Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

MNRAS **526,** 4557–4574 (2023) https://doi.org/10.1093/mnras/stad2940
Advance Access publication 2023 September 27

Machine learning uncovers the universe's hidden gems: A comprehensive catalogue of C IV absorption lines in SDSS DR12

Reza Monadi ,[1,2]★ Ming-Feng Ho ,[1] Kathy L. Cooksey[3] and Simeon Bird [1]★

[1] University of California, Riverside, CA 92521, USA
[2] California Polytechnic State University, San Luis Obispo, CA 93407, USA
[3] University of Hawai'i at Hilo, Hilo, HI 96720, USA

Accepted 2023 September 18. Received 2023 September 17; in original form 2023 April 28

ABSTRACT

We assemble the largest C IV absorption line catalogue to date, leveraging machine learning, specifically Gaussian processes, to remove the need for visual inspection for detecting C IV absorbers. The catalogue contains probabilities classifying the reliability of the absorption system within a quasar spectrum. Our training set was a sub-sample of DR7 spectra that had no detectable C IV absorption in a large visually inspected catalogue. We used Bayesian model selection to decide between our continuum model and our absorption-line models. Using a random hold-out sample of 1301 spectra from all of the 26 030 investigated spectra in DR7 C IV catalogue, we validated our pipeline and obtained an 87 per cent classification performance score. We found good purity and completeness values, both ~ 80 per cent, when a probability of ~ 95 per cent is used as the threshold. Our pipeline obtained similar C IV redshifts and rest equivalent widths to our training set. Applying our algorithm to 185 425 selected quasar spectra from SDSS DR12, we produce a catalogue of 113 775 C IV doublets with at least 95 per cent confidence. Our catalogue provides maximum a posteriori values and credible intervals for C IV redshift, column density, and Doppler velocity dispersion. We detect C IV absorption systems with a redshift range of 1.37–5.1, including 33 systems with a redshift larger than 5 and 549 absorbers systems with a rest equivalent width greater than 2 Å at more than 95 per cent confidence. Our catalogue can be used to investigate the physical properties of the circumgalactic and intergalactic media.

Key words: quasars: absorption lines – methods: statistical.

1 INTRODUCTION

Metals, elements heavier than helium, are formed in the hearts of massive stars and recycled into the interstellar medium (ISM) by supernovae and stellar winds. Ultimately some of these metals are transported into the circumgalactic medium (CGM) or even intergalactic medium (IGM). Studying metals in the CGM sheds light on the complex, interconnected processes of accretion, feedback, and continual recycling (Oppenheimer et al. 2019). Measurements of the abundance of metals in the universe over time allow us to study the cycling of baryons through galaxies and, thus, the formation and evolution of galaxies (Tumlinson, Peeples & Werk 2017; Péroux & Howk 2020).

Quasar absorption lines enable us to measure the abundance of elements and their ionization states within the intergalactic gas. Particularly useful is the C IV λλ1548, 1550 doublet. This doublet is caused by a strong transition of an abundant metal that redshifts into optical bands at $z \sim 1.5 - 5.2$, with lower redshifts observable in the UV (e.g. Cooksey et al. 2010; Shull, Danforth & Tilton 2014; Hasan et al. 2022) and higher in the IR (e.g. Simcoe et al. 2011; Ryan-Weber et al. 2009; Becker, Rauch & Sargent 2009; Davies et al. 2023). The rest wavelengths of the C IV doublet make it detectable outside the H 1 Ly α forest. Moreover, C IV has an unsaturated doublet ratio of 2: 1 for $W_{r,\,1548}$: $W_{r,\,1550}$, easing automated line detection methods (Churchill 2020).

C IV is a resonance line doublet that is useful for studying many physical properties of the IGM and CGM over cosmic time. It has been extensively studied; here we will provide an abbreviated overview, and the interested reader is referred to (and references therin Péroux & Howk 2020) for a more comprehensive review.

Studying statistical properties of C IV absorption systems, such as their rest equivalent width distribution, sheds light on all of the processes that contribute to the formation and propagation of this metal throughout the IGM and CGM (Songaila 2005; D'Odorico et al. 2010; Simcoe 2011; Hasan et al. 2020, 2022). The metallicity and enrichment history of the CGM was studied using the C IV/H I line ratio (Barlow & Tytler 1998; Ellison et al. 2000). The ratio of C IV to other metal lines can constrain the ionization state of the IGM (Boksenberg & Sargent 2015). Also the ratios of different carbon ions (C II C IV) can be used to infer the ionization state of the absorbing gas in the IGM at a redshift where neutral hydrogen absorption is saturated (Cooper et al. 2019). One can measure or constrain the temperature and kinematics of C IV absorbers to analyse the physics of the IGM (Rauch et al. 1996; Appleby et al. 2023). The study of the characteristics of metal lines, such as C IV, offers valuable information for developing models of contamination in baryon acoustic oscillation measurements of the Ly α forest (Yang et al. 2022). The autocorrelation (clustering) of C IV absorbers systems will constrain the IGM metallicity and enrichment topology

★ E-mail: reza.monadi@email.ucr.edu (RM); sbird@ucr.edu (SB)

Published by Oxford University Press on behalf of Royal Astronomical Society

(Sargent, Boksenberg & Steidel 1988; Petitjean & Bergeron 1994; Chen, Lanzetta & Webb 2001; Scannapieco et al. 2006; Tie et al. 2022). Close quasar–galaxy pairs connect C IV absorbers to galactic halos and provide a tool for studying galaxy evolution (Adelberger et al. 2005; Bordoloi et al. 2014; Rubin et al. 2015; Burchett et al. 2015, 2016).

C IV absorbers have been observed at $z > 5$, probing the tail end of the reionization epoch (Becker et al. 2009; Ryan-Weber et al. 2009; Simcoe et al. 2011; D'Odorico et al. 2013; Codoreanu et al. 2018; Doughty & Finlator 2023). Combining C IV ionization data with other ions like Si IV and C II at high redshift, one can study the reionization-epoch ultra violet background's slope (Finlator et al. 2016). Comparing C IV observational data with simulations shows some discrepancies in the production of carbon, motivating more detailed observations and improved theoretical models (Finlator et al. 2020).

Most relevant to our current work, Cooksey et al. (2013) detected strong C IV absorbers in the low-signal-to-noise spectra of the Sloan Digital Sky Survey (SDSS) (Abazajian et al. 2009; Eisenstein et al. 2011). [1] On the theory side, C IV has been associated with enriched gas surrounding galactic halos in cosmological simulations (Haehnelt, Steinmetz & Rauch 1996; Bird et al. 2016).

The above surveys and catalogues of C IV were assembled by visual inspection of quasar spectra by trained astronomers, sometimes supplemented by template fitting to discover candidate absorbers. However, this visual inspection is prohibitively time consuming with the large size of modern quasar surveys. The largest C IV catalogues are from SDSS: Cooksey et al. (2013) used Data Release (DR) 7 and Chen et al. (2014) used DR9. The visually inspected quasar catalogue of SDSS DR12 contains 185,541 quasars (Ross et al. 2012), which can potentially have C IV absorbers. The upcoming Dark Energy Spectroscopic Instrument (DESI Collaboration et al. 2016) will obtain spectra for more than 30 million galaxies and quasars. DESI will observe more than 10 times the number of galaxies observed by SDSS and $\sim 10^7$ quasars. Leveraging the increase in quasar spectra for C IV studies is best served by an automated detection algorithm. Neural networks are machine learning methods used for detecting different absorption lines in quasar spectra, for example C II (Xia et al. 2022), Ly α forest (Cheng, Cooke & Rudie 2022), C IV broad absorption lines in SDSS (Guo & Martini 2019), damped Ly α absorbers (DLAs, Parks et al. 2018; Wang et al. 2022), and Mg II (Zhao et al. 2019). Our catalogue is the first machine learning based C IV quasar absorption line catalogue. Our use of Gaussian processes instead of neural networks allow for reduced training times and easier estimation of uncertainty.

SDSS DR12 contains the largest extant quasar spectral catalogue with visually verified redshifts. However, it has a relatively low-spectroscopic resolution and a low median signal-to-noise ratio (SNR). This makes the detection of an absorption line, like the C IV doublet, quite challenging. However, our Bayesian approach based on Gaussian processes is capable of extracting reliable information even from noisy data.

Our automated C IV detection pipeline is based on the technique for detecting DLAs from Garnett et al. (2017), which was extended to multiple absorbers by Ho, Bird & Garnett (2020). A Gaussian process model with a bespoke learned kernel is built for the quasar spectrum in the absence of absorption, and Bayesian model selection is used to determine whether an absorber is preferred over the no-absorption (i.e. continuum) model given the quasar instrumental noise. The pipeline is built using a Bayesian framework, allowing us to make probabilistic statements even about the noisiest observed data. Detection probabilities can be used to further refine the catalogue to increase purity or completeness. Furthermore, as a fully Bayesian pipeline, it provides a posterior distribution for the column density, redshift, and Doppler velocity dispersion for each absorber.

The rest of this paper is structured as follows. In Section 2, we summarize the data we used for different stages in our pipeline. In Section 3, we detail the mathematical framework for obtaining our absorption models, our Gaussian process model for quasar emission, and our Bayesian approach to search for absorbers in the quasar spectra. We validate our approach by testing our algorithm in a hold-out sub-sample of our training set in Section 4. The resulting C IV catalogue is presented and discussed in Section 5. We summarize and discuss potential future applications of our catalogue in Section 6.

2 DATA

Our primary data set was SDSS quasar spectra; we followed Cooksey et al. (2013) in designating quasars with their spectroscopic modified Julian date, fibre identification number, and plate number. We trained our absorption-free model on a subset of SDSS DR7 (Abazajian et al. 2009) filtered to avoid C IV absorbers as detected by the so called 'Precious Metals' (PMs) catalogue (Cooksey et al. 2013).[2] The PM catalogue did not search for absorbers in spectra that did not meet certain criteria (see table 1 in Cooksey et al. (2013). The initial DR7 quasar catalogue contains 105 783 quasars, of which 26 030 were searched for C IV absorption. Our training set further excluded the 10 861 spectra which contain one or more C IV absorbers in the PM catalogue. Our null model was thus trained on 15 169 'C IV-free' spectra, meaning spectra that either were not found as a C IV candidate (as defined by Cooksey et al. (2013)) or did not pass the visual verification check.

Before training a continuum model on all of the 15 169 spectra in Cooksey et al. (2013), we train a number of candidate continuum models on 95 per cent of our training set and then validate these candidate continuum models on a random hold-out sample of 5 per cent of all searched spectra in the DR7 catalogue, which contains 1301 quasars. This is our validation set that we used as a tool to find the optimum values of the parameters needed to train a candidate continuum model. These tuning parameters include: flux normalization wavelength range, the minimum number of non-NaN pixels in a training spectrum, the dimension of the covariance matrix (see equation 10), etc. After applying our pipeline on the validation set, we assessed the performance of the classification (i.e. classifying a given spectrum as having C IV absorber(s) or otherwise) using the PM catalogue as a 'ground truth'. We found the best candidate continuum model by maximizing the classification score (see Section 4.2) and purity/completeness (see Section 4.3). At this point, we took the parameters of the best candidate continuum model and built our final model from all of the 15 169 'C IV-free' DR7 spectra investigated in the PM catalogue pipeline.

We applied our algorithm on a subset of the SDSS DR12 quasar catalogue (Alam et al. 2015) to build our new C IV catalogue. We chose our working quasar sample starting from the SDSS-DR12

[1] Chen et al. (2014) assembled a C IV catalogue from SDSS DR9 quasar spectra. However, we did not use their candidate absorbers as a detailed comparison to previous C IV catalogues was missing.

[2] We obtained the list of spectra from igmabsorbers.info and downloaded the spectra from http://das.sdss.org/spectro/1d_26

Table 1. For each sightline, identified by Column 1 and 2, we report the absorber's redshift (Column 3), column density in log (cm^{-2}) (Column 4), Doppler velocity dispersion in km s^{-1} (Column 5), rest equivalent width for 1548 Å W$_{r,\,1548}$ (Column 6), rest equivalent width for 1550 Å W$_{r,\,1550}$ (Column 7), the posterior probability of the C IV absorber P(M$_D$) (Column 8), and the posterior probability of the singlet absorber P(M$_S$) (Column 9). We show only absorbers with P(M$_D$) ≠NaN. This table demonstrates a portion of the full table for the first 10 rows. Note that those measurements with large errors are uncertain (i.e. low absorption model posterior probability). The full table with 445 765 rows is available at https://doi.org/10.5281/zenodo.7872725.

(1) QSO-ID	(2) $z_{\rm QSO}$	(3) $z_{\rm C\,IV}$	(4) log($N_{\rm C\,IV}$) log (cm^{-2})	(5) $\sigma_{\rm C\,IV}$ km s^{-1}	(6) W$_{r,\,1548}$ (Å)	(7) W$_{r,\,1550}$ (Å)	(8) P(M$_D$)	(9) P(M$_S$)
56238-6173-528	2.3091	1.91039 ± 0.00037	15.66 ± 0.52	52.24 ± 0.06	1.306 ± 0.960	1.178 ± 1.142	0.63	0.18
		2.21620 ± 0.00078	14.75 ± 0.24	104.74 ± 0.13	1.328 ± 1.129	0.838 ± 0.790	0.40	0.00
56268-6177-595	2.4979	2.11727 ± 0.00036	13.96 ± 0.22	58.93 ± 0.06	0.280 ± 0.469	0.152 ± 0.294	0.23	0.61
		1.99557 ± 0.00027	13.99 ± 0.26	39.18 ± 0.04	0.261 ± 0.455	0.148 ± 0.324	0.41	0.00
55810-4354-646	2.3280	1.90383 ± 0.00027	13.89 ± 0.11	48.29 ± 0.05	0.230 ± 0.144	0.125 ± 0.083	0.30	0.36
		1.94502 ± 0.00060	13.77 ± 0.27	55.32 ± 0.10	0.187 ± 0.330	0.099 ± 0.186	0.31	0.00
56565-6498-177	2.3770	1.95293 ± 0.00049	14.52 ± 0.34	49.38 ± 0.08	0.636 ± 0.803	0.418 ± 0.678	0.29	0.32
56268-6177-608	3.7120	3.33339 ± 0.00033	14.49 ± 0.05	75.29 ± 0.04	0.763 ± 0.174	0.460 ± 0.125	1.00	0.00
		3.51346 ± 0.00065	14.21 ± 0.17	96.87 ± 0.08	0.530 ± 0.564	0.289 ± 0.345	0.76	0.00
		3.22375 ± 0.00065	14.06 ± 0.23	62.16 ± 0.08	0.344 ± 0.363	0.189 ± 0.208	0.28	0.00

quasar catalogue.[3] We kept only quasars with rest-frame wavelength coverage between 1310 and 1548 Å, the region of potential C IV absorption (avoiding both the Ly α forest and the potential for false positives of C IV from O I λ1302 or Si II λ1304). This means quasars with redshifts satisfying 1310 Å(1 + $z_{\rm QSO}$)> 3650 Å (or $z_{\rm QSO} > 1.7$) and 1548 Å(1 + $z_{\rm QSO}$) < 10400 Å (or $z_{\rm QSO} < 5.7$). We removed detected broad absorption line quasars (BAL) using the SDSS BAL catalogue.[4] After these selections, we downloaded the list of quasar spectra from the SDSS-III Baryon Oscillation Spectroscopic Survey Science Archive Server.[5]

We converted all observed spectra to the emission rest-frame using the visually inspected quasar redshift estimate from the SDSS pipeline, which we assume to be exact.[6] Missing or otherwise masked flux values (e.g. from a bad pixel) were denoted by NaN and were not used in our pipeline.

3 METHOD

We modified the pipeline introduced in Garnett et al. (2017) and Ho et al. (2020) to look for C IV absorbers in SDSS DR12. We learnt an a priori distribution for the shape of the quasar emission spectra without C IV using SDSS DR7 spectra classified by the PM catalogue from Cooksey et al. (2013). The null model, M$_N$, was learned from SDSS DR7 spectra identified as 'non-detection' (i.e. no C IV candidate in the PM study). Each iteration, we did a Bayesian model selection between the null model, a model for a C IV doublet model (M$_D$), and a model for an 'interloper' singlet absorption line (M$_S$) to compute the posterior probability of C IV absorption. We searched for up to seven C IV absorbers in each spectrum, reporting probabilities for each. We stopped at seven absorbers as only 4 spectra among 26 030 investigated spectra in the DR7 C IV catalogue contained seven absorbers and none contained more.

