A&A 627, A32 (2019)
https://doi.org/10.1051/0004-6361/201935371

Astronomy & Astrophysics

Proximate molecular quasar absorbers

Excess of damped H_2 systems at $z_{abs} \approx z_{QSO}$ in SDSS DR14

P. Noterdaeme[1], S. Balashev[2], J.-K. Krogager[1], R. Srianand[3], H. Fathivavsari[1,4], P. Petitjean[1], and C. Ledoux[5]

[1] Institut d'astrophysique de Paris, CNRS-SU, UMR 7095, 98bis bd Arago, 75014 Paris, France
e-mail: noterdaeme@iap.fr
[2] Ioffe Institute, Polyteknicheskaya 26, 194021 Saint-Petersburg, Russia
e-mail: s.balashev@gmail.com
[3] Inter-University Centre for Astronomy and Astrophysics, Pune University Campus, Ganeshkhind, Pune 411007, India
[4] School of Astronomy, Institute for Research in Fundamental Sciences (IPM), PO Box 19395-5531, Tehran, Iran
[5] European Southern Observatory, Alonso de Córdova 3107, Vitacura, Casilla 19001, Santiago 19, Chile

Received 26 February 2019 / Accepted 3 May 2019

ABSTRACT

We present results from a search for strong H_2 absorption systems proximate to quasars ($z_{abs} \approx z_{em}$) in the Sloan Digital Sky Survey (SDSS) Data Release 14. The search is based on the Lyman-Werner band signature of damped H_2 absorption lines without any prior on the associated metal or neutral hydrogen content. This has resulted in the detection of 81 systems with $N(H_2) \sim 10^{19}-10^{20}$ cm^{-2} located within a few thousand km s^{-1} from the quasar. Compared to a control sample of intervening systems, this implies an excess of proximate H_2 systems by about a factor of 4 to 5. The incidence of H_2 systems increases steeply with decreasing relative velocity, reaching an order of magnitude higher than expected from intervening statistics at $\Delta v < 1000$ km s^{-1}. The most striking feature of the proximate systems compared to the intervening ones is the presence of Ly$-\alpha$ emission in the core of the associated damped H I absorption line in about half of the sample. This puts constraints on the relative projected sizes of the absorbing clouds to those of the quasar line emitting regions. Using the SDSS spectra, we estimate the H I, metal and dust content of the systems, which are found to have typical metallicities of one tenth Solar, albeit with a large spread among individual systems. We observe trends between the fraction of leaking Ly$-\alpha$ emission and the relative absorber-quasar velocity as well as with the excitation of several metal species, similar to what has been seen in metal-selected proximate DLAs. With the help of theoretical H I-H_2 transition relations, we show that the presence of H_2 helps to break the degeneracy between density and strength of the UV field as main sources of excitation and hence provides unique constraints on the possible origin and location of the absorbing clouds. We suggest that most of these systems originate from galaxies in the quasar group, although a small fraction of them could be located in the quasar host as well. We conclude that follow-up observations are still required to investigate the chemical and physical conditions in individual clouds and to assess the importance of AGN feedback for the formation and survival of H_2 clouds.

Key words. quasars: general – quasars: absorption lines – quasars: emission lines – ISM: molecules

1. Introduction

Damped Ly-α absorption systems (DLAs, see Wolfe et al. 2005) observed in the spectra of distant light sources belong to two main categories, intervening and associated, depending on their origin with respect to the background sources. Intervening DLAs are produced by neutral H I gas located by chance along the line of sight to the background sources without being related to the sources themselves. Using intervening absorption systems identified in large spectroscopic surveys (such as the Sloan Digital Sky Survey, hereafter SDSS), it is possible to conduct a census of the neutral gas in the Universe and study its evolution over cosmic time (e.g. Péroux et al. 2003; Prochaska et al. 2005; Noterdaeme et al. 2012). Moreover, DLAs are very useful probes of cosmic chemical evolution (e.g. Rafelski et al. 2012; De Cia et al. 2018), and the physical conditions of the absorbing medium can be probed by studying the excitation of various species, in particular molecular hydrogen (e.g. Srianand et al. 2005; Noterdaeme et al. 2007; Jorgenson et al. 2010; Balashev et al. 2017). Overall, intervening DLAs exhibit characteristics and a complexity indicating an origin from interstellar or circumgalactic gas. Indeed, a direct connection between intervening DLAs and galaxies is now emerging thanks to the detection of galaxies in emission at the absorption redshift (e.g. Krogager et al. 2017; Neeleman et al. 2019).

