









Towards an automatic approach to modelling the circumgalactic medium: new tools for mock making and fitting of metal profiles in large surveys

Alessia Longobardi ^{1,2}★ Matteo Fossati ^{1,2} Michele Fumagalli ^{1,3} Bhaskar Agarwal ^{1,4},
 Emma Lofthouse^{1,2}, Marta Galbiati ¹, Rajeshwari Dutta ^{1,2}, Trystyn A. M. Berg ^{1,2}
 and Louise A. Welsh ^{1,2}

¹Dipartimento di Fisica G. Occhialini, Università degli Studi di Milano-Bicocca, Piazza della Scienza 3, I-20126 Milano, Italy

²INAF – Osservatorio Astronomico di Brera, via Bianchi 46, I-23087 Merate (LC), Italy

³INAF – Osservatorio Astronomico di Trieste, via G. B. Tiepolo 11, I-34143 Trieste, Italy

⁴Cineca, Via Magnanelli, 6/3, I-40033 Casalecchio di Reno (BO), Italy

Accepted 2023 July 3. Received 2023 June 20; in original form 2023 May 4

ABSTRACT

We present two new tools for studying and modelling metal absorption lines in the circumgalactic medium. The first tool, dubbed ‘NMF Profile Maker’ (NMF–PM), uses a non-negative matrix factorization (NMF) method and provides a robust means to generate large libraries of realistic metal absorption profiles. The method is trained and tested on 650 unsaturated metal absorbers in the redshift interval $z = 0.9\text{--}4.2$ with column densities in the range of $11.2 \leq \log(N/\text{cm}^{-2}) \leq 16.3$, obtained from high-resolution ($R > 4000$) and high-signal-to-noise ratio ($S/N \geq 10$) quasar spectroscopy. To avoid spurious features, we train on infinite S/N Voigt models of the observed line profiles derived using the code ‘Monte-Carlo Absorption Line Fitter’ (MC–ALF), a novel automatic Bayesian fitting code that is the second tool we present in this work. MC–ALF is a Monte-Carlo code based on nested sampling that, without the need for any prior guess or human intervention, can decompose metal lines into individual Voigt components. Both MC–ALF and NMF–PM are made publicly available to allow the community to produce large libraries of synthetic metal profiles and to reconstruct Voigt models of absorption lines in an automatic fashion. Both tools contribute to the scientific effort of simulating and analysing metal absorbers in very large spectroscopic surveys of quasars like the ongoing Dark Energy Spectroscopic Instrument, the 4-m Multi-Object Spectroscopic Telescope, and the WHT Enhanced Area Velocity Explorer surveys.

Key words: Machine Learning – Data Methods – Astrophysics – Astrophysics of Galaxies – galaxies: evolution.

1 INTRODUCTION

In the current cold dark matter paradigm of structures’ formation and evolution, the modelling of baryons is representing a challenge due to the large array of physical processes affecting this baryonic component (e.g. Vogelsberger et al. 2020, and references therein). As these processes (e.g. gas cooling, star formation, stellar/active galactic nucleus feedback, and their interplay with gravity) combine to shape the morphological and physical properties of the galaxies as we observe them today, the detailed study of the gas environment has become a priority in the field of galaxies’ evolution.

The circumgalactic medium (CGM), i.e. the baryonic component that connects the galaxies’ interstellar medium with their large-scale environment (intergalactic medium), plays a key role in our understanding of how galaxies evolve, by allowing us to follow the ‘baryon cycle’ in which gas cycles into, out of, and through galaxies. Observations of this component both in emission and absorption allow us to trace the thermal and chemical evolution of

diffuse gas in the Universe as a function of redshift (e.g. Tumlinson, Peebles & Werk 2017, and references therein). In the last two decades, an increasing effort has been put into the study of the CGM physical properties (temperature, density, and metallicity) through the analysis of absorption lines in spectra of background sources like bright quasars, such as intervening metal absorbers, broad absorption line (BAL) quasars, HI selected systems like damped Ly α absorbers (DLAs), or Lyman limit systems (LLSs). Studies across these different populations enable access to a wide range of gas column densities, i.e. $\log(N/\text{cm}^{-2}) \geq 11$ (e.g. Sargent, Steidel & Boksenberg 1989; Schaye et al. 2000; Prochaska, Herbert-Fort & Wolfe 2005; Simcoe et al. 2011; Rafelski et al. 2012; Fumagalli et al. 2016; D’Odorico et al. 2022), which in turn trace varying degrees of overdensities.

With the advent of the Sloan Digital Sky Survey (SDSS; York et al. 2000) and the compilation of the SDSS quasar catalogues (final release by Lyke et al. 2020), astronomers now have access to just under a million spectroscopically confirmed quasars to statistically assess the properties of absorption line systems and trace the distribution and physical properties of the gaseous component around galaxies across the Universe (e.g. Noterdaeme et al. 2008; Prochaska, O’Meara &

* E-mail: alessia.longobardi@unimib.it

Worseck 2010; Lan, Ménard & Zhu 2014; Garnett et al. 2017; Anand, Nelson & Kauffmann 2021; Anand, Kauffmann & Nelson 2022). These statistical studies have further enabled follow-up observations with high-resolution spectrographs on 8- and 10-m class telescopes that opened the possibility of identifying characteristic features in absorption systems while simultaneously allowing us to map the correlations between the galaxies and the ambient gaseous haloes (e.g. Werk et al. 2016; Prochaska et al. 2017; Fossati et al. 2019; Mackenzie et al. 2019; Rudie et al. 2019; Dutta et al. 2020; Lofthouse et al. 2020, 2023; Wilde et al. 2021).

A boost to these studies is imminent thanks to large surveys at 4-m class telescopes, like the Dark Energy Spectroscopic Instrument (DESI; DESI Collaboration et al. 2016) survey, the 4-m Multi-Object Spectroscopic Telescope (*4MOST*; de Jong 2019) survey, and the WHT Enhanced Area Velocity Explorer (WEAVE, Dalton et al. 2012; Jin et al. 2023) survey. As the size and quality of the data increase, novel, fast, and efficient approaches for identifying and characterizing absorption lines are needed. This is why an increasing effort has been put into proposing efficient approaches to: (i) identify the different classes of quasars absorbers like metals (Cooksey et al. 2013; Zhu & Ménard 2013; Zou et al. 2021), BALs (Guo & Martini 2019), DLAs (Garnett et al. 2017), and LLSs (Fumagalli, Fotopoulou & Thomson 2020) and (ii) to derive models of the different line profiles that can accurately reproduce their behaviour in terms of maximum likelihood (e.g. Carswell & Webb 2014) or Bayesian estimates (Liang & Kravtsov 2017). Similarly, as samples increase, systematic errors overcome by far statistical uncertainties (e.g. Fumagalli et al. 2020). Tools are therefore required to create high-quality mocks to train and validate the science pipelines used to analyse the spectra and extract physical information on the absorption lines.

The objective of this paper is to contribute to the tools available in the field to tackle these challenges. Specifically, we present a new code, Monte-Carlo Absorption Line Fitter (MC-ALF), which provides an automatic reconstruction of the shape of line profiles and extracts posterior distributions of the relevant physical parameters (such as the number of components, column density, and Doppler parameter of each component). Moreover, we introduce a new method, called NMF Profile Maker (NMF-PM) that generates synthetic but realistic-looking line profiles following a given distribution of desired line widths. Combined, these new codes provide useful pre- and post-processing tools to aid the science exploitation of future wide-field surveys, contributing both to the simulation of quasar spectra with realistic absorption lines, and to the higher level analysis of downstream data products. In particular, our codes are tailored to the electronic transition lines of metal species (both low- and moderate-ion transitions, with ionization potential $IP \lesssim 30\text{--}40$ and $IP \sim 40\text{--}100$ eV, respectively) that can be used to study the multiphase nature of the CGM.

Our tools rely on advanced numerical techniques. Specifically, for MC-ALF, we adopt a Bayesian nested sampling approach to the line profile fitting. This method efficiently explores the full parameter space by slicing it into subvolumes and fitting nested N^{th} dimensional contours to identify the regions with a strong likelihood gradient where accurate sampling is required. For NMF-PM, instead, we rely on the non-negative matrix factorization (NMF), which is a subclass of multivariate analysis techniques often associated with pattern recognition and blind source separation (Lee & Seung 2000), also used within the astronomical community (e.g. Zhu & Ménard 2013; Hurley et al. 2014; Ren et al. 2018). Offering us a well-established statistical framework for carrying out the representation of positive and continuous signals, the NMF is an ideal algorithm

for summarizing the information contained in a large data set of metal absorbers and for carrying out robust modelling and prediction making.

One of the largest difficulties in the development of synthetic metals' profiles via machine learning (ML) methods is to have a sample of training data, which satisfies both quality and quantity requirements. However, as part of three large and complete galaxies' surveys in quasar fields – the MUSE Analysis of Gas around Galaxies (MAGG, Dutta et al. 2020; Lofthouse et al. 2020) survey, the Quasar Sightline and Galaxy Evolution (QSAGE, Bielby et al. 2019; Dutta et al. 2021) survey, and the MUSE Ultra Deep Field (MUDF; Fossati et al. 2019) – our team has assembled a library of ~ 700 metal absorption lines representative of moderately to highly overdense gas. Thanks to this data set, we are now able to train algorithms to generate metal profiles in quasar spectra with ML across a large variety of column densities, redshifts, and line widths.

The paper is structured as follows: In Section 2, we describe the spectroscopic surveys that provided the data to compile our library of metal systems. In Section 3, we present and test the Voigt fitting algorithms at the basis of MC-ALF, which we used to produce a set of unsaturated metals profiles with infinite signal-to-noise ratio (S/N). In Section 4, we introduce instead NMF-PM, presenting the NMF formulation useful to produce synthetic metal profiles. The latter are presented in Section 5 and are discussed in the context of the upcoming large surveys of background quasars. A summary is presented in Section 6.

2 LIBRARY OF ABSORPTION LINE SYSTEMS

2.1 Spectroscopic surveys adopted

For the purpose of developing, training, and testing our codes, we assemble a library of moderate-to-high S/N spectra of absorption lines in different ionization stages and at different redshifts. Next, we provide a brief description of the compilation of the spectroscopic campaigns that form the data set used in this work. We refer the reader to the listed references for additional details on data quality, and reduction techniques.

2.1.1 The MAGG survey

The MAGG survey (Lofthouse et al. 2020) is based upon a MUSE Large Programme (ID 197.A-0384; PI Fumagalli) of 28 quasar fields at redshift $3.2 \leq z \leq 4.5$ for which $S/N \geq 10$ and medium- (4000–10 000) or high-resolution (20 000–50 000) spectroscopy is available. High-resolution spectroscopy is a compilation of data from the Ultraviolet and Visual Echelle Spectrograph (UVES; Dekker et al. 2000), the High-Resolution Echelle Spectrometer (HIRES; Vogt et al. 1994), and the Magellan Inamori Kyocera Echelle instruments (Bernstein et al. 2003), while moderate resolution spectroscopy is from ESI (Sheinis et al. 2002) and X-SHOOTER (Spanò et al. 2006; Vernet et al. 2011). A total of 62 individual spectra were assembled for the 28 quasars.

Instrument-specific pipelines were used to carry out the data reduction, which included bias subtraction, flat-fielding, dark subtraction (where applicable), and wavelength calibration. Once one-dimensional (1D) spectra were extracted, and eventually combined if multiple exposures were present, the spectra were further flux-calibrated and continuum-normalized, when applicable (details are provided in Lofthouse et al. 2020).

The MAGG surveys led to the identification of a large variety of metals (low- and moderate-ions) associated with LLSs (Lofthouse

et al. 2023) or selected to be C IV and Si IV doublets at $3.0 \leq z \leq 4.2$ (Galbiati et al. 2023), and Mg II absorbers at $0.9 \leq z \leq 1.4$ (Dutta et al. 2020).

2.1.2 The QSAGE survey

The QSAGE survey (Bielby et al. 2019) is a *Hubble Space Telescope* (HST) Wide-Field Camera 3 survey of 12 quasar fields at redshift $1.2 \leq z \leq 2.4$ imaged in the near-infrared (90 per cent complete down to F140W ≈ 26 mag) and with HST Space Telescope Imaging Spectrograph (STIS; Kimble et al. 1998) high-resolution ($\approx 30\,000$) archival ultraviolet spectra. As for the MAGG survey, the QSAGE quasar fields were supplemented by additional spectroscopy. Additional medium-to-high-resolution ($\approx 12\,000$ – $24\,000$) far-ultraviolet and near-ultraviolet data were taken with the HST Cosmic Origins Spectrograph (COS, Osterman et al. 2011; Green et al. 2012) as part of the COS Absorption Survey of Baryon Harbors (e.g. Tripp et al. 2011). Together with the STIS data, COS data were reduced using the instrument-specific pipelines, which carried out overscan and bias subtraction, cosmic rays rejection, dark subtraction, flat fielding, spectroscopic wavelength, and flux calibration. Finally, supplementary optical high-resolution data ($\approx 40\,000$) were a compilation of HIRES and UVES spectra retrieved from the Keck Observatory Database of Ionized Absorption toward Quasars (O’Meara et al. 2015, 2017) and from the Spectral Quasar Absorption Database (Murphy et al. 2019), respectively (we refer the reader to Dutta et al. 2021 for further details on data reduction). The QSAGE survey provides us with Mg II systems across $z \approx 0.1$ – 1.3 and C IV systems across $z \approx 0.1$ – 2.4 as identified in $S/N \geq 10$ spectra by Dutta et al. (2021).

2.1.3 The MUDF survey

The MUDF survey is a MUSE large program (ID 1100.A-0528; PI Fumagalli) targeting a region on the sky containing two bright quasars at $z \approx 3.2$ (Lusso et al. 2019). As a part of the MUDF survey, the MUSE observations were complemented by ancillary UVES high-resolution spectroscopy (D’Odorico, Petitjean & Cristiani 2002; Fossati et al. 2019). As for this work, we make use of the data set relative to the brighter quasar, which provides us with $S/N \approx 25$ per pixel. Data were reduced with the UVES pipeline following a standard reduction process. The reduced spectra were then reformatted with a custom script and input to the ESPRESSO Data Analysis Software (Cupani et al. 2016) for the final operations of co-addition and continuum fitting. Further details on the data acquisition and data reduction of the MUDF data can be found in Fossati et al. (2019). The MUDF provides us with low- and moderate-ion absorbers in the redshift range $z \approx 0.9$ – 3.2 .

2.2 Statistical properties of the absorption line library

MAGG, QSAGE, and MUDF led to a total sample of 688 metal absorption lines. As these data and their fits set the basis for our NMF algorithm with which we aim at tracing and reproducing the lines’ intrinsic shapes (see Section 4), we restricted the sample to only unsaturated metal lines by excluding those profiles for which the continuum normalized flux reaches zero. For each ion, we then selected the strongest transition (highest oscillator strength). However, in case the corresponding profile was saturated, we then selected the weakest transition, subject to the constraint that it was not saturated as well. Less than 1 per cent of our library consists of

medium-to-low-resolution spectra (e.g. ESI and XSHOOTER data) of single transition lines for which hidden saturation may affect the associated velocity profiles. The remaining sample of medium-to-low-resolution absorbers consists of doublets and multiple transitions of the same ion for which the effect of hidden saturation is mitigated by MC–ALF, which fits together transitions with different oscillator strengths belonging to the same ion (see Section 3.4). Finally, since we decouple the column density from the line profile, hidden saturation should not affect building the profile generation library – provided that the shape of the line is not distorted in the core as in the case of evident saturation. This resulted in 650 profiles of which we show a small sample and the relative ion contribution in Fig. 1.