There were six main changes since Ho et al. (2020). First, the absorption profile was updated to model a C IV doublet, instead of a DLA. Secondly, a model for singlet line absorbers was introduced, which serves a similar role to the sub-DLA model in Ho et al. (2020). Without this singlet absorption line model, the pipeline produces excessive false positives, as it has no other way to match absorption except a C IV doublet. Thirdly, in addition to sampling absorber redshift and column density, we sampled the Doppler velocity dispersion, which allows more accurate fits. Fourth, we no longer model the Ly α forest in the null model, as it does not overlap our C IV absorption region. Fifth, instead of a fixed instrumental broadening profile, we used the reported Gaussian wavelength-dependent dispersion from SDSS.[7] Sixth, we report an individualized probability for each absorber we detect, rather than the joint probability of observing at least a certain number of absorbers, as in Ho et al. (2020). Fig. 1 shows an overview of our pipeline, as described in this Section.

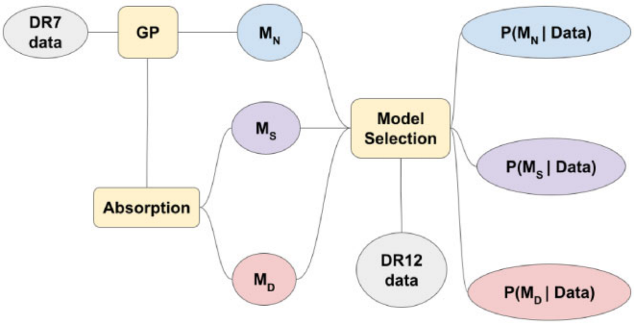


Figure 1. This is a flow chart for our pipeline. Training spectra from SDSS DR7 are used to train a Gaussian process kernel with which to model the quasar continuum (i.e. null model, M$_N$). Analytic Voigt profiles are used to construct models for absorption from a C IV doublet (M$_D$) or a generic singlet absorber (M$_S$). Conditioning on DR12 spectra produces a posterior probability estimate for each model that can be used to decide if there is a C IV absorber in the given spectrum or not. Moreover, for the absorber models, M$_D$ and M$_S$, we have a posterior distribution for each model parameter: absorber redshift, Doppler velocity dispersion for the absorption profile, and the absorber column density.

Section 2 described our initial training data, a subset of SDSS DR7. Section 3.1 explains the Voigt-profile model for any absorber detection. Section 3.2 summarizes our null (also known as absorption-free or continuum) model, M$_N$, for the quasar emission function, which

[3] http://data.sdss3.org/sas/dr12/boss/qso/DR12Q/DR12Q.fits

[4] http://data.sdss3.org/sas/dr12/boss/qso/DR12Q/DR12Q_BAL.fits

[5] https://data.sdss.org/sas/dr12/boss/spectro/redux/

[6] We used Z_VI, column 8 of SDSS DR12 quasar catalogue.

[7] Column 6 of the fits files of SDSS spectra, see the SDSS data model.

uses a bespoke Gaussian Process kernel. Section 3.3 describe two analytic absorption models, M_S and M_D, which are generated by convolving M_N with a singlet or doublet Voigt profile, respectively. In addition, we need model priors, $Pr(\mathrm{M})$, for each model (see Section 3.4). The model likelihood is discussed in Section 3.5. Section 3.6 explains our technique for deciding how many C IV absorbers to search for.

3.1 Absorption function

Voigt profiles are useful for modelling the absorption effect in the observed spectrum of an emitting source such as quasars (Churchill 2020). A Voigt profile is given by

$$\begin{aligned}\phi(\mathrm{v}, \sigma_{\mathrm{C\,IV}}, \gamma_{\ell \mathrm{u}}) \\ = \int \frac{d\mathrm{v}}{\sqrt{2\pi}\sigma_{\mathrm{C\,IV}}} \exp\left(-\mathrm{v}^2/2{\sigma_{\mathrm{C\,IV}}}^2\right) \\ \times \frac{4\gamma_{\ell u}}{16\pi^2[\nu - (1-\mathrm{v}/c)\nu_{\ell u}]^2 + \gamma_{\ell u}^2},\end{aligned} \tag{1}$$

which is a convolution between Lorentzian and Gaussian profiles. The former computes the natural broadening and the latter thermal broadening (Draine 2011). The velocity, v, in equation (1) is given by

$$\mathrm{v}(\lambda) = \mathrm{c}\left(\frac{\lambda}{\lambda_{\ell \mathrm{u}}(1 + \mathrm{z}_{\mathrm{C\,IV}})} - 1\right). \tag{2}$$

A negative (positive) velocity refers to a position in λ-space that is red-ward (blue-ward) of the observed C IV absorption in $\lambda_{\ell u}(1 + z_{\mathrm{C\,IV}})$. The Lorentzian broadening contribution is

$$\gamma_{\ell u} = \frac{\Gamma \lambda_{\ell u}}{4\pi}, \tag{3}$$

where Γ is the damping constant. The Doppler velocity dispersion for a C IV absorber, $\sigma_{\mathrm{C\,IV}}$, is

$$\sigma_{\mathrm{C\,IV}} = \sqrt{\frac{kT}{6m_p + 6m_n}}, \tag{4}$$

where k, T, m_p, and m_n are the Boltzmann constant, gas temperature, proton mass, and neutron mass, respectively. The Doppler velocity dispersion controls the width of the absorption profile as a function of temperature. For the C IV doublet at $\lambda = 1548$ Å, $\Gamma = 2.643 \times 10^8\ \mathrm{s}^{-1}$ and for $\lambda = 1550$ Å, $\Gamma = 2.628 \times 10^8\ \mathrm{s}^{-1}$. Lorentzian broadening is thus small ($\gamma_{\ell u}/\sigma_{\mathrm{C\,IV}} \sim 0.01$ for $T \sim 10^4$ K) and the Voigt profile is close to Gaussian.

The optical depth, τ, itself is a function of observed frequency ($\nu = c/\lambda$) given: absorber column density $N_{\mathrm{C\,IV}}$, which controls the depth of the profile, absorber redshift $z_{\mathrm{C\,IV}}$, which sets the wavelength where we observe the absorption, and Doppler velocity dispersion $\sigma_{\mathrm{C\,IV}}$. The optical depth is given by

$$\tau_{\ell u}(\lambda; z_{\mathrm{C\,IV}}, N_{\mathrm{C\,IV}}, \sigma_{\mathrm{C\,IV}}) = \frac{N_{\mathrm{C\,IV}} \pi e^2 f_{\ell u} \lambda_{\ell u}}{m_e c} \phi(\mathrm{v}(\lambda), \sigma_{\mathrm{C\,IV}}, \gamma), \tag{5}$$

where c is the speed of light, e is the elementary charge, m_e is the mass of the electron, and $\lambda_{\ell u}$ is the transition wavelength for the lower state (ℓ) and the upper state (u) and $f_{\ell u}$ is the oscillator strength of the transition. Using spectroscopic notation (Tennyson 2019), the 1548 Å absorption line is a transition from $2^2S_{\frac{1}{2}}$ to $2^2P^o_{\frac{1}{2}}$ and the 1550 Å absorption line is a transition from $2^2S_{\frac{1}{2}}$ to $2^2P^o_{\frac{3}{2}}$. The absorption profile is related to the optical depth via

$$a_{\ell u}(\lambda; z_{\mathrm{C\,IV}}, N_{\mathrm{C\,IV}}, \sigma_{\mathrm{C\,IV}}) = \exp\left(-\tau_{\ell u}(\lambda; z_{\mathrm{C\,IV}}, N_{\mathrm{C\,IV}}, \sigma_{\mathrm{C\,IV}})\right), \tag{6}$$

where the ℓu subscript can refer to either 1548 or 1550 Å transitions. The doublet model M_D will be built by convolving the null model with an absorption profile that considers both 1548 or 1550 Å. The singlet model M_S, on the other hand, only considers the 1548 Å transition.

SDSS resolution is insufficient for detailed modelling of C IV absorption systems as is done with high-resolution spectra (e.g. Hasan et al. 2020). Indeed, strong C IV absorption at SDSS resolution can be reasonably modelled by a single Voigt profile with appropriate choice of $z_{\mathrm{C\,IV}}$, $N_{\mathrm{C\,IV}}$, and $\sigma_{\mathrm{C\,IV}}$, as we do in this work (see Section 3.5). We acknowledge that the same absorption at higher resolution would reveal finer structure and require multiple Voigt profiles, with different combinations of $z_{\mathrm{C\,IV}}$, $N_{\mathrm{C\,IV}}$, and $\sigma_{\mathrm{C\,IV}}$, that would be strong constraints on the physical conditions of the gas giving rise to the absorption. The $N_{\mathrm{C\,IV}}$ and $\sigma_{\mathrm{C\,IV}}$ values returned by our algorithm may not be as tightly constrained as the $z_{\mathrm{C\,IV}}$ measurements. Remember that $N_{\mathrm{C\,IV}}$ and $\sigma_{\mathrm{C\,IV}}$ control the Voigt profile shape in our absorption model that is compared to the observed flux deficit in the SDSS spectra (see Table 1).

3.2 Quasar emission function

The physics of quasar emission is not fully understood, and there is considerable variety in observed quasar spectra. Thus we used an empirical model for the quasar emission function (also known as continuum) based on the observed spectra. We modelled the emission function of a quasar, f, in the absence of any absorption (including C IV absorption) using Gaussian processes that generate a distribution over functions. Gaussian Processes result from a generalization of a multivariate Gaussian distribution to infinite domains (Rasmussen & Williams 2006). As the standard library of kernels is insufficiently flexible to model the complicated correlations between different emission lines in a quasar spectrum, we used a customized kernel learned directly from the training set.[8] We described the training set in Section 2.

Here, we briefly summarize the technique. Our model was similar to Garnett et al. (2017) where the process of obtaining a Gaussian process model for quasar emission spectra is described in more detail. However, unlike Garnett et al. (2017), we did not model the Ly α forest as we were looking for C IV absorbers outside of the forest. We trained a C IV-free model between 1310 and 1555 Å, which produced the best results during the validation phase. This range is close to the rest-frame C IV absorption wavelength searched in the PM catalogue. Fig. 2 shows an example learned quasar continuum together with the observed flux and noise.

Even from low-SNR spectra, our method extracts some statistical information, so we do not enforce a minimum SNR in our search. Our pipeline naturally gives low likelihoods to low-SNR spectra during the training. We can completely specify a Gaussian process by its mean and correlation functions (analogous to the first two moments of a Gaussian distribution). We specify the mean function μ by

$$\mu(\lambda) = \langle y(\lambda) \rangle, \tag{7}$$

where λ is the rest-frame wavelength and $y(\lambda)$ is the observed rest-frame flux for the training-set spectra, after applying a mask for missing pixels; angle brackets ($\langle\rangle$) denote an average over wavelengths. Before computing this average over the training set, we have normalized the quasar flux and the flux variance so that

[8] Our training set consisted of all of the spectra investigated in the PM C IV catalogue and classified as not containing C IV absorbers.

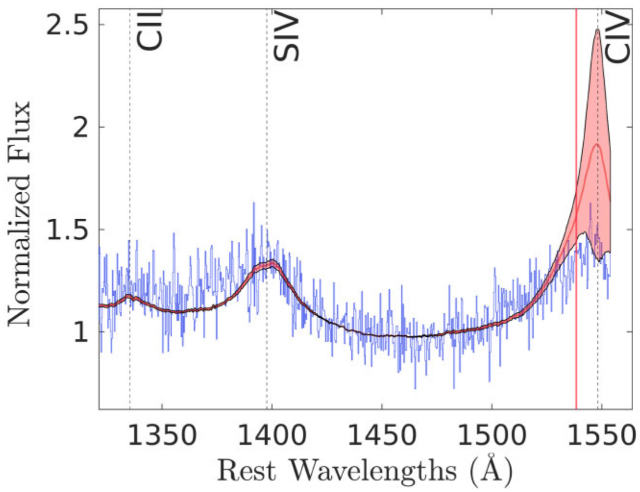


Figure 2. An example learned quasar emission function (red curve) with the normalized observed smoothed flux (blue curve). The shaded red region shows 1σ uncertainties. The SDSS DR7 quasar has QSO-ID: 51630-0266-280 and redshift 2.57. Note that we search for absorbers starting 3000 km s^{-1} red-ward of the quasar's redshift (shown by the solid red vertical line), so the moderate failure to match the quasar C IV emission line in this case does not lead to an artificial preference for C IV absorption. Prominent emission lines are marked by dashed vertical lines.

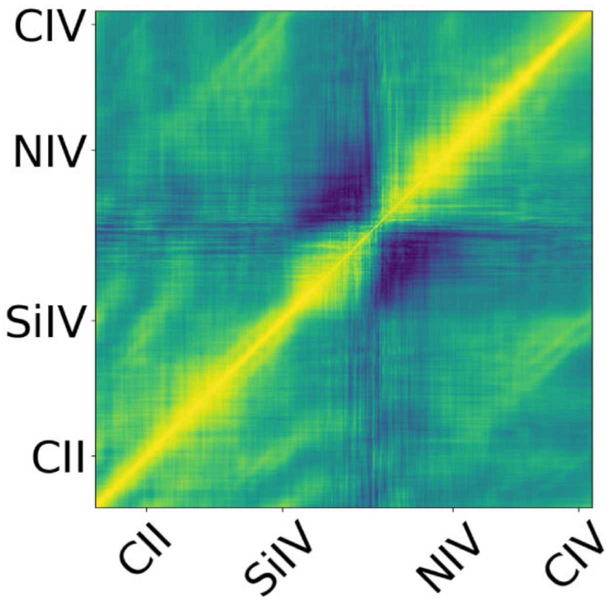


Figure 3. Learned covariance matrix **K** (see equation 8 and equation 10) for our null (continuum) model. This matrix is built up by considering the observed flux and noise from our C IV-free training set (see Section 2). Brighter pixels show stronger correlations and darker regions weaker ones. The wavelengths of prominent emission lines are labelled. The bright diagonal implies stronger correlations between pixels at smaller wavelength separation.

they have a median value of unity in the normalization range. This normalization was needed so that the Gaussian process model is insensitive to variations in (observed) quasar brightness. We chose the range from 1420 to 1475 Å as it contains no prominent emission lines (Zhu & Ménard 2013; Hamann et al. 2017; Monadi & Bird 2022). We also confirmed empirically that this normalization range produces the best score when applied to our validation set. We remind the reader that the validation set is a random subset (1301 spectra) of all candidate DR7 spectra in the PM catalogue (see Section 2).

The Gaussian process covariance function describes the correlation between flux values at two separate wavelengths, λ and λ'. Most applications of Gaussian processes assume a simple kernel for the covariance, such as the exponential squared kernel (Rasmussen & Williams 2006). However, the complex correlation between features in quasar continua is hard to describe using the simple/standard covariance functions like the radial basis function. Instead, our algorithm directly learned a covariance function:

$$K(\lambda, \lambda') = \mathrm{cov}[f(\lambda), f(\lambda')], \tag{8}$$

by considering all of the cumulative information contained in the spectra of our training set: all of the flux measurements and noise measurements given at the observed wavelengths.[9] We need to maximize the joint likelihood of generating the whole training set given that the underlining model is the null model (i.e. absorption-free). We assume our observations (i.e. flux and noise given at each observed wavelength in the training set) are independent and drawn from a Gaussian distribution with width corresponding to the observed noise of the SDSS pipeline. Next we maximize the likelihood (see section 5.3 of (Garnett et al. 2017) for details) and learn the quasar mean function (equation 7) and quasar covariance function (equation 8). Optimizing this joint likelihood function was done using `minFunc`: a Matlab function for unconstrained optimization of differentiable real-valued multivariate functions using line-search methods.[10]

We binned quasar spectra linearly in wavelength, from 1310 to 1555 Å, with a bin size of $\Delta\lambda$. This gave us the number of bins as

$$N_{\rm bin} = \frac{1555 - 1310}{\Delta\lambda}. \tag{9}$$

If we input the binned wavelength grid, $\boldsymbol{\lambda}$, to equation (7) we get the learned mean vector $\boldsymbol{\mu}$, with $N_{\rm bin}$ elements. The covariance matrix, $\mathbf{K}$, an $N_{\rm bin} \times N_{\rm bin}$ matrix, is calculated on two discretized wavelength grids, $\boldsymbol{\lambda}$ and $\boldsymbol{\lambda}'$, using equation (8). A very fine $\Delta\lambda$ is not desirable because it increases the size of $\boldsymbol{\mu}$ and $\mathbf{K}$ and thus is more computationally expensive. Alternatively, a coarse $\Delta\lambda$ cannot capture enough information from the quasar spectra. The optimum $\Delta\lambda$ in Garnett et al. (2017) and Ho et al. (2020) was 0.25 Å. We empirically found that $\Delta\lambda = 0.5$ Å is the optimum value for the redder spectral region we examine here which gives us $N_{\rm bin} = 490$. Without further structural assumptions on $\mathbf{K}$, our algorithm would have to learn a matrix of $N_{\rm bin}^2 \sim 2.4 \times 10^5$ elements. To circumvent this, we used a low-rank decomposition

$$\mathbf{K} = \mathbf{M}\mathbf{M}^\top, \tag{10}$$

where $\mathbf{M}$ is a $N_{\rm bin} \times k$ matrix, for any positive integer k. Larger-k models allow for higher fidelity modelling of $\mathbf{K}$. Following Garnett et al. (2017), we set $k = 20$. We also checked $k = 19$, 21, and 22, finding that our results were insensitive to this choice. Fig. 3 shows the learned covariance matrix. This covariance matrix describes how likely the quasar emission spectrum is to vary around the mean spectrum. It encodes the information contained in the spectra of our training set, the 'C IV-free' spectra from the PM catalogue.

