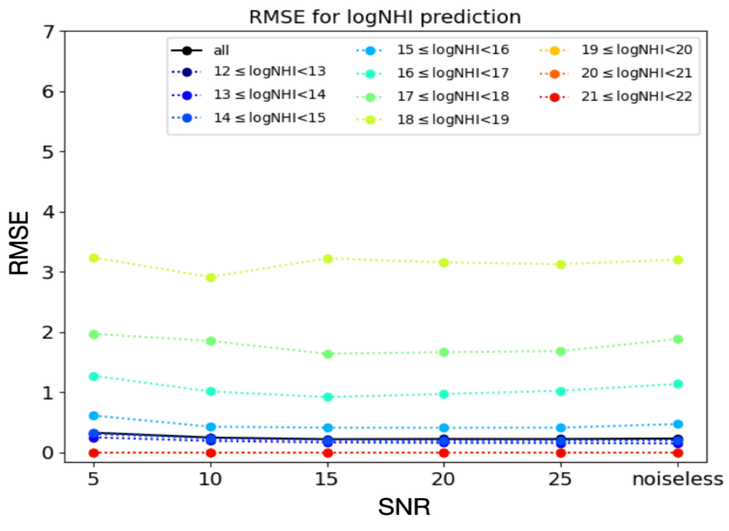
Lyman alpha (Lyα) forest records how light of a quasar interacts with neutral hydrogen atoms in intergalactic medium (IGM). These systems provide much information to cosmological studies. However, the conventional line detection methods are restricted to spectra with high signal-to-noise ratio (SNR) and are time-consuming to apply in analysis. In this work, we apply convolution neural networks (CNN) to scan through quasar spectra for Lyα absorption line detection and simultaneously predict physical properties such as column densities (logNHI), redshifts (z), and Doppler widths (b) of the identified Lyα systems. For systems with logNHI ≤ 18.0, our CNN, trained with simulated mock spectra, provides a state-of-art prediction in physical properties. We apply this model to build a catalogue of Lyα absorption systems for 300 quasar spectra from High Resolution Echelle Spectrometer (HIRES).

Methodology — Our CNN model is trained with several segments of quasar spectra. We scan through a spectrum and use the labels (logNHI, z, b) at the central pixel in a segmentation to train the CNN (as shown below).

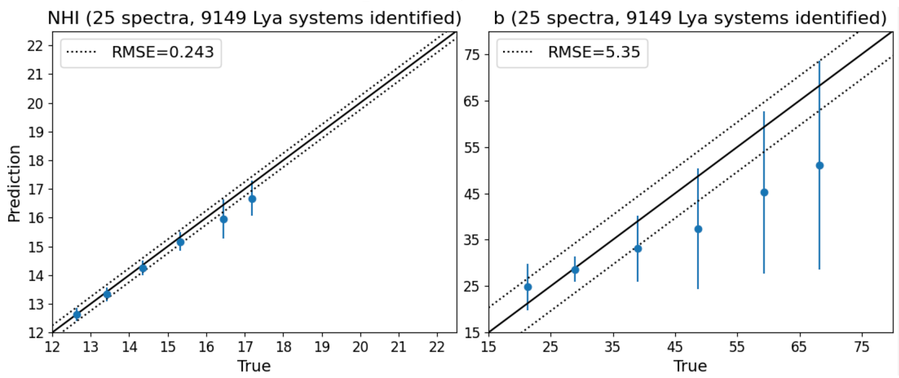


Mock spectra (for training) — Mock spectra for training is created using PYIGM[1] with a column density distribution function (CDDF) from Prochaska et al. (2014) and FWHM=7.0 km/s. Labels of physical properties (logNHI, b) from simulated spectra are used in the training process. Additional SNR is added and this stabilises the CNN predictions when applying to spectra with different SNR (figure shown as below).

