



A comparative study of intervening and associated H I 21-cm absorption profiles in redshifted galaxies

S. J. Curran,¹★ S. W. Duchesne,¹ A. Divoli² and J. R. Allison³

¹*School of Chemical and Physical Sciences, Victoria University of Wellington, PO Box 600, Wellington 6140, New Zealand*

²*Pingar, 55 Anzac Ave, Auckland 1010, New Zealand*

³*CSIRO Astronomy and Space Science, PO Box 76, Epping, NSW 1710, Australia*

Accepted 2016 August 2. Received 2016 July 19; in original form 2016 April 4

ABSTRACT

The star-forming reservoir in the distant Universe can be detected through H I 21-cm absorption arising from either cool gas associated with a radio source or from within a galaxy intervening the sightline to the continuum source. In order to test whether the nature of the absorber can be predicted from the profile shape, we have compiled and analysed all of the known redshifted ($z \geq 0.1$) H I 21-cm absorption profiles. Although between individual spectra there is too much variation to assign a typical spectral profile, we confirm that associated absorption profiles are, on average, wider than their intervening counterparts. It is widely hypothesized that this is due to high-velocity nuclear gas feeding the central engine, absent in the more quiescent intervening absorbers. Modelling the column density distribution of the mean associated and intervening spectra, we confirm that the additional low optical depth, wide dispersion component, typical of associated absorbers, arises from gas within the inner parsec. With regard to the potential of predicting the absorber type in the absence of optical spectroscopy, we have implemented machine learning techniques to the 55 associated and 43 intervening spectra, with each of the tested models giving a $\gtrsim 80$ per cent accuracy in the prediction of the absorber type. Given the impracticability of follow-up optical spectroscopy of the large number of 21-cm detections expected from the next generation of large radio telescopes, this could provide a powerful new technique with which to determine the nature of the absorbing galaxy.

Key words: methods: data analysis – galaxies: active – galaxies: high redshift – galaxies: ISM – quasars: absorption lines – radio lines: galaxies.

1 INTRODUCTION

Cool neutral hydrogen (H I), the raw material for star formation, is traced through the absorption of radio continuum radiation by the atoms undergoing the 21-cm spin-flip transition. The continuum can be intercepted by gas associated with the host galaxy of the radio source or within a galaxy intervening the line of sight to a more distant source.¹ As well as tracing the star formation history of the Universe (e.g. Lagos et al. 2014) back to $z = 0$, observations of H I give insight into the mass assembly and distribution of galaxies (e.g. Rawlings et al. 2004), a means of detecting the Epoch of

Re-ionization (e.g. Carilli et al. 2004), in addition to the potential to obtain highly accurate measurements of the fundamental constants of nature at large look-back times (e.g. Curran, Kanekar & Darling 2004).

The detection of distant galaxies through 21-cm absorption is a science goal of the forthcoming Square Kilometre Array (SKA; Morganti, Sadler & Curran 2015), which, through its large instantaneous bandwidth and large field of view, will avoid observational biases introduced by the conventional requirement of an optical redshift to which to tune the receiver:

(i) For the associated systems, the optical pre-selection biases towards high UV luminosities in the source rest frame, which can be sufficient to ionize all of the neutral gas within the host galaxy (Curran & Whiting 2012), causing the observed paucity of associated 21-cm absorption at high redshift (Curran et al. 2008b, 2011a, 2013a,b, 2016a; Grasha & Darling 2011; Allison et al. 2012; Geréb et al. 2015; Aditya, Kanekar & Kurapati 2016).

(ii) For the intervening absorbers, pre-selection using optical redshift biases against optically obscured sightlines (e.g. Ellison, Hall

* E-mail: scurran.astro@gmail.com

¹ Intervening absorption usually within galaxies which exhibit damped Lyman α absorption, which occurs in the ultraviolet (UV) band and is redshifted into the optical band at $z \gtrsim 1.7$. A damped Lyman α absorber (DLA) is defined as having a neutral hydrogen column density exceeding $N_{\text{HI}} = 2 \times 10^{20}$ atoms per cm^{-2} and DLAs could account for more than 80 per cent of the neutral gas content in the Universe (Prochaska, Herbert-Fort & Wolfe 2005).

& Lira 2005), as well as absorbers rich in molecular gas (Curran et al. 2006, 2011c). These are of particular interest since molecular lines provide excellent probes of the physical and chemical conditions of the gas (e.g. Muller et al. 2013), as well as the potential to obtain accurate measurements of the fundamental constants from a single species (the OH radical), thus eliminating line-of-sight effects which could mimic an apparent change in the constants (Darling 2003).

In order to obtain an unbiased census of the distribution and abundance of the cool neutral gas along each sightline, it is therefore necessary to dispense with the optical pre-selection of targets which has dominated previous 21-cm absorption searches, in favour of using wide instantaneous bandwidths free of frequency interference (RFI). Coverage of the whole redshift space to $z \sim 1$ is already possible with the Australian Square Kilometre Array Pathfinder (Allison et al. 2016), although dispensing with optical spectroscopy does present an obstacle in determining the nature of the absorber (see Allison et al. 2015).

In both the nearby and redshifted active galaxies, 21-cm absorption profiles are often found to be broad ($\gtrsim 150 \text{ km s}^{-1}$; Conway & Blanco 1995), due either to more than one deep component (Conway & Blanco 1995; Mundell et al. 1995; Carilli et al. 1998; Pihlström et al. 1999; Taylor et al. 1999; van Langevelde et al. 2000; Taylor et al. 2002; Morganti et al. 2005, 2008; Morganti, Emonts & Oosterloo 2009) or additional broad, shallow ‘wings’ to either side of the main component (Mirabel 1989; Morganti et al. 2005; Salter et al. 2010; Struve & Conway 2010; Morganti et al. 2011; Allison et al. 2012, 2013). This broadening of the additional shallow component is believed to arise from cold gas in the sub-pc, fast rotating central black hole accretion disc/obscuring torus, invoked by unified schemes of active galactic nuclei (AGN; e.g. Antonucci 1993; Urry & Padovani 1995). The fast rotating/disturbed gas in the narrow-line region can lead to the broadening of the H I profile in AGN (e.g. Holt, Tadhunter & Morganti 2008), whereas its absence in the more quiescent intervening absorbers results in generally narrower and less complex profiles (e.g. Gupta et al. 2009a). In this paper, we investigate the potential of using these differences in the associated and intervening profiles to classify the nature of a newly identified absorption in the absence of complementary optical data. Such a method could provide an invaluable tool in forthcoming spectral scans with the next generation of large radio telescopes.

2 DIFFERENCES IN ASSOCIATED AND INTERVENING 21-CM ABSORPTION SPECTRA

2.1 The sample

As mentioned above, the use of optical redshifts can bias against the detection of 21-cm absorption and this, in conjunction with radio flux limits, the narrow bandwidths free of RFI and the limited frequency coverage of current telescopes, means that the detection of 21-cm absorption at $z \geq 0.1$ (look-back times of $\gtrsim 1$ Gyr) is currently relatively rare, limited to 57 associated and 50 intervening cases (Fig. 1). However, we limit our analysis to these redshifts since we are interested in developing a technique to classify absorbers for future redshifted 21-cm surveys. Exclusion of low redshift absorption also has several advantages.

(i) Minimizing contamination of the sample introduced by differences between the nearby and distant absorption profiles. Even if there are no intrinsic evolutionary differences, the inclusion of spatially resolved absorption/emission, giving several sightlines,

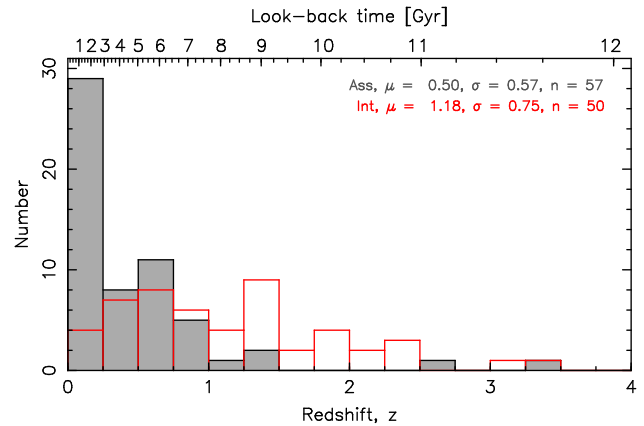


Figure 1. The redshift distribution of detected associated (filled histogram) and intervening (unfilled histogram) $z \geq 0.1$ H I 21-cm absorbers. The mean redshift, μ , and the standard deviation, σ , for each type is shown.

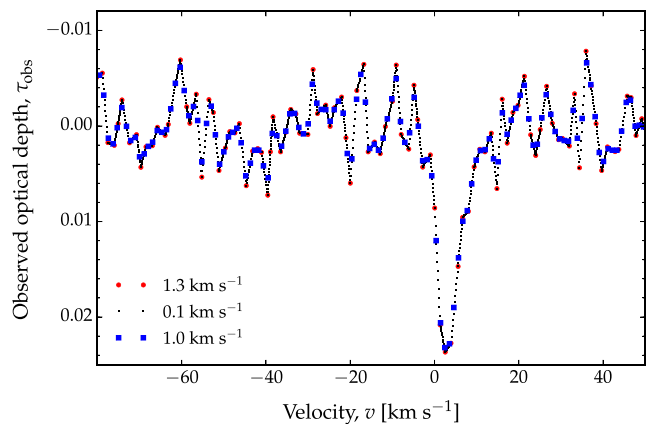


Figure 2. An example of a raw spectrum (with a spectral resolution of 1.3 km s^{-1} , shown as circles) interpolated to 0.1 km s^{-1} (small points) and re-sampled to 1 km s^{-1} (square markers), used for the averaging (Section 2.3).

is expected to confuse the analysis of the unresolved sources of interest.

(ii) Dilution of the absorption profile by 21-cm emission, which is very faint beyond these redshifts (e.g. Catinella & Cortese 2015). This effect will be minimal where the spatial resolution is finest, although, again, the mixture of unresolved and resolved features could add confusion.

(iii) Due, at least in part, to the optical pre-selection of 21-cm absorption searches, nearby intervening absorption is limited to 90 sightlines (see Curran et al. 2016b and reference therein), whereas the nearby associated absorbers have not been compiled. Limiting the analysis to $z \gtrsim 0.1$ ensures approximately equal numbers of associated and intervening absorbers, which is important for class recognition by machine learning models (see Section 3.3).