The moderate-ions (447 profiles) are dominated by C IV and Si IV absorbers, while the majority of the low-ions (203 profiles) is represented by Mg II absorbers. We also compare the distributions of redshifts, column densities, and ΔV_{90} values, i.e. the velocity range within which the velocity distribution encompasses 90 per cent of the optical depth of the line, relative to the low- and moderate-ions classes (cyan and red sample in Fig. 2). On average the low-ion distribution peaks at lower redshifts, $z_l = 2.4 \pm 1.1$, than the distribution traced by the moderate-ions, $z_m = 3.4 \pm 0.4$. On the other hand, the column density distributions peak at similar values although they are characterized by a different dispersion, i.e. $\log(N/\text{cm}^{-2})_l = 13.1 \pm 0.9$ and $\log(N/\text{cm}^{-2})_m = 13.4 \pm 0.7$. Finally, when the ΔV_{90} distribution is plotted separately for the low- and intermediate-ions, the highly ionized species can show broader line widths.

3 MC–ALF: THE MC–ALF CODE

To extract the wealth of information on the kinematic, chemical, and ionization conditions of the gas probed by absorption line systems, it becomes necessary to model the spectral features. Albeit non-parametric techniques exist (e.g. apparent optical depth; Savage & Sembach 1991), Voigt fitting has become the main modelling technique to extract the lines’ physical properties. Decomposing absorption line profiles into Voigt components can be an expensive task, particularly because of the degree of subjectivity in setting initial conditions. Alternative approaches to Voigt profile fitting that used χ^2 -based codes (e.g. Fontana & Ballester 1995; Davé et al. 1997; Carswell & Webb 2014; Cooke et al. 2014; Krogager 2018) have been sought, e.g. by using Bayesian techniques. These techniques have the advantage of being relatively less computationally expensive in cases where multiple absorption component fitting is needed. Moreover, by sampling the posterior distribution of the parameters’ values their uncertainties and degeneracies can be better constrained.

Within this framework, one example is BayesVP (Liang & Kravtsov 2017) which models Voigt profiles and generates posterior distributions for the column density, Doppler parameter, and redshifts of the corresponding absorber. However, it is based on an affine-invariant Markov chain Monte Carlo (MCMC) sampler that does not easily converge in a high-dimensional parameter space, thus resulting in computationally expensive runs when the initial conditions or the number of free parameters are not known. To obviate this problem, one can resort to nested sampling (Skilling 2006) which provides complete statistical information and makes it possible to efficiently carry out model comparison via the Bayesian evidence.

In this work, we present the technical details and quality assurance tests of a new Bayesian fitting code first introduced in Fossati et al. (2019) and dubbed the MC–ALF. MC–ALF has four innovative features compared to other absorption line fitting codes: (i) it requires minimal input from the user as no initial conditions are given but only

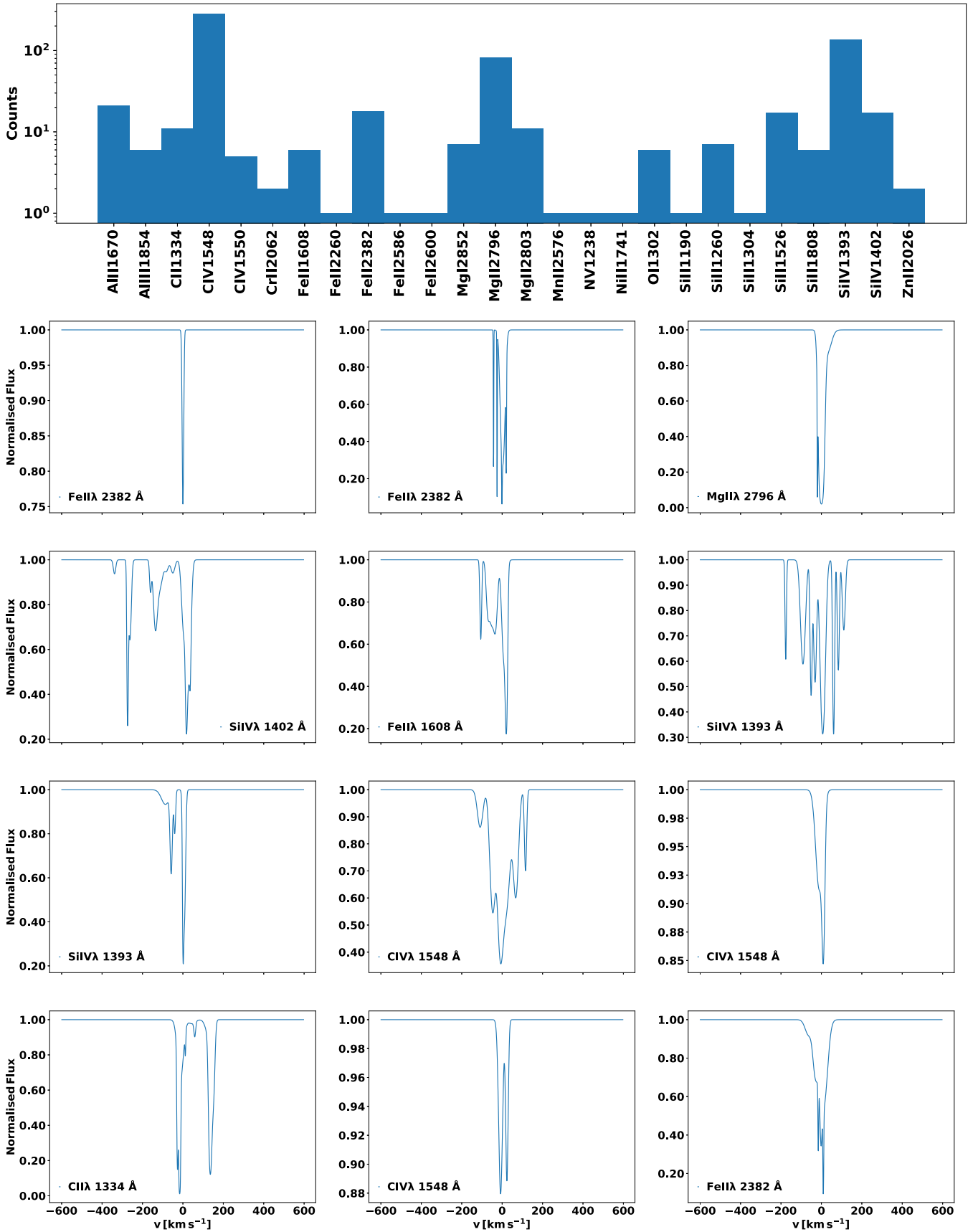


Figure 1. Top panel: 650 absorbers in our sample – ~ 70 per cent are moderate-ions, the remaining ~ 30 per cent is represented by low-ions. Bottom panels: Subsample of data fitted with MC-ALF. The sample gathers a large variety of profiles in terms of the number of components in each profile and ΔV_{90} distribution.

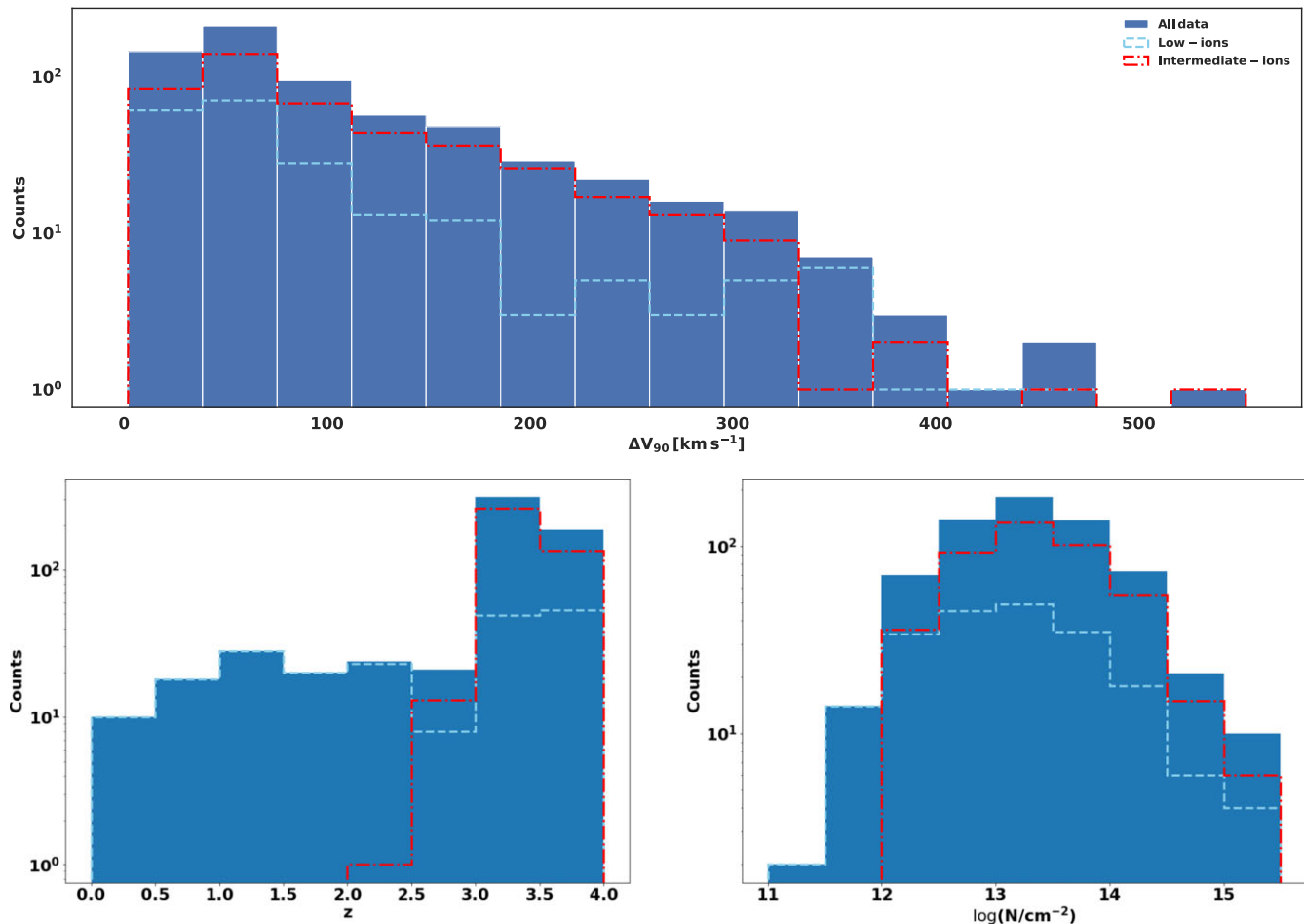


Figure 2. Top panel: Distribution of ΔV_{90} values in our sample (dark blue). Dashed cyan and dot-dashed red histograms show the ΔV_{90} values for the low- and moderate-ions, respectively. Bottom panels: Same as top panel; however, this time the histograms are relative to the redshift (left-hand panel) and column density (right-hand panel) distributions.

the allowed range of the parameters is required; (ii) using Bayesian statistics the final result provides the full posterior distribution for each parameter and their covariance matrix, leading to an optimal statistical description of the data; (iii) it samples the multidimensional likelihood space using POLYCHORD (Handley, Hobson & Lasenby 2015), a nested sampling algorithm that has the best performance for high-dimensional parameter spaces with multiple degeneracies between parameters, as it is the case of the multicomponent Voigt parametrization of complex absorption profiles; (iv) it is naturally adaptive, shown to accurately retrieve all the information of both high- and low-resolution profiles with execution times that scale with the complexity of the profile (typically related to the instrument resolution and data S/N . See Sections 3.4 and 3.6 for more details). All these characteristics make MC-ALF ideally suited to study in an automatic fashion big-data samples from large spectroscopic surveys where the combination of moderate resolution and S/N reduces the need for complex (and expensive to compute) absorption models.

3.1 Formalism for Voigt profile fitting

The analytic model at the basis of MC-ALF is the canonical combination of multiple Voigt functions that are used to describe line profiles of any complexity. The absorption line arising from a transition i of an ion can be described by the optical depth of the

transition, $\tau_i(\nu)$, which is determined by the column density of the ion, N , along with a set of atomic parameters describing the line strength, f_i , the damping constant, Γ_i , and the resonance frequency, ν_i , i.e.

$$\tau_i(\nu) = N s_i \phi_i(\nu), \quad (1)$$

where s_i is the frequency-integrated absorption cross-section given by

$$s_i = \frac{\pi e^2}{m_e c} f_i, \quad (2)$$

with e the electron charge, m_e the electron mass, and c the speed of light. Finally, the frequency-dependent line profile for a single component is

$$\phi_i(\nu) = \frac{H(u_i, a_i)}{\Delta \nu_i \sqrt{\pi}}, \quad (3)$$

with the Voigt function

$$H(u_i, a_i) = \frac{a_i}{\pi} \int_{-\infty}^{+\infty} dy \frac{\exp(-y^2)}{(u_i - y)^2 + a_i^2}, \quad (4)$$

where $a_i = \Gamma_i/4\pi \Delta \nu_i$, $u_i = (\nu - \nu_i)/\Delta \nu_i$ is the re-scaled frequency, $y = \nu/b$ is the velocity in units of the Doppler parameter, $b = (2kT/m + \xi^2)^{1/2}$, which is given by the gas temperature, T , the element mass, m , and the turbulent velocity, ξ . Finally, $\Delta \nu_i = \nu_{i,0} b/c$, with

$\nu_{i,0}$ the transition rest-frame frequency. Thus, the Voigt function is the convolution of the Gaussian line broadening due to thermal and turbulent motion with the Lorentzian contribution from natural line broadening and it can be separated from the normalization factors, N and s_i , intervening in Equation (1) to model the optical depth of the considered transition.

By summing over all the N_V Voigt components, the resulting transmitted flux, $I_i(\nu)$, of a background source with intensity, I_0 , is given as

$$I_i(\nu) = I_0 e^{-\sum_j^{N_V} \tau_{ij}(\nu)}. \quad (5)$$

Equation (5) defines our model and its free parameters are the column density, the Doppler parameter, the redshift, and the number of components, i.e. $\theta = \{N, b, z, N_V\}$.

3.2 Bayesian inference and posterior sampling

Within the Bayesian inference framework the posterior distribution, \mathcal{P} , is proportional to the product between the likelihood, \mathcal{L} and the prior probability distribution functions (PDFs), π , so that

$$\mathcal{P}(\theta) \propto \mathcal{L}(\theta) \times \pi(\theta) \quad (6)$$

is proportional to the product between the probability of observing the data given a specific set of parameters and their prior distributions. Specialized to our case, the data are represented by the measured flux, F , in a spectral interval. Then we can write

$$\mathcal{L}(\theta) = \prod_i l_i(\theta|F_i), \quad (7)$$

where l_i is the likelihood relative to an individual pixel i , i.e.

$$l_i = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left[-\frac{\bar{F}_i(\theta) - F_i^2}{2\sigma_i^2}\right], \quad (8)$$

with σ_i the flux error.

The likelihood space is sampled via the nested sampling algorithm POLYCHORD. The sampling starts with a large number of live points (n_{live}) within a region of the parameter space sampled by the prior distribution. These points are sequentially updated so that those with the smallest value of the posterior density are eliminated (termed dead points) and are replaced by a new live point, again drawn from the prior, whose likelihood is larger than that of the point that was discarded. To generate new points POLYCHORD uses the so-called slice sampling where new live points are generated by taking a random slice through the parameter space that includes the current live point, and randomly generating new points until one with a higher likelihood is found. The process is then repeated with the new point and a slice in a new random direction, for a user-defined number of repetitions (n_{repeat}). The length of this chain of repetitions should be large enough so that the final live point is decorrelated from the start point.