[9] The third column of the SDSS fits tables for observed spectra contains inverse noise variance $(\sigma(\lambda))^{-2}$.

[10] https://www.cs.ubc.ca/~schmidtm/Software/minFunc.html

Having learned the mean quasar vector $\boldsymbol{\mu} = \mu(\boldsymbol{\lambda})$ (see equation 7) and the lower rank decomposition matrix $\mathbf{M}$ in equation (10) which gives us the covariance matrix $\mathbf{K}$ (equation 8), we can write the Gaussian processes model for the quasar emission function, trained on the observed spectra, as a multivariate Gaussian distribution

$$p(f(\boldsymbol{\lambda})) = GP(\mu(\boldsymbol{\lambda}), K(\boldsymbol{\lambda}, \boldsymbol{\lambda}')) = N(f(\boldsymbol{\lambda}); \mu(\boldsymbol{\lambda}), K(\boldsymbol{\lambda}, \boldsymbol{\lambda}')), \quad (11)$$

where GP denotes a Gaussian process. We remind the reader that a Gaussian process is a Gaussian distribution over functions. Therefore, we can write the Gaussian process for the quasar emission function f given our learned mean vector $\boldsymbol{\mu}$ and covariance matrix $\mathbf{K}$ as

$$N(f; \boldsymbol{\mu}, \mathbf{K}) = \frac{1}{\sqrt{(2\pi)^d \det(\mathbf{K})}} \exp\left(-\frac{1}{2}(f - \boldsymbol{\mu})^\top \mathbf{K}^{-1}(f - \boldsymbol{\mu})\right), \quad (12)$$

where d is the dimension of the quasar emission function f.

3.3 Absorption line models

We want to find the probability of a C IV doublet in the observed spectrum of a quasar given the observed rest-frame flux $\boldsymbol{y}(\boldsymbol{\lambda})$, under our null (also known as absorption-free or continuum) GP model M_N. Our data were composed of the observed wavelengths $\boldsymbol{\lambda}$, their corresponding observed quasar flux $\boldsymbol{y}(\boldsymbol{\lambda})$, and their corresponding observed noise $\boldsymbol{\sigma}(\boldsymbol{\lambda})$. We define the data as

$$D = \{\boldsymbol{\lambda}; \boldsymbol{y}(\boldsymbol{\lambda}), \boldsymbol{\sigma}(\boldsymbol{\lambda})\}. \quad (13)$$

Bayes' rule gives the model posterior, the probability of each model given the data

$$P(\mathrm{M}_i|D) = \frac{P(D|\mathrm{M}_i)Pr(\mathrm{M}_i)}{\sum_j P(D|\mathrm{M}_j)Pr(\mathrm{M}_j),}. \quad (14)$$

We defined three models

(i) M_N models the quasar continuum without absorption (equation 11).

(ii) M_D is a model containing exactly one C IV doublet. M_D is built by convolving M_N with the absorption function (equation 6) for all observed wavelengths.

$$\mathrm{M}_D \rightarrow \mathrm{convolve}(a_{1548,1550}(\boldsymbol{\lambda}), \mathrm{M}_N) \quad (15)$$

(iii) M_S is a singlet model containing exactly one generic singlet absorption line. For simplicity, we implemented M_S using the same Voigt profile as M_D but including only the 1548 Å absorption line.

$$\mathrm{M}_S \rightarrow \mathrm{convolve}(a_{1548}(\boldsymbol{\lambda}), \mathrm{M}_N) \quad (16)$$

We added this singlet model, in addition to the C IV-free and C IV-doublet models, so that our Bayesian framework is not forced to give a high probability of a C IV doublet if there is a strong singlet line in the spectrum and nearby noise happens to be similar to a C IV doublet. For example, a broad singlet line like Si II 1526, Fe II λ1608, or Al II λ1670, can be misidentified as a C IV doublet, if we have only two models (i.e. M_N and M_D). The singlet model, M_S, provides an alternative to both M_N and M_D for such lines.

Fig. 4 shows an example, the application of our pipeline to QSO-ID: 51608-0267-264 with $z_{\mathrm{QSO}} = 1.89$. Here a noise fluctuation and a strong line happen to have a velocity separation similar to a C IV doublet. For this spectrum, we have

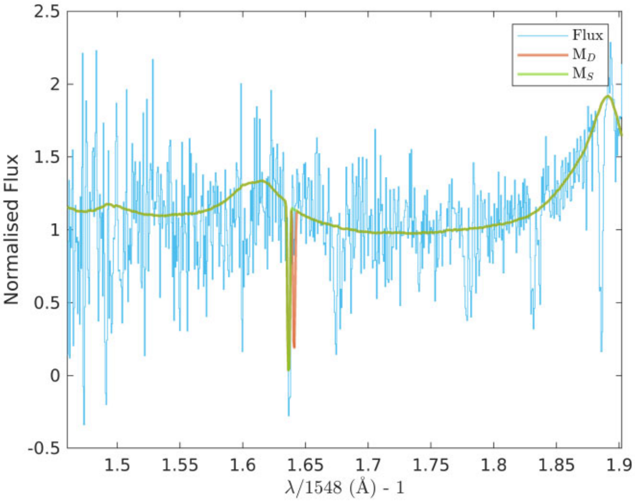


Figure 4. The figure shows the spectrum of QSO-ID: 51608-0267-264 with $z_{\mathrm{QSO}} = 1.89$ (blue) where the singlet model (green) is preferred over the C IV doublet model (red), which is in turn preferred over the null model. If we did not have M_S, our pipeline would have incorrectly detected a C IV absorber at $z_{\mathrm{C\,IV}} = 1.635$.

$$\log(P(\mathrm{M}_N|D)) = -297.8216,$$
$$\log(P(\mathrm{M}_S|D)) = -250.1609, \text{ and}$$
$$\log(P(\mathrm{M}_D|D)) = -257.1906.$$

Although the doublet model is not a very good fit, the null model is even worse. Thus without the singlet model, M_S, our pipeline would incorrectly prefer the doublet model and detect a C IV absorber. Note that these log-likelihoods are not normalized. We fit to the whole quasar spectrum so the number of data points (degrees of freedom) is large compared to the local change in the spectrum around one absorber system. In addition, this particular spectrum has a second absorber visible at $z = 1.82$, which reduces the likelihood during the first absorber search. What is important here is the difference between these log-likelihoods. In the process of Bayesian model selection we need the Bayes's factor[11] which is proportional to the difference between these log-likelihoods (See equation 14).

A sampling problem arises due to the low resolution of the SDSS spectrograph. Real spectrographs measure the total integrated flux across the spectral pixel. A simple estimate for this is to evaluate the Voigt profile at the centre of the pixel. However, at the low resolution of the SDSS spectrograph, this can be a poor estimate, leading to unphysical doublet ratios. For this reason we compute the integrated flux by first evaluating the Voigt profile on a grid of pixels which is finer than the grid in the SDSS spectrum by a factor of n_{ave}. We found by experiment that the model accuracy does not improve for $n_{\mathrm{ave}} > 20$ sub-samples.

3.4 Model priors

To calculate the model posterior (equation 14), we need model priors, $Pr(\mathrm{M})$, for each of the three models. We set priors for the C IV doublet, M_D, using population statistics from our training set, the PM catalogue of Cooksey et al. (2013). We counted the fraction

[11] Bayes factor is the ratio between posterior probability of two models.

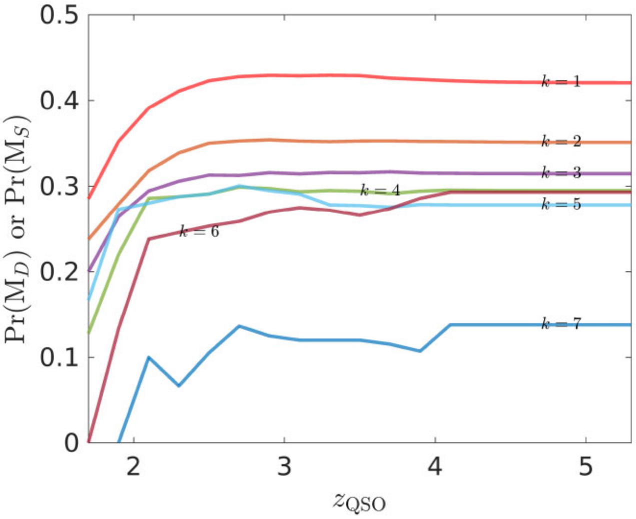


Figure 5. Prior probability for a spectrum containing k C IV absorbers as a function of quasar redshift, for $k = 1 - 7$. We use the average number of absorbers in the PM spectrum in our wavelength search range. C IV is a priori more likely as $z_{\rm QSO}$ increases but reaches a plateau at $z_{\rm QSO} \sim 2.5 - 3$. This is because the C IV wavelength coverage is shorter for low $z_{\rm QSO}$ as the 1548 Å emission line pushes to the blue-end of the SDSS spectral range. Note that we assume the same prior for the singlet model for $k = 1 - 7$.

of spectra with absorbers at $z_{\rm C\,IV} < z_{\rm QSO} - 30000/c$, where c is the speed of light in km s^{-1}. This small decrease in our upper limit for the absorption redshift accounts for any possible error in estimating the redshift from the SDSS pipeline. For simplicity, we used the same prior for the singlet and doublet line models, that is $Pr(\mathrm{M}_S) = Pr(\mathrm{M}_D)$. There are no single-line catalogues for these data, and using equal priors ensures that whichever model is the best fit will be used.

The prior for the C IV-free model can be obtained by

$$Pr(\mathrm{M}_N(k\,\mathrm{C\,IV})) = 1 - Pr(\mathrm{M}_D(k\,\mathrm{C\,IV})), \tag{17}$$

where 'k C IV' denotes some integer number k of C IV systems. We did not include $Pr(\mathrm{M}_S)$ in equation (17) to enable a pointwise model comparison between M_D, M_S, and M_N. Especially when searching for multiple absorbers, our main purpose is deciding the probability of detection or non-detection of C IV absorbers in a spectrum. Furthermore, the small shift in the normalization of model priors is several orders of magnitude smaller than the effect of normalizing the model posteriors in equation (14).

When searching for additional absorbers in spectra where there is already a detection, we use the prior probability of spectra with $(k-1)$ C IV absorbers having k absorbers

$$\begin{aligned} Pr(k\,\mathrm{C\,IV}) &= P(k\,\mathrm{C\,IV}|(k-1)\,\mathrm{C\,IV}) \\ &= \frac{P((k-1)\,\mathrm{C\,IV} \cap k\,\mathrm{C\,IV})}{P((k-1)\,\mathrm{C\,IV})} \\ &= \frac{P(k\,\mathrm{C\,IV})}{P((k-1)\,\mathrm{C\,IV})}. \end{aligned} \tag{18}$$

The equality follows as the intersection between the set with k C IV and the set with $(k-1)$ C IV will be the set of quasars with k C IV absorbers. $Pr(k\,\mathrm{C\,IV})$ is guaranteed to be less than 1, because there are always fewer spectra with more absorption systems.

Fig. 5 shows the M_D priors we used for different searches as a function of $z_{\rm QSO}$. When the redshift increases, all of the priors reach a plateau after $z_{\rm QSO} \sim 3$. There is a decrease from $Pr(1\,\mathrm{C\,IV})$ to the subsequent priors so that $Pr(7\,\mathrm{C\,IV}) < 15$ per cent. C IV absorbers cluster (eg. Boksenberg, Sargent & Rauch 2015), so the prior for detecting k absorbers in a spectrum given a redshift is larger than $Pr(1\,\mathrm{C\,IV})^k$: when there is no clustering and the absorbers are perfectly independent.

3.5 Model likelihood

The model likelihood, $P(D|\mathrm{M})$, in equation (14) is the probability that the observed data, D, have been generated by a considered model, M, after marginalizing the model parameters. We do the marginalization over a prior distribution for each parameter in the model

$$P(D|\mathrm{M}) = \int P(D|\mathrm{M}, \theta) P(\theta|\mathrm{M}) d\theta. \tag{19}$$

Here $P(D|\mathrm{M}, \theta)$ is the likelihood of the spectra being generated by model M, if the model has a certain set of parameters θ. We use the prior probability distribution of $P(\theta|\mathrm{M})$ from equation (19) to integrate out all of the possible θ and obtain a parameter-independent model likelihood. The null model M_N has no free parameters, but M_D and M_S have three free parameters each: (i) absorption redshift ($z_{\rm C\,IV}$), (ii) column density of C IV ($N_{\rm C\,IV}$), and (iii) Doppler velocity dispersion ($\sigma_{\rm C\,IV}$). As mentioned in Section 3.1, it is sufficient for our purposes to model an absorption line at SDSS resolution with a single Voigt profile (defined by $z_{\rm C\,IV}$, $N_{\rm C\,IV}$, and $\sigma_{\rm C\,IV}$) and use these values to measure a rest equivalent width; however, only redshift are well constrained by the data.

Our algorithm fits a single Voigt profile, instead of a combination of Voigt profiles with low-velocity separations, even when we have a blended/complex absorption system. To obtain a reasonable fit with a single Voigt profile, we include large Doppler broadening velocities, which do not necessarily reflect the temperature of the absorbing gas. This keeps our algorithm simpler by sacrificing the reliability of Doppler velocity dispersion measurements. In a follow-up work, we will modify the algorithm so that it can measure the temperature as well.

We need to have priors for each of these parameters to perform the integral in equation (19). A parameter prior is a probability distribution which we know a priori might be true for given possible values of a parameter in a model. In implementations of the Bayesian approach for detecting DLAs in quasar spectra (Ho et al. 2020; Ho, Bird & Garnett 2021), the prior distribution for column density was learned from previous DLA catalogues, and they used a uniform absorber red-shift prior distribution.

One of the input parameters in the Voigt profile[12] is the absorber column density, $N_{\rm C\,IV}$. Following Garnett et al. (2017) and Ho et al. (2020), we need to sample a column density distribution to perform the integral in equation (19) and obtain the model likelihood.

The column density range detected by Cooksey et al. (2013) was $\log N_{\rm C\,IV} \approx 13$ to >15.8. After some experimentation we chose a slightly larger range: $12.5 < \log_{10} N_{\rm C\,IV} < 16.1$, which maximized the performance on our validation set. We probed larger column densities than PM catalogue because: first, their column densities are often lower limits as they used the apparent optical depth method (Savage & Sembach 1991) and a lot of the absorption systems were saturated. Secondly, the larger size of SDSS DR12 gives us a longer survey pathlength which increases our chances of finding the exponentially rare strong systems. We searched for lower column densities than the PM catalogue since our catalogue could

[12] See `voigt_IP.c` in https://github.com/rezamonadi/GaussianProcessCIV

potentially be more sensitive to weaker absorbers (see Fig. 18 for the posterior distribution of column densities). We thus used a mixture probability density function consisting of (i) the $N_{\rm C\,IV}$ probability density function (obtained by kernel density estimation) from the reported values in the PM catalogue and (ii) a uniform probability density function in the same range. We have also confirmed that our column density prior sample reproduces a rest equivalent width ($W_{r,\,1548}$) distribution in reasonable agreement with the PM catalogue for the 1548 Å line.