Associated systems, in contrast, originate from gas belonging to the close environment of the background sources. As such, they provide unique information about the sources themselves or their environment. For example, in the case of long-duration γ-ray burst (GRB) afterglows, strong DLAs are almost systematically detected. While the so-called GRB-DLAs may not necessarily be associated to the GRB explosion site itself (which is thought to be associated to the death of a massive star), they still likely probe the gas in the GRB host galaxy, as evidenced by a N(H I)-distribution skewed to high column densities (Fynbo et al. 2009). The luminous and rapidly varying afterglow also leads to specific effects such a time-varying UV-pumping of excited levels of atomic species (Vreeswijk et al. 2007) or the presence of vibrationally excited H_2 (Sheffer et al. 2009).

In the case of quasars, associated DLAs may arise from infalling or outflowing gas, gas in the quasar host, or from

nearby galaxies in the group environment, all of which possibly affected by the quasar via radiation or mechanical feedback. For example, quasar activity can result in quenching of star formation in the quasar host due to gas consumption or gas ejection from the galaxy through powerful winds (so-called negative feedback). However, quasar activity may also lead to positive feedback on star formation through compression of the gas (e.g. Zubovas et al. 2013). The presence of a quasar may also affect the gas in nearby galaxies, and consequently their star formation. Moreover, the feeding of quasars with infalling gas is one of the most challenging problems in the field and lacks direct observational evidence. Finally, while outflows driven by the quasar are ubiquitously observed in various states, from highly ionised, atomic phases to molecular phases, detecting these in absorption will provide unique clues as to their physical and chemical states.

The various possible origins for the associated DLAs suggest that the frequency of these could be in excess compared to intervening systems, and that associated DLAs may exhibit different characteristics. However, it is not trivial to distinguish between intervening and associated systems through observations. The most direct piece of information regarding the respective location of the intervening and associated systems is the apparent velocity difference. Noting that 1000 km s^{-1} in the Hubble flow correspond to about 3 Mpc proper distance at $z \sim 3$, systems with apparent velocity differences larger than a few thousand kilometres per second are generally considered as intervening since peculiar motions are unlikely to reach such values. Nonetheless, it cannot be excluded that outflowing winds may produce DLAs with large velocities. For velocity differences less than a few thousands of kilometres per second, the absorber can either be associated (including the various possible origins discussed above) or still unrelated to the source environment (i.e. intervening). Such systems are therefore dubbed "proximate" until further information is available.

Based on the CORALS survey, Ellison et al. (2002) reported a factor of ∼4 excess of proximate DLAs (PDLAs) compared to intervening ones. From a systematic search of SDSS data release 5 (DR5), Prochaska et al. (2008a) later reported an excess of only a factor of ∼2 at redshift $z \sim 3$, but no statistically significant excess at $z < 2.5$ and $z > 3.5$. Studies of metal lines in both composite SDSS spectra (Ellison et al. 2010) and individual high resolution spectra (Ellison et al. 2011) suggest that PDLAs have properties that are only marginally different from those of intervening DLAs; On average the former have higher metallicities (although spreading a wide range) and stronger high-ionisation lines.

A more striking difference between PDLAs and intervening DLAs is the existence of a population of PDLAs that do not fully cover the Ly-α emission region of the background quasar (Finley et al. 2013). This results in an additional flux in the core of the DLA, which complicates their identification as DLAs. The system will appear as a *coronagraphic* DLA when the broad line region (BLR) of the quasar is fully covered by the absorbing cloud but the narrow line region (NLR) is not. Depending on the relative strength and width of the emission compared to that of the DLA absorption, there exists a continuous range of situations, starting from DLAs where some emission is seen in the core to systems where the damping wings are barely visible due to strong Ly-α emission (Jiang et al. 2016). We note that part of the emission can also be due to Ly-α photons originating from the quasar host galaxy or from Ly-α photons scattered out to very large distances (tens of kpc, e.g. Courbin et al. 2008; Cantalupo et al. 2014; Borisova et al. 2016; North et al. 2017). However, the total flux of such kpc-scale Ly-α emission is significantly smaller than that from the NLR (Fathivavsari et al. 2016), yet sometimes becoming comparable to the later (e.g. Fathivavsari et al. 2015). In some extreme cases (called *ghostly* DLAs by Fathivavsari et al. 2017), the BLR is not fully covered either and the absorption system is only witnessed by its Ly-β, Ly-γ and higher series H I lines as well as low-ionisation metal lines that indicate the presence of neutral gas along the line of sight. Based on an observed relation between the strength of leaking Ly-α emission and the fine-structure excitation of metal species, Fathivavsari et al. (2018a) suggested that systems with strong Ly-α emission could be located closer to the quasar where mechanical compression of the gas would be at play. We note that the enhanced UV flux may then also play a role in the excitation of the metal species.