3.3 Model comparison

Metal absorbers can be characterized by complex profiles where line blending can make it difficult to retrieve the number of Voigt components that better define the observed profile. This, together with the fact that we aim at sampling a large space of parameters, yields the necessity of choosing between competing models. In turn, this capability obviates the need for the user to specify a set of initial conditions or strong priors for the parameters. In a pre-release version of MC-ALF code used by Fossati et al. (2019), the number of

Voigt components was kept as a fixed parameter at each fit iteration and multiple fits with an increasing number of components were performed to decide on the best decomposition model using the Akaike Information Criterion (Akaike 1974). To improve the code performance, a non-negligible aspect for deploying this code in large surveys, we have refactored the code to include the number of components in the likelihood calculation, so that a single fit can be performed keeping the number of components as a free parameter. Thus, the algorithm is terminated once the improvement in the likelihood is some small fraction of the currently calculated one. Moreover, this version has the added value of providing posterior distributions of the number of components which can be useful in the case of highly complex profiles.

3.4 MC-ALF configuration file

A MC-ALF configuration file has three main blocks: input, components, and psettings, with which the user defines the input information, the parameters for the components to be fitted, and the setting of the POLYCHORD solver through their attributes. In input, the main information the code requires are the spectral data to fit. This is an ascii table with three columns providing the wavelength in Å, the continuum normalized flux, and its error. There is no preferential order with which the columns must be organized as long as this information is provided in the `coldef` attribute. The user will then have to specify the transitions to fit (only atomic transitions belonging to the same ion can be fit together), following a naming convention that sees the ion name followed by its rest-wavelength in Å and separated by a white space.

Next, the user will provide the wavelength range (or disjoint ranges) to fit the data with Voigt components. These are described by their column density (N in log units of cm^{-2}), the Doppler parameter (b in km s^{-1}), and the redshift of the transition. The number of Voigt components is an additional free parameter in the fit and the user will specify the range to be explored via the `ncomp` attribute. Similarly, the range of b -parameter values to be fitted can be passed as `brange`. If required, it is possible to include a user-defined number of ‘filler’ Voigt profiles designed to describe absorption lines arising from blends of different ions at different redshifts in the wavelength range being fit and controlled by the `nfill` attribute. The range of column density, and b -parameter values for the ‘fillers’ is then passed via `Nrangefill` and `brangefill`, respectively. Note that, while the dynamic range of each of the free parameters can be specified in the code configuration file, reasonable default values are provided to the code.

Finally, the user can control the parameters of the POLYCHORD algorithm directly in the `psettings` block, defining the number of live points (`nlive`) and the number of slices (`num_repeats`) at each iteration, therefore, balancing execution time and the likelihood accuracy. A more detailed description of these parameters can be found in Handley et al. (2015). An example of an MC-ALF configuration file is shown in Fig. 3.

We note that the code’s upgrade of including the number of Voigt components as free parameters (see Section 3.3) has improved the execution time by a factor of 5–10 so that, in its default configuration (`ncomp = 1–15`, `nlive = 500`, and `num_repeats = 50`) MC-ALF takes ~ 1.3 total CPU hours to run on recent Intel processor and to model multiple-components, high-resolution spectra. As the models are expected to be simpler for lower resolution and lower S/N data, the code can be optimized for speed by reducing the interval of components to be considered and by reducing the `nlive` and `num_repeats` (for typical WEAVE-like data we set these to `ncomp`

```

#Can handle multiple comma-separated regions to fit
#Specfile, linelist and wavefit are mandatory
[input]
specfile = J020944.61+051713.6_UVES_spec.txt
linelist = CII 1334
wavefit = 6488,6494
coldef = Wave,Flux,Err
specres = 7,10

# ncomp can be a range, nfill is currently a fixed value
# If not specified assuming one component and zero fillers
[components]
ncomp = 1,5
nfill = 0
contval = 0.95,1.04
Nrange = 13,15
Nrangefill = 11,18
brange = 1,30
brangefill = 1,30

# Define directories, chaindir and plotdir are subdirs
# of outdir
[pathing]
datadir = /Input_dir/
outdir = /Output_dir/
chainfmt = pc_output_name
chaindir = pc_Fits/
plotdir = pc_Plots/

# Settings of the polychord solver
[pcsettings]
nlive = 2000
num_repeats = 350
precision_criterion = 0.001
feedback = 1
do_clustering = False
equals = True
read_resume = False
write_resume = False
write_live = False
write_dead = False
write_prior = False
posteriors = False
cluster_posteriors = False

```

Figure 3. Example of an MC-ALF configuration file. In this example, the ion to fit is a C II ion in a high-resolution spectrum as specified in the `linelist` and `specres` attributes in the input block. The range of the fitted parameters, i.e. the column density (N in log units), the Doppler parameter (b in km s^{-1}), and the redshift of the transition are specified in the components block. If they are not provided, the code assumes default values. The pcsetting block controls the parameters of the POLYCHORD algorithm (see text for more details).

= 1–3, `nlive` = 350, and `num_repeats` = 50, having the profiles fully analysed in tens of seconds on a single core machine).

3.5 MC-ALF output

When the fit has converged, the MC-ALF output is saved in an ascii file that will list the following columns for each posterior sample: the components' weight, the total evidence, the best likelihood, the fitted continuum, and spectral resolution (if they are free parameters), and the values of the fitted parameters, i.e. the column density, the redshift, and the b -parameter. As an example, Figs 4 and 5 show the model fits obtained with MC-ALF. The fit is run on a C II $\lambda 1334 \text{ \AA}$ absorber (high-resolution spectrum, 8 km s^{-1}) with a value for the column density $\log(N/\text{cm}^{-2}) = 14.3$. We show that the best model reproduces the data mainly with three Voigt components, although there is a non-zero probability that a four-component model can also be compatible with the data (Fig. 4). Moreover, the full posterior distribution is saved and can be used for further processing and analysis by the users. For example, the corner plot in Fig. 5 shows the

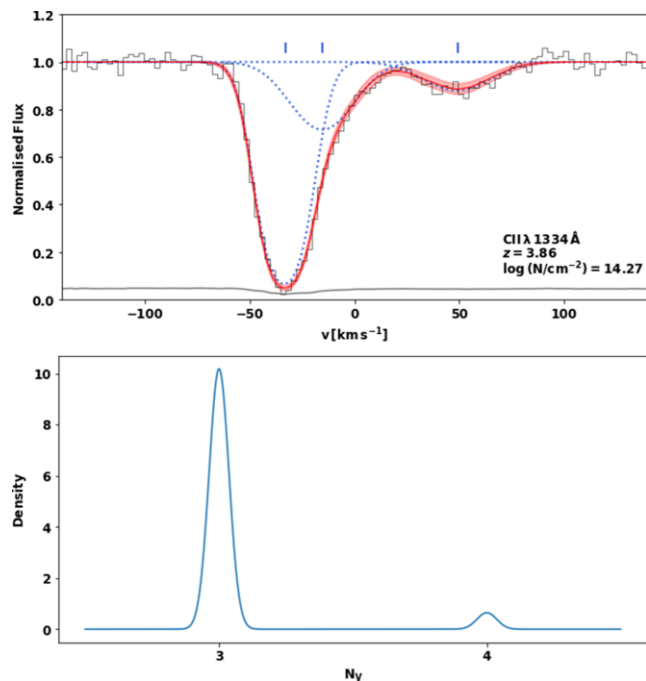


Figure 4. Top panel: The results of the fit procedure on a C II absorber. The solid red line is the median model of all the posterior samples, while the red shaded region represents the $\pm 1\sigma$ uncertainty on the model profile. The blue dotted lines represent individual Voigt components, centred at the velocities highlighted by the blue ticks. Bottom panel: The kernel density distribution of the number of Voigt components in the posterior samples. Nearly all the samples fit the data with three Voigt components. Note that the number of components is an integer in our fitting model.

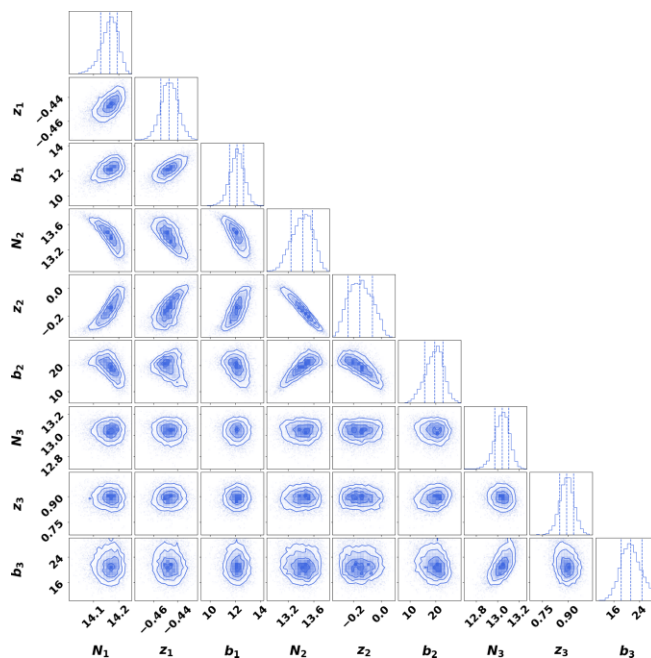


Figure 5. Corner plot of the posterior samples for the MC-ALF fits with three Voigt components of a C II absorber at $z \approx 3.86$. The contour panels show the posterior distribution of pairs of free parameters, while the histograms show the 1D posteriors of individual parameters. The dashed vertical lines correspond to the 16th, 50th, and 84th percentiles of the distributions, respectively. Units for b -parameters are km s^{-1} , column densities are in $\log(N/\text{cm}^{-2})$.

covariance and distribution of Voigt parameters for each component (N , z , and b).

3.6 Code validation and quality assurance tests

We test MC–ALF capabilities by analysing a set of five synthetic profiles mimicking Mg II, Si IV, and NV systems, characterized by different total column densities, $13.8 \leq \log(N/\text{cm}^{-2}) \leq 16.0$, and number of Voigt components, $6 \leq N_v \leq 15$, at different redshifts, and for which we have a priori knowledge of the Doppler parameter and column density values relative to each Voigt component (Fig. 6). The profiles have been created to reproduce the performance of UVES/VLT (Very Large Telescope) spectra, i.e. they are characterized by a resolution of 8 km s^{-1} with a pixel sampling of 2.5 km s^{-1} . Finally, a Poisson noise component is added in each profile, $\sigma = \sqrt{\sigma_{\text{source}}^2 + \sigma_{\text{sky}}^2}$, with σ_{source} so that the MC–ALF performances could be tested as a function of different S/N ratio (per pixel) with respect to the continuum of the background source, namely $S/N = 5, 10, 20, 30, 50, 100$, and 500 , and the sky noise, σ_{sky} , so that the continuum dominates by a factor 4 on the sky signal already at $S/N = 10$.

Next, we model these profiles with MC–ALF and the results are shown in Fig. 7 where we compare the probability densities of the input and retrieved distributions of b -parameters and column densities as a function of the different S/N of the input spectra. The Kolmogorov–Smirnov (KS) test p -value scores (given on top of the respective distributions) show that Doppler parameter estimates are more sensitive to the quality of the analysed spectra compared to that of the column density estimates. Nonetheless, MC–ALF is able to recover the total input distributions (p -value > 0.1) of both the b -parameters and column densities already at $S/N = 10$. At lower S/N ratios, some discrepancies are found when analysing individual components because it becomes increasingly difficult to accurately match, in a statistical sense, single input-versus-retrieved Voigt components (leading to p -value < 0.1). Hence, these discrepancies reflect a mismatch in components rather than inaccuracies in MC–ALF to recover values. We find that the retrieved fraction of the N_v components is on average $\langle f^{N_v} \rangle = 0.77, 0.83, 0.88$, and 0.88 for $S/N = 5, 10, 20$, and 30 , respectively, while at higher S/N , MC–ALF retrieves all the input components (except for single high-density components that MC–ALF may split in two; see below). Thus, in the low S/N regime, it is better to test the code capabilities relative to the total (integrated) values of the fitted parameters. When these are considered, MC–ALF successfully retrieves the input total distribution with a mean relative error of $\langle \delta^{\log N} \rangle = 0.12, 0.09, 0.11$, and 0.08 at $S/N = 5, 10, 20$, and 30 . In Section 6, we provide additional tests for thousands of simulated profiles in the low S/N , low-resolution regime.

Focusing on the $S/N = 500$ test, we compare the recovered values of b -parameters, column densities [in $\log(N/\text{cm}^{-2})$], and redshifts (the latter being converted to $\Delta V = c \frac{z_{\text{Ret}} - z_{\text{Input}}}{1 + z_{\text{Input}}}$, with c the speed of light) against their input values (Fig. 8). For the three distributions, we find mean relative errors of $\langle \delta \rangle = 0.4, 0.007$, and 0.62×10^{-6} . We note that the errors for the b -parameters and column densities are dominated by the errors associated with single Voigt components with $\log(N/\text{cm}^{-2}) > 15.0$. When these are excluded from our computation the relative errors for the two distributions drop to $\langle \delta \rangle = 0.03, 0.002$, respectively. These high-column density components are also responsible for MC–ALF to find in output one additional Voigt profile for the Mg II absorber with $\log(N/\text{cm}^{-2}) = 15.97$ (second-row panels in Fig. 6). This phenomenon is the result of the

decomposition of the single Voigt profile in a narrow component that best describes the high-optical depth regime and a broad component that best fits the wings. In Fig. 8, this spurious detection is responsible for the most discrepant Δb value.

Finally, in Fig. 8, we mark as red-empty dots the saturated Voigt components, i.e. with flux density levels reaching zero in the normalized spectra, for which the column density value may no longer be estimated with a few percent accuracy. We then additionally tested the performances of MC–ALF in presence of saturation. As before, the test is carried out on a UVES/VLT-like synthetic profile. The absorber is a single Voigt component of C II at $\lambda 1334 \text{ \AA}$ with a b -parameter fixed at $b = 15 \text{ km s}^{-1}$ and column density values $14 \leq \log(N/\text{cm}^{-2}) \leq 18.5$. The profiles, shown in Fig. 9 in the case of $S/N = 20$, are saturated for $\log(N/\text{cm}^{-2}) \geq 14.5$ and for $\log(N/\text{cm}^{-2}) = 18.5$ (an extreme value useful for testing) the damping wings of the Lorentzian become significant compared to those of the Gaussian contribution (see Equation 4).

The impact of saturation on the fits is shown in Fig. 10 as b versus $\log(N/\text{cm}^{-2})$ plot. In this figure, the blue dots trace the full posterior samples provided in output by MC–ALF and the dotted grey lines show the simulated b and column density values. When the line is not saturated, in our example for $\log(N/\text{cm}^{-2}) = 14.0$, the retrieved columns density and b -parameter is a sensitive measure of their true values. Moving towards the saturated regime, the column density estimate is a lower limit with an average relative error of $\langle \delta^{\log N} \rangle = 0.006$. The b -parameter is yet well constrained with an average relative error of $\langle \delta^b \rangle = 0.03$. Finally, for $\log(N/\text{cm}^{-2}) = 18.6$ the optical depth in the damping wings becomes significant and the fit returns accurate estimates of the column density, as expected.

4 NMF–PM: THE NMF–PM CODE

We now introduce the second tool we present in this paper, NMF–PM. In what follows we first outline the data standardization steps we followed to prepare our library of metal profiles for the NMF analysis. Afterwards, we present a brief overview of the NMF formalism and outline the details of how we built a statistically robust process of NMF reconstruction and simulation. Finally, we show how the results of this analysis are used to build the NMF–PM python module, a metal absorber profile maker which we make publicly available.

4.1 Data standardization

NMF is an alternative approach to dimensionality reduction (e.g. to principal component analysis) where it is assumed that the data can be decomposed (or transformed) into non-negative components. Despite its desirable properties (it automatically extracts sparse and meaningful features from a set of non-negative data vectors), the NMF fitting requires the data to be normalized and regularized for an unbiased decomposition. By using MC–ALF on our library of absorbers, we obtain infinite S/N Voigt models of absorption profiles of different strengths and with a range of velocity distributions (see Fig. 1). In particular, our library consists of 650 unsaturated metal profiles of which we aim at reproducing their intrinsic shape (Section 2.2). To prepare and standardize the data for the NMF decomposition, we adopt the following two-step procedure.