We also need a prior for the Doppler velocity dispersion, σ_v. The typical temperature for the IGM is $\sim 10^4 - 10^5$ K, which gives a $\sigma_v \sim 2.6 - 8.3\,{\rm km\,s^{-1}}$ for C IV. However, at the low resolution of the SDSS spectra ($\sim 150\,{\rm km\,s^{-1}}$), it is impossible to detect an absorption line with this velocity dispersion. Fortunately, C IV absorbers cluster (Boksenberg et al. 2015) and blend into a broader absorption profile with larger effective $\sigma_{\rm C\,IV}$. By experimenting with different ranges for $\sigma_{\rm C\,IV}$ we chose lower and upper bounds for $\sigma_{\rm C\,IV}$ to be 35 and $115\,{\rm km\,s^{-1}}$, respectively. Note that we used the hold-out sample described in Section 2 and tried different $\sigma_{\rm C\,IV}$ ranges to obtain the highest classification performance and purity/completeness (see Section 4). Moreover, we ensured that this range enables our process to be sensitive to similar rest equivalent widths as the PM catalogue.

We imposed a uniform prior distribution on the absorber redshift, $z_{\rm C\,IV}$. The lower limit is the redshift at which the 1548 Å line is observed at $1310(1 + z_{\rm QSO})$,[13] or the blue end of our input spectrum whichever is larger. Therefore

$$1 + z_{\rm min} = \max\left[\frac{\min(\lambda_{\rm obs})}{1548}, \frac{1310(1 + z_{\rm QSO})}{1548}\right]. \tag{20}$$

We also require a small velocity separation between the absorber and the quasar, to ensure that we are not finding intrinsic C IV absorbers around the host galaxy of the quasar:

$$z_{\rm max} = z_{\rm QSO} - \frac{\delta {\rm v}}{c}(1 + z_{\rm QSO}). \tag{21}$$

We considered $\delta{\rm v} = 1000 - 5000\,{\rm km\,s^{-1}}$, and achieved the best validation performance when $\delta{\rm v} = 3000\,{\rm km\,s^{-1}}$, which matches the minimum velocity separation between quasar and absorbers in the PM catalogue.

We assumed that $N_{\rm C\,IV}$ and $\sigma_{\rm C\,IV}$ are independent from $z_{\rm QSO}$, although $z_{\rm C\,IV}$ depends on $z_{\rm QSO}$ as described in equation (20) and equation (21). We calculated the marginalized model likelihood by integrating the absorption-model priors $\mathrm{M}_{D/S}$[14] as

$$P(\theta | z_{\rm QSO}) \propto P(z_{\rm C\,IV} | z_{\rm QSO}) P(N_{\rm C\,IV}) P(\sigma_{\rm C\,IV}). \tag{22}$$

Then we performed the integral for our absorption models, M_D and M_S, in equation (19):

$$P(D | z_{\rm QSO}) \propto \int P(\vec{y} | \theta, z_{\rm QSO}) P(\theta | z_{\rm QSO}) {\rm d}\theta. \tag{23}$$

However, equation (23) is intractable, so we approximated it with a quasi-Monte Carlo method. This method selected 10 000 samples of $\{N_{\rm C\,IV}, \sigma_{\rm C\,IV}, z_{\rm C\,IV}\}$ at which to calculate the model likelihood. The samples were drawn from a Halton sequence to ensure an approximately uniform spatial distribution. We approximate the model evidence by the sample mean

$$P(D | \mathrm{M}_{D/S}, z_{\rm QSO}) \simeq \frac{1}{N} \sum_{i=1}^{N} P(D | \theta_i, z_{\rm QSO}, \mathrm{M}_{D/S}). \tag{24}$$

We integrated out the parameters, $\theta = \{z_{\rm C\,IV}, N_{\rm C\,IV}, \sigma_{\rm C\,IV}\}$, with a given parameter prior $P(\theta | z_{\rm QSO}, \mathrm{M}_{D/S})$. We use 10 000 samples: lower sample sizes under-sample the likelihood function, while larger sample sizes cause the code to run slower. We considered 10 000–50 000 samples in the validation phase and found that increasing the number of samples did not significantly improve the validation performance. Note that using more samples increases the run-time cost of processing a quasar. In calculating the model evidence for the singlet model, M_S, we used a single-component Voigt profile centred on 1548 Å (equation 16) while for calculating the model evidence for the doublet model, M_D, we use a double-component Voigt profile centred at 1550 and 1548 Å (equation 15). We used the same parameter priors for both the singlet and doublet models for simplicity.

3.6 Multiple absorber search

In this paper, instead of reporting probabilities for multiple C IV absorbers as Ho et al. (2020) did for DLAs, we simplified and reported the probability that there is an absorber at a given redshift. For example, the posterior probability for the $k = 3$ model in Ho et al. (2020) does not indicate which of these three absorbers in $\mathrm{M}_{DLA(3)}$ is most probable, instead reporting the probability that a given spectrum contains some combination of three absorbers.

Here, we wish to find multiple absorbers in a spectrum. We proceed iteratively, noting that at any point the best fit may be a singlet or a doublet, and mask out the most likely absorber each time. We mask $350\,{\rm km\,s^{-1}}$ around 1548 Å(MAP($z_{\rm C\,IV}$) + 1) and $350\,{\rm km\,s^{-1}}$ around 1550 Å(MAP($z_{\rm C\,IV}$) + 1), where MAP($z_{\rm C\,IV}$) is the maximum a posteriori value for $z_{\rm C\,IV}$. For single-line absorbers, we mask $350\,{\rm km\,s^{-1}}$ around 1548 Å, again at MAP($z_{\rm C\,IV}$). Our procedure is as follows

(i) Fit our three models $\mathrm{M}_{N/S/D}$ on an observed spectrum.

(ii) If M_N (the null, C IV-free, model) has the highest posterior for any search, there is no C IV absorption in the given spectrum. Stop any further searches. Otherwise go to step 3.

(iii) If either M_S or M_D has the highest posterior, mask the spectral region around the most probable absorption profile. Return to step 1 to search for subsequent absorbers if no more than seven searches previously have been done. Otherwise stop any further searches.

Fig. 6 shows an example quasar spectrum (SDSS DR7 QSO-ID: 51608-0267-264 and $z_{\rm QSO} = 1.89$) within which both the PM and GP pipelines find three absorbers. Moreover, the GP pipeline finds an absorber at $z_{\rm C\,IV} = 1.489$ (probability 92 per cent) that was not detected by PM, due to noise in this part of the spectrum. Specifically, the 1550 line was not automatically detected with their parameters, thus the doublet was not visually inspected.

4 VALIDATION

For validation, we trained a C IV-free model, M_N, on a reduced training set of 95 per cent of the inspected spectra in the PM catalogue Cooksey et al. (2013). We then validated our algorithm with the remaining 5 per cent of the inspected (1301) spectra in the PM catalogue to check the agreement between the PM catalogue and our method. Note that when we applied our algorithm to DR12 spectra,

[13] To avoid possible confusion with any O I, Si II absorption pairs, see Cooksey et al. (2013).

[14] Either the doublet model or the singlet model.

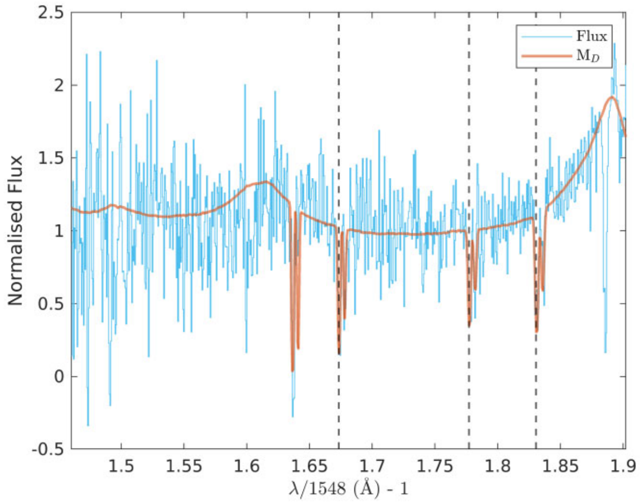


Figure 6. Example SDSS DR7 spectrum with QSO-ID: 51608-0267-264 and $z_{\rm QSO} = 1.89$. Both PM and our pipeline find three absorbers between $z_{\rm C\,IV} = 1.65$–1.85. We also find an absorber at $z_{\rm C\,IV} = 1.489$ (probability 92 per cent) that was not detected by PM, due to noise in this part of the spectrum (specifically, the 1550 line was not automatically detected with their parameters, thus the doublet was not visually inspected). The probabilities that our pipeline provides for the existence of the first, second, third, and fourth C IV absorber are $P(\rm C\,IV) = [1.00, 1.00, 1.00, 0.92]$, respectively, our maximum a posteriori absorber redshift values are $z_{\rm C\,IV} = [1.829, 1.672, 1.775, 1.489]$, and our rest equivalent widths from Voigt profile integration (see equation 30) are $W^{\rm GP}_{r,1548} = [1.37, 0.87, 0.90, 0.79]$ Å. In the PM-catalogue the absorber redshifts are $z_{\rm PM} = [1.831, 1.673, 1.777]$ with corresponding $W^{\rm PM}_{r,1548} = [1.21 \pm 0.18, 1.40 \pm 0.20, 0.94 \pm 0.19]$ Å.

we retrained our model using all SDSS DR7 spectra inspected in the PM catalogue without a reliable C IV absorber.

Our model is compared to the C IV absorbers as rated in the PM catalogue. Cooksey et al. 2013 rated their automatically detected C IV candidates from 0 (definitely not C IV), 1, 2, and 3 (definitely C IV), thus providing a rough estimate of confidence in an absorber. Absorbers with a ranking ≥2 are considered real C IV absorbers in the PM catalogue. We construct a 'ground truth' sample of the PM C IV with rating ≥2. Within a spectrum, we enforce that our GP-detected absorber is within 350 km s^{-1} of a PM-detected system to be considered as a 'match' between catalogues (see Fig. 6 for examples of matched absorbers); this cutoff is roughly $3 \times \max(\sigma_{\rm C\,IV})$ (where $\sigma_{\rm C\,IV}$ is measured by the GP),[15] which ensures we are not detecting a complex/blended system in two successive iterations (see Section 3.6). Moreover, we obtained a better purity/completeness (see Section 4.3) with a 350 km s^{-1} masking window.

4.1 Velocity separation

The velocity separation between absorbers detected in both the GP and PM catalogues is

$$\delta v_{\rm PM,GP} = \frac{z^{\rm PM}_{\rm C\,IV} - z^{\rm GP}_{\rm C\,IV}}{1 + z^{\rm PM}_{\rm C\,IV}} c. \qquad (25)$$

Fig. 7 shows that absorber redshifts obtained by our pipeline in the validation set are almost always consistent with the PM catalogue at the level of the SDSS spectral resolution, that is, $|\delta v_{\rm GP,\,PM}| \lesssim 150$ km s^{-1}. Very few points in Fig. 7 lie outside of the ±150 km s^{-1} horizontal lines. Our pipeline produces $z_{\rm C\,IV}$ on average slightly greater/redder than the PM catalogue, with a median offset $\delta v^{\rm med}_{\rm PM,GP}$≈-50 km s^{-1}. This is not a significant difference; by comparison an SDSS pixel is 69 km s^{-1}. In the PM catalogue, the redshift of the 1548 Å line was sometimes underestimated, as redshift estimation is weighted by flux-centred centroid where lower redshift wavelength pixels than 1548 Å have higher flux values. This can be more common at higher redshifts.

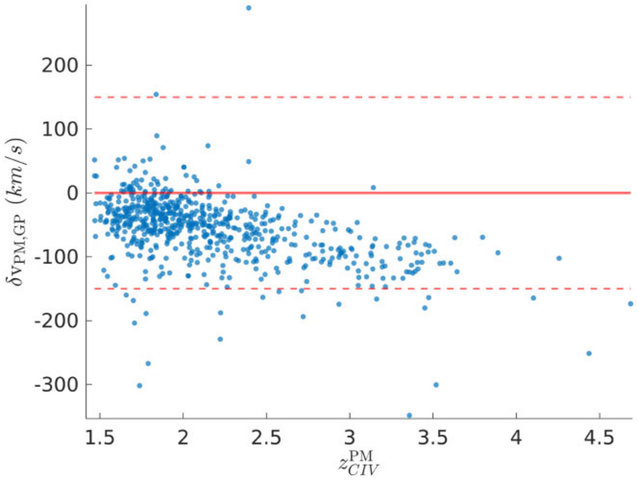


Figure 7. Velocity difference between the detected absorbers in the GP pipeline with P(M_D) ≥ 0.95 in the validation set and the absorbers in the PM catalogue. Only absorber pairs closer than 350 km s^{-1} are shown. The thick red line shows $\delta v_{\rm PM,\,GP} = 0$ and the dashed lines are $\delta v_{\rm PM,\,GP} = \pm 150$ km s^{-1} (the SDSS spectral resolution). The median offset is $\delta v^{\rm med}_{\rm PM,GP}$ ≈-50 km s^{-1}, which is less than an SDSS pixel (69 km s^{-1}).

We visually inspected the 9 spectra in our validation set of 1301 spectra with $\delta v_{\rm PM,\,GP} \geq 50$ km s^{-1}: most of them were in a complex/blend system and some of them were close to the QSO where the GP continuum was not perfect. We also investigated the 14 spectra in the validation set that show $\delta v_{\rm PM,\,GP} \leq -150$ km s^{-1}: most of them belong to a complex system or even a mini-BAL system. In some cases the GP continuum fit is not good. As a reference, we investigated 17 spectra with $\delta v_{\rm PM,\,GP} \sim -50$ km s^{-1}: these spectra are usually high-SNR and/or the GP continuum fit is very good, especially around the detected absorption system. Moreover, there is no significant correlation between the strength of the absorber systems and PM-GP velocity separation (equation 25) as shown by Fig. 8.

4.2 ROC curve

We use the ROC curve (Fig. 9), which is the true positive rate versus false positive rate for any classification threshold: $0 \leq P(M_D) \leq 1$ to obtain a score out of 1 for the performance of our classification (no C IV absorber versus C IV absorbers). The y-axis of the ROC curve in Fig. 9, the true positive rate, is the ratio of the number of C IV absorbers in our catalogue to the total number of of absorbers in the PM catalogue with a ranking ≥2. C IV absorbers in our catalogue are defined to be those with posterior probability greater than a threshold, P(M_D), between 0 and 1. They must also be less than 350 km s^{-1} apart from an absorber in the PM catalogue with a ranking≥2. The *x*-axis of the ROC curve in Fig. 9, the false positive rate, is the ratio of C IV absorbers in our catalogue that do not have any matching absorber

[15] For reference, in Cooksey et al. (2013), C IV absorbers were grouped into a single system if they were within 250 km s^{-1} of each other.

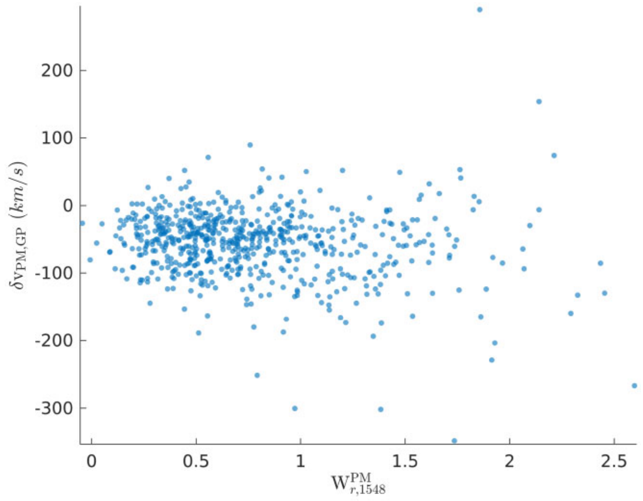


Figure 8. Velocity separation (equation 25) between GP and PM detected C IV absorption systems is shown versus the reported rest equivalent width values for 1548 Å in the PM catalogue ($\mathrm{W}^{\mathrm{PM}}_{r,1548}$). There is no correlation between the velocity separation and the strength of detected absorbers.