2.2 The spectra

Unlike at lower redshifts, most of the $z \geq 0.1$ detections are already compiled (in Tables 1 and 2, which are updated from Curran & Whiting 2010 and Curran 2010, respectively). However, the raw data were generally unavailable and so the spectra were acquired from the literature by digitizing the available figures. For this, we

Table 1. The features of the $z \geq 0.1$ associated 21-cm absorbers. The first column gives the IAU name, followed by the mean-weighted absorption redshift (see Section 2.3) and the number of Gaussian components required to fit the spectrum, n_g . The following columns give the full width at zero intensity, the peak observed optical depths, the average offset of the components from z_{weight} and the full width at half-maxima, respectively (see Fig. 7). The last column gives the reference for the 21-cm absorption.

| IAU | z_{weight} | n_g | FWZI | τ_{peak} | | | $\overline{\Delta v}$ (km s $^{-1}$) | $\overline{\Delta v}/\text{FWZI}$ (km s $^{-1}$) | FWHM (km s $^{-1}$) | | | Ref. |
|------------|---------------------|-------|------|---|-----------|-----------|--|--|----------------------|-----|-----|------|
| | | | | Ave | Max | Min | | | Ave | Max | Min | |
| B0023–26 | 0.321 409 | 2 | 482 | 0.005 73 | 0.009 44 | 0.002 01 | –39 | –0.081 | 149 | 173 | 126 | V03 |
| B0108+388 | 0.668 475 | 2 | 426 | 0.0423 | 0.0516 | 0.0329 | –5 | –0.011 | 107 | 139 | 76 | O06 |
| J0141+1353 | 0.620 390 | 2 | 92 | 0.0155 | 0.0169 | 0.0141 | 3 | 0.031 | 21 | 35 | 8 | V03 |
| B0316+16 | 0.906 947 | 3 | 243 | 0.0131 | 0.0273 | 0.002 26 | –7 | –0.031 | 42 | 95 | 10 | S10 |
| J0410+7656 | 0.598 867 | 3 | 785 | 0.006 56 | 0.0141 | 0.002 68 | –154 | –0.20 | 80 | 104 | 61 | V03 |
| J0414+0534 | 2.636 414 | 2 | 575 | 0.0131 | 0.0151 | 0.0111 | 29 | 0.051 | 147 | 154 | 141 | M99 |
| B0428+20 | 0.220 358 | 2 | 951 | 0.003 37 | 0.004 67 | 0.002 07 | 69 | 0.073 | 283 | 299 | 268 | V03 |
| B0500+019 | 0.584 693 | 2 | 224 | 0.0288 | 0.036 | 0.0216 | –16 | –0.071 | 54 | 63 | 45 | C98 |
| B0758+143 | 1.194 147 | 5 | 686 | 0.004 61 | 0.009 44 | 0.002 68 | 25 | 0.036 | 112 | 188 | 63 | I03 |
| J0834+5534 | 0.240 669 | 1 | 443 | 0.002 81 | 0.002 81 | 0.002 81 | 0 | – | 202 | 202 | 202 | V03 |
| B0839+458 | 0.192 255 | 2 | 325 | 0.178 | 0.299 | 0.0574 | –18.4 | –0.057 | 79 | 84 | 74 | G15 |
| B0859+032 | 0.288 116 | 2 | 547 | 0.0474 | 0.0560 | 0.0387 | 14 | 0.025 | 73 | 92 | 54 | Y16 |
| J0901+2901 | 0.193 870 | 1 | 179 | 0.000 53 | 0.000 53 | 0.000 53 | 0 | – | 117 | 117 | 117 | V03 |
| B0902+34 | 3.396 779 | 1 | 700 | 0.008 49 | 0.008 49 | 0.008 49 | 0 | – | 277 | 277 | 277 | U91 |
| J0909+4253 | 0.670 341 | 2 | 276 | 0.006 47 | 0.0102 | 0.002 73 | 55 | 0.20 | 61 | 101 | 20 | V03 |
| B20917+27B | 0.206 698 | 2 | 96 | 0.0612 | 0.0819 | 0.0405 | –0.9 | –0.0089 | 230 | 33 | 13 | Y16 |
| J0942+0623 | 0.123 206 | 3 | 191 | 0.723 | 0.896 | 0.381 | –13 | –0.068 | 30 | 45 | 14 | S15 |
| B1003+35 | 0.099 742 | 3 | 558 | 0.0139 | 0.0203 | 0.005 06 | 42 | 0.075 | 117 | 203 | 65 | V89 |
| B1107–187 | 0.491 705 | 2 | 76 | 0.0161 | 0.0189 | 0.0133 | 0 | 0.0046 | 19 | 22 | 17 | C11a |
| J1120+2736 | 0.111 720 | 1 | 221 | 0.159 | 0.159 | 0.159 | 0 | – | 67 | 67 | 67 | G15 |
| J1124+1919 | 0.165 161 | 6 | 77 | 0.0407 | 0.101 | 0.005 49 | 12 | 0.15 | 9 | 13 | 7 | G06 |
| B1147+557 | 0.138 297 | 3 | 234 | 0.0217 | 0.0494 | 0.0204 | 9 | 0.038 | 34 | 42 | 22 | C11 |
| B1126+569 | 0.891 604 | 2 | 464 | 0.0170 | 0.0276 | 0.0063 | –25 | –0.053 | 137 | 170 | 104 | Y16 |
| B1142+052 | 1.343 073 | 3 | 157 | 0.005 24 | 0.007 07 | 0.003 52 | 6 | 0.037 | 31 | 39 | 15 | K09 |
| J1202+1637 | 0.118 568 | 2 | 622 | 0.0213 | 0.0279 | 0.0147 | 6 | 0.0091 | 179 | 231 | 127 | G15 |
| B1203+645 | 0.371 883 | 1 | 405 | 0.003 85 | 0.003 85 | 0.003 85 | 0 | – | 177 | 177 | 177 | V03 |
| B1206+469 | – | – | – | Could not be fit/spectrum of too poor quality | | | | | | | – | G15 |
| B1244+49 | 0.205 956 | 4 | 686 | 0.000 591 | 0.000 657 | 0.000 463 | 22 | 0.032 | 192 | 260 | 119 | G15 |
| J1254+1856 | – | – | – | Could not be fit/spectrum of too poor quality | | | | | | | – | G15 |
| J1301+4634 | 0.205 041 | 3 | 901 | 0.007 77 | 0.0108 | 0.004 65 | 7 | 0.0081 | 198 | 298 | 111 | G15 |
| J1326+3154 | 0.368 430 | 1 | 459 | 0.001 70 | 0.001 70 | 0.001 70 | 0 | – | 227 | 227 | 227 | V03 |
| J1347+1217 | 0.121 924 | 3 | 336 | 0.004 93 | 0.0107 | 0.0018 | 58 | 0.150 | 97 | 130 | 23 | G06 |
| B1355+441 | 0.645 449 | 1 | 1085 | 0.0537 | 0.0537 | 0.0537 | 0 | – | 359 | 359 | 359 | V03 |
| J1357+0046 | 0.796 663 | 2 | 295 | 0.0109 | 0.013 | 0.0088 | 25 | 0.085 | 76 | 80 | 71 | Y16 |
| J1400+6210 | 0.430 137 | 1 | 411 | 0.006 11 | 0.006 11 | 0.006 11 | 0 | – | 169 | 169 | 169 | V03 |
| J1409+3604 | 0.148 418 | 2 | 249 | 0.034 | 0.0373 | 0.0317 | –10 | –0.038 | 57 | 66 | 50 | C11 |
| B1413+135 | 0.246 079 | 2 | 72 | 0.0288 | 0.036 | 0.0216 | –16 | –0.22 | 54 | 63 | 45 | C92 |
| J1422+2105 | 0.190 425 | 1 | 544 | 0.0434 | 0.0434 | 0.0434 | 0 | – | 183 | 183 | 183 | G15 |
| B1504+37 | 0.672 634 | 3 | 258 | 0.205 | 0.335 | 0.0759 | –8 | –0.031 | 45 | 85 | 17 | C98 |
| B1543+480 | 1.277 005 | 2 | 532 | 0.0425 | 0.0533 | 0.0317 | –20 | –0.038 | 109 | 129 | 89 | C13 |
| B1549–79 | 1.518 581 | 3 | 330 | 0.009 37 | 0.0174 | 0.004 39 | 12 | 0.035 | 72 | 140 | 34 | M01 |
| B1601+5252 | 0.105 545 | 1 | 243 | 0.009 74 | 0.009 74 | 0.009 74 | 0 | – | 95 | 95 | 95 | C11 |
| B1603+609 | 0.559 129 | 1 | 21 | 0.0142 | 0.0142 | 0.0142 | 0 | – | 8 | 8 | 8 | Y16 |
| B1614+26 | 0.755 466 | 1 | 1129 | 0.0084 | 0.0084 | 0.0084 | 0 | – | 447 | 447 | 447 | Y16 |
| B1649–062 | 0.236 387 | 1 | 503 | 0.0238 | 0.0238 | 0.0238 | 0 | – | 179 | 179 | 179 | C11b |
| B1717+547 | 0.147 402 | 2 | 351 | 0.0276 | 0.0326 | 0.0225 | –7 | –0.020 | 100 | 149 | 51 | C11 |
| B1814+34 | 0.243 994 | 1 | 231 | 0.0396 | 0.0396 | 0.0396 | 0 | – | 79 | 79 | 79 | P00 |
| J1815+6127 | 0.596 812 | 1 | 327 | 0.0207 | 0.0207 | 0.0207 | 0 | – | 118 | 118 | 118 | V03 |
| J1821+3942 | 0.795 323 | 2 | 259 | 0.008 84 | 0.01 | 0.007 68 | –2 | –0.0077 | 53 | 61 | 44 | V03 |
| J1944+5448 | 0.258 259 | 1 | 795 | 0.008 64 | 0.008 64 | 0.008 64 | 0 | – | 313 | 313 | 313 | V03 |
| J1945+7055 | 0.100 209 | 1 | 306 | 0.55 | 0.55 | 0.55 | 0 | – | 87 | 87 | 87 | P99 |
| B2050+36 | 0.354 67 | 2 | 109 | 0.125 | 0.204 | 0.0462 | –6 | –0.055 | 24 | 32 | 16 | V03 |
| B2121+248 | 0.107 209 | 3 | 527 | 0.002 17 | 0.004 32 | 0.000 999 | 1 | 0.002 | 124 | 200 | 70 | M89 |
| B2252–089 | 0.607 037 | 2 | 286 | 0.132 | 0.181 | 0.0839 | 0.028 8 | – | 55 | 92 | 18 | C11a |
| J2255+1313 | 0.543 101 | 1 | 281 | 0.001 62 | 0.001 62 | 0.001 62 | 0 | – | 140 | 140 | 140 | V03 |
| J2316+0405 | 0.219 135 | 1 | 289 | 0.003 05 | 0.003 05 | 0.003 05 | 0 | – | 130 | 130 | 130 | V03 |
| J2355+4950 | 0.237 905 | 2 | 272 | 0.0145 | 0.0176 | 0.0114 | 59 | 0.22 | 47 | 81 | 13 | V03 |

Notes. References: M89 – Mirabel (1989), V89 – van Gorkom et al. (1989), U91 – Uson, Bagri & Cornwell (1991), C92 – Carilli, Perlman & Stocke (1992), C98 – Carilli et al. (1998), M99 – Moore, Carilli & Menten (1999), P99 – Peck, Taylor & Conway (1999), P00 – Peck et al. (2000), M01 – Morganti et al. (2001), I03 – Ishwara-Chandra, Dwarakanath, & Anantharamaiah (2003), V03 – Vermeulen et al. (2003), G06 – Gupta et al. (2006), O06 – Orienti, Morganti & Dallacasa (2006), K09 – Kanekar et al. (2009), S10 – Salter et al. (2010), C11 – Chandola, Sirothia & Saikia (2011), C11a – Curran et al. (2011a), C11b – Curran et al. (2011b), C13 – Curran et al. (2013b), G15 – Geréb et al. (2015), S15 – Srianand et al. (2015), Y16 – Yan et al. (2016).