As a first step, we need to determine the rest-frame velocity of the profiles and shift them to a common velocity frame. For this, we can use the model Voigt components to re-sample the profiles at a resolution of 1 km s^{-1} and to transform them to a common rest-frame velocity centred at 0 km s^{-1} . As the profiles may exhibit several Voigt components of different strength, here we define the zero velocity as

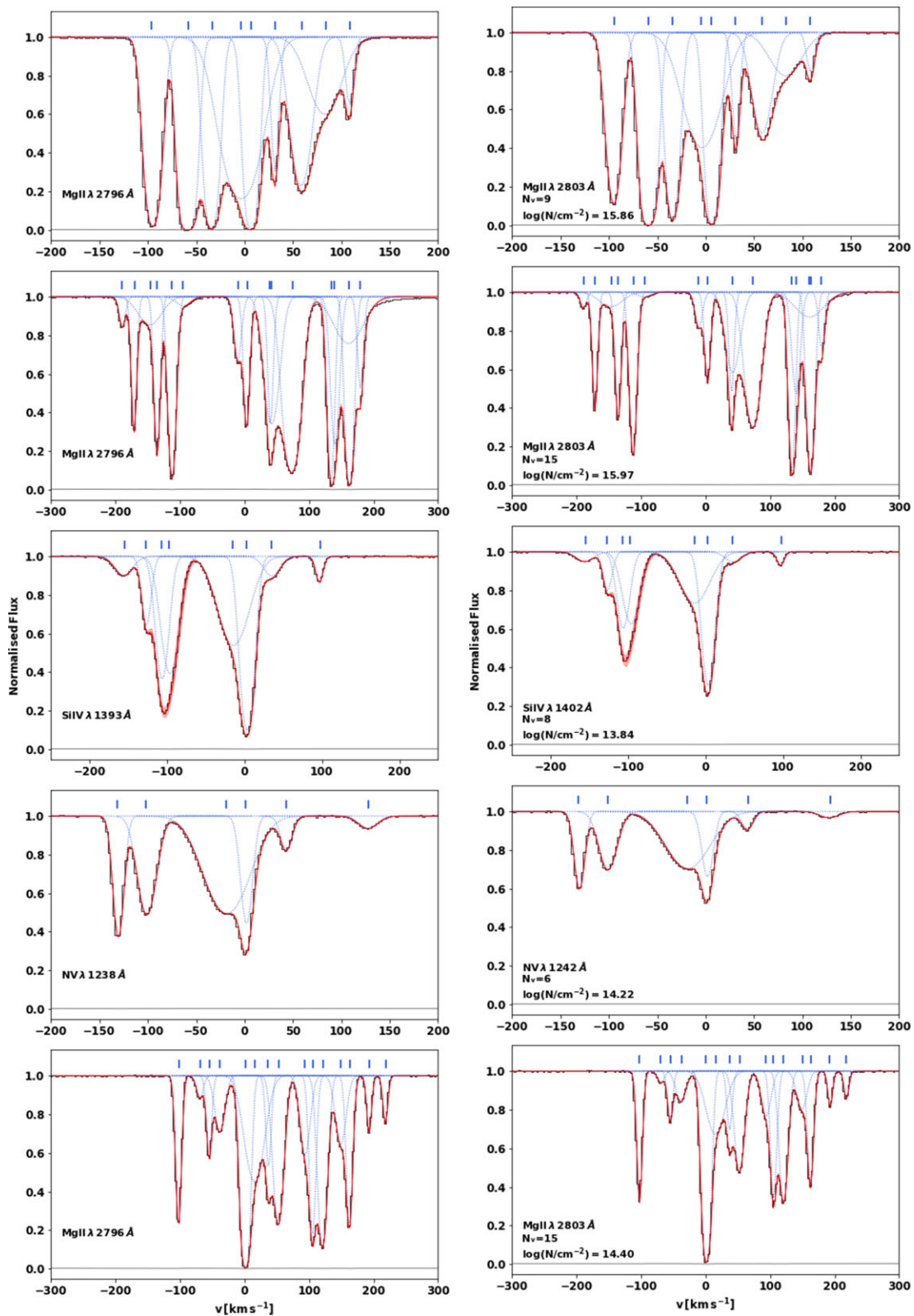


Figure 6. Simulated profiles (black) used to test the Voigt Component fitting routine. The profiles are characterized by different column densities and number of Voigt components as given in the legend. For simplicity, we only show the case for $S/N = 500$, with the 1σ sigma array in grey. For each profile, the MC-ALF fit is shown as a solid red line with $\pm 1\sigma$ uncertainty as a shaded area. The dotted blue lines represent individual Voigt components centred at the velocities highlighted by the blue ticks.

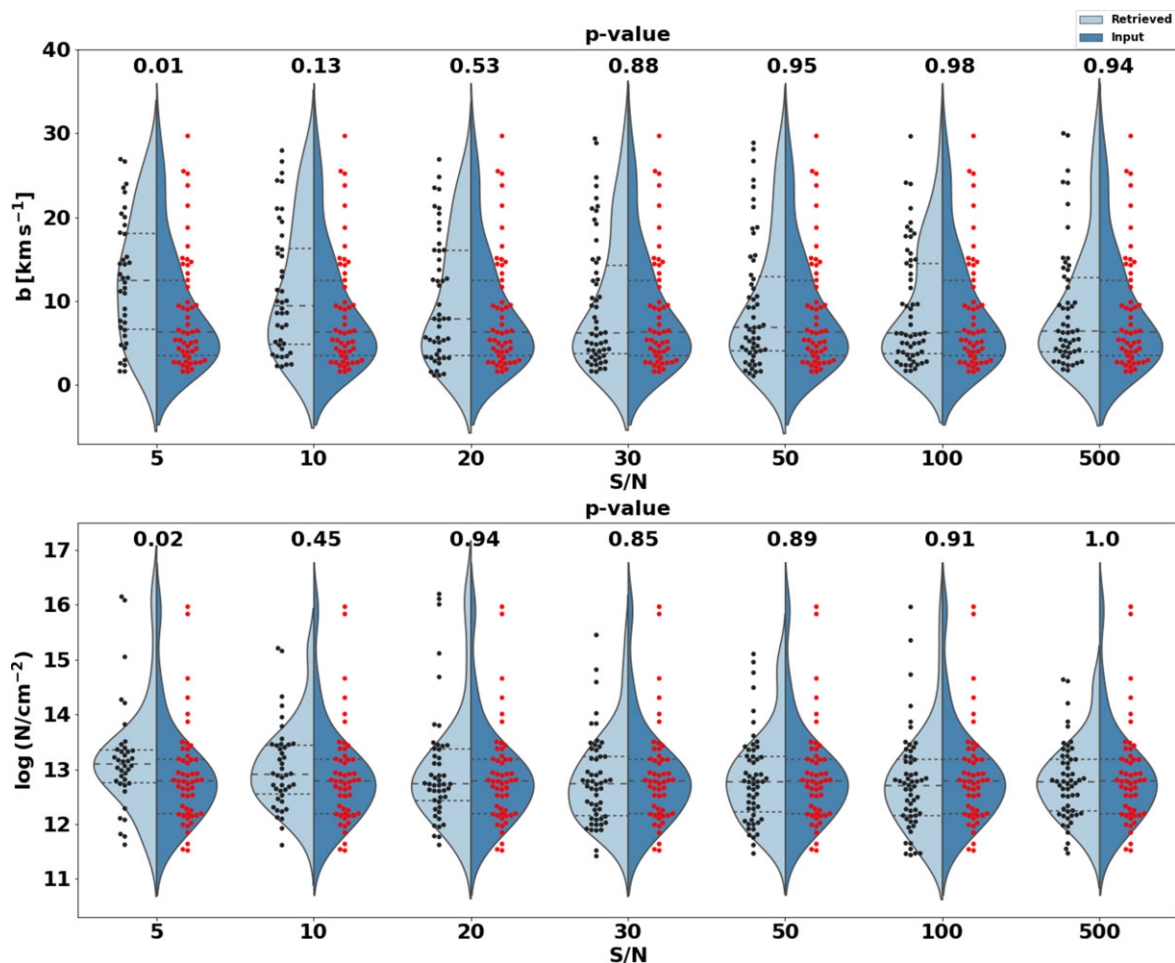


Figure 7. Violin plots comparing the probability density of b -parameters and column densities as traced by the input (light blue) and retrieved (dark blue) samples as a function of the S/N of the input spectra. In each violin, the horizontal central dashed line is the median and the dotted lines are the 25 per cent and 75 per cent quartiles. The distributions are determined from the entire sample of Voigt components as depicted by their relative swarm plots (dots). MC-ALF recovers the total input distributions (p -value > 0.1) for $S/N \geq 10$. At lower S/N , the code must be tested relative to the integrated values of the fitted parameters.

ΔV_{50} , i.e. the value at which the velocity distribution encompasses 50 per cent of the optical depth of the line.

The second step involves the recovery of the optical depth and the normalization of the line profiles. Rather than considering the transmitted flux, we elect to describe profiles in terms of their optical depth, $\tau(\nu)$, for which non-negativity is inherent to the data being considered and the normalization step is more straightforward. As recalled in Equation (1), $\tau(\nu)$ is the product of the ion column density, N , the frequency integrated absorption cross-section, s , and of the velocity profile. Thus, the normalization of the optical depth by the product $N \times s$ allows us to retrieve the line intrinsic profiles, $\phi(\nu)$, without carrying the added complexity of individual column densities and oscillator strengths of different absorbing ions, and to focus on the line shape as the only general property we wish to describe and reproduce in the mock-making step. Once velocity profiles are generated, the full absorption line systems can then be recovered by multiplying back the desired column density and strength of an ion. The imperfect approximation we are introducing at this step is to separate the correlation between an ion and its Doppler parameter, due to the atomic mass dependence. This approximation fails in the limit of single lines that are thermally broadened, but it holds for the majority of the profiles where turbulence and the combination of multiple components determine the line shape. Finally, as the profiles

used to train and test the NMF algorithm are models obtained via runs of MC-ALF on our set of observed data, very small structure is lost at moderate resolutions compared to high-resolution modes, so our modelling performs best for resolutions that are comparable to the lowest one in our library, i.e. X-shooter-like, and caution should be taken when applying this model to particularly high spectral resolutions.

4.2 Application of the NMF method

4.2.1 Overview of the formalism and general concepts

The NMF formalism assumes that a non-negative data set of n samples and ν features can be approximated by the (dot) product of two non-negative matrices.

$$D \approx XC, \quad (9)$$

where $D \in \mathbb{R}^{n \times \nu}$ is the matrix representation of the original data. Matrix X has the shape $n \times m$, where m is the number of reduced features in NMF space. The matrix C has the shape $m \times \nu$ and represents the coefficient matrix of the m reduced features, or in other words a representation of the new reduced features in the original feature space. Thus, via NMF, we generate a low-dimensional

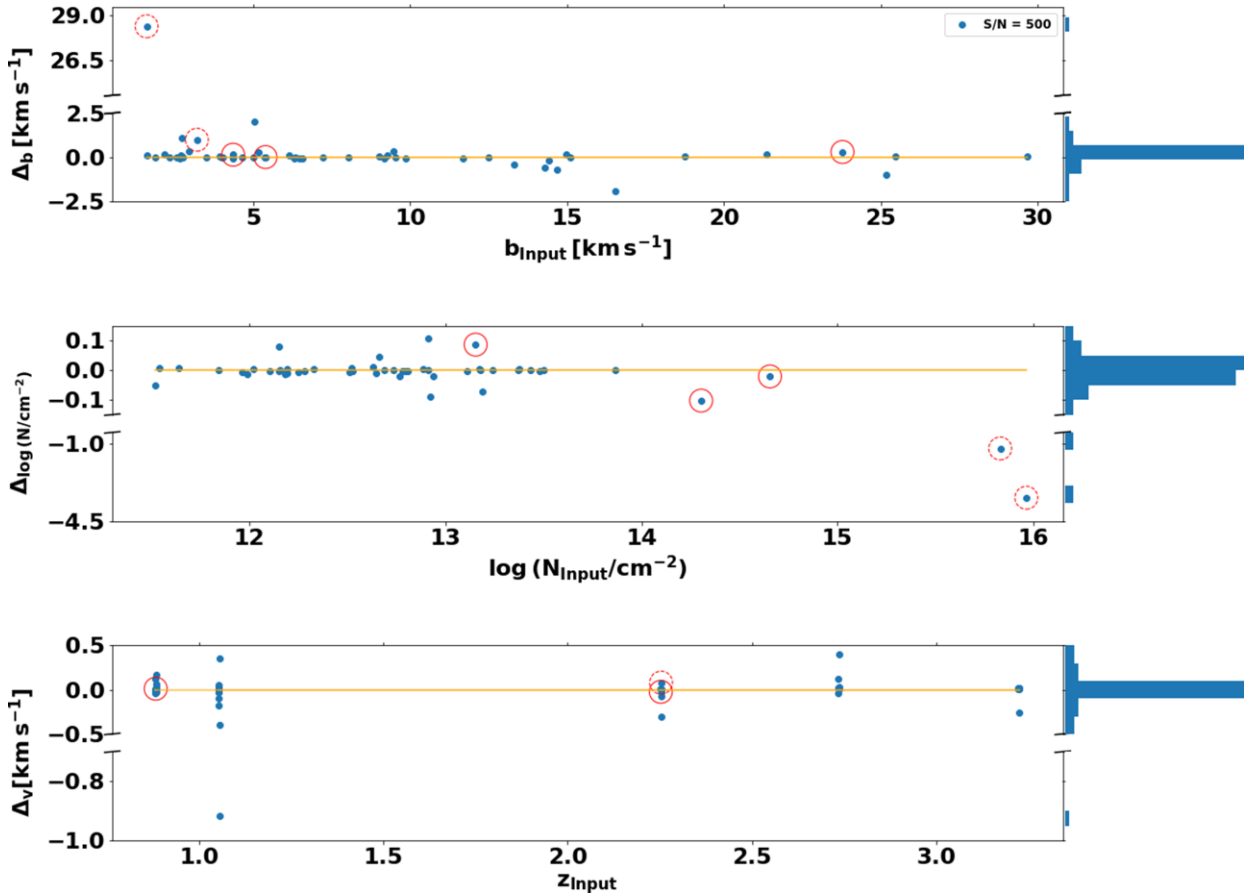


Figure 8. b -Parameters, column densities, and velocity accuracy for the retrieved sample of Voigt components when the profiles with $S/N = 500$ are analysed. Continuous and dashed red circles identify saturated lines and lines with $\log(N/\text{cm}^{-2}) > 15$, respectively. The largest deviating point in the retrieved b -parameter corresponds to one of these lines. The histograms of the residuals are plotted on the right.

encoding of a high-dimensional space. From Equation (9), it follows that each row in the matrix \mathbf{D} (each sample) is a linear combination of the row vector in the matrix \mathbf{X} with coefficient vectors supplied by the matrix \mathbf{C} , i.e.

$$d_i = \sum_{j=1}^m x_j c_{jv}. \quad (10)$$

Thus, NMF recasts an original vector onto new component axes of *latent features*, x_j , and the projections onto such an NMF space are given by the vectors in \mathbf{C} .

Specialized to our application, we have n line profiles characterized by v velocities, which collectively can be represented by a matrix \mathbf{Q} of dimension $n \times v$. We wish to reduce the dimensionality of the problem and assume that each of these profiles can be represented by m features where $m < v$. We apply NMF to \mathbf{Q} and obtain two matrices \mathbf{X} and \mathbf{C} whose matrix multiplication is represented by \mathbf{R} :

$$\mathbf{Q} \approx \mathbf{R} = \mathbf{X}\mathbf{C}. \quad (11)$$

We find \mathbf{R} such that it is the closest representation of \mathbf{Q} . The decomposition works by minimizing the squared Frobenius norm (i.e. a generalization of the Euclidean norm to matrix algebra) between \mathbf{Q} and the matrix product $\mathbf{X}\mathbf{C}$. In particular, our NMF fit implements a coordinate descent solver, i.e. an iterative process that successively updates the fitted parameters until convergence is reached.