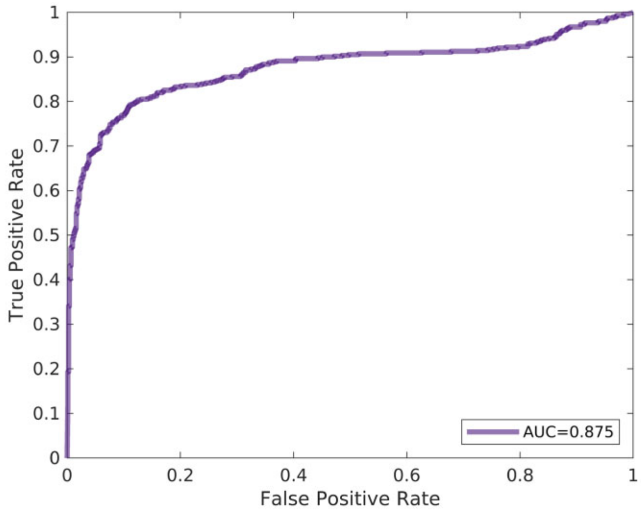


Figure 9. Receiver Operator Characteristic (ROC) curve for our DR7 validation. True Positive Rate is plotted versus False Positive Rate. True positives are C IV systems in our catalogue at least 350 km s^{-1} apart from an absorber in the PM catalogue with ranking ≥2 given any P(M_D) threshold between 0 and 1. False positives are those absorbers in our catalogue that do not have any matching absorber in the PM catalogue; though they may be real C IV absorbers (see Fig. 6). Above a relatively small False Positive Rate (∼0.2), our algorithm procedure obtains True Positive Rate above 80 per cent and, hence, is a successful way to identify C IV absorbers. The area under the ROC curve (AUC) is a quantitative metric for the equality of the GP algorithm; we get AUC = 0.87, a reasonable value compared to an ideal classification that gives AUC = 1.00.

with ranking ≥2 in the PM catalogue (given any P(M_D) threshold between 0 and 1) to the total number of absorbers in the PM catalogue with a ranking ≥2.

A higher classification performance (i.e. in each search run over a spectrum we classify it as C IV-free or having a C IV absorber) is reflected in a larger area under the curve (AUC) for the ROC curve. We obtain a quite reasonable AUC = 0.87. Note that here ‘true positive’ refers to a PM C IV absorber recovered by the GP algorithm in the training set, and ‘false positive’ is a GP C IV absorber not in the PM catalogue. However, as seen in Fig. 6, the GP procedure *can* find real/true C IV absorption not identified in the PM survey; hence, ‘false positives’ may be better considered ‘GP unique’. This also means that the classification performance (AUC = 0.87) we obtained here might underestimate the true performance.

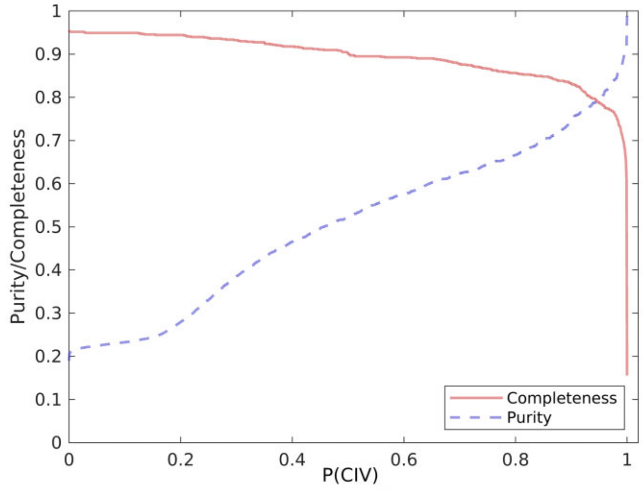


Figure 10. Purity (equation 26) and completeness (equation 27) of the GP catalogue compared to the PM catalogue for different C IV posterior probability (equation 14) thresholds. The maximum allowed velocity separation between our catalogue and the PM catalogue absorbers is 350 km s^{-1}. The intersection of the purity (dashed blue curve) and completeness (solid red curve) at a threshold of ∼ 95 per cent gives us a balanced purity/completeness of ∼ 80 per cent.

4.3 Purity and completeness

We assessed our algorithm’s performance by comparing individual absorption systems. We can compare our GP catalogue for various C IV posterior probabilities to the ‘ground truth’ sample of the PM catalogue. We define the purity of our GP catalogue as the fraction of the GP catalogue also in the PM catalogue:

$$\mathrm{Purity} = \frac{\mathrm{GP} \cap \mathrm{PM}}{\mathrm{GP}}. \tag{26}$$

The completeness is the fraction of the PM catalogue also in the GP catalogue:

$$\mathrm{Completeness} = \frac{\mathrm{GP} \cap \mathrm{PM}}{\mathrm{PM}}. \tag{27}$$

Fig. 10 shows completeness and purity as a function of threshold value. One should choose a threshold that gives the best possible combination of purity and completeness, around the point where the curves intersect. We thus choose a threshold of 95 per cent, which Fig. 10 shows produces purity and completeness of ∼ 80 per cent in a roughly equal balance. However, our catalogue reports posterior probabilities, so the user may choose a different threshold as desired for their application.

One of the strengths of our catalogue is the freedom that the user has for choosing the absorbers based on the C IV absorption model posterior probability P(M_D). The user can sacrifice the purity of absorbers for their completeness or vice versa. By sacrificing the purity, we are accepting absorbers with lower P(M_D). This will increase the number of accepted absorbers. However, it increases the chance of misidentifying some absorption features as C IV ones.

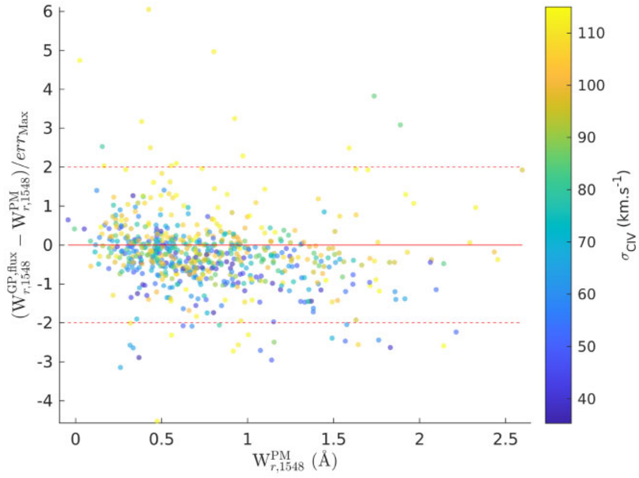


Figure 11. The ratio of the difference between rest equivalent width from our pipeline with boxcar flux summation ($W^{GP,flux}_{r,1548}$) and rest equivalent width from the PM catalogue ($W^{PM}_{r,1548}$) to the total error (see equation 28) from the PM catalogue and our pipeline for $W_{r,\,1548}$. The data points here are those absorption systems in the validation set where our pipeline reports an absorber with $P(M_D) \geq 0.95$ and for which there is an absorber with ranking ≥2 in the PM catalogue at a redshift offset less than 350 km s^{-1} (GP and PM in Section 4.4). As the colour bar shows, there is a trend towards larger maximum a posteriori $\sigma_{\rm C\,IV}$ when the GP rest equivalent width is larger than the rest equivalent width from the PM catalogue.

By sacrificing the completeness, we are restricting our catalogue to absorbers with higher $P(M_D)$. This will lower the number of accepted absorbers but will boost the purity of our catalogue.

4.4 Rest equivalent width comparison

We can evaluate our algorithm by comparing 1548 Å rest equivalent width $W^{GP,flux}_r$ between the GP and PM catalogues. $W^{GP,flux}_r$ is obtained by integrating the normalized flux deficit from our GP continuum (M_N) in a wavelength integration window corresponding to $4 \times \sigma_{\rm C\,IV}$ around the maximum a posteriori $z_{\rm C\,IV}$ for the 1548 Å line. We impose that the flux integration window does not exceed the midpoint of the 1550 and 1548 Å lines.

Fig. 11 shows the difference ratio in our validation set between the rest equivalent width of the 1548 Å line in the GP catalogue and the PM catalogue, scaled by the maximum error (because the rest equivalent width errors from the PM and GP catalogues are highly correlated):

$$\frac{W^{GP,flux}_{r,1548} - W^{PM}_{r,1548}}{err_{max}}.$$

The maximum error, err_{max}, is obtained by comparing the rest equivalent width error from the GP pipeline to that from the PM catalogue:

$$err_{max} = \max\{err(W^{GP,flux}_{r,1548}), err(W^{PM}_{r,1548})\}. \tag{28}$$

We obtained $err(W^{GP,flux}_{r,1548})$ by considering the observed noise in each pixel included in the integration window described above.

Around 94 per cent of the data points in Fig. 11 have $|\left(W^{GP,flux}_{r,1548} - W^{PM}_{r,1548}\right)/err_{max}| \leq 2$, which shows a reasonable consistency between the GP and PM rest equivalent widths. We visually inspected all of the 21 absorbers with $\left(W^{GP,flux}_{r,1548} - W^{PM}_{r,1548}\right)/err_{max} < -2$: they mostly have continuum issues and a low GP continuum. For QSO 53886-1823-377, a triplet[16] C IV system at $z_{\rm C\,IV} = 1.838$ caused a lower 1548 rest equivalent width in the GP catalogue than in the PM catalogue. In the spectrum of QSO 53083-1757-529 our algorithm finds an absorber at the end of the spectrum, which also yields a lower rest equivalent width when compared to the PM catalogue. Looking at 13 absorbers with $\left(W^{GP,flux}_{r,1548} - W^{PM}_{r,1548}\right)/err_{max} > 2$, we realized that most of these absorbers belong to a triplet C IV system. As a reference we also checked absorbers with $\left|W^{GP,flux}_{r,1548} - W^{PM}_{r,1548}\right|/err_{max} < 0.025$: these spectra were mostly high-SNR and the GP continuum fit the observed quasar very well.

The colour bar in Fig. 11 shows the maximum a posteriori $\sigma_{\rm C\,IV}$ that the GP algorithm produces for each absorber. There is a correlation between larger maximum a posteriori $\sigma_{\rm C\,IV}$ and absorbers where the GP rest equivalent width, $W^{GP,flux}_{r,1548}$, is larger than the PM rest equivalent width, $W^{PM}_{r,1548}$. We visually inspected these systems and found that many of them are triplet or mini-BAL systems, for which the GP is more likely to give a large $\sigma_{\rm C\,IV}$.

The difference in rest equivalent widths of the 1550 Å line between the GP and PM catalogues behaves similarly. We find that 518 (86 per cent) of GP absorbers with a PM absorber system at a redshift offset less than 350 km s^{-1} away showed $\left|W^{GP,flux}_{r,1500} - W^{PM}_{r,1500}\right|/err_{max} \leq 2$. The 1550 Å line is weaker than the 1548 Å line, leading to a generally lower detection significance. However, for strong absorbers it is useful because it is less saturated.

The GP pipeline finds 822 absorbers in the validation set spectra with $P(M_D) \geq 95$ per cent. In the PM catalogue the validation set spectra contain 829 absorbers with a ranking ≥2. We can divide these absorbers into four different categories with the following statistics:

(i) PM and GP: absorbers with a ranking≥2 in the PM catalogue, $P(M_D) \geq 0.95$ in the GP catalogue, and $\delta v_{\rm PM,\,GP} \leq 350$ km s^{-1}. This category contains 647 absorbers, ∼ 78 per cent of the PM absorbers and ∼ 79 per cent of the GP absorbers among the 1301 spectra in the validation set.

(ii) GP only: absorbers with $P(M_D) \geq 0.95$ but no absorber in the PM catalogue with ranking ≥2 and a velocity offset less than 350 km s^{-1}. This category includes 175 absorbers (21 per cent of the GP absorbers in the validation set). Some of these absorbers are true C IV which fell beneath the sensitivity of the candidate search in Cooksey et al. (2013), and some are other doublet lines which our GP model has incorrectly classified as C IV.

(iii) GP uncertain: absorbers with a ranking ≥2 in the PM catalogue, and an absorber from the GP catalogue with a velocity offset less than 350 km s^{-1} but $P(M_D) < 95$ per cent. There are 142 of these absorbers, ∼ 17 per cent of the PM absorbers in the validation set spectra. Note that 85 (∼ 60 per cent) of these GP uncertain absorbers have $P(M_D) \geq 50$ per cent in the GP catalogue and that the authors of the PM catalogue resolved ambiguous absorbers by inspecting other metal lines from the same system.

(iv) PM only: 40 (4.8 per cent) of 829 absorbers with ranking ≥2 in the PM catalogue validation set had no GP absorber candidates within 350 km s^{-1} in the GP catalogue. The GP pipeline thus misses these absorbers in its successive searches of the validation set spectra. Two absorbers were assigned to this category because there were two PM absorbers in the spectrum closer than 350 km s^{-1} to each other. The GP catalogue found one, and the region containing the second was masked. Note that these absorbers are the reason why Fig. 10

[16] When a lower redshift absorber's 1550 Å line blends with the 1548 Å line of the higher redshift absorber.

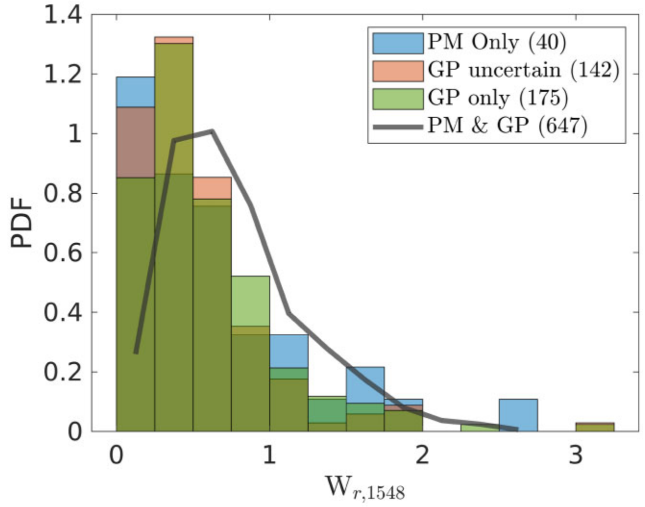


Figure 12. Distribution of W$_{r,\,1548}$ for absorbers in four categories described in Section 4.4: detected in both the GP and PM catalogues (thick black line), in the GP uncertain (brown), in GP only (green), in the PM catalogue only (blue). The rest equivalent width distribution is similar for all categories. There are some strong absorbers with (W$_{r,\,1548}$ > 1.2 Å) classified as 'PM only'. Visual inspection of the spectra of these systems indicates that they are part of a triplet/complex absorber or a broad mini-BAL system.

does not show a completeness of 1, even with a threshold of P(M$_D$) = 0.

Fig. 12 shows the distribution of rest equivalent widths for 40 PM only absorbers, 142 GP uncertain absorbers, 175 GP only absorbers, and 647 PM and GP absorbers in the four categories described earlier. Fig. 12 also demonstrates that weak absorbers ($W_{r,\,1548}$<0.3 Å) are detected less often than other categories. Considering that there are intrinsically more weak absorbers in the population, it is likely that each algorithm (i.e. GP and PM) captures a different subset of weak absorbers. However, for stronger absorbers, the four categories are more or less consistent considering the uncertainties in measuring rest equivalent width.

There are 17 strong absorbers (W$_{r,\,1548}$ ≥ 1.2 Å) in the PM only or GP uncertain categories. 11 of these absorbers are triplet/mini-BAL systems where the GP pipeline gives P(M$_S$) ∼ P(M$_D$). The complex shape of these absorption systems are not a good match to either model, so the GP pipeline is not able to distinguish between them. Two absorbers have P(M$_D$) ≥ 0.95 but a velocity separation more than 350 km s^{-1}. One of these is also close to a complex absorber system. Five absorbers are detected by the GP catalogue as a singlet, and so have P(M$_S$) ≫ P(M$_D$).

Thus most missed strong absorbers are caused either by C IV triplets or by the GP pipeline preferring a singlet fit to a doublet in cases where the doublet structure is not well resolved. Note that when conducting visual inspection of C IV absorbers, an observer may resolve ambiguous lines using information from other metal line transitions associated with the same system, whereas our GP pipeline uses only information from the C IV transition.