Table 2. As Table 1 but for the $z \geq 0.1$ intervening 21-cm absorbers.

| IAU | z_{weight} | n_g | FWZI | Ave | τ_{peak} Max | Min | $\overline{\Delta v}$ (km s ⁻¹) | $\overline{\Delta v}/\text{FWZI}$ (km s ⁻¹) | FWHM (km s ⁻¹) | | | Ref. |
|------------|---------------------|-------|------|----------|-----------------------------|----------|--|--|----------------------------|-----|------|------|
| | | | | | | | | | Ave | Max | Min | |
| J0108–0037 | 1.370 985 | 1 | 52 | 0.0731 | 0.0731 | 0.0731 | 0 | – | 17 | 17 | 17 | G09b |
| B0132–097 | 0.764 436 | 2 | 519 | 0.0294 | 0.0299 | 0.0288 | 0 | –0.069 | 110 | 145 | 74 | K03a |
| B0201+113 | 3.386 789 | 2 | 143 | 0.0129 | 0.017 | 0.0088 | –2 | –0.015 | 29 | 37 | 22 | K14a |
| B0218+35 | 0.684 651 | 1 | 143 | 0.0454 | 0.0454 | 0.0454 | 0 | – | 49 | 49 | 49 | C93 |
| B0235+164 | – | – | | | | | Could not be fit/spectrum of too poor quality | | | | | R76 |
| B0237–233 | – | – | | | | | Could not be fit/spectrum of too poor quality | | | | | K09 |
| B0248+430 | 0.394 153 | 3 | 49 | 0.175 | 0.222 | 0.126 | 0 | 0.0041 | 6 | 7 | 4 | L01 |
| B0311+430 | 2.289 521 | 2 | 145 | 0.0101 | 0.0146 | 0.0055 | –14 | –0.099 | 37 | 40 | 34 | Y07 |
| J0414+0534 | 0.959 790 | 3 | 127 | 0.0113 | 0.0174 | 0.003 48 | 13 | 0.11 | 26 | 31 | 19 | C07a |
| B0438–436 | 2.347 469 | 2 | 83 | 0.004 52 | 0.0059 | 0.00 314 | 7 | 0.087 | 21 | 24 | 18 | K14a |
| B0454–234 | 0.891 324 | 1 | 39 | 0.0130 | 0.0130 | 0.0130 | 0 | – | 15 | 15 | 15 | G12 |
| B0458–020 | 1.560 516 | 1 | 17 | 0.0233 | 0.0233 | 0.0233 | 0 | – | 7 | 7 | 7 | K09 |
| – | 2.039 484 | 2 | 54 | 0.237 | 0.334 | 0.139 | 4 | 0.069 | 12 | 13 | 10 | K14a |
| B0738+313 | 0.220 999 | 2 | 14 | 0.0475 | 0.0634 | 0.0316 | 0 | 0.0071 | 4 | 5 | 3 | K01b |
| J0804+3012 | 1.190 890 | 2 | 198 | 0.002 97 | 0.003 92 | 0.002 01 | –9 | –0.045 | 66 | 89 | 43 | G09b |
| J0808+4950 | 1.407 309 | 1 | 28 | 0.007 66 | 0.007 66 | 0.007 66 | 0 | – | 11 | 11 | 11 | G09b |
| B0809+483 | 0.436 899 | 2 | 347 | 0.0101 | 0.0167 | 0.003 54 | 47 | 0.13 | 56 | 72 | 39 | B01 |
| B0827+243 | 0.524 762 | 1 | 89 | 0.005 78 | 0.005 78 | 0.005 78 | 0 | – | 37 | 37 | 37 | K01 |
| J0849+5108 | 0.311 991 | 2 | 43 | 0.0415 | 0.0602 | 0.0302 | –5 | –0.11 | 9 | 12 | 5 | G13 |
| J0850+5159 | 1.326 818 | 3 | 150 | 0.274 | 0.441 | 0.126 | –1 | –0.0053 | 28 | 47 | 13 | G09b |
| J0852+3435 | 1.309 508 | 2 | 193 | 0.0868 | 0.0885 | 0.0851 | –8 | –0.040 | 42 | 62 | 22 | G09b |
| B0927+469 | 0.621 550 | 1 | 22 | 0.0323 | 0.0323 | 0.0323 | 0 | – | 9 | 9 | 9 | Z15 |
| B0952+179 | 0.237 808 | 1 | 19 | 0.013 | 0.013 | 0.013 | 0 | – | 8 | 8 | 8 | K01a |
| B1055+499 | 1.211 757 | 2 | 69 | 0.0121 | 0.0188 | 0.005 38 | –4 | –0.059 | 21 | 25 | 16 | G09b |
| B1127–145 | 0.313 012 | 7 | 123 | 0.0522 | 0.127 | 0.003 27 | 1 | 0.0070 | 111 | 15 | 6 | C00 |
| B1157+014 | 1.943 628 | 2 | 33 | 0.0564 | 0.0698 | 0.0429 | 2 | 0.053 | 8 | 8 | 7 | K14a |
| B1229–0207 | – | – | | | | | Could not be fit/spectrum of too poor quality | | | | | L01 |
| B1243–072 | 0.437 217 | 1 | 39 | 0.0684 | 0.0684 | 0.0684 | 0 | – | 13 | 13 | 13 | L01 |
| B1252+4427 | 0.911 272 | 3 | 229 | 0.0183 | 0.0277 | 0.009 95 | 2 | 0.0083 | 41 | 61 | 27 | G12 |
| B1328+307 | 0.692 150 | 1 | 8 | 0.093 | 0.093 | 0.093 | 0 | – | 9 | 9 | 9 | D78 |
| B1331+17 | – | – | | | | | Could not be fit/spectrum of too poor quality | | | | | B83 |
| J1337+3152 | 3.174 47 | 1 | 17 | 0.611 | 0.611 | 0.611 | 0 | – | 5 | 5 | 5 | S12 |
| B1406–076 | 1.274 564 | 1 | 71 | 0.0160 | 0.0303 | 0.006 01 | –1 | –0.0089 | 11 | 12 | 10 | G12 |
| B1430–178 | 1.326 455 | 2 | 119 | 0.001 58 | 0.002 15 | 0.001 01 | –6 | –0.054 | 37 | 57 | 16 | K09 |
| J1431+3952 | 0.601 849 | 2 | 39 | 0.226 | 0.227 | 0.224 | 1 | 0.026 | 7 | 8 | 6 | E12 |
| J1443+0214 | 0.371 540 | 2 | 42 | 0.256 | 0.402 | 0.109 | 0 | –0.0060 | 10 | 14 | 6 | G13 |
| B1621+047 | 1.335 695 | 2 | 65 | 0.0322 | 0.0413 | 0.0231 | 5 | 0.068 | 15 | 17 | 12 | G09a |
| B1622+238 | 0.655 943 | 1 | 500 | 0.008 77 | 0.008 77 | 0.008 77 | 0 | – | 233 | 233 | 23 | C07b |
| B1629+120 | 0.531 764 | 2 | 44 | 0.0211 | 0.0261 | 0.0161 | 1 | 0.027 | 13 | 21 | 6 | K03b |
| B1755+758 | 1.970 875 | 1 | 137 | 0.0227 | 0.0227 | 0.0227 | 0 | – | 49 | 49 | 49 | K14b |
| B1830–21 | 0.885 469 | 2 | 574 | 0.0338 | 0.0368 | 0.0307 | –18 | –0.032 | 132 | 197 | 67 | C99 |
| – | 0.192 504 | 3 | 124 | 0.0116 | 0.0134 | 0.009 54 | 2 | 0.017 | 24 | 28 | 15 | L96 |
| B1850+402 | 1.989 599 | 2 | 88 | 0.0784 | 0.0956 | 0.0612 | 5 | 0.058 | 18 | 20 | 16 | K14b |
| B2003–025 | 1.410 732 | 2 | 64 | 0.004 98 | 0.005 87 | 0.004 08 | –5 | –0.082 | 18 | 24 | 12 | K09 |
| B2029+121 | 1.115 735 | 2 | 266 | 0.0156 | 0.0186 | 0.0125 | –3 | –0.012 | 70 | 78 | 61.4 | G12 |
| B2039+187 | 2.191 798 | 1 | 38 | 0.0320 | 0.0320 | 0.0320 | 0 | – | 13 | 13 | 13 | K12 |
| J2340–0053 | 1.360 890 | 3 | 6 | 0.400 | 0.82 | 0.182 | 0 | 0.022 | 3 | 6 | 1 | G09b |
| B2351+456 | – | – | | | | | Could not be fit/spectrum of too poor quality | | | | | D04 |
| B2355–106 | 1.173 038 | 1 | 17 | 0.0333 | 0.0333 | 0.0333 | 0 | – | 6 | 6 | 6 | G09b |

Notes. References: R76 – Roberts et al. (1976), D78 – Davis & May (1978), B83 – Briggs & Wolfe (1983), C93 – Carilli, Rupen & Yanny (1993), L96 – Lovell et al. (1996), C99 – Chengalur, de Bruyn & Narasimha (1999), B01 – Briggs, de Bruyn & Vermeulen (2001), L01 – Lane & Briggs (2001), K01a – Kanekar & Chengalur (2001), K01b – Kanekar, Ghosh & Chengalur (2001), K03a – Kanekar & Briggs (2003), K03b – Kanekar & Chengalur (2003), D04 – Darling et al. (2004), C07a – Curran et al. (2007a), C07b – Curran et al. (2007b), Y07 – York et al. (2007), G09a – Gupta et al. (2009b), G09b – Gupta et al. (2009a), K09 – Kanekar et al. (2009), E12 – Ellison et al. (2012), G12 – Gupta et al. (2012), K12 – Kanekar et al. (2013), S12 – Srianand et al. (2012), G13 – Gupta et al. (2013), K14a – Kanekar et al. (2014), K14b – Kanekar (2014), Z15 – Zwaan et al. (2015).