Once \mathbf{R} is obtained we can further create a synthetic set of profiles by randomly assigning NMF latent features from their retrieved distributions in \mathbf{X} and then carrying out the linear combination as in Equation (10), i.e.

$$s_i = \sum_{j=1}^m \bar{x}_j c_{jv}, \quad (12)$$

where \mathbf{s}_i is the i th simulated vector in the matrix \mathbf{S} of dimension $n \times v$, and \bar{x}_j is a random sampling of the NMF features in \mathbf{X} relative to the j th NMF component. This is the main concept on which this work is based. In what follows, we show how we decompose a line profile, \mathbf{q}_i , into its low-dimensional representation, \mathbf{r}_i , and then use the resulting NMF decomposition to create a set of synthetic spectra, \mathbf{s}_i .

4.2.2 Implementation and tests of the NMF reconstruction

We apply the formalism set above to our library of absorption line profiles, with which we compute the low-dimensional representation needed for profile generation. Key to this process is to determine how well the reconstructed values fit the observed ones. It is also important to quantitatively assess how reliable the new synthetic data are with respect to the observed spectra. These considerations set a twofold testing process to quantify the ability of the NMF in: (i) reconstructing the profiles, and (ii) producing a new set of synthetic

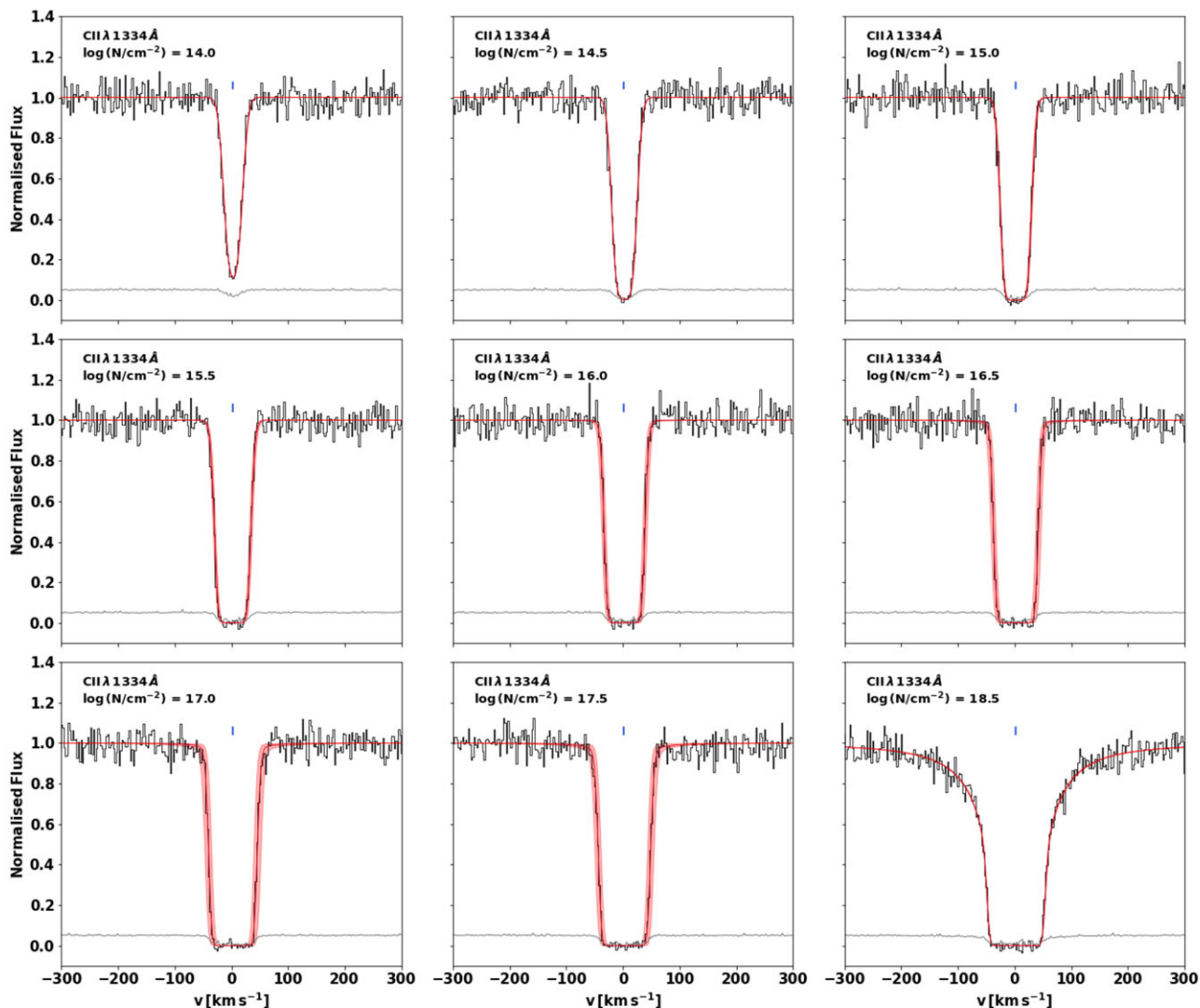


Figure 9. Simulated C II profiles (normalized flux in black and error in grey) with $b = 15 \text{ km s}^{-1}$ and $S/N = 20$. From top to right the profile is shown with higher column density values as given in the legend. For $\log(N/\text{cm}^{-2}) \geq 14.5$ the profiles are saturated and for $\log(N/\text{cm}^{-2}) = 18.5$ the damping wings of the Lorentzian become significant compared to those of the Gaussian contribution. For each profile, the MC-ALF fit is shown as a solid red line with $\pm 1\sigma$ uncertainty as a shaded area and centred at the velocities highlighted by the blue ticks.

data. For the first task, we use the residual variance, σ^2 , defined as the sum of the squares of the difference between the input profile, \mathbf{q}_i , and its reconstructed counterpart, \mathbf{r}_i . We also test the model accuracy by carrying out KS tests of the ΔV_{90} distributions as traced by the input and reconstructed data. For the second task, we again use the KS test, setting as a requirement that the synthetic profiles must be characterized by a ΔV_{90} distribution that is statistically consistent with that of the original profiles. These tests define our key performance indicators (KPIs).

When applying this method to our library, we noticed that the variety of profiles in our sample, which is described by the large range in ΔV_{90} , affected the goodness of the NMF fitting. Running NMF on the entire sample resulted in a model with a high degree of complexity (high number of NMF components, or equivalently a high-dimensional NMF space). This model ended up producing synthetic data not always similar to the observed profile shapes, thus failing to reproduce the input ΔV_{90} distribution. To obviate this issue, we designed an algorithm that applies the NMF on subsets of profiles

in smaller bins of ΔV_{90} , where the bins are selected adaptively by optimizing the two KPIs defined above.

We now describe the step-by-step procedure followed in designing this algorithm.

(i) Definition of ΔV_{90} bins: We define bins of ΔV_{90} of increasing size varying as $s_k = b_{\text{edge}} + k \times 20 \text{ km s}^{-1}$, where b_{edge} represents the lower bound edge of the considered ΔV_{90} bins and k in the range $1 \leq k \leq 4$. We select the subsample of profiles, satisfying the condition $\Delta V_{90i} \leq s_k$ and on this, we perform multiple runs of NMF fitting, each with an increasing number of NMF components, m , specifically $2 \leq m \leq 30$.

(ii) NMF fits: For each ΔV_{90} bin, the NMF analysis returns the feature vectors \mathbf{x}_j and their coefficient vectors in \mathbf{C} . The reconstructed profiles, \mathbf{r}_i , are then computed following Equation (10). Mock profiles are created by randomly sampling latent feature components from \mathbf{X} (i.e. from each column, see Equation 12) to create the new latent feature matrix. As the number of artificial profiles created

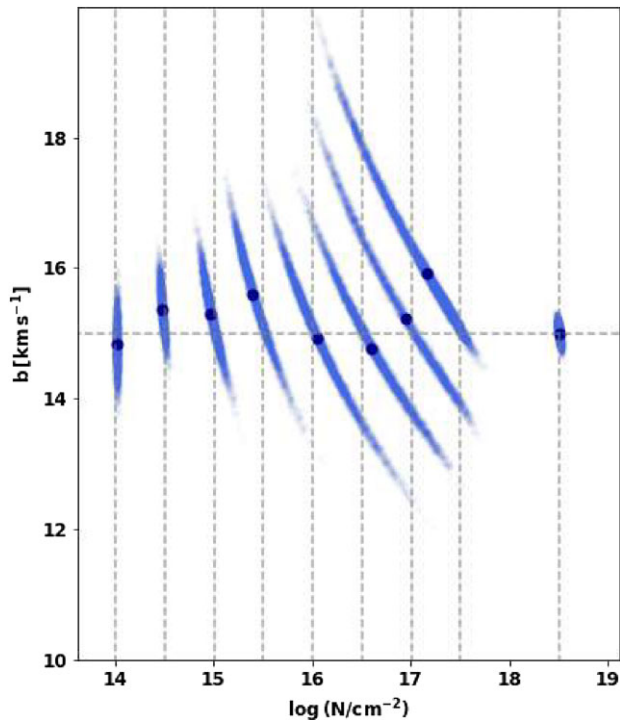


Figure 10. b versus $\log(N/\text{cm}^{-2})$ plot to test MC-ALF fits on a saturated C II absorber. Blue dots trace the full posterior samples provided in output by MC-ALF. The dotted grey lines show the simulated b and column density values. In presence of saturation, however, when the damping wings are not yet significant, the column density estimate is on average a lower limit with a relative error of $< \delta^{\log N} > = 0.006$. When the damping wings contribution becomes significant, in our example for $\log(N/\text{cm}^{-2}) > 18$, the fit returns accurate estimates of the column density.

in this manner cannot exceed the size of the input data sample in that specific ΔV_{90} bin, we carry out 100 different realizations of the simulation, each time sampling 66 per cent of the size of the input data.

(iii) KPIs analysis: The reconstructed and simulated samples are analysed in terms of ΔV_{90} distributions and finally statistically compared with the distribution of ΔV_{90} values of the input data sample using the p -values returned by the KS test. The performance of the reconstruction process is additionally tested by computing the mean residual variance between the input and reconstructed profiles, namely $< \sigma_i^2 > = \frac{\sum_{i=1}^n (q_i - r_i)^2}{n}$, with n the number of input data.

(iv) NMF model selection: To determine the optimal NMF model, we consider all the NMF representations that simultaneously satisfy the condition p -value > 0.1 in both the data-versus-reconstructed and data-versus-simulated KS tests and for these we compare their σ^2 distributions as a function of the number of NMF components, m . A reasonable expectation is that σ^2 decreases in value as m increases. However, an overestimation of m would include noise in the simulated profiles. As a solution, we consider the relative $< \sigma^2 >$ improvement and select the optimal NMF model as the one for which we first measure an improvement larger than 40 per cent. In case the p -value condition is satisfied in multiple ΔV_{90} bins, the optimal NMF model in each bin is selected as above, and finally the optimal bin size is chosen as the one in which the $< \sigma^2 >$ value is the lowest. At the end of this step, the bin edge value, b_{edge} , is updated to be the upper bound of the current step, and the process is repeated till the entire sample of data is analysed.

(v) Problematic ΔV_{90} bins: The procedure outlined above also identifies ΔV_{90} bins in which the condition p -value > 0.1 is never met. These are bins of ΔV_{90} values in the ranges 80–100, 100–120, and 120–140 km s^{-1} . As one would expect, the most dominant factor that can cause the NMF to fail is the diversity in the complexity of the input data, which we can parametrize with the number of Voigt components. Thus, we further divided the data falling in such problematic bins into low and high number of Voigt components subsets. We stress that low/high number of Voigt components does not imply low/high ΔV_{90} values as it can be seen from Fig. 1 (bottom panel), where profiles with similar velocity widths are characterized by significantly different numbers of Voigt components. Thus, we split the two categories such that each contains roughly 50 per cent of the total profiles in the bin. Once the division is done, we repeat the procedure outlined earlier to find the optimal NMF decomposition.

Examples of the procedure described above are shown in Figs 11 and 12 (the KPI analysis run over the entire sample of data is provided in the online material) where, our KPIs for the NMF fitting and modelling are presented for a subsample of data falling in increasing size of ΔV_{90} bins used in the iterative process that determines the final bin to select and, within this, the optimal NMF model. The profiles characterized by low values of ΔV_{90} are always preferred to be grouped together in the smallest bin size of 20 km s^{-1} . For example, the NMF decomposition on all the profiles with $\Delta V_{90} \leq 80 \text{ km s}^{-1}$ would succeed in the reconstruction step (i.e. p -values > 0.1), but it would fail in generating synthetic profiles with the targeted ΔV_{90} distribution (i.e. at least one NMF model for which p -values > 0.1 is present). On the other hand, running the procedure on the sample of data for which $\Delta V_{90} \leq 20 \text{ km s}^{-1}$ identifies multiple NMF models (shown as black framed in the top-left panel of Fig. 11) that simultaneously satisfy the condition p -value > 0.1 in both the data versus reconstructed and data versus simulated KS tests. Thus, the algorithm selects the NMF model with a $< \sigma^2 >$ improvement closest to 40 per cent (bottom panel in Fig. 11), i.e. the model with $m = 6$ NMF components (dotted-white frame in the top-right panel of Fig. 11).

At larger ΔV_{90} values ($> 180 \text{ km s}^{-1}$), the NMF fitting is less sensitive to the bin size. A clear case is shown in Fig. 12, where both the reconstructed and simulated profiles are statistically consistent in following the same ΔV_{90} distributions as the one traced by the input data in bins of size 20, 40, 60, and 80 km s^{-1} . As described above, by analysing the variation of the relative $< \sigma^2 >$ improvement we are able to avoid NMF overfitting, and finally the optimal bin size is chosen as the one in which the difference between the input and the reconstructed profiles is the lowest (lowest value of $< \sigma^2 >$ as shown in ‘data – reconstruction residual variance’ plot). In our example (Fig. 12), the optimal NMF fit is obtained for profiles with $200 < \Delta V_{90}/\text{km s}^{-1} \leq 220$ with $m = 11$ NMF components.

Fig. 13 shows the NMF analysis relative to the profiles for which the analysis based on a simple division in bins of ΔV_{90} is not possible due to the complexity of their shapes. These are profiles with a spread in velocities mostly falling within the range $100 < \Delta V_{90}/\text{km s}^{-1} < 140$. For these profiles we take advantage of the information we have on the number of Voigt components used for their decomposition (see Section 3 for more details) and the NMF fitting is run separately on two different samples, namely the low- and high-Voigt components samples, defined such that each contains roughly 50 per cent of the total profiles.