4.5 Example absorbers

In this section we examine example absorbers from the GP only, GP uncertain, and PM only categories discussed earlier. Fig. 13 shows an example of a spectrum where the GP catalogue shows two absorbers with probability more than 95 per cent, but the PM catalogue has zero detections. In this case, the GP C IV at z = 2.288 is actually Al II λ1670 from a strong, multicomponent system at z = 2.05 with Mg II λλ2796, 2803; Fe II λλλ2344, 2374, 2384, and λλ2586, 2600; and Al III λλ1854, 1862. The latter was flagged by the GP algorithm as $z_{\mathrm{C\,IV}}$ = 2.650. The z = 2.288 'C IV' was detected as a candidate in Cooksey et al. (2013) but visual inspection revealed its true identification; the z = 2.650 'C IV' was not even a candidate in Cooksey et al. (2013) because the would-be 1550 line was not detected by the automated candidate finder (i.e. it fell below their sensitivity threshold). Thus some of the absorbers in the GP only category are simply missed by the PM pipeline and some are false detections of other doublets.

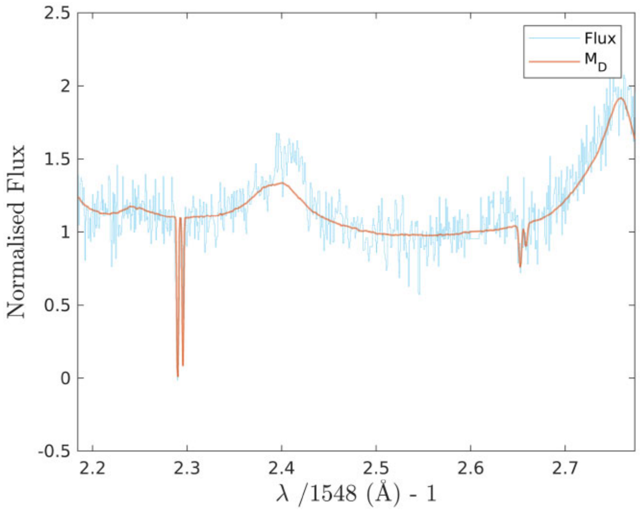


Figure 13. Example spectrum with two C IV absorbers found by GP with high confidence but not included in the PM catalogue. The QSO-ID is 51994-0309-592 and z_{QSO} = 2.76. Posterior probabilities for the two searches are P(M$_D$) = [1.00, 0.98]. The maximum a posteriori absorption redshifts are $z_{\mathrm{C\,IV}}$ = [2.288, 2.650], and the rest equivalent widths are $\mathrm{W}^{\mathrm{GP,flux}}_{r,1548}$ = [0.90, 0.32] Å. These two 'C IV' systems are actually non-C IV absorption lines from a strong, complex system at lower redshift. The PM pipeline identified the z = 2.288 lines as a C IV candidate but ranked it zero; the z = 2.650 'C IV 1550 Å' line fell below the PM detection threshold.

Fig. 14 shows an example spectrum where the GP pipeline is uncertain about an absorber detected in the PM catalogue. Both pipelines find the C IV absorber at ${z_{\mathrm{C\,IV}}}^{\mathrm{GP}}$ = 1.827. However, the PM catalogue identifies a second absorber at ${z_{\mathrm{C\,IV}}}^{\mathrm{PM}}$ = 1.822. This absorber is also detected by the GP pipeline. However, the GP pipeline is unable to distinguish between the doublet and singlet models as the 1550 Å line is blended with the higher redshift absorber. It thus assigns both models equal probability, hence P(M$_D$) = 49 per cent.

Fig. 15 illustrates QSO-ID: 51943-0300-475, which contains an example of an absorber in the PM only category. The PM catalogue contains two absorbers at $z^{\mathrm{PM}}_{\mathrm{C\,IV}}$ = [3.5309, 3.5389]. The GP pipeline finds an absorber in the first C IV-search at $z^{GP}_{\mathrm{C\,IV}}$ = 3.540574 with a posterior probability of 1 for the doublet model. According to our multi-absorber finding procedure (see Section 3.6) we mask the observed flux 350 km s^{-1} around the found absorber and do the next search. However, since the other reported absorber in the PM catalogue ($z^{\mathrm{PM}}_{\mathrm{C\,IV}}$ = 3.5309) is offset only 110 km s^{-1} from the absorber found in the first GP search, it is in a masked region and not identified by the GP pipeline in the second search.

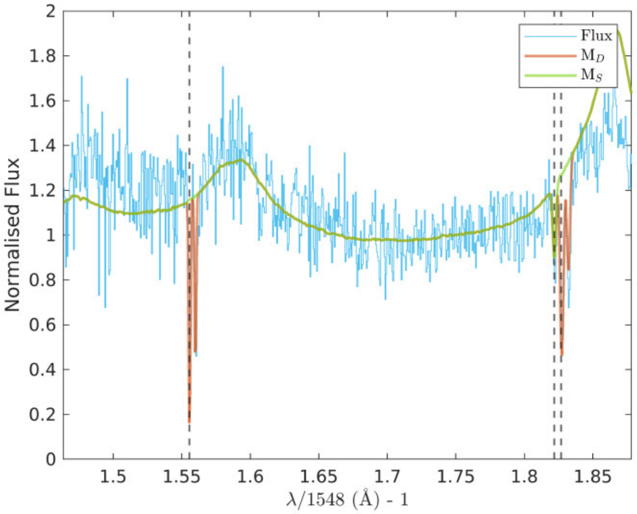


Figure 14. Example of an absorber at ${z_{\rm C\,IV}}^{\rm PM} = 1.822$ detected by the PM catalogue, but assigned a relatively low probability (P(M_D) = 49 per cent) by the GP catalogue. The QSO-ID for this spectrum is 52367-0332-585, and the quasar redshift is 1.87. The vertical dashed lines show the position of PM absorbers. The posterior absorption probabilities are P(M_D) = [1.00, 1.00, 0.49, 0.15], with maximum a posterior absorber redshifts of $z_{\rm C\,IV}$ = [1.556, 1.827, 1.822, 1.693], and the rest equivalent widths are $W_{r,1548}^{\rm GP,flux}$ = [0.528 ± 0.37, 1.21 ± 0.25, 0.55 ± 0.30, 0.05 ± 0.35] Å. The PM catalogue reported absorbers at $z_{\rm C\,IV}^{\rm PM}$ = [1.556, 1.827, 1.822] with $W_{r,1548}^{\rm PM}$ = [0.88 ± 0.12, 0.88 ± 0.08, 0.40 ± 0.10] Å.

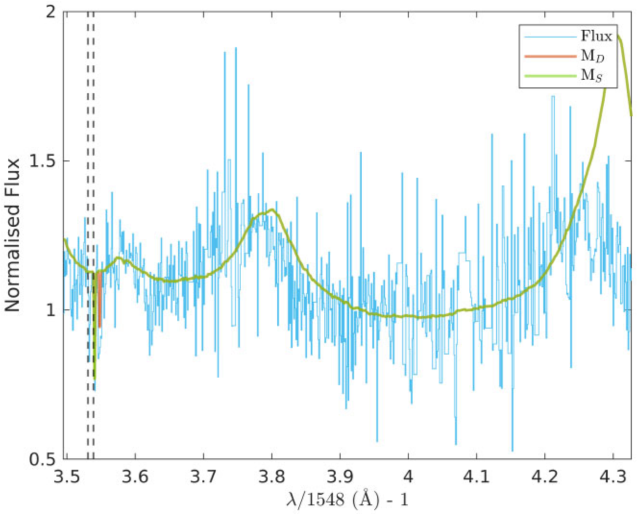


Figure 15. Example spectrum containing a PM only absorber for QSO-ID: 51943-0300-475 and $z_{\rm QSO} = 4.31$ where $z_{\rm C\,IV}^{\rm PM}$ = [3.5309, 3.5389] (vertical dashed lines). GP assigns P(M_D) = 1 to $z_{\rm C\,IV}^{GP}$ = 3.540574 which is offset by only 110 km s^{-1} from $z_{\rm C\,IV}^{\rm PM}$ = 3.5389. Before the second search, we mask 350 km s^{-1} around the first absorber and thus are unable to detect the second PM catalogue absorber.

5 RESULTS FOR SDSS DR12

We applied our model to find C IV absorbers in a subset of the SDSS DR12 quasar catalogue. We searched quasars with rest-frame wavelength coverage between 1310 and 1548 Å (1.7 < $z_{\rm QSO}$ < 5.7), and without detected BALs. This leaves 185 425 quasar spectra (see Section 2). For each spectrum, the GP pipeline provides (shown as columns in Table 1): posterior probability of C IV absorption, maximum a posteriori values for our absorption model parameters ($z_{\rm C\,IV}$, $N_{\rm C\,IV}$, and $\sigma_{\rm C\,IV}$), together with their 95 per cent confidence intervals, and rest equivalent widths (for 1548 and 1550 Å) and their 95 per cent confidence intervals. Maximum a posteriori values and 95 per cent confidence intervals for our absorption model parameters summarize the likelihood distribution, $P(D|\theta_i, z_{\rm QSO}, {\rm M}_D)$, of our 10 000 parameter samples (see equation 24). Each of these results are contained in a 185 425 × 7 array. If the search terminated finding fewer than seven absorbers, we report a `NaN` value for the columns associated with all further absorbers. Table 1 shows a snapshot of our search results for the first 10 absorbers with P(M_D) ≥ 0.

Figs 16a – 16d illustrate the four C IV searches done by the GP pipeline on QSO-56265-6151-936 with $z_{\rm QSO}$ = 2.4811, and we briefly explain these iterations here. We found a C IV absorber at $z_{\rm C\,IV}$ = 2.13682. In the first search, the null model had $P({\rm M}_N)$ = 0.0, the single line model $P({\rm M}_S)$ = 0.0, and the C IV doublet model $P({\rm M}_D)$ = 1.0. We thus masked the C IV doublet model 350 km s^{-1} around the C IV absorber at $z_{\rm C\,IV}$ = 2.13682 in the first C IV search and commenced the second search, shown in Fig. 16b. Our second search found an absorber at $z_{\rm C\,IV}$ = 2.15132 with $P({\rm M}_D)$ = 1.0, $P({\rm M}_S)$ = 0.0, and $P({\rm M}_N)$ = 0.0. For the third C IV search we masked 350 km s^{-1} around each of the absorbers found in the previous steps and found a third absorber at $z_{\rm C\,IV}$ = 2.42670, again with with $P({\rm M}_D)$ = 1.0, $P({\rm M}_S)$ = 0.0, and $P({\rm M}_N)$ = 0.0. The fourth search, with regions around all three previous absorbers masked, found P(M_D) = 0.27, P(M_S) = 0.0 and P(M_N) = 0.73. Since the null model now had the largest model posterior, this was the final C IV absorber search in this spectrum (see Section 3.6).

Looking at the distribution of each model posterior probability gives us an insight into how the spectra have been classified. Fig. 17 shows the distribution of doublet model posterior probabilities for the first four C IV absorber searches in the DR12 spectra. For each search, we see a peak in the posterior probability distribution around 30 per cent, stronger for earlier searches. We also examined P(M_S) and P(M_N) for the first search and found that many of these ambiguous C IV absorbers also have P(M_S) ∼ 0.3. In addition, these absorbers are generally in lower S/N spectra. Thus this peak occurs when the spectra are weakly constraining and the posterior absorber probability is dominated by the model priors.

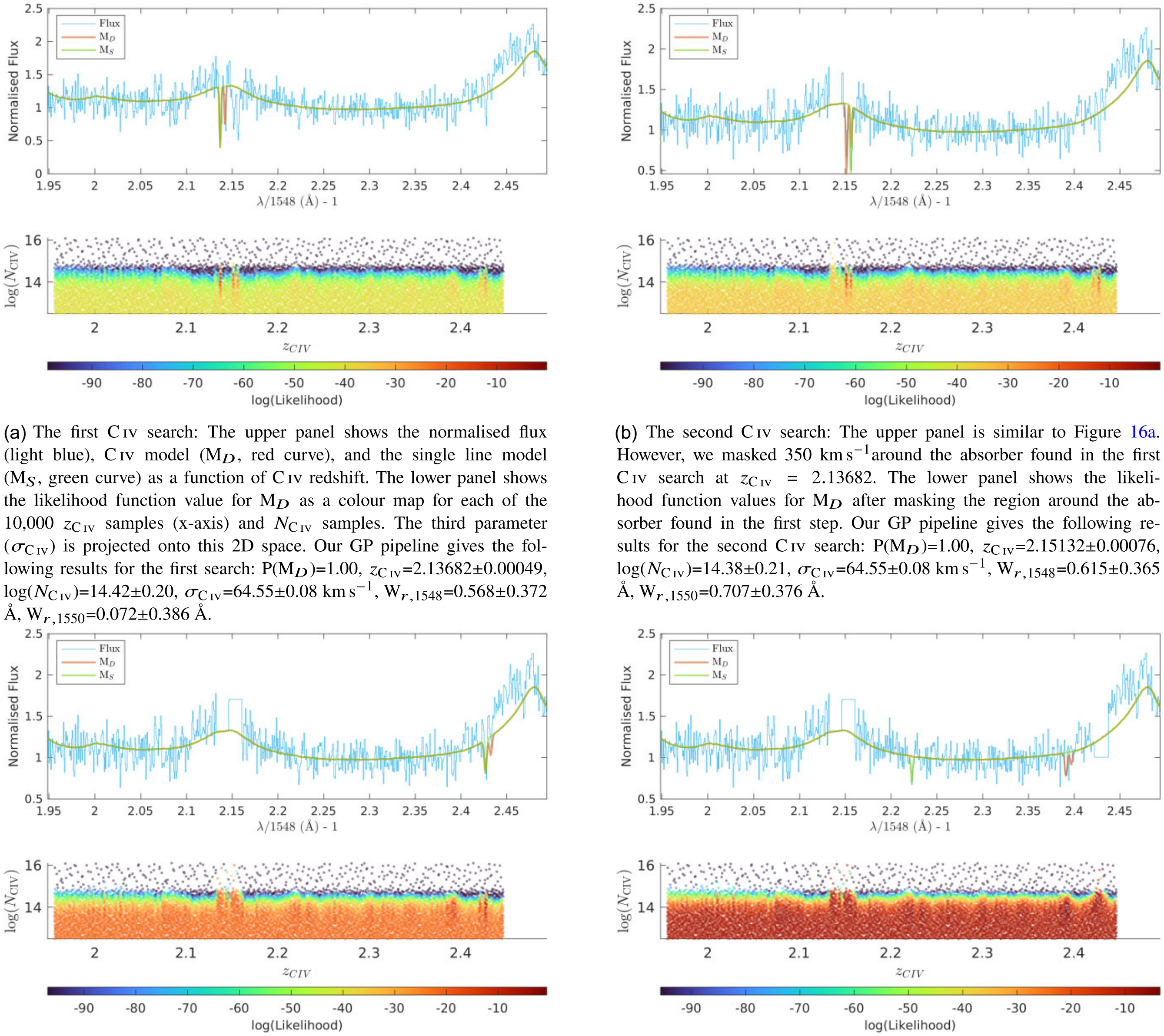
Table 2 summarizes the reported posterior probabilities for our catalogue. Around 66 per cent of spectra have no C IV absorbers detectable at more than 85 per cent confidence. Around 15 per cent of spectra have one doublet and around 8 per cent two doublets, each with a confidence more than 85 per cent. The probability for detecting two independent absorbers in a spectrum is

$$\text{P(2 C IV)} = \text{P(1 C IV)} \times \text{P(1 C IV)} = 0.15^2 \sim 2.2 \text{ per cent.} \qquad (29)$$

The actual probability of two C IV absorbers in a spectrum is higher, ∼ 8 per cent, demonstrating that absorbers are not independent but strongly correlated. Furthermore, we find five or more doublets at > 85 per cent confidence in 3.1 per cent of spectra.

We detected a single-line absorber in ∼10 per cent of sightlines. If single-line absorbers were independent, we would expect 0.1^2 or 1 per cent of spectra to contain two singlet line absorbers. The actual probability of finding two singlets in a single sightline was ∼ 2 per cent, so the correlation between single-line absorbers is much weaker than for C IV.