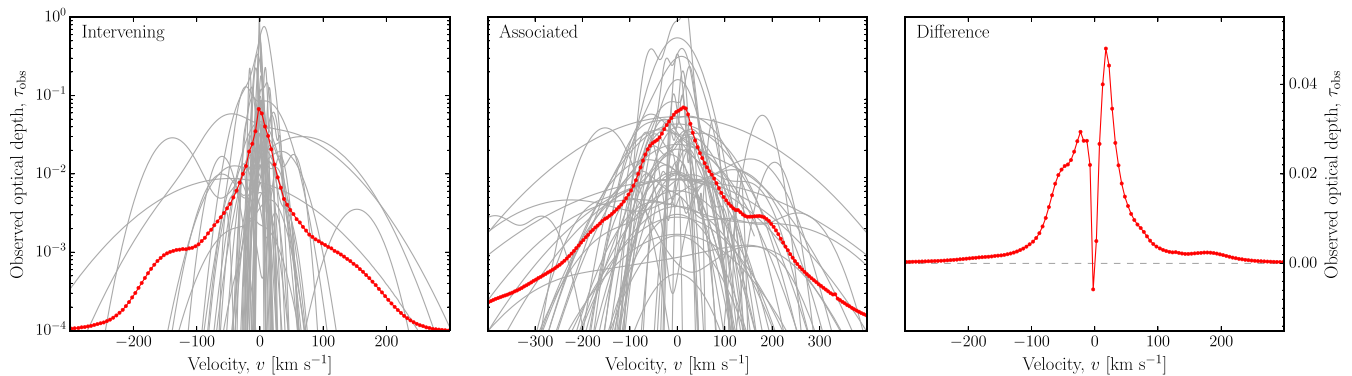


Figure 3. The intervening (left), associated (middle) spectra and the mean associated minus the mean intervening spectrum (right, demonstrating the additional ‘wings’). The thin traces show the individual spectra and the thick traces the averages of these.

used the GETDATA GRAPH DIGITIZER² package for all the spectra, except those in Srikanth et al. (2015) and Yan et al. (2016), which were constructed from Gaussian parameters presented. This process was successful for 55 associated and 43 intervening spectra, with unsuccessful acquisition resulting from noisy data or low-quality figures. The axes were then normalized by converting the ordinate to observed optical depth (Section 2.3) and the abscissa to velocity dispersion, which was defined relative to the optical depth weighted mean velocity of the absorption profile, v_{wm} . For a spectrum sampled over i components, this is

$$v_{\text{wm}} = \frac{\sum_i \tau_i v_i}{\sum_i \tau_i}, \quad (1)$$

which is used to shift the abscissa so that $v_{\text{wm}} = 0$.

In order to allow the spectra to be inter-compared and averaged, each was oversampled and then re-binned to a common spectral resolution (see Fig. 2). Since the spectra were shown over a variety of velocity ranges, generally proportional to the linewidth, in order to ensure that the full mean velocity range was evenly weighted, we added the typical noise level of $\tau_{\text{obs}} \sim 10^{-4}$ to each end of each spectrum to give each the same velocity range.

Half of the original spectra were presented as single or multiple Gaussian fits and so, in order to include these, Gaussian fits were applied to all of the oversampled spectra. Representing the spectra by Gaussians was also useful in parametrizing the spectra for comparison through machine learning (Section 3.2.1). Generally, the number of Gaussian components quoted by the authors was used and when this was not given, we used the minimum number necessary for a fit. Fitting was done using the FITYK 0.9.8³ package which utilizes the Levenberg–Marquardt algorithm, a least-squares fitting routine developed in particular for non-linear fitting.

In Fig. 3, we show the digitized spectra,⁴ from which we see a large variation in profile shapes, although the intervening tend to have narrower velocity dispersions. Highlighting this, in Fig. 4, we show the distribution of the effective profile widths of the individual spectra, from which we see that the associated are, on average, three times wider than the intervening profiles. A Kolmogorov–Smirnov test gives a probability of $P(\text{KS}) = 4.17 \times 10^{-10}$ that the

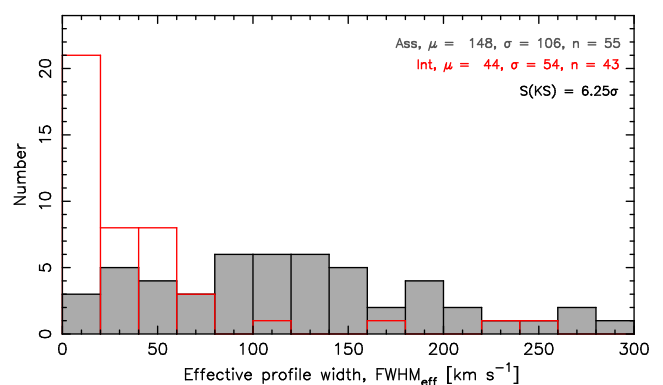


Figure 4. The effective profile width, for the associated (filled histogram) and the intervening (unfilled histogram) absorbers. This is defined by $\text{FWHM}_{\text{eff}} = \int \tau dv / \tau_{\text{peak}}$ (e.g. Dickey 1982; Allison et al. 2013), since for complex profiles (multiple or non-Gaussian), the full width at half-maxima are difficult to measure in a consistent manner.

associated and intervening velocity distributions are drawn from the same population, which is significant at $S(\text{KS}) = 6.25\sigma$, assuming Gaussian statistics.

It is therefore clear that the profile width is related to the absorption type: the narrower profile widths of the intervening absorbers being possibly due to the absence of the high-dispersion component, in addition to the possibility that intervening absorption may be more favourably detected in galactic discs of high inclination (close to face-on; see Curran et al. 2016b). However, the associated absorbers are seen to be roughly equally distributed over the whole range of widths, which makes it difficult to predict the nature of the absorption based upon the full width at half-maximum (FWHM) distribution alone.

2.3 Mean spectral properties

Although the profile widths of the individual spectra are too varied to effectively predict the absorber type, we can average the spectra in order to explore any strong statistical difference between the types. Since the peak optical depth is not necessarily the centroid of the absorption, we average the spectra by the mean-weighted absorption, where the zero-velocity offset is defined by the median of the velocity-integrated optical depth (equation 1), giving a more consistent measure of redshift (e.g. Tzanavaris et al. 2007).

² <http://www.getdata-graph-digitizer.com/>

³ <http://fityk.nieto.pl/>

⁴ These are presented individually in Duchesne (2015) and machine readable versions of these will be made available in a forthcoming online catalogue of the properties of the known redshifted H I 21-cm absorbers.

From the mean spectra (shown by the thick traces in Fig. 3), it is clear the associated spectrum has additional low-optical depth gas at large-velocity dispersions ($|\Delta V| \gtrsim 200 \text{ km s}^{-1}$). As discussed in Section 1, it has been hypothesized that this is due to additional fast moving neutral gas close to the nucleus, probably associated with the obscuring torus/accretion disc. In order to model this additional absorption, as a starting point in Fig. 5 (top left), we show the mean profiles where the blue and redshifted components have been averaged. We then convert each optical depth–dispersion distribution to a column density–radial distribution, by quantifying how the velocity of the gas is expected to vary with galactocentric radius. In Fig. 5 (top right), we fit a polynomial to the velocity distribution of the Milky Way (Bhattacharjee et al. 2014). Being a large spiral, this may not represent an accurate depiction of the large-scale rotation curve of an AGN host. However, although early-type galaxies exhibit a variety of rotation curves, many exhibit similar curves to that of the Milky Way (Noordermeer et al. 2007). Furthermore, since we are interested in comparing the associated and intervening absorbers (which themselves may arise in a wide variety of galaxy types; e.g. Wolfe et al. 1986; Matteucci, Molaro & Vladilo 1997; Prochaska & Wolfe 1997; Haehnelt, Steinmetz & Rauch 1998; Jimenez, Bowen & Matteucci 1999), the Milky Way provides a classic example of the rapid velocity increase within the central $\lesssim 100 \text{ pc}$, before reaching velocities of $200\text{--}300 \text{ km s}^{-1}$ at $\gtrsim 100 \text{ pc}$ (Noordermeer et al. 2007).