Investigating the presence of molecular gas (in particular H_2) in PDLAs could bring new clues to the overall picture since the production and destruction of molecules is very sensitive to the physical conditions of the gas. In cold neutral gas, the molecular hydrogen fraction is governed by the equilibrium between the formation of H_2 on the surface of dust grains and photo-dissociation by UV photons through line absorption in the Lyman and Werner bands (see e.g. Wakelam et al. 2017). The proximity of the central engine not only increases the photo-dissociation rate but may also lead to complex effects such as an increase of the dust temperature that decreases the formation efficiency of H_2 on the surface of grains. On the other hand, the fragmentation of dust due to strong UV radiation increases the grain surface-to-mass ratio, which could increase the H_2 formation, but at the same time, the grains fragments will also be heated. It is therefore not obvious what the net effect on the H_2 formation rate would be. Additionally, mechanical feedback from the quasar may result in an increase in the number density, n_H, and thus a significant increase in the H_2 production rate, which scales as n_H^2. More generally, it is crucial to investigate how H_2 clouds can survive or form in harsh environments and thereby how star formation is affected close to the quasar.

The presence of molecular hydrogen proximate to the quasar was first shown by Levshakov & Varshalovich (1985) who detected H_2 with $N(H_2) \sim 10^{18}$ cm^{-2} at $z_{abs} = 2.811$ towards PKS 0528−250 ($z_{em} = 2.77$). This was later confirmed by Foltz et al. (1988) who also discussed the possible reasons for the existence of H_2 gas when the extinction measured towards the quasar is low. The authors suggested that the formation rate could be more efficient than seen locally, that the incident UV flux could actually be low, or that H_2 could be formed in non-equilibrium in cooling zones behind shocks. Levshakov & Foltz (1988) discussed the transverse size of the associated atomic gas from the complete absorption of Ly-α+N V emission by the DLA and Klimenko et al. (2015) demonstrated that the emission regions were not fully covered by the molecular cloud. A detailed investigation of physical conditions in this system from the excitation of various species is still to be done (Balashev et al., in prep.).

It is also remarkable that the proximate H_2 system from Levshakov & Varshalovich (1985) also represents the first detection of molecules in absorption at high-redshift. Since then, several systematic searches have been performed to search for intervening H_2 towards quasars (Ledoux et al. 2003; Noterdaeme et al. 2008, 2018; Jorgenson et al. 2014; Balashev et al. 2014), but no systematic search has been performed for systems proximate to the quasar, for which the available pathlength is actually much smaller. Considering the very large number of quasar spectra now available in the SDSS, we

initiated a campaign to study molecular gas absorbers proximate to quasars. In this paper, we present our results based on a automated search of H_2 in the SDSS quasar catalogue. The SDSS is indeed a gold mine for such studies since strong H_2 absorption systems can be efficiently identified in the SDSS spectra, as demonstrated by Balashev et al. (2014). We present the search of strong H_2 systems proximate to the quasar without any other prior in Sect. 2 and build a sample of about 80 such systems. In Sect. 3, we study the excess of such systems compared to what could be expected from intervening statistics. We then investigate the main properties of the systems, as can be derived from SDSS data in Sect. 4. In Sect. 5, we discuss our results within a theoretical frame for the transition from atomic to molecular gas, and lastly, we offer a summary of our main findings in Sect. 6.

2. Detection of proximate H_2 absorbers

2.1. Parent sample

We searched for H_2 lines at the redshift of the quasars in the SDSS DR14 catalogue (Pâris et al. 2018). A total of 103 320 quasars have emission redshifts $z > 2.5$ and are therefore suitable to search for H_2 bands in their SDSS spectra. In case several spectra are available for a given quasar, we used the combined spectrum that consists of the co-addition of all exposures of that object. We then rejected spectra with median signal-to-noise ratio (S/N) per pixel lower than 2 in the 1400−1500 Å region in the rest-frame of the quasar, yielding a parent sample of 82 564 quasars (including also quasars with broad absorption line features) whose spectra were effectively searched for strong proximate H_2 absorption.