To validate the procedure, we use the NMF algorithm described above to generate artificial profiles, and reproduce the input data in

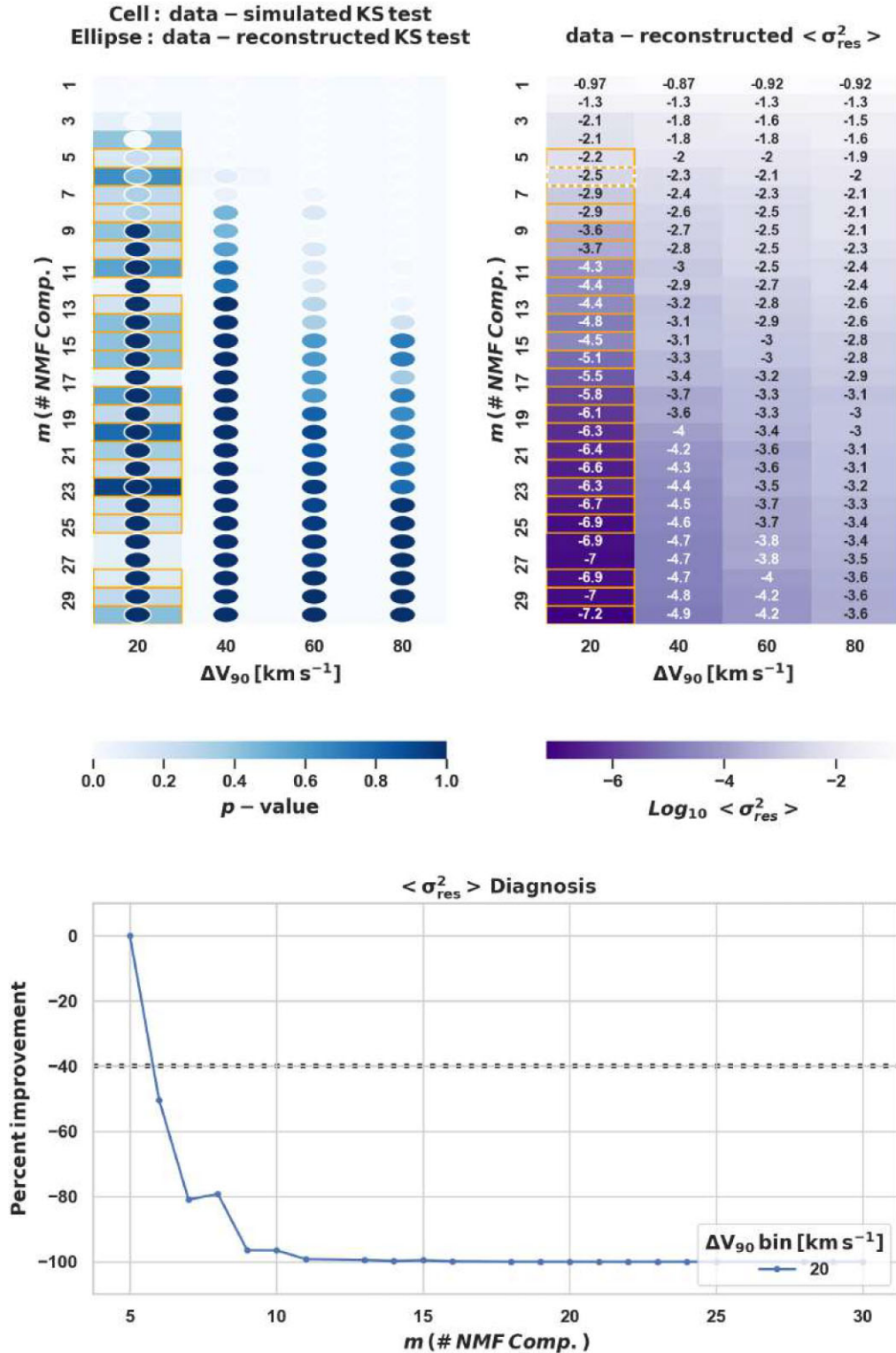


Figure 11. Examples of the performance metrics (KPIs) in the NMF reconstruction and simulation process for profiles characterized by small values of velocity widths ($\Delta V_{90} \leq 80 \text{ km s}^{-1}$). Top-left panel: Data-versus-reconstructed (grid cells) and data-versus-simulated (ellipses) KS tests of the ΔV_{90} distributions as a function of the number of NMF components, m , and ΔV_{90} bins of the input data. The colour code follows the KS test p -value statistics, with p -value > 0.1 the threshold we use for statistical significance. Orange framed regions are where p -value > 0.1 for both data-versus-reconstructed and data-versus-simulated distributions. Top-right panel: Same as the top-left panel, with the map coloured by the mean residual variance ($\log \langle \sigma^2 \rangle$) between the input and reconstructed profiles. The white dotted frame identifies the selected NMF model. Bottom panel: Relative $\langle \sigma^2 \rangle$ improvement between NMF models with an increasing number of NMF components (blue dots with line) for the ΔV_{90} bin where the condition p -value > 0.1 is satisfied as given in the legend. The grey horizontal line identifies the 40 per cent threshold in $\langle \sigma_{res}^2 \rangle$ improvement we use to avoid NMF overfitting.

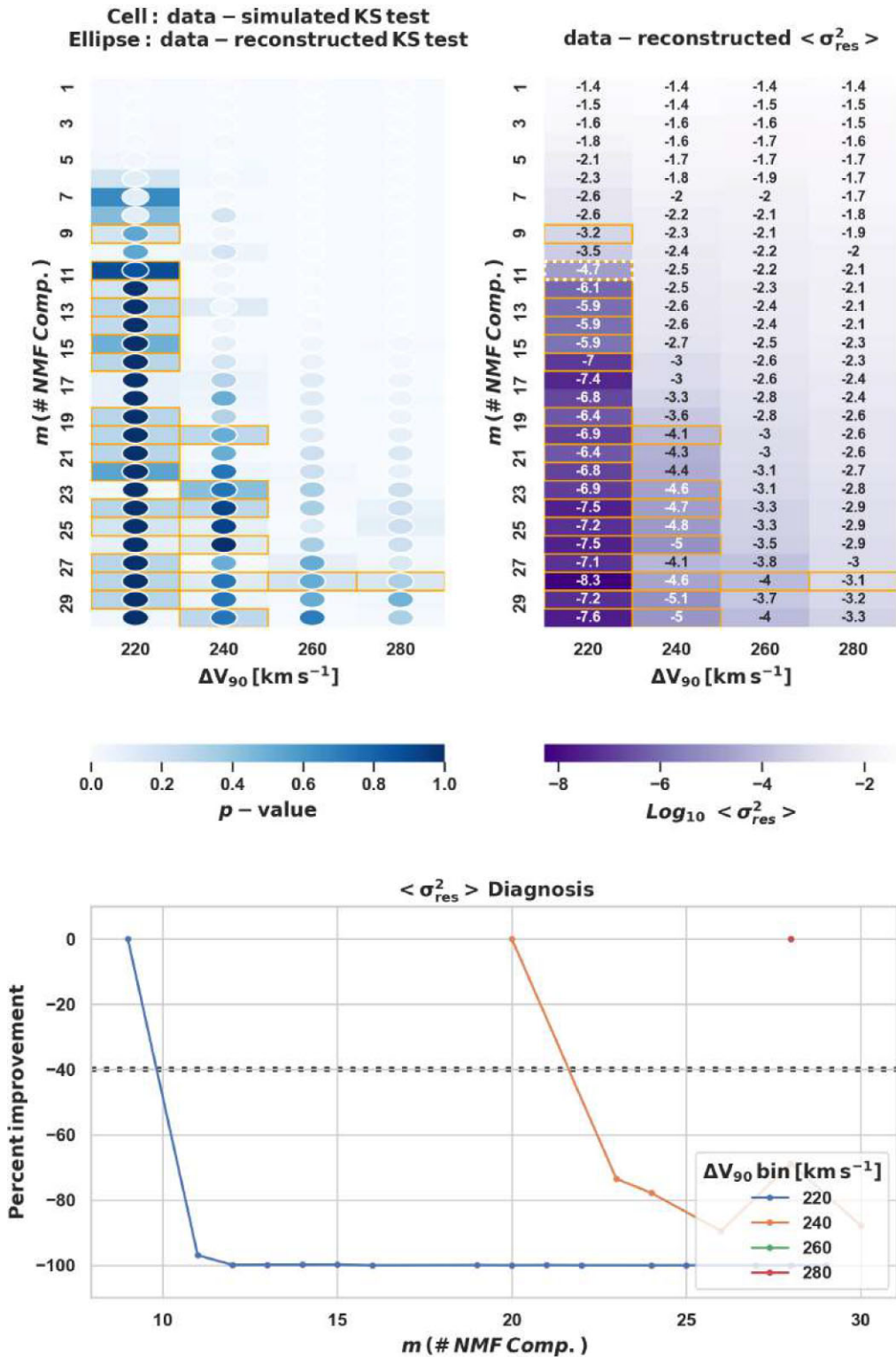


Figure 12. Same as Fig. 11 but for data characterized by larger velocity widths, $200 < \Delta V_{90} \leq 280 \text{ km s}^{-1}$. In this example, the p -value condition is satisfied in multiple ΔV_{90} bins (as given in the legend). Dots with lines show the relative $\langle \sigma^2 \rangle$ improvement between NMF models with increasing number of NMF components. Relative to the last two ΔV_{90} bins only one NMF configuration satisfies our criteria, i.e. $m = 27$, resulting in a single value in the $\langle \sigma_{res}^2 \rangle$ diagnosis plot and the two points, red and green, overlapping.

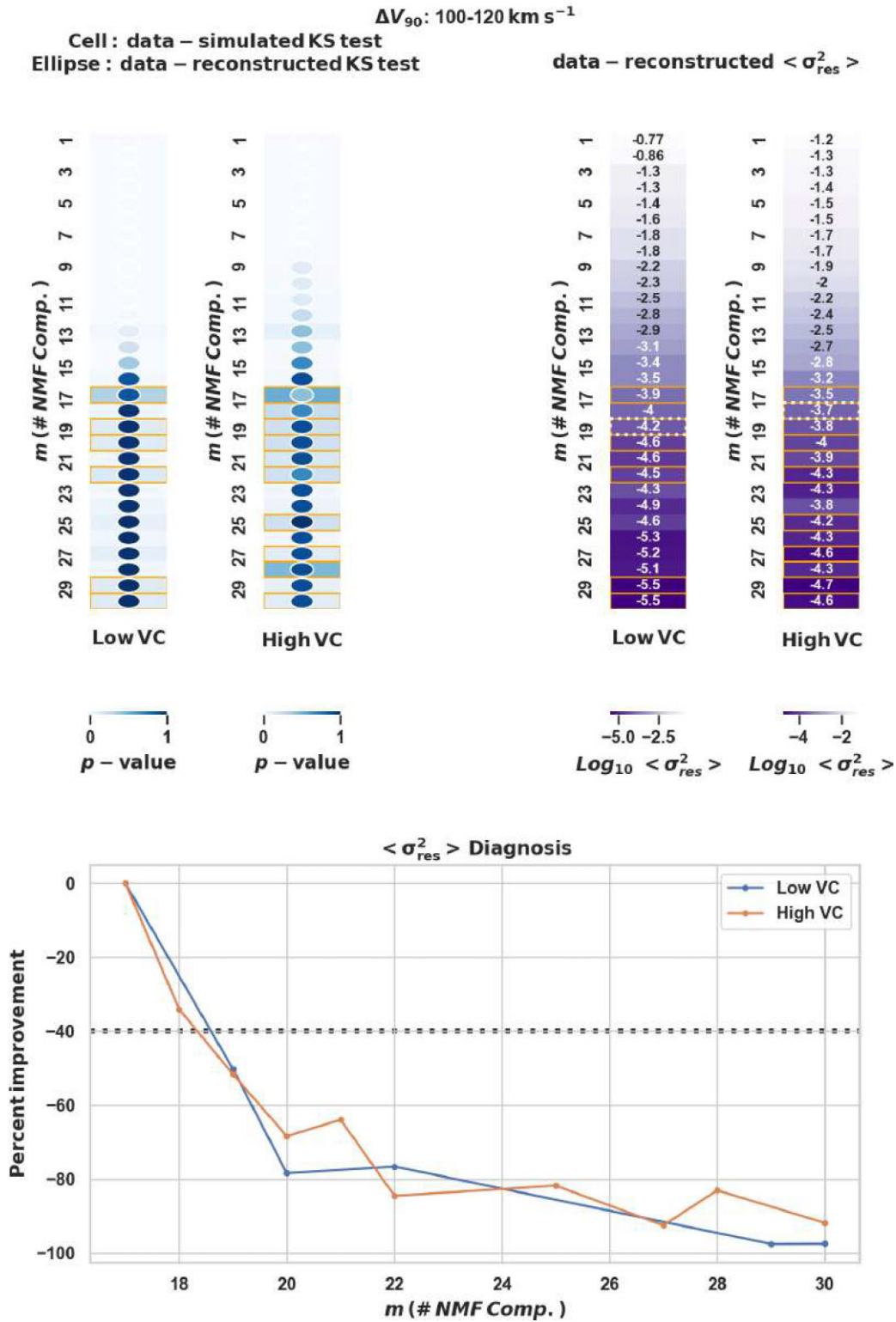


Figure 13. Same as Figs 11–12, but for input data falling in the problematic bin with $100 < \Delta V_{90} \leq 120 \text{ km s}^{-1}$. KS tests (top-left panel) and residual variance estimations (top-right panel) are carried out on the low and high number of Voigt components.

size, resolution (1 km s^{-1}), and quality (infinite S/N). The goal is to compare synthetic and input data qualitatively and quantitatively by comparing their optical depths and ΔV_{90} distributions, respectively. As we have shown in Section 4.2.2 this is achieved by randomly sampling the NMF coefficients from their respective distributions of

PDFs and finally using them as a new set of projections onto the NMF axes. The results of our simulations are presented in Fig. 14, where we show the comparison between synthetic (gold) and observed (blue) profiles, in terms of their optical depth distribution after having re-sampled them at a resolution of 8 km s^{-1} (from left to right and

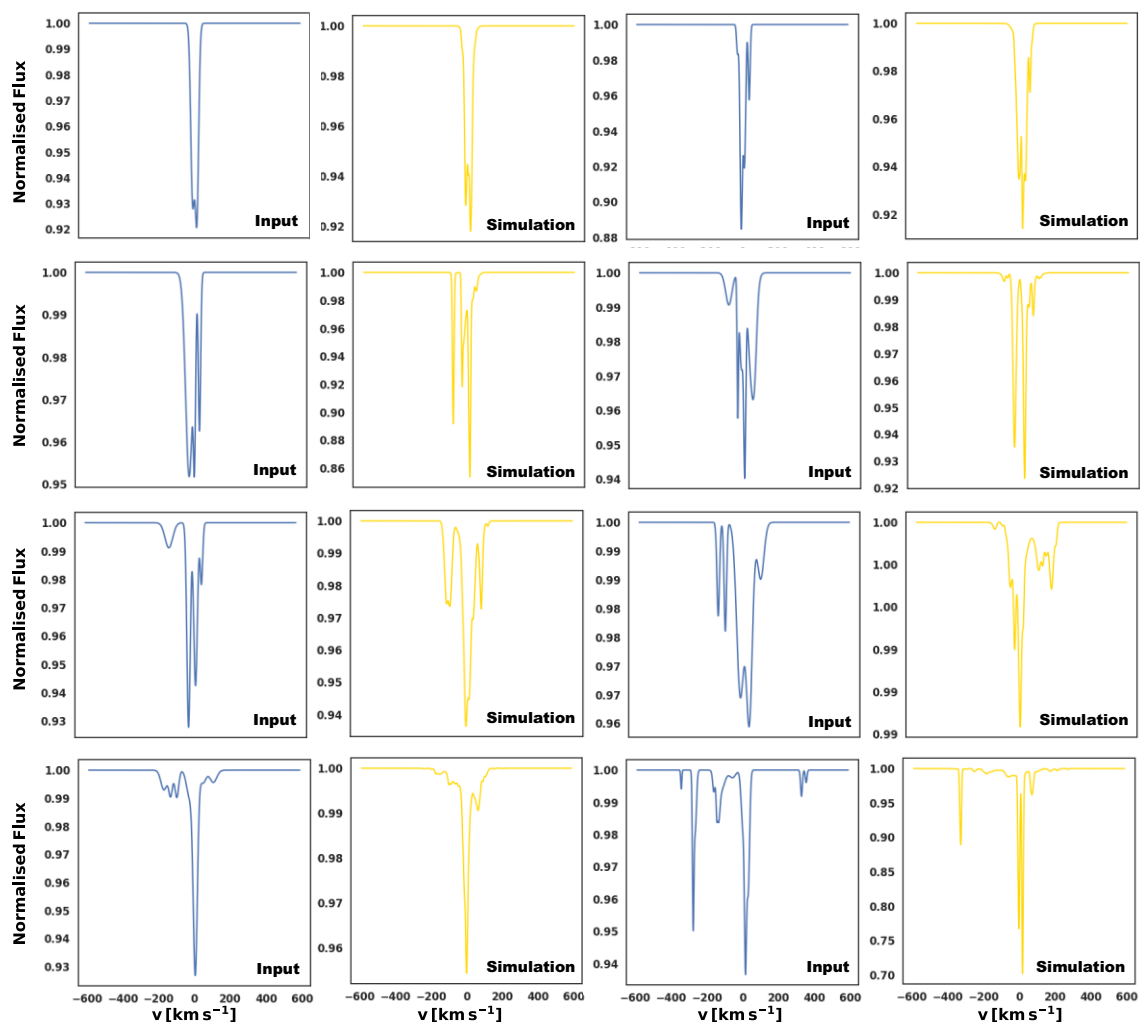


Figure 14. Comparison between input (blue) and simulated (gold) velocity profiles for a fixed column density of $\log(N/\text{cm}^{-2}) = 12$.