Since the rotation curve of the Milky Way is only well mapped beyond $r \gtrsim 200 \text{ pc}$, we supplement this with data from the Circinus galaxy, a nearby spiral in which the rotation curve at inner radii is readily available. Circinus is known to host a Seyfert 2 nucleus (Moorwood & Oliva 1990; Oliva et al. 1994) and so may at least provide a reasonable model of the inner regions of the associated absorbers. In order to match the velocities between the two galaxies, each of the Circinus values have been upsampled by a factor of 1.7, which is close to the value expected based upon the scaling ratio between the nuclear black hole and host galaxy mass (e.g. Ferrarese & Merritt 2000; Bennert et al. 2015).⁵

Once the velocity is mapped, the column density is obtained from

$$N_{\text{H I}} = 1.823 \times 10^{18} T_{\text{spin}} \int \tau dv, \quad (2)$$

where T_{spin} is the spin temperature of the gas, which is a measure of the excitation from the lower hyperfine level (Purcell & Field 1956; Field 1959), and $\int \tau dv$ is the velocity-integrated optical depth of the absorption. The observed optical depth is related to this via

$$\tau \equiv -\ln\left(1 - \frac{\tau_{\text{obs}}}{f}\right) \approx \frac{\tau_{\text{obs}}}{f}, \quad \text{for } \tau_{\text{obs}} \equiv \frac{\Delta S}{S_{\text{obs}}} \lesssim 0.3, \quad (3)$$

where the covering factor, f , is a measure of the fraction of observed background flux (S_{obs}) intercepted by the absorber. In the optically thin regime (where $\tau_{\text{obs}} \lesssim 0.3$), equation (2) can be rewritten as $N_{\text{H I}} \approx 1.06 \times 1.823 \times 10^{18} (T_{\text{spin}}/f) \tau_{\text{peak}} \Delta V$, where ΔV is the dispersion of the absorption. Assuming that the peak of the absorption occurs, on average, in the centre of the galaxy and that the gas is dynamically coupled to the sub-kpc rotation, the dispersion is related to the rotational velocity via $\Delta V = v_{\text{rot}} \cos i$. To obtain the

column density, we assume a constant $T_{\text{spin}} = 500 \text{ K}$ across the disc, since this is a constant $250\text{--}400 \text{ K}$ across the Milky Way (Dickey et al. 2009) and a constant $T_{\text{spin}}/f \sim 1000 \text{ K}$ ($T_{\text{spin}} \lesssim 1000 \text{ K}$) across external galaxies (Curran et al. 2016b), with a mean of $T_{\text{spin}}/f \approx 2000 \text{ K}$ ($T_{\text{spin}} \lesssim 2000 \text{ K}$) at higher redshift (in DLAs). Higher spin temperatures close to the AGN would lead to higher column densities in the associated absorbers and so those derived (Fig. 5, bottom left) should be considered lower limits. For the covering factor, we assume full coverage ($f = 1$) for the mean face-on ($i = 90^\circ$) disc, so that $\tau = \tau_{\text{obs}}/\sin^2 i$ (Curran 2012).

Assuming that the gas remains sufficiently cool and neutral to exhibit detectable 21-cm absorption within the inner $\sim 1 \text{ pc}$, where the kinematics are dominated by Keplerian rotation around a massive compact object (the central black hole), this simple model does indeed suggest that additional high-dispersion gas in the associated absorbers arises from a central component, rather than being dominated by orientation effects (Section 2.1).

Furthermore, the high column densities for the low-inclination associated absorbers are consistent with what we expect from the Milky Way, in which the volume density of the neutral gas exhibits an exponential decrease with galactocentric radius according to $n = n_0 e^{-r/R}$, where $n_0 = 13.4 \text{ cm}^{-3}$ and the scalelength $R = 3 \text{ kpc}$ (Kalberla et al. 2007). From $N_{\text{H I}} \equiv \int n dl$, the column density has a maximum at $i = 0^\circ$, given by $N_{\text{H I}} = n_0 \int_0^\infty e^{-r/R} dr = n_0 R = 1.2 \times 10^{23} \text{ cm}^{-2}$, which remains high to large radii, since this is the total volume density-integrated over the path length.

2.4 Redshift evolution

In addition to any possible differences in the intervening and associated profiles due to the presence of gas in close proximity to the AGN, it is also possible that differences could be introduced by redshift evolution. Unfortunately, the associated absorbers are predominantly at $z \lesssim 1$, due to higher redshifts preferentially selecting the most UV luminous sources (Section 1). For the intervening absorbers, however, the sample is split in half at $z \sim 1$, which, at a look-back time of $\text{LBT} \approx 8 \text{ Gyr}$, is close to half the age of the Universe, which allows us to compare the profiles between these two epochs.

From the distribution of profile width with redshift (Fig. 6), we see no evidence of any evolution. If the thermal broadening is comparable to that introduced by the gas kinematics (Section 2.3), this would suggest no mean evolution in the kinetic temperature of the gas, T_{kin} . The intervening absorbers are not subject to the same broadening mechanisms as the associated absorbers, where gas kinematics, turbulence and radiative heating can be significant, leading to possible line broadening (see Fig. 4). That is, in thermodynamic equilibrium the spin temperature is coupled to the kinetic temperature (e.g. Lane & Briggs 2001; Roy, Chengalur & Srianand 2006), which is given by

$$T_{\text{kin}} \leq \frac{\Delta V^2}{8 \ln(2)} \frac{m_{\text{H}}}{k_{\text{B}}} [\text{K}] \lesssim 22 \Delta V^2 [\text{K for } \Delta V \text{ in km s}^{-1}],$$

where m_{H} is the mass of the hydrogen atom and k_{B} is the Boltzmann constant. From the binned line-widths (Fig. 6), we obtain $T_{\text{kin}} \lesssim 22000 \text{ K}$ at LBTs $\lesssim 8 \text{ Gyr}$ and $T_{\text{kin}} \lesssim 9000 \text{ K}$ at LBTs $\gtrsim 8 \text{ Gyr}$. The limit arises due to other possible broadening mechanisms, although the absence of any increase in profile width with redshift does not support the argument that there is an increase in temperature with redshift (Kanekar & Chengalur 2003, where the

⁵ For the Milky Way, $M_{\text{BH}} \approx 4 \times 10^6 M_{\odot}$ (e.g. Reid & Brunthaler 2004), cf. $1.7 \times 10^6 M_{\odot}$ for Circinus (Greenhill et al. 2003). Assuming circular orbits, this gives $v_{\text{MW}}/v_{\text{Circinus}} \approx \sqrt{4/1.7} \approx 1.5$.

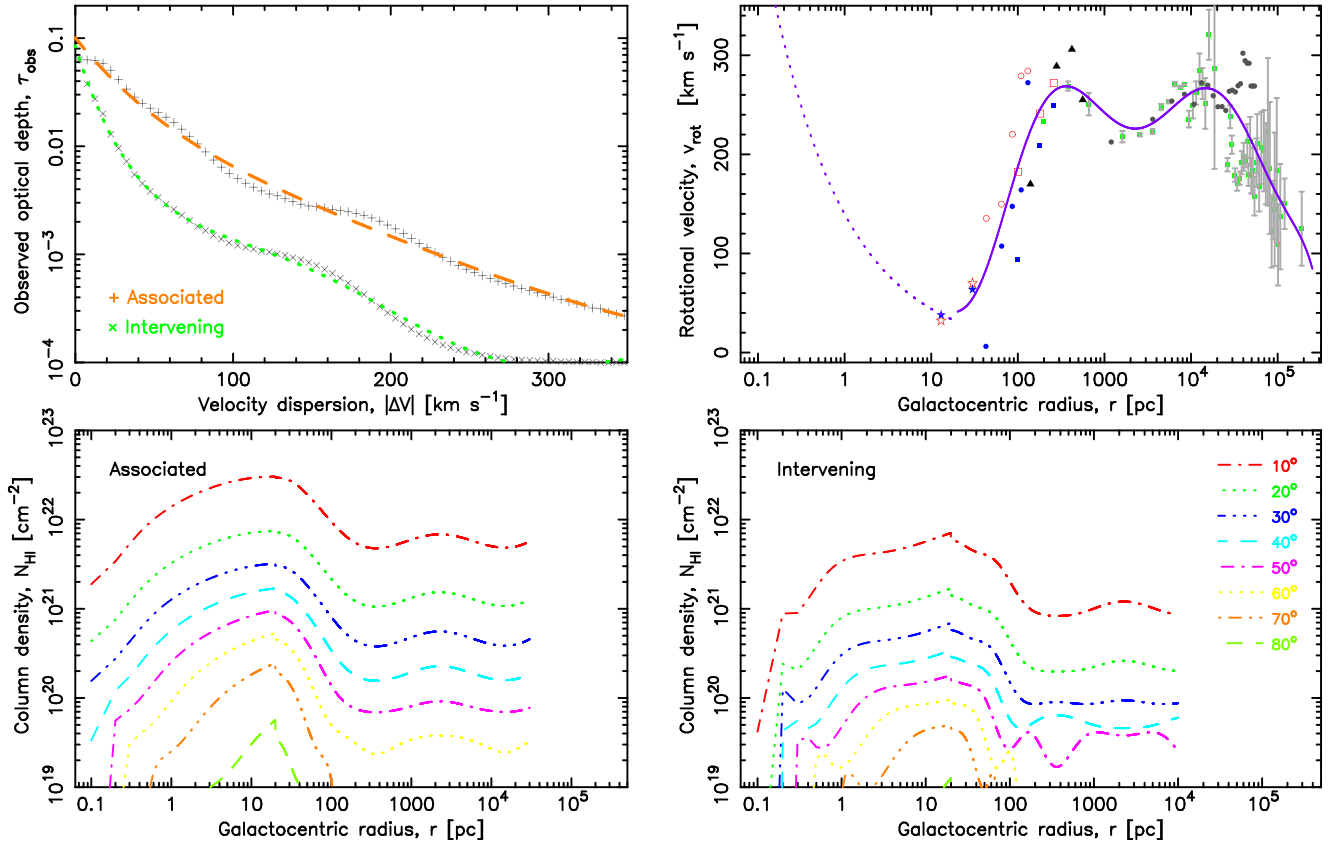


Figure 5. Top left: the mean associated and intervening spectra (where the receding and approaching data are averaged), overlain with low-order polynomial fits. Top right: the rotation curve of a galaxy, based upon data from the Milky Way (shown with error bars; Bhattacharjee, Chaudhury & Kundu 2014) and the Circinus galaxy (compiled from Oliva et al. 1994; Curran et al. 1998; Davies et al. 1998; Maiolino et al. 1998, see Curran, Koribalski & Bains 2008a). The full curve shows a polynomial fit to the large-scale Milky Way and Circinus data and the broken curve shows the scaled Keplerian orbit of the H₂O masers within the central 0.4 pc of Circinus (Greenhill et al. 2003), extrapolated to 20 pc to provide continuity. Bottom left: the derived column density distribution at various disc inclinations for the associated absorbers. Bottom right: for the intervening absorbers.

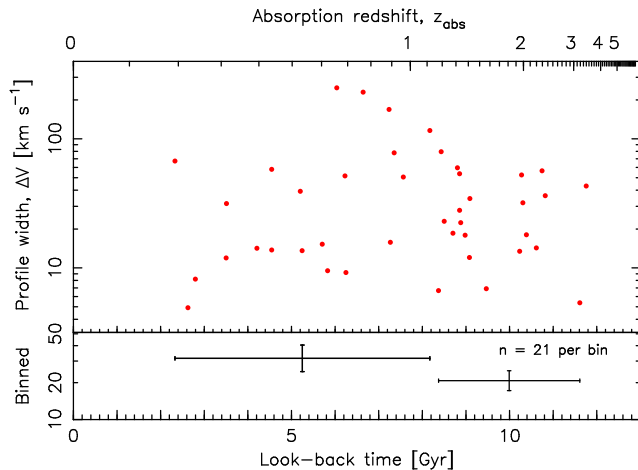


Figure 6. The effective profile width versus look-back time for the intervening absorbers.