We provided the distribution of maximum a posteriori column densities for with P(M_D) > 0.95 in Fig. 18 for different redshift bins. Blue histograms in each panel correspond to the PDF from the GP pipeline while the red histograms show the column densities in the PM catalogue. Note that column density values measured by



(a) The first C IV search: The upper panel shows the normalised flux (light blue), C IV model (M_D, red curve), and the single line model (M_S, green curve) as a function of C IV redshift. The lower panel shows the likelihood function value for M_D as a colour map for each of the 10,000 $z_{\rm C\,IV}$ samples (x-axis) and $N_{\rm C\,IV}$ samples. The third parameter ($\sigma_{\rm C\,IV}$) is projected onto this 2D space. Our GP pipeline gives the following results for the first search: P(M_D)=1.00, $z_{\rm C\,IV}$=2.13682±0.00049, $\log(N_{\rm C\,IV})$=14.42±0.20, $\sigma_{\rm C\,IV}$=64.55±0.08 km s^{-1}, $W_{r,1548}$=0.568±0.372 Å, $W_{r,1550}$=0.072±0.386 Å.

(b) The second C IV search: The upper panel is similar to Figure 16a. However, we masked 350 km s^{-1}around the absorber found in the first C IV search at $z_{\rm C\,IV}$ = 2.13682. The lower panel shows the likelihood function values for M_D after masking the region around the absorber found in the first step. Our GP pipeline gives the following results for the second C IV search: P(M_D)=1.00, $z_{\rm C\,IV}$=2.15132±0.00076, $\log(N_{\rm C\,IV})$=14.38±0.21, $\sigma_{\rm C\,IV}$=64.55±0.08 km s^{-1}, $W_{r,1548}$=0.615±0.365 Å, $W_{r,1550}$=0.707±0.376 Å.

(c) The third C IV search: The upper panel is similar to Figure 16b but with 350 km s^{-1}around the two absorbers found in the first and second C IV searches at $z_{\rm C\,IV}$ = 2.13682 and $z_{\rm C\,IV}$ = 2.15132 masked. The lower panel shows the likelihood function value as a colour map after masking both absorbers. Our GP pipeline gives the following results for the third search: P(M_D)=1, $z_{\rm C\,IV}$=2.42670±0.00006, $\log(N_{\rm C\,IV})$=14.17±0.02, $\sigma_{\rm C\,IV}$=111.81±0.01 km s^{-1}, $W_{r,1548}$=0.164±0.407 Å, $W_{r,1550}$=-0.602±.396 Å.

(d) The fourth and final C IV search. The upper panel is similar to Figure 16c but with 350 km s^{-1}around the absorbers found by the previous three searches masked. The lower panel shows the likelihood function value as a colour map for each of the 10,000 $z_{\rm C\,IV}$ samples (x-axis) and $N_{\rm C\,IV}$ samples. Our GP pipeline gives the following results for the final search: P(M_D)=0.27, $z_{\rm C\,IV}$=2.39100± 0.00341, $\log(N_{\rm C\,IV})$=14.04±0.82, $\sigma_{\rm C\,IV}$=105.99±0.56 km s^{-1}, $W_{r,1548}$=0.248±0.434, $W_{r,1550}$=-0.085±0.424. Note that since the highest probability in the fourth search was P(M_N)=0.73, the algorithm performs no further searches (see Section 3.6).

Figure 16. Panels (a) through (d), show four subsequent searches for C IV absorption on the spectrum of QSO-56265-6151-936 with a redshift of 2.4811.

our pipeline are often lower limits because given the low-resolution spectra of SDSS, they are partially to completely saturated.

C IV absorbers have a power-law distribution (Ellison et al. 2000), so that we should observe more absorbers for lower equivalent widths. However, weaker absorbers are harder to detect due to observational limitations. That is why the distribution of column densities in all of the panels of Fig. 18 starts to drop for $\log(N_{\rm C\,IV}) \lesssim 14$. This turning point allows a rough estimate for the completeness limit of our catalogue for detecting weak absorbers. The turning point is at lower column densities for the blue histograms, indicating that the GP pipeline is slightly more sensitive than the PM catalogue.

One can do a completeness test for the detection of column density or rest equivalent width by randomly injecting synthetic absorbers into absorption-free quasar spectrum and assessing the probability of recovering those artificial absorbers (Cooksey et al. 2013). Panels of Fig. 18 shows again that the completeness for column density

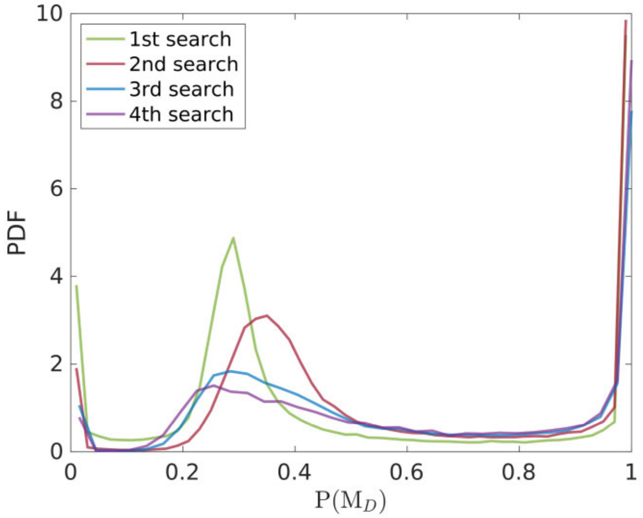


Figure 17. Distribution of P(M$_D$) for the first to fourth searches. We do not show the fifth to seventh searches as they find very few absorbers (see Table 2). The peak around P(M$_D$) ∼ 0.3 comes from low-SNR spectra where the posterior probabilities of our three models are dominated by their priors.

Table 2. The number of spectra containing different numbers of C IV absorbers for various doublet model probability thresholds, P(M$_D$). The first column shows the number of C IV absorbers found within each spectrum (see Section 3.6). The second through fourth columns show probability thresholds of > 65 per cent, 85 per cent, and 95 per cent, respectively. Cells show the number of quasar spectra falling in each category, together with the corresponding percentage of the 185 425 spectra in our SDSS DR12 sample.

N(C IV)	P(M$_D$)>0.65	P(M$_D$)>0.85	P(M$_D$)>0.95
0	113142 61.0 per cent	123994 66.8 per cent	131767 71.0 per cent
1	31163 16.8 per cent	27733 14.9 per cent	24981 13.5 per cent
2	17526 9.4 per cent	14533 7.8 per cent	12777 6.9 per cent
3	10020 5.4 per cent	8424 4.5 per cent	7218 3.9 per cent
4	6112 3.3 per cent	4960 2.6 per cent	4176 2.3 per cent
5	4155 2.2 per cent	3342 1.8 per cent	2656 1.4 per cent
6	2426 1.3 per cent	1771 0.9 per cent	1348 0.7 per cent
7	881 0.5 per cent	668 0.4 per cent	502 0.3 per cent

is around log($N_{\rm C\,IV}$) ∼ 14. In a follow-up project, we will perform a detailed analysis of the completeness of our technique, including any possible redshift or signal-to-noise dependence.

Fig. 19 shows the distribution of (maximum a posteriori) absorber redshifts. There is a peak around z ∼ 2, mirroring the distribution of quasar redshifts. Overall, we have detected an order of magnitude more absorbers than the PM catalogue, reflecting the larger size of our sightline sample. There are 33 absorbers in DR12 with P(M$_D$) > 0.85 and a redshift higher than 4.68, the maximum reported $z_{\rm C\,IV}$ in the PM catalogue.

In Fig. 20 we show the distribution of maximum a posteriori Doppler velocity dispersion, $\sigma_{\rm C\,IV}$, for the absorbers detected in the DR12 spectra. While the adopted prior for $\sigma_{\rm C\,IV}$ was a flat distribution between 35 and 115 km s^{-1}, we see that the posterior distribution is moderately bimodal. The peak at larger $\sigma_{\rm C\,IV}$ values is connected with larger column densities. We examined the spectra of detected absorbers with log $N_{\rm C\,IV}$ > 16 and $\sigma_{\rm C\,IV}$ > 110 km s^{-1}. Most of these spectra are noisy and in some cases the line is heavily blended. The mean S/N in the region of C IV search for absorbers with probability larger than 85 per cent is 6.4 pix^{-1}, compared to a mean S/N of 5 pix^{-1} for the search region of C IV absorbers in the spectra that contain absorbers with $\sigma_{\rm C\,IV}$ > 80 km s^{-1} and log $N_{\rm C\,IV}$ > 15.

Fig. 21 shows the 1548 Å rest equivalent widths from our SDSS DR12 catalogue. We use rest equivalent widths derived from the

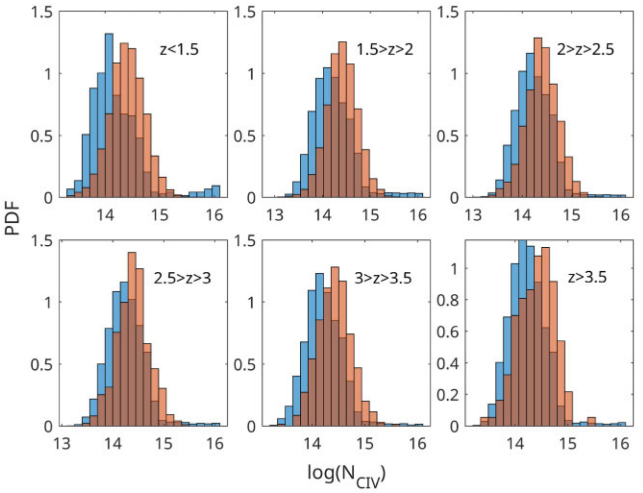


Figure 18. Column density statistics for our GP results (blue histograms) compared to the training C IV catalogue from SDSS DR7 (red histograms). Each panel is showing a specific redshift bin. Y-axes show the normalized probability distribution function (PDF). Please note that column density values should be considered as lower limits ((Cooksey et al. 2013)) because they lines are partially to completely saturated so we can only measure lower limits.

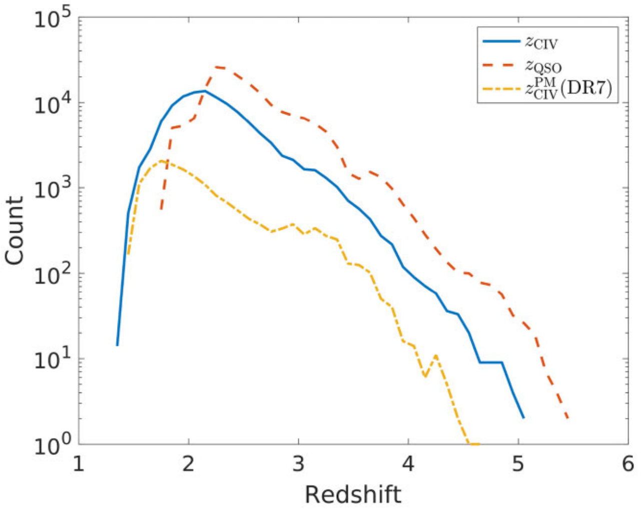


Figure 19. The redshift distributions of DR12 quasars (red dashed), high-probability (P(M$_D$) ≥ 0.95) DR12 GP C IV absorbers (blue solid), and DR7 PM C IV absorbers with ranking≥2 (dot-dashed yellow). The quasar redshift is offset towards redder values than the absorber redshift, as expected, since absorbers cannot be more redshifted than the quasar. The GP catalogue finds absorbers outside of the absorber redshift range reported in the PM catalogue.

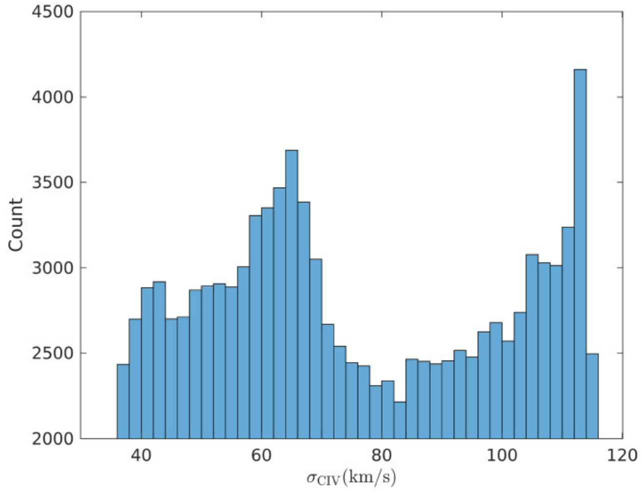


Figure 20. Distribution of the maximum a posteriori Doppler velocity dispersion values for absorbers detected in SDSS DR12 with P(M_D) ≥ 0.95. Our prior distribution for Doppler velocity dispersion was uniform between 35 and 115 km s^{-1} but the posterior distribution is bimodal. The larger $\sigma_{\rm C\,IV}$ posterior values are mostly associated with C IV absorbers found near low SNR pixels.

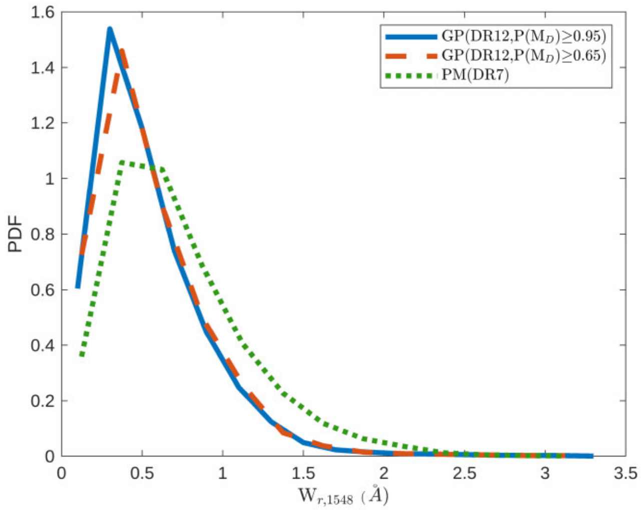


Figure 21. The distribution of the rest equivalent width of the 1548 Å ($W^{\rm GP,Voigt}_{r,1548}$) line obtained by Voigt profile integration (equation 30). We show $W^{\rm GP,Voigt}_{r,1548}$ for all detected DR12 absorbers with P(M_D) ≥ 0.95 (blue curve) and P(M_D) ≥ 0.65 (dashed red curve). We also show for comparison $W^{\rm PM}_{r,1548}$ values from the PM (DR7) catalogue in the dotted green curve.

parameters of the Voigt profile doublet model, $W^{\rm GP,Voigt}_{r,1548}$, which are computed using

$$W^{\rm GP,Voigt}_{r,1548} = \int (1 - a_{1548}) d\lambda. \tag{30}$$

Here a_{1548} is the absorption function for our 1548 Å line, and we compute the rest equivalent width from the maximum a posteriori values of $z_{\rm C\,IV}$, $N_{\rm C\,IV}$, and $\sigma_{\rm C\,IV}$ under M_D. Fig. 21 also shows the DR7 PM catalogue rest equivalent width distribution for comparison.

The larger sample of spectra in SDSS DR12 allows us to probe higher rest equivalent widths, with 110 absorbers at larger rest equivalent widths than 3.15 Å, the highest rest equivalent width in the PM catalogue. The 50 per cent completeness limit for $W_{r,\,1548}$ in the PM catalogue is 0.6 Å (Cooksey et al. 2013). Fig. 21 suggests that our catalogue is reasonably complete for $W_{r,\,1548} > 0.4$ Å. We will perform a detailed statistical analysis, including the completeness for rest equivalent widths, in a follow-up paper.

6 CONCLUSIONS

We trained a quasar continuum model to detect C IV absorbers using a Gaussian process. The training was done on a sample of DR7 spectra which were labelled as C IV free in the PMs catalogue of Cooksey et al. (2013). We used Bayesian model selection to compare our continuum model to models containing one to seven C IV doublets. We added an extra model for single-line absorbers, to avoid confusion from interloping metal lines. The prior distribution was taken from our training catalogue and flat parameter priors were used for the C IV redshift, $z_{\rm C\,IV}$ and Doppler velocity dispersion $\sigma_{\rm C\,IV}$. We searched for up to 7 absorbers in each tested spectrum and provide a comprehensive catalogue containing C IV detection probability, as well as maximum a posteriori values and credible intervals for $z_{\rm QSO}$, $N_{\rm C\,IV}$ and $\sigma_{\rm C\,IV}$. We validated our pipeline by applying it to a hold-out sample of 1301 spectra from the PM catalogue. Our pipeline produced similar results to the PM catalogue and has good purity and completeness. Generally the two catalogues produced similar C IV redshifts and rest equivalent widths.