1: https://pyigm.readthedocs.io/en/latest/#

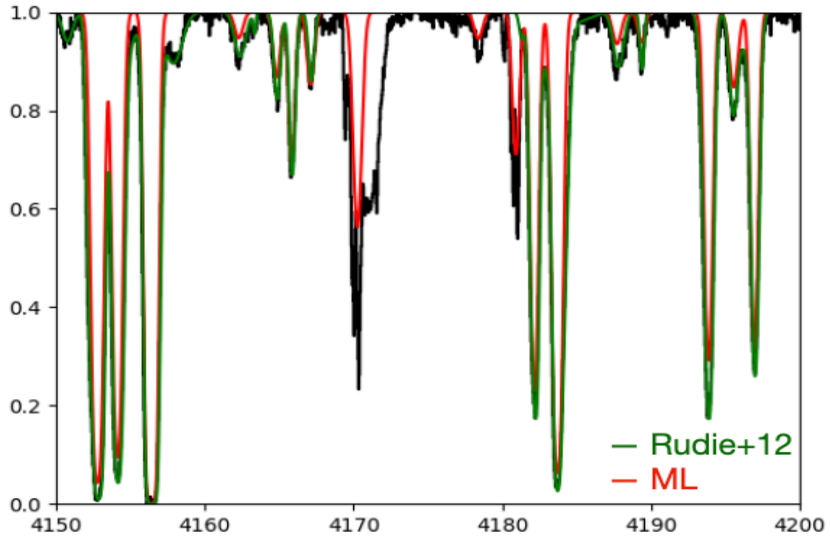


Results with mock spectra — The CNN model identifies Lyα absorption lines with logNHI < 18.0 from a mock quasar spectrum with an additional SNR of 10.0. An accuracy of over 0.85 for the identification is reached when applying a probability threshold of 0.5 to the predictions. The prediction of logNHI and b has a root-mean-square error (RMSE) of 0.243 cm^{-2} and 5.35 km/s, respectively.

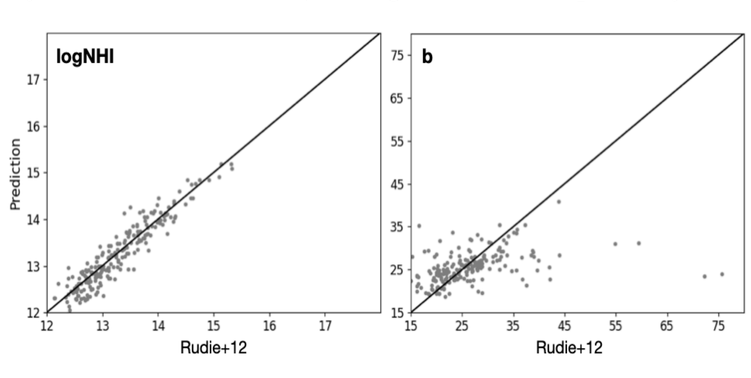


Observational spectra — Spectroscopic data from the High Resolution Echelle Spectrometer (HIRES) has R ~ 45,000 (FWHM=7.0 km/s). We apply our CNN model to them and compare our identification and predictions with a catalogue of Lyα absorption systems from Rudie et al. (2012).

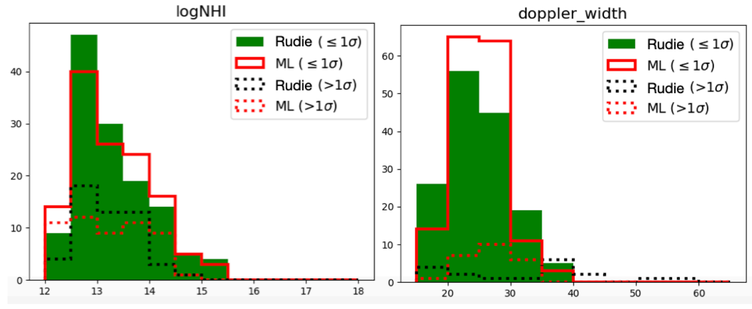
Results — Our CNN model trained with simulated mock spectra picked up corresponding Lyα systems listed in Rudie+12 (examples shown as below).



Results — We match the catalogue of Rudie+12 with our CNN model with a probability threshold of 0.5 using a redshift difference of 0.0001 (~6.52 pixels). The CNN shows a comparable result to the one using simulated mock spectra with a RMSE of 0.251 cm^{-2} and 8.89 km/s (5.37 if excluding systems with b>45 km/s) for logNHI and b, respectively.



The above figure combined with the histogram of logNHI and b, the CNN model shows great consistency with the measurement in Rudie+12. The mis-prediction (>1σ) in b occurs when a system has a broad feature (i.e. large b), our machine tends to under-estimate b for this kind of system.



Summary & Future Plan — The CNN model trained with simulated mock spectra shows a state-of-art predicting ability to observational data. This takes ~5 mins to generate a catalogue of Lyα forest for one quasar spectrum. We apply this technique to build a catalogue of 300 HIRES spectra. The paper of this work is in preparation. This technique can be easily adapted to other spectroscopic surveys. More importantly, with the generated Lyα forest catalogue, several studies of IGM can be carried out such as the thermal state of IGM. Follow-up works are in progress.