$T_{\text{spin}} = T_{\text{kin}} = 22\Delta V^2$ assumption is used). This is consistent with the argument that the larger T_{spin}/f ratios measured at high redshift are dominated by lower covering factors (Curran et al. 2005) through the geometry effects of an expanding Universe (Curran 2012).

3 EXPLORING MACHINE LEARNING MODELS FOR CLASSIFICATION

3.1 Motivation for using machine learning

The main motivation for this study was to determine whether the absorber type can be predicted from the profile properties without a priori knowledge of the emission redshift from an optical spectrum. As seen in Fig 3, however, the individual spectra are too varied to permit this, with the associated absorbers spanning the full range of linewidths (Fig. 4). We therefore apply machine learning techniques with the aim of building a classifier which can be used to make such a prediction. While the data set is quite limited, with $\lesssim 50$ useful spectra in each class, we can explore how feasible machine learning models are and the potential for prediction, particularly as more data are added. We use the WEKA package (Hall et al. 2009), a suite of machine learning algorithms designed for data mining tasks.

3.2 Models

A model is the result of the training data, the features selected to represent the data, the algorithm used, as well as a number of parameters set for the algorithm.

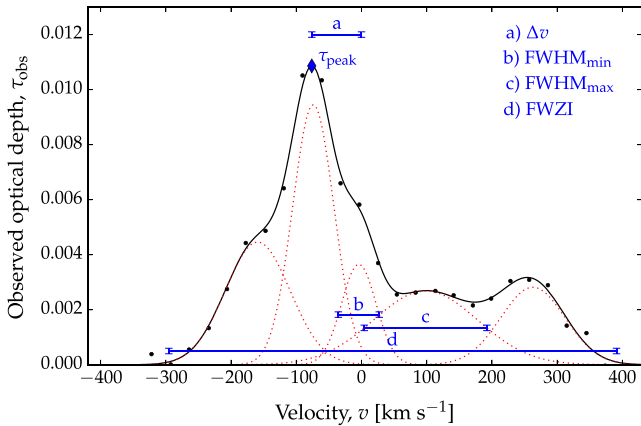


Figure 7. An example of a complex ($n_g = 5$) absorption profile, labelled with the features used for the machine learning: Δv the offset of the component from z_{weight} (a), the minimum (b) and maximum (c) FWHM of the profile and the full width at zero intensity (d, where $\tau_{\text{obs}} < 10^{-4}$).

3.2.1 Feature selection

A crucial step in machine learning is the selection of features. These represent each datum and therefore need to be discriminating and informative. The feature set forms the input given to the selected algorithm and it is the combination of the features which allow different algorithms to discriminate between classes. Here we identified a mixture of such features for each spectrum, illustrated in Fig. 7 and listed in Tables 1 and 2.

Although each spectrum can be quantified through standard parameters such as the number of Gaussians (n_g), the velocity offset (Δv), the peak optical depth (τ_{peak}) and the FWHM, this will present an issue for the machine learning, in that, for $n_g > 1$, there will be several parameters classified as Δv , τ_{peak} and the FWHM, which must be independently flagged as such while also being compared to other spectra. For example, if both spectra **A** and **B** each have two components, then the first component of **A**, FWHM_1^{A} , must be compared with both FWHM_1^{B} and FWHM_2^{B} , as must FWHM_2^{A} , while all retaining their identity as representing the linewidth of an unspecified component. In machine learning, it is common practice to compare global properties, which in this instance would be the total or average FWHM. The chosen features represent unique (or a combination of unique) properties of the spectrum, with no prior expectations of how they will perform. Therefore, in addition to features analogous to those standard in quantifying the spectral properties, such as the linewidth and optical depth, we include the average offset of the components from z_{weight} . This is included solely for the reason that it is an additional property which can be extracted, although there is the possibility that this could be non-zero for the associated spectra, where there is additional in or outflowing gas.

In the final models, we excluded z_{weight} itself, although it may be a powerful feature (see Fig. 1), and indeed, upon testing, can increase the precision of the models (see Section 3.3). However, we believe that the tendency for the associated absorbers to be detected at $z \lesssim 1$ is due to their optical pre-selection, where at higher redshifts, the faint optical light observed is intense UV in the rest frame of the source, ionizing the gas in the high-redshift objects (Curran & Whiting 2012). Since, the SKA and its pathfinders will not be reliant upon optical redshifts, through full-redshift spectral scans (Section 1), we expect higher redshift detections to be forthcoming giving a z_{weight} distribution more similar to that of the intervening absorbers.

Table 3. The rankings of the features for the whole sample and excluding the optically thick absorbers (in descending order for the whole sample).

| Feature | Pearson's correlation | |
|----------------------------------|-----------------------|---|
| | Whole | Excluding $\tau_{\text{peak}} \geq 0.3$ |
| FWZI | 0.5552 | 0.5423 |
| FWHM max | 0.5447 | 0.5367 |
| FWHM ave | 0.5017 | 0.4943 |
| FWHM min | 0.4444 | 0.4384 |
| $\tau_{\text{peak}} \text{ max}$ | 0.0907 | 0.0321 |
| $\tau_{\text{peak}} \text{ min}$ | 0.0904 | 0.1580 |
| $\tau_{\text{peak}} \text{ ave}$ | 0.0782 | 0.0900 |
| Δv | 0.0540 | 0.0654 |
| n_g | 0.0420 | 0.0494 |
| $\Delta v/\text{FWZI}$ | 0.0336 | 0.0648 |

3.2.2 Comparison of predictive power of different features

WEKA also offers the ability to explore the predictive power of the different features via the *Select Attributes* function. This is useful in removing non-contributing features, particularly for large data sets where computational power is an issue. This does not apply to our small sample, although it is of interest to find which features contribute most to the predictive power of the models.

In Table 3, we show the rankings returned by the attribute evaluator. From this, we see that all features contribute to the classification. Those related to the profile width contribute the most, while other features make relatively little contribution, including, surprisingly, the number of Gaussians.

3.2.3 Training data

Since the data set is too small to split into training and testing sets, we use all data, 55 associated and 43 intervening spectra, for training. The models were run for both the whole sample and with the exclusion of spectra exhibiting optically thick components (equation 3).

3.2.4 Algorithms

Since there is no single algorithm that is more suited to all cases (e.g. Wopler 1996), nor any previous application of machine learning to 21-cm absorption profiles in the literature, we experimented with different algorithms in WEKA. As this is a small data set, we kept the default parameters for each algorithm. For the same reason, we considered all features with equal weight. Several algorithms performed comparably well, here we report five of which are classifiers for categorical prediction, as opposed to classifiers for numeric prediction (Witten & Frank 2011).

(i) Bayesian network from the ‘Bayes’ group (Bouckaert 2004). This is a probabilistic model, which through the learning of Bayesian nets, represents a set of random variables which may be observable quantities, latent variables, unknown parameters or hypotheses (Bouckaert 2008).

(ii) Sequential minimal optimization from the ‘functions’ group. This algorithm solves the quadratic programming problem, arising from the training of support vector machines (Platt 1998).

(iii) Classification via regression from the ‘meta’ group. This algorithm uses regression methods, where one regression model is built for each class value (Frank et al. 1998).

(iv) Logistic model tree from the ‘trees’ group. This combines logistic regression and decision tree learning, based upon an earlier

Table 4. The models, without the z_{weight} feature, and their performance for the whole sample and excluding the optically thick absorbers. $P = t_p/(t_p + f_p)$ is the precision (the positive predictive value), where t_p is the number of true positives and f_p the number of false positives. $R = t_p/(t_p + f_n)$ is the recall (the fraction of relevant instances that are retrieved), where f_n is the number of false negatives and the F -measure, $F = 2PR/(P + R)$, is the harmonic mean of precision and recall. The accuracy, A , is the fraction of correctly classified instances. The mean absolute error, $|\bar{\sigma}|$, is a measure of how close the predictions are to the eventual outcomes. The κ -statistic is the chance-corrected measure of agreement between the classifications and the true classes – $\kappa > 0$ signifies that the classifier is doing better than predicting by chance and $\kappa = 1$ signifies completely accurate prediction.

| Algorithm | P (%) | R (%) | F (%) | A (%) | $ \bar{\sigma} $ | κ | P (%) | R (%) | F (%) | A (%) | $ \bar{\sigma} $ | κ |
|---------------------------------|--------------------------------------|---------|---------|---------|------------------|----------|--|---------|---------|---------|------------------|----------|
| | Whole (55 associated/43 intervening) | | | | | | Excluding $\tau_{\text{peak}} \geq 0.3$ (52 associated/39 intervening) | | | | | |
| Bayesian network | 81.2 | 80.6 | 80.7 | 80.6 | 0.194 | 0.611 | 83.4 | 83.3 | 83.4 | 83.3 | 0.182 | 0.660 |
| Sequential minimal optimization | 80.6 | 78.6 | 78.6 | 78.6 | 0.214 | 0.577 | 78.5 | 76.7 | 76.8 | 76.97 | 0.233 | 0.537 |
| Classification via regression | 80.0 | 79.6 | 79.7 | 79.6 | 0.302 | 0.590 | 75.6 | 75.6 | 75.6 | 76.9 | 0.333 | 0.499 |
| Logistic model tree | 80.9 | 80.6 | 80.7 | 80.6 | 0.346 | 0.610 | 78.3 | 77.8 | 77.9 | 77.8 | 0.353 | 0.551 |
| Random forest | 81.6 | 81.6 | 81.6 | 81.6 | 0.305 | 0.422 | 81.1 | 81.1 | 81.1 | 81.1 | 0.312 | 0.612 |

Table 5. The confusion matrices for the models in Table 4.

| | Whole sample | Excluding $\tau_{\text{peak}} \geq 0.3$ |
|---------------------------------|---|--|
| Bayesian network | $\begin{bmatrix} 43 & 12 \\ 7 & 36 \end{bmatrix}$ | $\begin{bmatrix} 44 & 8 \\ 7 & 31 \end{bmatrix}$ |
| Sequential minimal optimization | $\begin{bmatrix} 39 & 16 \\ 5 & 38 \end{bmatrix}$ | $\begin{bmatrix} 37 & 15 \\ 6 & 32 \end{bmatrix}$ |
| Classification via regression | $\begin{bmatrix} 43 & 12 \\ 8 & 35 \end{bmatrix}$ | $\begin{bmatrix} 41 & 11 \\ 11 & 27 \end{bmatrix}$ |
| Logistic model tree | $\begin{bmatrix} 44 & 11 \\ 8 & 35 \end{bmatrix}$ | $\begin{bmatrix} 40 & 12 \\ 8 & 30 \end{bmatrix}$ |
| Random forest | $\begin{bmatrix} 46 & 9 \\ 9 & 34 \end{bmatrix}$ | $\begin{bmatrix} 44 & 8 \\ 9 & 29 \end{bmatrix}$ |

version of the tree. Each ‘leaf’ in the tree represents a model and the logistic variant produces a regression model at every node in the tree, which is then split (Landwehr, Hall & Frank 2005).