Thus validated, we applied our model to SDSS DR12, and produced the largest C IV absorption catalogue yet seen. Among the total 185 425 selected quasar spectra in SDSS DR12, we found 113 775 C IV doublets with > 95 per cent confidence. Note that the user may pick the desired confidence threshold in our catalogue, thanks to our reported posterior probabilities for each absorber. We detected C IV absorption up to $z \sim 5$, including 33 systems at higher redshift than seen in DR7. We also detect 110 absorbers in DR12 with a rest equivalent width larger than the maximum in the DR7 catalogue.

This paper presents the first machine learning approach for detecting C IV absorption lines in quasar spectra. Our GP pipeline models a quasar spectrum by learning the covariance between different parts of quasar spectra, using an absorber-free training set. Then analytic Voigt profiles are used to find the absorbers. Other machine learning methods, such as convolutional neural networks, focus on modelling the absorbers without modelling the quasar continuum (eg. Parks et al. (2018)). Moreover, methods based on Convolutional Neural Networks have a considerable number of hyperparameters, scaling with the number of layers in the network. While Convolutional Neural Networks are very promising, Gaussian Processes are easier to interpret and come with a natural estimate of uncertainty. GP pipeline proved that is able to detect DLAs in a Bayesian manner by considering quasar redshift uncertainty which is usually considered as an exact parameter (Fauber et al. 2020). This represents a distinct advantage of the GP model, as current neural network-based absorber finders do not possess a straightforward approach for marginalizing QSO redshift uncertainty.

Our method is good for detecting unblended C IV absorbers. Note however, that our GP pipeline contains models only for singlet and doublet absorbers, and so may misclassify blended or mini-BAL systems, which do not strongly resemble the models it uses. In these rare cases, when the absorption systems are blended together, the line may be a poor match to both the singlet and doublet models. In these cases, our pipeline is sometimes unable to distinguish between genuine C IV doublet absorption and other interloper metal lines. It is possible that two single lines, with the same rest-frame wavelength separation as C IV co-exist in a spectrum, causing our algorithm to misidentify them as a C IV doublet. An example for this case is O I 1302 Å and Si II 1304 Å pair. To avoid this specific example, we start our C IV search redward of 1310 Å. Our algorithm could potentially be confused by other similar situations still in our wavelength search region. However, since our model uses a doublet Voigt profile designated for C IV absorption, it does not easily mistake doublet metal lines with other wavelength separations as a C IV doublet, such as Si IV $\lambda\lambda$ 1393 Å, 1402 Å.

We could have considered a mixture of various absorbers and then properly marginalize different types of absorbers in the Bayesian model selection. However, our primary objective in this paper is to efficiently identify C IV candidates for a large number of spectra. Introducing this level of complexity into the search pipeline would significantly slow down the process. Nevertheless, we have provided the catalogue containing all posterior samples and all the files to reproduce the inference. This allows interested users the flexibility to pinpoint specific C IV absorber candidates of interest and subsequently recalculate the uncertainty associated with other types of metal lines as needed. An advantage of our pipeline is its modularity, enabling users to seamlessly incorporate new absorbers into the code by adding the corresponding Voigt profile.

Here we used a single Voigt profile to fit the absorption systems. The Doppler velocity dispersion prior distribution for our absorption models ranges from 35 to 115 km s^{-1}, which translates to a temperature range of $\gtrsim 10^5 - \gtrsim 10^6$ K. We use a wide prior on the IGM/CGM temperature to (i) be able to represent blended/complex absorption systems that could have been modelled by a combination of Voigt profiles as a single Voigt Profile. (ii) Keep our method simple. In the future work, we will modify the algorithm so that it can better model blended systems and so reliably estimate the temperature of the absorbing environment.

Potential applications of our catalogue include (i) finding targets for high-resolution follow-up of complex C IV systems (e.g. Galbiati et al. 2023), (ii) cross-matching with galaxy catalogues to find the properties of the galactic CGM within which our C IV absorbers lie (Gontcho et. al. 2018), and (iii) cross-matching with a damped Ly α catalogue to investigate the relationship between the highly ionized carbon and neutral hydrogen in the CGM (Danforth C. W. et. al. 2008).

Finally, the statistical properties of our catalogue can be computed and compared to the outputs of cosmological simulations to test and improve models for galactic feedba (Bird et. al 2017).

Our technique can also be applied to later, larger quasar catalogues such as those from the SDSS DR16 and the upcoming DESI quasar survey.

ACKNOWLEDGEMENTS

RM was supported by Higher Education Emergency Relief Funds. RM thanks Fred Hamann for supporting him for part of this work from NSF grant number AST-1911066. RM also appreciates Farhanul Hasan's time for his discussions about various topics throughout this project. SB was supported by NASA ATP 80NSSC22K1897. We used the HPCC cluster at UC Riverside and AWS credits provided under an amazon machine learning research award. MH was supported by NASA FINESST grant number 80NSSC21K1840. KLC acknowledges partial support from NSF AST-1615296. Funding for the SDSS and SDSS-II has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Science Foundation, the U.S. Department of Energy, the National Aeronautics and Space Administration, the Japanese Monbukagakusho, the Max Planck Society, and the Higher Education Funding Council for England. The SDSS Web Site is http://www.sdss.org/.

The SDSS is managed by the Astrophysical Research Consortium for the participating institutions. The participating institutions are the American Museum of Natural History, Astrophysical Institute Potsdam, University of Basel, University of Cambridge, Case Western Reserve University, University of Chicago, Drexel University, Fermilab, the Institute for Advanced Study, the Japan Participation Group, Johns Hopkins University, the Joint Institute for Nuclear Astrophysics, the Kavli Institute for Particle Astrophysics and Cosmology, the Korean Scientist Group, the Chinese Academy of Sciences (LAMOST), Los Alamos National Laboratory, the Max-Planck-Institute for Astronomy (MPIA), the Max-Planck-Institute for Astrophysics (MPA), New Mexico State University, Ohio State University, University of Pittsburgh, University of Portsmouth, Princeton University, the United States Naval Observatory, and the University of Washington.

7 DATA AVAILABILITY

All of our codes are available publicly in GitHub[17] and our final catalogue can be found in Zenodo. [18]

REFERENCES

Abazajian K. N. et al., 2009, ApJS, 182, 543
Adelberger K. L., Shapley A. E., Steidel C. C., Pettini M., Erb D. K., Reddy N. A., 2005, ApJ, 629, L636
Ahumada R. et al., 2020, ApJS, 249, 3
Alam S. et al., 2015, ApJS, 219, 12
Appleby S., Davé R., Sorini D., Cui W., Christiansen J., 2023, MNRAS, 519, 5514
Barlow T. A., Tytler D., 1998, AJ, 115, 1725
Becker G. D., Rauch M., Sargent W. L. W., 2009, ApJ, 698, L1010
Bird S., Rubin K. H. R., Suresh J., Hernquist L., 2016, MNRAS, 462, 307
Bird S., Garnett R., Ho S., 2017, MNRAS, 466, 2111
Boksenberg A., Sargent W. L. W., 2015, ApJS, 218, 7
Bordoloi R. et al., 2014, ApJ, 796, L136
Burchett J. N. et al., 2016, ApJ, 832, L124
Burchett J. N. et al., 2015, ApJ, 815, L91
Chen Z.-F., Qin Y.-P., Pan C.-J., Huang W.-R., Qin M., Wu H.-N., 2014, ApJS, 210, 7
Chen H.-W., Lanzetta K. M., Webb J. K., 2001, ApJ, 556, L158
Cheng T.-Y., Cooke R. J., Rudie G., 2022, MNRAS, 517, 755
Guo Z., Martini P., 2019, ApJ, 879, 72
Gontcho A Gontcho S., Miralda-Escudé J., Font-Ribera A., Blomqvist M., Busca N. G., Rich J., 2018, MNRAS, 480, 610
Churchill C. W., 2020, Cosmological absorption line spectroscopy
Codoreanu A., Ryan-Weber E. V., García L. Á., Crighton N. H. M., Becker G., Pettini M., Madau P., Bram V., 2018, MNRAS, 481, 4940
Cooksey K. L., Thom C., Prochaska J. X., Chen H.-W., 2010, ApJ, 708, L868
Cooksey K. L., Kao M. M., Simcoe R. A., O'Meara J. M., Prochaska J. X., 2013, ApJ, 763, L37
Cooper T. J., Simcoe R. A., Cooksey K. L., Bordoloi R., Miller D. R., Furesz G., Turner M. L., Eduardo B., 2019, ApJ, 882, L77
Danforth C. W., Shull J. M., 2008, ApJ, 679, L194
Davies R. L., et al., 2023, MNRAS, 521, 314
DESI Collaboration, Aghamousa A., Aguilar J., Ahlen S., Alam S., Allen L. E., Allende Prieto C., et al., 2016, preprint (arXiv:1611.00036)
D'Odorico V., Calura F., Cristiani S., Viel M., 2010, MNRAS, 401, 2715

[17] https://github.com/rezamonadi/GaussianProcessCIV
[18] https://doi.org/10.5281/zenodo.7872725

D'Odorico V. et al., 2013, MNRAS, 435, 1198
Doughty C. C., Finlator K. M., 2023, MNRAS, 518, 4159
Draine B. T., 2011, Physics of the interstellar and intergalactic medium. Princeton University Press
Eisenstein D. J. et al., 2011, AJ, 142, 72
Ellison S. L., Songaila A., Schaye J., Pettini M., 2000, AJ, 120, 1175
Fauber L., Ho M.-F., Bird S., Shelton C. R., Garnett R., Korde I., 2020, MNRAS, 498, 5227
Finlator K., Oppenheimer B. D., Davé R., Zackrisson E., Thompson R., Huang S., 2016, MNRAS, 459, 2299
Finlator K., Doughty C., Cai Z., Díaz G., 2020, MNRAS, 493, 3223
Galbiati M. et al., 2023, MNRAS, 524, 3474
Garnett R., Ho S., Bird S., Schneider J., 2017, MNRAS, 472, 1850
Haehnelt M. G., Steinmetz M., Rauch M., 1996, ApJL, 465, L95
Hamann F. et al., 2017, MNRAS, 464, 3431
Hasan F. et al., 2020, ApJ, 904, L44
Hasan F., Churchill C. W., Stemock B., Nielsen N. M., Kacprzak G. G., Croom M., Murphy M. T., 2022, ApJ, 924, L12
Ho M.-F., Bird S., Garnett R., 2020, MNRAS, 496, 5436
Ho M.-F., Bird S., Garnett R., 2021, MNRAS, 507, 704
Karaçaylı N. G. et al., 2023, preprint (arXiv:2306.06316)
Monadi R., Bird S., 2022, MNRAS, 511, 3501
Oppenheimer B. et al., 2019, BAAS, 51, 280
Parks D., Prochaska J. X., Dong S., Cai Z., 2018, MNRAS, 476, 1151
Péroux C., Howk J. C., 2020, ARA&A, 58, 363
Petitjean P., Bergeron J., 1994, A&A, 283, 759
Rasmussen C. E., Williams C. K. I., 2006, Gaussian Processes for Machine Learning. MIT Press, Cambridge
Rauch M., Sargent W. L. W., Womble D. S., Barlow T. A., 1996, ApJL, 467, L5
Ross N. P. et al., 2012, ApJS, 199, 3
Rubin K. H. R. et al., 2015, ApJ, 808, L38
Ryan-Weber E. V., Pettini M., Madau P., Zych B. J., 2009, MNRAS, 395, 1476
Ross N. P. et al., 2012, ApJS, 199, 3
Sargent W. L. W., Boksenberg A., Steidel C. C., 1988, ApJS, 68, 539
Savage B. D., Sembach K. R., 1991, ApJ, 379, L245
Scannapieco E. et al., 2006, MNRAS, 365, 615
Shen Y. et al., 2011, ApJS, 194, 45
Shull J. M., Danforth C. W., Tilton E. M., 2014, ApJ, 796, L49
Simcoe R. A. et al., 2011, ApJ, 743, L21
Simcoe R. A., 2011, ApJ, 738, L159
Songaila A., 2005, AJ, 130, 1996
Tennyson J., 2019, Astronomical Spectroscopy: An Introduction to the Atomic, Molecular Physics of Astronomical Spectroscopy. World Scientific, London
Tie S. S., Hennawi J. F., Kakiichi K., Bosman S. E. I., 2022, MNRAS, 515, 3656
Tumlinson J., Peeples M. S., Werk J. K., 2017, ARA&A, 55, 389
Wang B. et al., 2022, ApJS, 259, 28
Xia I., Ge J., Willis K., Zhao Y., 2022, MNRAS, 517, 4902
Yang L. et al., 2022, ApJ, 935, L121
Zhao Y., Ge J., Yuan X., Zhao T., Wang C., Li X., 2019, MNRAS, 487, 801
Zhu G., Ménard B., 2013, ApJ, 770, L130

APPENDIX A: REST EQUIVALENT WIDTH ESTIMATES

As a consistency check, we compared the rest equivalent width from the maximum a posteriori values of our model fit, using equation (30), to the rest equivalent width from integrating the flux around the absorber, as in the validation phase. Fig. A1 shows the difference between the two rest equivalent width estimates, normalized by the error estimate. Fig. A1 shows a symmetric unit Gaussian distribution centred at zero, demonstrating that our model parameters are both approximately unbiased and have well-calibrated error estimates.

In the validation phase (see Section 4.4), we calculated the rest equivalent width by integrating the flux around the absorber, in order to compare to the rest equivalent widths from the PM catalogue. However, we prefer to estimate rest equivalent widths for our SDSS DR12 catalogue directly from our maximum a posteriori model parameters, as these are less sensitive to noisy pixels in the integration range.

Table A1 shows the number of candidate absorbers for the single-line absorber model in SDSS DR12. Note that our training set does not label these absorbers and so we have not validated the potential detections.

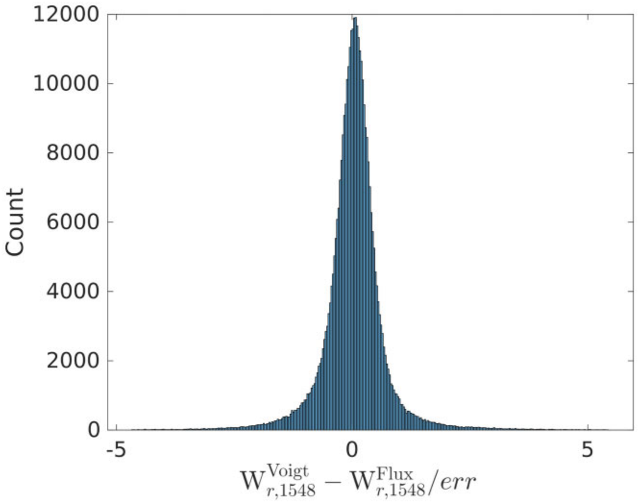


Figure A1. The difference between the two rest equivalent width estimates for the 1548 Å line explained in the text. These are using the maximum a posteriori model parameters and integrating the flux around the detected C IV absorber. Differences are normalized by the expected error from the model parameter posteriors, and show the expected Gaussian distribution.

Table A1. Table of probabilities P(M_S): first column shows the number of single-line absorbers. The 2nd through 4th columns show the number of single absorbers with probabilities > 65 per cent, 85 per cent, and 95 per cent, respectively.

C IV	P(M_S) > 0.65	P(M_S) > 0.85	P(M_S) > 0.95
0	155905 (84.0 per cent)	162533 (87.6 per cent)	166675 (89.9 per cent)
1	25441 (13.7 per cent)	19159 (10.3 per cent)	15366 (8.3 per cent)
2	3210 (1.7 per cent)	2914 (1.6 per cent)	2626 (1.4 per cent)
3	675 (0.4 per cent)	637 (0.3 per cent)	583 (0.3 per cent)
4	161 (0.09 per cent)	152 (0.08 per cent)	147 (0.08 per cent)
5	29 (0.02 per cent)	26 (0.01 per cent)	24 (0.01 per cent)
6	4 (0.00 per cent)	4 (0.00 per cent)	4 (0.00 per cent)

This paper has been typeset from a TEX/LATEX file prepared by the author.