top to bottom, the profiles are characterized by increasing ΔV_{90} values). Our main result is that the diversity of the line profiles is well reproduced in the simulated sample despite the complexity of the input information. We evaluate the accuracy of our model by carrying out KS tests between the ΔV_{90} distributions of the input versus simulated and reconstructed profiles, as shown in Fig. 15. We find that the simulated profiles have a ΔV_{90} distribution statistically close to the one traced by the reconstructed (p -value = 0.9) and input (p -value = 0.8) samples. Finally, in Fig. 16, we show that the individual components of the synthetic profiles have representative b -parameters and column densities when compared to the distributions traced by the observed absorbers (p -values > 0.9 for both samples). The evaluation for the b -parameters is carried out by running MC-ALF on 100 simulated CIV profiles that are characterized by a resolution of 8 km s^{-1} , ideal $S/N = 500$ values, and total column densities that follow the same distribution we measure for our input sample (see Fig. 2 bottom panel). The check on the column density values, instead, is run on a smaller sample of 20 profiles, simulated with a realistic noise component, such that $15 < S/N < 30$ and moderately strong, i.e. with total column densities in the range $13 \leq \log(N/\text{cm}^{-2}) \leq 13.5$. Such a choice is for the results to be the least affected by uncertainties related to profile fitting. Indeed, for lower column densities it may become difficult to match single

input-versus-retrieved Voigt components, while for higher values, single components may reach saturation and the column densities may no longer be estimated with a few per cent accuracy (see Fig. 7 central panel). The comparison is then carried out with a sample of real CIV profiles with similar characteristics.

4.3 The NMF-PM python package

To enable the use of this tool by the community, we inserted the NMF feature matrix, \mathbf{X} , and coefficient matrix, \mathbf{C} in a python module, which we dubbed the NMF-PM. To run NMF-PM, the user will have to specify the number of simulated profiles to obtain in output via the parameter `nsim`, the ions to simulate, together with their rest-frame wavelengths and column density values passed via the parameters `ion`, `trans.wl`, and `ion.logN`, respectively. As discussed in Section 2.2, when considering different families of ions, i.e. moderate- and low-ion families, the absorbers may be characterized by different ΔV_{90} distributions. Thus one feature of the NMF-PM is to allow the user to specify which class of ion they are simulating via the parameter `ion.family`. This can be set to `'moderate'` or `'low'` for the simulated profiles to follow a ΔV_{90} distribution as the one we measure for our samples of moderate- and low-ions (see Fig. 2, top panel), or the user can feed their own ΔV_{90}

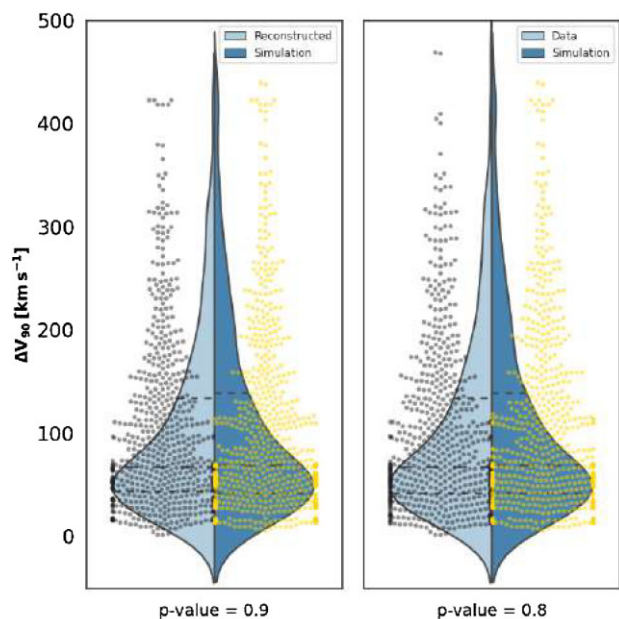


Figure 15. Left-hand panel: Violin plots comparing the probability density of the ΔV_{90} distributions as traced by the reconstructed (light blue) and simulated (dark blue) profiles, where the central dashed line is the median and the dotted lines are the first and third quartiles. The entire distribution is shown as a swarm plot (dots). Right-hand panel: Same as the left-hand panel, but where the comparison is carried out between the input and simulated profiles. The p -value scores show that the simulated profiles are characterized by values statistically close to those traced by the reconstructed and input data.

PDF. NMF-PM also allows for the creation of ion doublets (e.g. C IV and Mg II) by setting to ‘true’ doublets and by providing a value for the `dbl_ratio` and `dbl_dvel` parameters for a given oscillator strength ratio and velocity shift (in km s^{-1}) for the second line.

With this configuration, NMF-PM simulates absorber profiles characterized by 1 km s^{-1} resolution and with no noise. However, the user can further: (i) convolve the profiles with a Gaussian kernel switching to ‘true’ the `convolved` parameter and consequently providing the resolution (full width at half-maximum; FWHM) via `res`; (ii) add a random Gaussian noise component by providing a value for the desired S/N (per pixel) via `SNR`¹ and (iii) carry out a profile re-sampling providing a value for the `px_scale` parameter. The re-sampling is implemented such that it conserves overall the integrated flux. Via its attribute, NMF-PM will return the convolved and re-sampled synthetic metal profiles with noise, and the associated noise and wavelength arrays. It will also return the original flux and wavelength arrays at a resolution of 1 km s^{-1} not convolved nor re-sampled. The NMF-PM class with its parameters and attributes is shown in Fig. 17.

NMF-PM has been optimized to efficiently generate synthetic metal profiles: Even performing the convolution step, the addition of Gaussian noise, and pixel re-sampling, the NMF-PM method runs in a matter of minutes on a single core computer to generate a library of 10^5 objects.

¹This represents the S/N ratio (per pixel) with respect to the continuum of the background source. To add a noise component relative to the sky signal the parameter `sigma_sky` can be used.

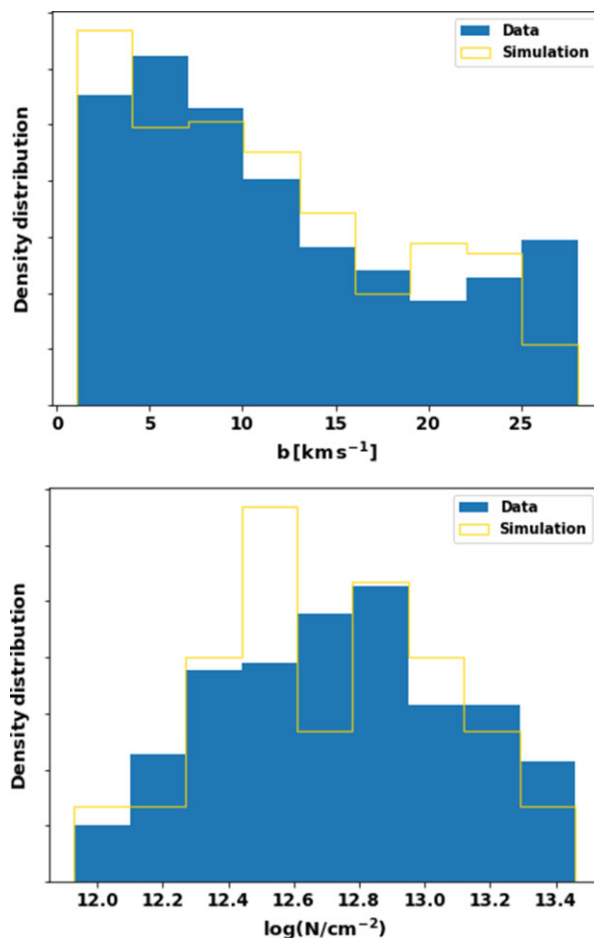


Figure 16. Top panel: Normalized distributions of b -parameter values for observed (blue) and simulated (gold) velocity profiles. The simulations trace a family of high-resolution C IV profiles, with $S/N = 500$. Bottom panel: Same as the top panel; however, this time the distributions are relative to column densities values. Here, the simulations trace high-resolution C IV profiles, with $15 < S/N < 30$, and $13 \leq \log(N/\text{cm}^{-2}) \leq 13.5$. p -Value scores larger than 0.9 for both distributions show that the synthetic profiles have representative b -parameters and column densities values.

5 SIMULATED PROFILES IN LARGE SURVEYS

The automated tools we have described in this work open to the opportunity of testing our capability of detecting and analysing absorption features in spectra of different data quality in large surveys. To showcase the capabilities and further test the performance of our code for these applications, we run a library of 10^6 synthetic metal profiles mimicking C IV absorbers using the moderate velocity distribution (see the red histogram in the top panel of Fig. 2). Profiles are generated in a flat distribution of column density in the interval $10^{13} - 10^{15.5} \text{ cm}^{-2}$, while the S/N is uniformly distributed in the interval 2.5–15. We mock a WEAVE-like survey (Jin et al. 2023) by setting the pixel scale to 16 km s^{-1} and the resolution to 60 km s^{-1} .

Fig. 18 (left-hand panel) shows the measured equivalent width (EW) for the stronger line of the doublet measured on noisy profiles in comparison with the intrinsic value derived from the noise-free simulated profiles. To better capture the intrinsic scatter in the distribution rather than a possible bias, we compute the average of the absolute discrepancy, normalized by the true value. Values less than one in this metric identify EW values retrieved with high precision. These plots reveal the expected trend of increasing precision in the

nmf_profile_maker.NMFPM

```
class nmf_profile_maker.NMFPM(NMF_dct, nsim=None, ion_family='moderate', filename_ion_family=None, ion_logN=[14.0], ion=None,
trans_wl=None, filename_ion_list=None, convolved=False, res=8, px_scale=None, SN=[None], sigma_sky=None, doublet=False, dbl_dvel=0)
```

Non-Negative Matrix Factorization - Profile Maker (NMFPM)

Generate profiles from two non-negative matrices (X,C), whose product approximates the matrix Q of observed metals in quasar spectra. These profiles can be used to generate large libraries of realistic metal absorption profiles

Parameters: **NMF_dct:** Dictionary containing the information about the non-negative matrices (X,C).
X: NMF_dct['X'] ndarray of shape $n \times m$, where m is the number of reduced features in the NMF space
C: NMF_dct['C'] ndarray of shape $m \times v$. It represents the coefficient matrix of the m reduced features

nsim: int, default = 1
Number of profiles to be generated.

ion_family: {'moderate', 'low', 'user'}, default='moderate'
Ions families to be considered. Valid options:

- 'moderate': If 'moderate' the profiles will follow a DeltaV₉₀ distribution typical of moderate ions transitions.
- 'low': If 'low' the profiles will follow a DeltaV₉₀ distribution typical of low ions transitions
- 'user': If 'user' the profiles will follow a DeltaV₉₀ distribution provided by the user with filename_ion_familiy

filename_ion_familiy: str, default=None
User filename for DeltaV₉₀ pdf if ion_family = 'user'

ion_logN: ndarray of shape (nsim,), default=[14.0]
log Ion Column Density in cm^{-2}

ion: ndarray of shape (nsim,), default=[CIV]
Ion transition to be simulated

trans_wl: ndarray of shape (nsim,), default=[1548.2040]
Ion transition wavelength in Angstrom

filename_ion_list: str, default = None
User filename for lines' physical parameters

convolved: Boolean, default = True
Allow for the generated profile to be convolved with a Gaussian kernel

res: float, default = 8
Resolution of the generated profiles in km/s

px_scale: float, default = None
If not None resample while preserving flux the final profiles using px_scale as pixel sampling

SN: float, default = [None]
Signal-to-Noise ratio of the continuum signal used to compute the Gaussian noise to be added

sigma_sky = int, default = None
RMS value of the sky signal. If not None the sky noise is computed as a random distribution centred on 0 and with dispersion sigma_sky

doublet: Boolean, default = False
Enables creation of doublets

dbl_fratio: float, default = 0
If doublet True, create a second line with oscillator strength $f_{\text{line}_2} = \text{dbl_fratio} * f_{\text{line}_1}$

dbl_dvel: float, default = 0
If doublet True, create a second line with center shifted in velocity by dbl_dvel [km/s]

Attributes: **flux:** nsim synthetic spectra, with noise if so desired
flux_no noise: nsim synthetic spectra, no noise
noise: associated noise values
wave: wavelenath values for the nsim svnthetic spectra

Figure 17. NMF-PM python class with parameters and attributes.

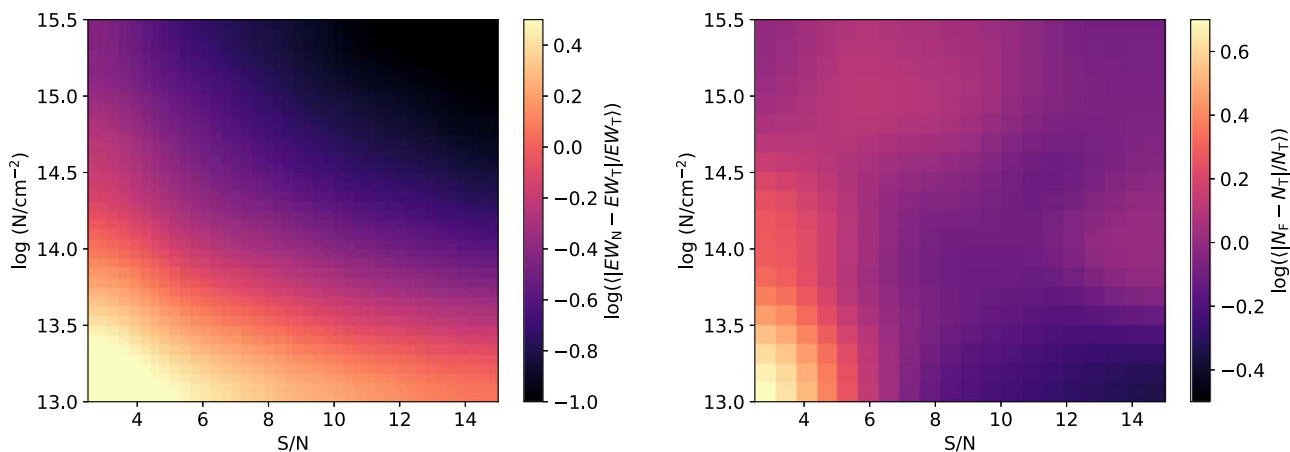


Figure 18. Test of detection and analysis of C IV absorption features of different quality in WEAVE-like spectra. Left-hand panel: Mean EW relative errors (on a logarithmic scale) as a function of S/N and input column density values for 10^6 test profiles. For clarity reasons, the metric has been smoothed with a 2×2 Gaussian kernel. Right-hand panel: Same as the left-hand panel; however, this time the mean relative error statistics is shown for the column densities and for a subsample of 10^4 profiles. Given the lower number statistics, the smoothing is carried out on a 3×3 kernel window. Values less or equal than zero identify the regions in the $\log(N/\text{cm}^{-2})$ versus S/N plane of high precision in the fitted values.

measurement as both column density and S/N increase. At moderate S/N , or $\lesssim 4$, C IV can be measured reliably only for column density $\gtrsim 10^{14} \text{ cm}^{-2}$. Next, we proceed and fit a subsample of 10^4 profiles with MC-ALF to study the accuracy in retrieving the column density using the same metric we used for the previous test. The results are shown in Fig. 18 (right-hand panel): Excluding the bins characterized by both $3 \leq S/N \leq 5$ and $\log(N/\text{cm}^{-2}) = 13.5$, for which the quality of the data prevent the fit to run correctly, MC-ALF can retrieve the input information at all S/N and column density values considered. In particular, fits of mildly saturated and unsaturated profiles with $\log(N/\text{cm}^{-2}) \leq 14.5$ are less sensitive to variations in S/N for $S/N > 7$. For $S/N \leq 7$, the accuracy decreases, with the lowest values (~ 40 per cent) measured for $S/N \sim 3$. For heavily saturated lines (in our example for $\log(N/\text{cm}^{-2}) > 14.5$), MC-ALF fits have larger uncertainties (~ 20 per cent for $S/N < 10$), although the accuracy increases as a function of the S/N . The effect of saturation on low-resolution spectra is further analysed by repeating the test presented in Fig. 10 for C II profiles this time convolved with a FWHM of $\sim 60 \text{ km s}^{-1}$. The results (Fig. 19) show that, by exploring the full underlying posterior, MC-ALF is able to recover the degeneracy between the b -parameter and the column density estimates due to hidden saturation resulting in broad and degenerate posterior distributions. For saturated profiles, in the regime when the damping wings are not yet significant (in this example for $\log(N/\text{cm}^{-2}) < 18$), the full posterior PDF should be used for an accurate propagation of uncertainties.