(v) Random forest also from the ‘trees’ group. These are an ensemble of learning methods for classification, which operate via a multitude of decision trees (Breiman 2001).

3.3 Results

We summarize the results in Table 4. We report on 10-fold cross validation performance, where the data are split into 10 sets, each of which will contain $(55 + 43)/10 \approx 10$ spectra. Nine of the data sets are used to train the model, with the resulting classifier used to test the one remaining data set. This process is randomized and repeated 10 times with the mean accuracy being reported. 10-fold cross validation is typical practice for small data sets, where there are not enough data to split for training and testing.

Table 5, shows the confusion matrices associated with the models. These are in the format,

| | Predicted ass | Predicted int |
|--------------------|---------------|---------------|
| Actual associated | TA | FI |
| Actual intervening | FA | TI |

where TA – true associated, FA – false associated, TI – true intervening, FI – false intervening. For example, for the confusion matrix $\begin{bmatrix} 43 & 12 \\ 7 & 36 \end{bmatrix}$, out of $43 + 12$ associated spectra, the model correctly identifies 43 as associated and 12 erroneously as intervening and out of $7 + 36$ intervening spectra, 36 are correctly identified as intervening and 7 erroneously as associated. The data in the confusion matrices form the raw data used for reporting several evaluation measures – the precision, recall, F -measure and accuracy (Table 4). Those in

Table 5 show that the two classes are balanced in terms of training and classification. For instance, a matrix such as $\begin{bmatrix} 90 & 0 \\ 10 & 0 \end{bmatrix}$ would return a 90 per cent accuracy, but in this case, the model would have a 100 per cent recognition rate for one class and 0 per cent for the other.

3.4 DISCUSSION

Most of the models return a precision, recall, F -measure and accuracy of $\gtrsim 80$ per cent, making us optimistic that as more data are collected by the community, machine learning can provide a useful tool for classification of redshifted 21-cm absorption spectra.

During our feature selection experimentation, we found that the accuracy of prediction is dominated by the profile width – models trained on the full width at zero intensity alone return accuracies just a few per cent below those reported in Table 4. Models trained on the three FWHM features give accuracies ≈ 5 per cent lower. The removal of all line-width information features, reduces the accuracy to ≈ 50 per cent, i.e. what we would obtain from chance. The number of components, n_g , contributes relatively little to the prediction, which is surprising since we expect the associated profiles to be more complex. However, the combination of all features do provide the better performing models, in agreement with the attribute evaluator analysis (Table 3).

Another feature we considered is z_{weight} . Some experimentation with classifiers trained using the z_{weight} as an additional feature improved the precision in some cases (e.g. up to 5.2 per cent, from 78.5 per cent to 83.7 per cent for the sequential minimal optimization algorithm, excluding the $\tau_{\text{peak}} \geq 0.3$ set), while being detrimental in others (e.g. Bayesian network). However, since the purpose of this work is to classify the absorber type without the use of an optical redshift, the pre-selection of which may introduce a bias (Section 1), we did not consider z_{weight} in the final models (Table 4).

The machine learning results are very encouraging and, as more data are added, we expect the predictive power of such classifiers to improve. This will prove invaluable as the number of new 21-cm detections becomes too large to feasibly follow-up with optical observations.

4 CONCLUSIONS

Forthcoming spectral lines surveys with the next generation of large radio telescopes are expected to detect large numbers of new redshifted H I 21-cm absorbers. Although the measured redshifts of the absorbing galaxies will be extremely accurate, due to the phase-locking of radio receivers, without follow-up optical-band

observations, it is generally not possible to determine whether the absorption arises within the source host or from a galaxy intervening the sightline to a more distant radio source. Given these large numbers, and the possibility that optical pre-selection biases against the detection of cool, neutral gas, it would be of great value to be able to determine the nature of the absorber based upon the radio data alone.

To this end, we have compiled and digitized the known $z \geq 0.1$ H I 21-cm absorbers, converted these to consistent dimensions (optical depth and velocity) and re-sampled to a common spectral resolution. However, the normalized spectra in each of the associated and intervening classes exhibit a wide range of profile shapes, not making it possible to manually ascertain a typical profile shape. By applying machine learning algorithms, we find that, even for our limited sample of $\lesssim 50$ of each type of absorber, the type can be predicted with $\gtrsim 80$ per cent accuracy. As new detections are made, follow-up optical-band observations will allow us to improve the classifier in preparation for forthcoming H I 21-cm surveys with the SKA and its precursors.

Although machine learning was invoked to classify the individual spectra, by averaging all those in each class in order to examine the bulk differences, we find:

(i) That the mean associated profile is wider than the mean intervening profile, with a Kolmogorov–Smirnov of the individual widths giving a 4×10^{-10} probability that the associated and intervening velocity distributions are drawn from the same population. This is consistent with the profile width being the one single feature which lowers the predictive power of the classifier to that of chance when removed.

(ii) From a simple model of the H I column density distribution, that the high-velocity wings often observed in associated absorption, arise from the sub-pc gas, which appears to be absent in the intervening absorbers. This supports the widely proposed conjecture that the additional component of the associated absorption is due to the dense circum-nuclear torus, invoked by unified schemes of AGN. This is also consistent with hypothesis that associated absorption arises in AGN (radio galaxies and quasars), where significant amounts of gas are accreted on to the central supermassive black hole, whereas intervening absorption arises in more quiescent galaxies.

(iii) The consistency in the mean intervening profile widths to either side of $z \sim 1$ (where the sample is split in half), indicates no kinematical or thermal evolution with redshift. While the H I column density model is consistent with the bulk of profile broadening being kinematical, rather than thermal, in nature, this result does not support the proposed increase in the spin temperature of the gas with redshift in DLAs.

ACKNOWLEDGEMENTS

We wish to thank the anonymous referee for their detailed comments and Nathan Holmberg for his advice. This research has made use of the NASA/IPAC Extragalactic Database (NED) which is operated by the Jet Propulsion Laboratory, California Institute of Technology, under contract with the National Aeronautics and Space Administration and NASA's Astrophysics Data System Bibliographic Service. This research has also made use of NASA's Astrophysics Data System Bibliographic Service.

REFERENCES

- Aditya J. N. H. S., Kanekar N., Kurapati S., 2016, *MNRAS*, 455, 4000
Allison J. R. et al., 2012, *MNRAS*, 423, 2601
Allison J. R., Curran S. J., Sadler E. M., Reeves S. N., 2013, *MNRAS*, 430, 157
Allison J. R. et al., 2015, *MNRAS*, 453, 1249
Allison J. R. et al., 2016, *Astron. Nachr.*, 337, 175
Antonucci R. R. J., 1993, *ARA&A*, 31, 473
Bennert V. N. et al., 2015, *ApJ*, 809, 20
Bhattacharjee P., Chaudhury S., Kundu S., 2014, *ApJ*, 785, 63
Bouckaert R. R., 2004, Technical Report, Bayesian Networks in Weka. Univ. Waikato
Bouckaert R. R., 2008, Technical Report, Bayesian Network Classifiers in Weka for Version 3-5-7. Univ. Waikato
Breiman L., 2001, *Mach. Learn.*, 45, 5
Briggs F. H., Wolfe A. M., 1983, *ApJ*, 268, 76
Briggs F. H., de Bruyn A. G., Vermeulen R. C., 2001, *A&A*, 373, 113
Carilli C. L., Perlman E. S., Stocke J. T., 1992, *ApJ*, 400, L13
Carilli C. L., Rupen M. P., Yanny B., 1993, *ApJ*, 412, L59
Carilli C. L., Menten K. M., Reid M. J., Rupen M. P., Yun M. S., 1998, *ApJ*, 494, 175
Carilli C. L., Gnedin N., Furlanetto S., Owen F., 2004, *New Astron. Rev.*, 48, 1053
Catinella B., Cortese L., 2015, *MNRAS*, 446, 3526
Chandola Y., Sirothia S. K., Saikia D. J., 2011, *MNRAS*, 418, 1787
Chengalur J. N., de Bruyn A. G., Narasimha D., 1999, *A&A*, 343, L79
Conway J. E., Blanco P. R., 1995, *ApJ*, 449, L131
Curran S. J., 2010, *MNRAS*, 402, 2657
Curran S. J., 2012, *ApJ*, 748, L18
Curran S. J., Whiting M. T., 2010, *ApJ*, 712, 303
Curran S. J., Whiting M. T., 2012, *ApJ*, 759, 117
Curran S. J., Johansson L. E. B., Rydbeck G., Booth R. S., 1998, *A&A*, 338, 863
Curran S. J., Kanekar N., Darling J. K., 2004, *New Astron. Rev.*, 48, 1095
Curran S. J., Murphy M. T., Pihlström Y. M., Webb J. K., Purcell C. R., 2005, *MNRAS*, 356, 1509
Curran S. J., Whiting M. T., Murphy M. T., Webb J. K., Longmore S. N., Pihlström Y. M., Athreya R., Blake C., 2006, *MNRAS*, 371, 431
Curran S. J., Darling J. K., Bolatto A. D., Whiting M. T., Bignell C., Webb J. K., 2007a, *MNRAS*, 382, L11
Curran S. J., Tzanavaris P., Pihlström Y. M., Webb J. K., 2007b, *MNRAS*, 382, 1331
Curran S. J., Koribalski B. S., Bains I., 2008a, *MNRAS*, 389, 63
Curran S. J., Whiting M. T., Wiklind T., Webb J. K., Murphy M. T., Purcell C. R., 2008b, *MNRAS*, 391, 765
Curran S. J. et al., 2011a, *MNRAS*, 413, 1165
Curran S. J., Whiting M. T., Webb J. K., Athreya A., 2011b, *MNRAS*, 414, L26
Curran S. J. et al., 2011c, *MNRAS*, 416, 2143
Curran S. J., Whiting M. T., Sadler E. M., Bignell C., 2013a, *MNRAS*, 428, 2053
Curran S. J., Whiting M. T., Tanna A., Sadler E. M., Pracy M. B., Athreya R., 2013b, *MNRAS*, 429, 3402
Curran S. J., Allison J. R., Whiting M. T., Sadler E. M., Combes F., Pracy M. B., Bignell C., Athreya R., 2016a, *MNRAS*, 457, 3666
Curran S. J., Reeves S. N., Allison R., Sadler E. M., 2016b, *MNRAS*, 459, 4136
Darling J. K., 2003, *Phys. Rev. Lett.*, 91, 011301
Darling J., Giovanelli R., Haynes M. P., Bower G. C., Bolatto A. D., 2004, *ApJ*, 613, L101
Davies R. I. et al., 1998, *MNRAS*, 293, 189
Davis M. M., May L. S., 1978, *ApJ*, 219, 1
Dickey J. M., 1982, *ApJ*, 263, 87

- Dickey J. M., Strasser S., Gaensler B. M., Haverkorn M., Kavars D., McClure-Griffiths N. M., Stil J., Taylor A. R., 2009, *ApJ*, 693, 1250
- Duchesne S. W., 2015, *Star-Forming Gas in the Distant Universe: Distinguishing Between Intervening and Gas Associated with the Radio Source for the Next Generation of Large Radio Telescopes*. Tech. rep., Victoria Univ. Wellington
- Ellison S. L., Hall P. B., Lira P., 2005, *AJ*, 130, 1345
- Ellison S., Kanekar N., Prochaska J. X., Momjian E., Worseck G., 2012, *MNRAS*, 424, 293
- Ferrarese L., Merritt D., 2000, *ApJ*, 539, L9
- Field G. B., 1959, *ApJ*, 129, 536
- Frank E., Wang Y., Inglis S., Holmes G., Witten I., 1998, *Mach. Learn.*, 32, 63
- Geréb K., Maccagni F. M., Morganti R., Oosterloo T. A., 2015, *A&A*, 575, 44
- Grasha K., Darling J., 2011, *Am. Astron. Soc. Meeting Abstr.*, 43, 345.02
- Greenhill L. J. et al., 2003, *ApJ*, 590, 162
- Gupta N., Salter C. J., Saikia D. J., Ghosh T., Jeyakumar S., 2006, *MNRAS*, 373, 972
- Gupta N., Srianand R., Petitjean P., Noterdaeme P., Saikia D. J., 2009a, *MNRAS*, 398, 201
- Gupta N., Srianand R., Petitjean P., Noterdaeme P., Saikia D. J., 2009b, in Saikia D. J., Green D. A., Gupta Y., Venturi T., eds, *ASP Conf. Ser. Vol. 407, The Low-Frequency Radio Universe*. Astron. Soc. Pac., San Francisco, p. 67
- Gupta N., Srianand R., Petitjean P., Bergeron J., Noterdaeme P., Muzahid S., 2012, *A&A*, 544, 21
- Gupta N., Srianand R., Noterdaeme P., Petitjean P., Muzahid S., 2013, *A&A*, 558, A84
- Haehnelt M. G., Steinmetz M., Rauch M., 1998, *ApJ*, 495, 647
- Hall M., Frank E., Holmes G., Pfahringer B., Reutemann P., Witten I., 2009, *SIGKDD Explor.*, 11, 10
- Holt J., Tadhunter C. N., Morganti R., 2008, *MNRAS*, 387, 639
- Ishwara-Chandra C. H., Dwarakanath K. S., Anantharamaiah K. R., 2003, *JA&A*, 24, 37
- Jimenez R., Bowen D. V., Matteucci F., 1999, *ApJ*, 514, L83
- Kalberla P. M. W., Dedes L., Kerp J., Haud U., 2007, *A&A*, 469, 511
- Kanekar N., 2014, *ApJ*, 797, L20
- Kanekar N., Briggs F. H., 2003, *A&A*, 412, L29
- Kanekar N., Chengalur J. N., 2001, *MNRAS*, 325, 631
- Kanekar N., Chengalur J. N., 2003, *A&A*, 399, 857
- Kanekar N., Ghosh T., Chengalur J. N., 2001, *A&A*, 373, 394
- Kanekar N., Prochaska J. X., Ellison S. L., Chengalur J. N., 2009, *MNRAS*, 396, 385
- Kanekar N., Ellison S. L., Momjian E., York B. A., Pettini M., 2013, *MNRAS*, 428, 532
- Kanekar N. et al., 2014, *MNRAS*, 438, 2131
- Lagos C. D. P., Baugh C. M., Zwaan M. A., Lacey C. G., Gonzalez-Perez V., Power C., Swinbank A. M., van Kampen E., 2014, *MNRAS*, 440, 920
- Landwehr N., 2003, Master's thesis, Univ. Freiburg
- Landwehr N., Hall M., Frank E., 2005, *Mach. Learn.*, 95, 161
- Lane W. M., Briggs F. H., 2001, *ApJ*, 561, L27
- Lovell J. E. J. et al., 1996, *ApJ*, 472, L5
- Maiolino R., Krabbe A., Thatte N., Genzel R., 1998, *ApJ*, 493, 650
- Matteucci F., Molaro P., Vladilo G., 1997, *A&A*, 321, 45
- Mirabel I. F., 1989, *ApJ*, 340, L13
- Moore C. B., Carilli C. L., Menten K. M., 1999, *ApJ*, 510, L87
- Moorwood A. F. M., Oliva E., 1990, *A&A*, 239, 78
- Morganti R., Oosterloo T. A., Tadhunter C. N., van Moorsel G., Killeen N., Wills K. A., 2001, *MNRAS*, 323, 331
- Morganti R., Oosterloo T. A., Tadhunter C. N., van Moorsel G., Emonts B., 2005, *A&A*, 439, 521
- Morganti R., Oosterloo T., Struve C., Saripalli L., 2008, *A&A*, 485, L5
- Morganti R., Emonts B., Oosterloo T., 2009, *A&A*, 496, L9
- Morganti R., Holt J., Tadhunter C., Ramos Almeida C., Dicken D., Inskip K., Oosterloo T., Tzioumis T., 2011, *A&A*, 535, A97
- Morganti R., Sadler E. M., Curran S., 2015, *Proc. Sci., Advancing Astrophysics with the Square Kilometre Array (AASKA14)*. SISSA, Trieste, PoS#134
- Muller S. et al., 2013, *A&A*, 551, A109
- Mundell C. G., Pedlar A., Baum S. A., O'Dea C. P., Gallimore J. F., Brinks E., 1995, *MNRAS*, 272, 355
- Noordermeer E., van der Hulst J. M., Sancisi R., Swaters R. S., van Albada T. S., 2007, *MNRAS*, 376, 1513
- Oliva E., Salvati M., Moorwood A. F. M., Marconi A., 1994, *A&A*, 288, 457
- Orienti M., Morganti R., Dallacasa D., 2006, *A&A*, 457, 531
- Peck A. B., Taylor G. B., Conway J. E., 1999, *ApJ*, 521, 103
- Peck A. B., Taylor G. B., Fassnacht C. D., Readhead A. C. S., Vermeulen R. C., 2000, *ApJ*, 534, 104
- Pihlström Y. M., Vermeulen R. C., Taylor G. B., Conway J. E., 1999, *ApJ*, 525, L13
- Platt J. C., 1998, in Schoelkopf B., Burges C., Smola A., eds, *Fast Training of Support Vector Machines using Sequential Minimal Optimization*. MIT Press. Available at: <http://www.cs.utsa.edu/~bylander/cs6243/smo-book.pdf>
- Prochaska J. X., Wolfe A. M., 1997, *ApJ*, 487, 73
- Prochaska J. X., Herbert-Fort S., Wolfe A. M., 2005, *ApJ*, 635, 123
- Purcell E. M., Field G. B., 1956, *ApJ*, 124, 542
- Rawlings S., Abdalla F. B., Bridle S. L., Blake C. A., Baugh C. M., Greenhill L. J., van der Hulst J. M., 2004, *New Astron. Rev.*, 48, 1013
- Reid M. J., Brunthaler A., 2004, *ApJ*, 616, 872
- Roberts M. S., Brown R. L., Brundage W. D., Rots A. H., Haynes M. P., Wolfe A. M., 1976, *AJ*, 81, 293
- Roy N., Chengalur J. N., Srianand R., 2006, *MNRAS*, 365, L1
- Salter C. J., Saikia D. J., Minchin R., Ghosh T., Chandola Y., 2010, *ApJ*, 715, L117
- Srianand R., Gupta N., Petitjean P., Noterdaeme P., Ledoux C., Salter C. J., Saikia D. J., 2012, *MNRAS*, 421, 651
- Srianand R., Gupta N., Momjian E., Vivek M., 2015, *MNRAS*, 451, 917
- Struve C., Conway J. E., 2010, *A&A*, 513, A10
- Taylor G. B., O'Dea C. P., Peck A. B., Koekemoer A. M., 1999, *ApJ*, 512, L27
- Taylor G. B., Peck A. B., Henkel C., Falcke H., Mundell C. G., O'Dea C. P., Baum S. A., Gallimore J. F., 2002, *ApJ*, 574, 88
- Tzanavaris P., Murphy M. T., Webb J. K., Flambaum V. V., Curran S. J., 2007, *MNRAS*, 374, 634
- Urry C. M., Padovani P., 1995, *PASP*, 107, 803
- Uson J. M., Bagri D. S., Cornwell T. J., 1991, *Phys. Rev. Lett.*, 67, 3328
- van Gorkom J. H., Knapp G. R., Ekers R. D., Ekers D. D., Laing R. A., Polk K. S., 1989, *AJ*, 97, 708
- van Langevelde H. J., Pihlström Y. M., Conway J. E., Jaffe W., Schilizzi R. T., 2000, *A&A*, 345, L45
- Vermeulen R. C. et al., 2003, *A&A*, 404, 861
- Witten I. H., Frank E., 2011, *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, San Francisco
- Wolfe A. M., Turnshek D. A., Smith H. E., Cohen R. D., 1986, *ApJS*, 61, 249
- Wopler D. H., 1996, *Neural Comput.*, 8, 1341
- Yan T., Stocke J. T., Darling J. K., Momjian E., Sharma S., Kanekar N., 2016, *AJ*, 151, 74
- York B. A., Kanekar N., Ellison S. L., Pettini M., 2007, *MNRAS*, 382, L53
- Zwaan M. A., Liske J., Péroux C., Murphy M. T., Bouché N., Curran S. J., Biggs A. D., 2015, *MNRAS*, 453, 1268

This paper has been typeset from a \LaTeX file prepared by the author.