6 SUMMARY AND CONCLUSIONS

In this work, we present two new tools for studying and modelling metal absorption lines in the CGM: MC-ALF to automatically reconstruct the physical parameters of the absorbers and NMF-PM to generate synthetic but realistic-looking line profiles following a given distribution of desired line width.

The observational data we used for developing, training, and testing our codes come from a compilation of spectroscopical campaigns, which collected high-resolution, high- S/N spectra of 42 quasar fields at redshifts $1.2 \leq z \leq 4.5$ (Section 2). These surveys identified a family of ~ 1000 moderate- and low-ion absorbers along the quasars' line of sight. By considering only unsaturated profiles, we selected

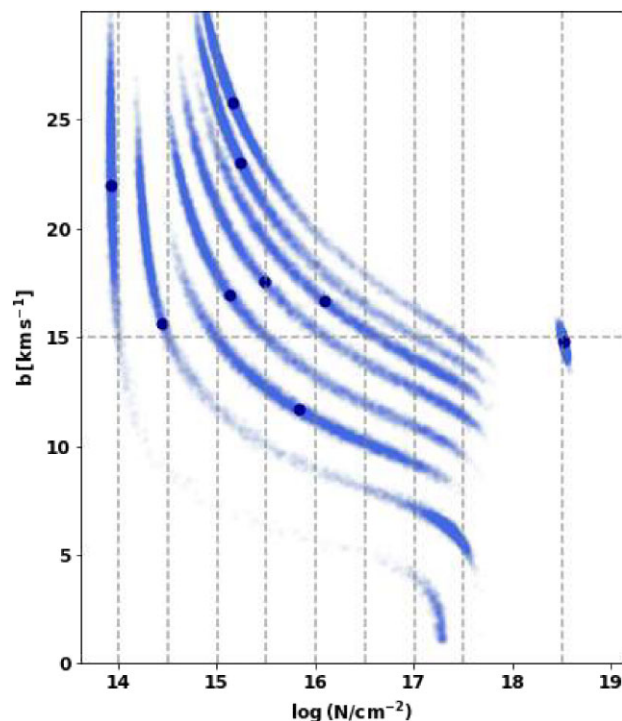


Figure 19. Same as Fig. 10; however, this time the b versus $\log(N/\text{cm}^{-2})$ plot is shown for saturated profiles at a resolution of 60 km s^{-1} . The broader range of fitted parameters is an indication of hidden saturation at play, resulting in a larger relative error for the column density estimates ($\langle \delta^{\log N} \rangle = 0.05$).

a sample of 650 absorbers, with redshifts in the range of $z = 0.9 - 4.2$ and column densities in the range $11.2 \leq \log(N/\text{cm}^{-2}) \leq 16.3$. These represent our library of absorption line systems, which gathers a large variety of profiles in terms of shape and line widths. Our tools rely on advanced numerical techniques. MC-ALF uses a Bayesian approach to absorption line fitting, which, with minimal human intervention, can decompose metal lines into individual Voigt components providing the posterior distributions of the line parameters such as the column density, the Doppler parameter, and the redshift. Moreover,

as the likelihood space is sampled via the nested sampling algorithm POLYCHORD, MC–ALF is highly efficient in discriminating among competing models for profiles of different complexity (typically related to the instrument resolution and data S/N). Quality assurance tests on simulated UVES-like profiles demonstrate that MC–ALF is able to recover the input information with small relative errors: For the b -parameters, column densities, and redshifts distributions we find mean relative errors of $< \delta \geq 0.03, 0.002, \text{ and } 0.62 \times 10^{-6}$, respectively.

We next showed that NMF methods offer a straightforward statistical framework for physically relevant predictions of non-negative, continuous signals after the data have been properly standardized (Section 4.1). Moreover, as outliers can significantly impact NMF, we build a statistical framework to select the most appropriate bin in ΔV_{90} to perform the fitting. The results are evaluated in terms of residual variance, σ^2 , of the difference between the input profile and its reconstructed counterpart and KS tests among the ΔV_{90} distributions as traced by the input, reconstructed, and simulated data. We then inserted the NMF feature matrix, \mathbf{X} , and coefficient matrix, \mathbf{C} in the NMF–PM python module with which the user can simulate 10^6 metal profiles following a given distribution of desired line width in approximately 10 min (on a one core machine).

Upcoming wide-field surveys, like DESI, 4MOST, and WEAVE, are taking the challenge of observing an unprecedented sample (around a million) of quasar spectra to detail the properties and the evolution of the galaxies' CGM across the Universe. This work aims at contributing to the scientific effort of simulating, testing the detection, and calibrating the observations of metal absorbers in large quasar surveys. In particular, we have shown that our tools will make it possible to reliably simulate, identify and characterize both weak and strong metal absorption lines even in a low-resolution regime. This will, in turn, enable the study of a large sample of lower and higher column density and/or higher redshift systems to resolve small- and large-scale CGM effects and their relation with the surrounding larger-scale environment (e.g. Dutta et al. 2020; Lofthouse et al. 2020, 2023) and to target regions in the Universe at a key epoch for galaxy formation and evolution. On the basis of making our modelling easily accessible to the large astronomical community, we make publicly available MC–ALF and NMF–PM that will allow any user to produce a library of synthetic profiles and analyse them with a simple click of a key. MC–ALF and NMF–PM are available on the github pages provided in the Data Availability section.

ACKNOWLEDGEMENTS

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 757535) and by Fondazione Cariplo (grant No. 2018-2329).

DATA AVAILABILITY

The authors provide the KPI analysis run over the entire sample of data (see Section 4.2.2) as online material, while the two tools MC–ALF and NMF–PM are publicly available at the github pages <https://github.com/matteofox/MC-ALF/> and <https://github.com/alongobardi/NMFPM/>, respectively. Finally, the library of $\sim 10^6$ synthetic WEAVE-like profiles will be shared upon request to the corresponding author.

REFERENCES

- Akaike H., 1974, *IEEE Trans. Autom. Control*, 19, 716
- Anand A., Nelson D., Kauffmann G., 2021, *MNRAS*, 504, 65
- Anand A., Kauffmann G., Nelson D., 2022, *MNRAS*, 513, 3210
- Bernstein R., Shtetman S. A., Gunnels S. M., Mochnacki S., Athey A. E., 2003, in Iye M., Moorwood A. F. M., eds, Proc. SPIE Conf. Ser. Vol. 4841, Instrument Design and Performance for Optical/Infrared Ground-based Telescopes. SPIE, Bellingham, p. 1694
- Bielby R. M. et al., 2019, *MNRAS*, 486, 21
- Carswell R. F., Webb J. K., 2014, Astrophysics Source Code Library, record ascl:1408.015
- Cooke R. J., Pettini M., Jorgenson R. A., Murphy M. T., Steidel C. C., 2014, *ApJ*, 781, 31
- Cooksey K. L., Kao M. M., Simcoe R. A., O'Meara J. M., Prochaska J. X., 2013, *ApJ*, 763, 37
- Cupani G. et al., 2016, in Chiozzi G., Guzman J. C., eds, Proc. SPIE Conf. Ser. Vol. 9913, Software and Cyberinfrastructure for Astronomy IV. SPIE, Bellingham, p. 99131T
- D'Odorico V., Petitjean P., Cristiani S., 2002, *A&A*, 390, 13
- D'Odorico V. et al., 2022, *MNRAS*, 512, 2389
- Dalton G. et al., 2012, in McLean I. S., Ramsay S. K., Takami H., eds, Proc. SPIE Conf. Ser. Vol. 8446, Ground-based and Airborne Instrumentation for Astronomy IV. SPIE, Bellingham, p. 84460P
- Davé R., Hernquist L., Weinberg D. H., Katz N., 1997, *ApJ*, 477, 21
- de Jong R. S., 2012, *Nat. Astron.*, 3, 574
- Dekker H., D'Odorico S., Kaufer A., Delabre B., Kotzłowski H., 2000, in Iye M., Moorwood A. F., eds, Proc. SPIE Conf. Ser. Vol. 4008, Optical and IR Telescope Instrumentation and Detectors. SPIE, Bellingham, p. 534
- DESI Collaboration, 2016, preprint (arXiv:1611.00036)
- Dutta R. et al., 2020, *MNRAS*, 499, 5022
- Dutta R. et al., 2021, *MNRAS*, 508, 4573
- Fontana A., Ballester P., 1995, *The Messenger*, 80, 37
- Fossati M. et al., 2019, *MNRAS*, 490, 1451
- Fumagalli M., Cantalupo S., Dekel A., Morris S. L., O'Meara J. M., Prochaska J. X., Theuns T., 2016, *MNRAS*, 462, 1978
- Fumagalli M., Fotopoulou S., Thomson L., 2020, *MNRAS*, 498, 1951
- Garnett R., Ho S., Bird S., Schneider J., 2017, *MNRAS*, 472, 1850
- Galbati M., Fumagalli M., Fossati M., Lofthouse E. K., Dutta R., Prochaska J. X., Murphy M. T., Cantalupo S., 2023, *MNRAS*, 524, 3474
- Green J. C. et al., 2012, *ApJ*, 744, 60
- Guo Z., Martini P., 2019, *ApJ*, 879, 72
- Handley W. J., Hobson M. P., Lasenby A. N., 2015, *MNRAS*, 453, 4384
- Hurley P. D., Oliver S., Farrah D., Lebouteiller V., Spoon H. W. W., 2014, *MNRAS*, 437, 241
- Jin S. et al., 2023, *MNRAS*, in press
- Kimble R. A. et al., 1998, in Bely P. Y., Breckinridge J. B., eds, Proc. SPIE Conf. Ser. Vol. 3356, Space Telescopes and Instruments V. SPIE, Bellingham, p. 188
- Krogager J.-K., 2018, preprint (arXiv:1803.01187)
- Lan T.-W., Ménard B., Zhu G., 2014, *ApJ*, 795, 31
- Lee D., Seung H. S., 2000, in Leen T., Dietterich T., Tresp V., eds, Advances in Neural Information Processing Systems, Vol. 13. MIT Press
- Liang C., Kravtsov A., 2017, preprint (arXiv:1710.09852)
- Lofthouse E. K. et al., 2020, *MNRAS*, 491, 2057
- Lofthouse E. K. et al., 2023, *MNRAS*, 518, 305
- Lusso E. et al., 2019, *MNRAS*, 485, L62
- Lyke B. W. et al., 2020, *ApJS*, 250, 8
- Mackenzie R. et al., 2019, *MNRAS*, 487, 5070
- Murphy M. T., Kacprzak G. G., Savorgnan G. A. D., Carswell R. F., 2019, *MNRAS*, 482, 3458
- Noterdaeme P., Ledoux C., Petitjean P., Srianand R., 2008, *A&A*, 481, 327
- O'Meara J. M. et al., 2015, *AJ*, 150, 111
- O'Meara J. M., Lehner N., Howk J. C., Prochaska J. X., Fox A. J., Peebles M. S., Tumlinson J., O'Shea B. W., 2017, *AJ*, 154, 114
- Osterman S. et al., 2011, *Ap&SS*, 335, 257
- Prochaska J. X., Herbert-Fort S., Wolfe A. M., 2005, *ApJ*, 635, 123
- Prochaska J. X., O'Meara J. M., Worsack G., 2010, *ApJ*, 718, 392

- Prochaska J. X. et al., 2017, *ApJ*, 837, 169
- Rafelski M., Wolfe A. M., Prochaska J. X., Neeleman M., Mendez A. J., 2012, *ApJ*, 755, 89
- Ren B., Pueyo L., Zhu G. B., Debes J., Duchêne G., 2018, *ApJ*, 852, 104
- Rudie G. C., Steidel C. C., Pettini M., Trainor R. F., Strom A. L., Hummels C. B., Reddy N. A., Shapley A. E., 2019, *ApJ*, 885, 61
- Sargent W. L. W., Steidel C. C., Boksenberg A., 1989, *ApJS*, 69, 703
- Savage B. D., Sembach K. R., 1991, *ApJ*, 379, 245
- Schaye J., Theuns T., Rauch M., Efstathiou G., Sargent W. L. W., 2000, *MNRAS*, 318, 817
- Sheinis A. I., Bolte M., Epps H. W., Kibrick R. I., Miller J. S., Radovan M. V., Bigelow B. C., Sutin B. M., 2002, *PASP*, 114, 851
- Simcoe R. A. et al., 2011, *ApJ*, 743, 21
- Skilling J., 2006, *Bayesian Analysis*, 1, 833
- Spanò P. et al., 2006, in McLean I. S., Iye M., eds, Proc. SPIE Conf. Ser. Vol. 6269, Ground-based and Airborne Instrumentation for Astronomy. SPIE, Bellingham, p. 62692X
- Tripp T. M. et al., 2011, *Science*, 334, 952
- Tumlinson J., Peebles M. S., Werk J. K., 2017, *ARA&A*, 55, 389
- Vernet J. et al., 2011, *A&A*, 536, A105
- Vogelsberger M., Marinacci F., Torrey P., Puchwein E., 2020, *Nat. Rev. Phys.*, 2, 42
- Vogt S. S. et al., 1994, in Crawford D. L., Craine E. R., eds, Proc. SPIE Conf. Ser. Vol. 2198, Instrumentation in Astronomy VIII. SPIE, Bellingham, p. 362
- Werk J. K. et al., 2016, *ApJ*, 833, 54
- Wilde M. C. et al., 2021, *ApJ*, 912, 9
- York D. G. et al., 2000, *AJ*, 120, 1579
- Zhu G., Ménard B., 2013, *ApJ*, 770, 130
- Zou S. et al., 2021, *ApJ*, 906, 32

SUPPORTING INFORMATION

Supplementary data are available at [RASTAI](#) online.

supp_data

Please note: Oxford University Press is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

This paper has been typeset from a \TeX/L\AA\TeX file prepared by the author.