2.2. Searching procedure

We used a Spearman's rank correlation analysis to search for strong H_2 lines by correlating the observed data with a synthetic H_2 profile. We used a synthetic H_2 template built considering a total column density $N(H_2) = 10^{20}\,cm^{-2}$ that is distributed over the first three rotational levels, assuming an excitation temperature $T_{0,1,2} = 100\,K$ (as typically seen for H_2 clouds in absorption). This theoretical profile was convolved with the SDSS instrumental line-spread function (corresponding to a resolving power of $R = 1500$ in the blue) and re-binned to the same grid, that is, with a constant $\log(\lambda)$ pixel-spacing of 10^{-4} dex, or equivalently $69\,km\,s^{-1}$. We note that our procedure is little sensitive to the exact column density and excitation temperature since the lines we are looking for are intrinsically saturated and because the rank correlation is mostly sensitive to the global "comb-like" shape of the H_2 absorption profile and not on their actual strength. Nevertheless, we tested that changing the column density (by a factor of ten either upwards or downwards) and excitation temperature in the template has no effect on the detection of strong H_2 systems. Since we do not know a priori the exact velocity shift between any H_2 absorber and the quasar redshift and because the later is not known to high accuracy, we first cross-correlated the template with the data over a velocity interval that encompasses the pipeline and visual redshift estimates and extends by $2000\,km\,s^{-1}$ on each side. We then calculate the significance of the Spearman's correlation coefficient at the redshift of the maximum cross-correlation. The significance of the deviation from zero is expressed in terms of a probability which we call P. A small P-value indicates a significant correlation. The Spearman's correlation test is performed over the regions of H_2 bands (from $v' = 0$ up to $v' = 9$), avoiding L(6-0), which is blended with Ly-β and restricting to $\lambda_{obs} > 3650$ Å because of the significantly increased noise level and frequent data issues. In order to ascertain the presence of strong H_2 lines, we also measure the median ratio of the flux at the expected position of the H_2 lines with respect to the flux in-between the lines. In other words, this parameter provides a measurement of the contrast. In what follows, this "flux ratio" parameter is denoted FR. An example of a quasar spectrum with H_2 detection is shown along with the comparison template in Fig. 1.

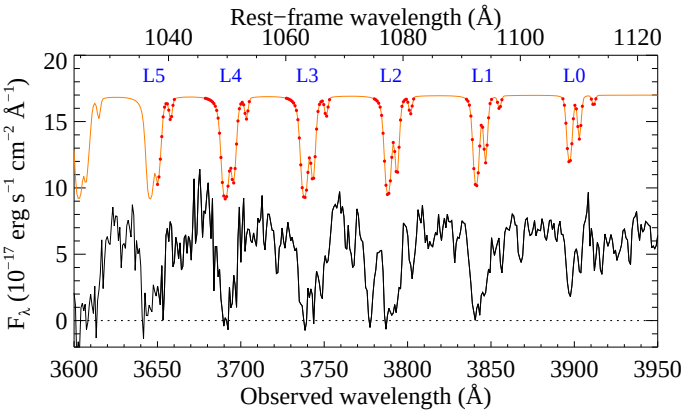

Fig. 1. Portion of the SDSS spectrum of quasar J 1031+2240 (black) with detected H_2 lines. The H_2 template is shown in orange arbitrarily scaled and shifted above the observed spectrum for visual clarity. The pixels used here to calculate the Spearman's correlation are highlighted by red dots. The blue label on the top of each Lyman (L) band indicates the vibrational level of the upper-state of that band.

2.3. Selection of the H_2 candidates and visual inspection

The distribution of the parameters P and FR for all the quasar spectra is shown in Fig. 2. The presence of a strong H_2 system in the search window is expected to result in small values for both P (i.e. high correlation significance) and FR (decrease in flux at expected position of H_2). The corresponding points also naturally appear as outliers compared to the main locus. Based on these considerations, we used two approaches to select the candidate H_2 absorbers. For the first approach (selection #1), we isolate all candidates (170) that have $\log P < -7$ and $FR < 0.75$ (dashed lines on Fig. 2), noting that beyond these values, it is generally hard to confirm or reject any putative H_2 system. We call this sample: $\mathcal{S}^P_{c1}$. This selection has the advantage of simplicity, but the number of candidates also increases quickly when both P and FR values increase, while the fraction of them being confirmed visually decreases. The second approach (selection #2) is based on a detection of outliers from the main locus of points in the (P, FR) parameter space. The selected candidates (188) are those found beyond the contour containing 99.73% of the points (equivalent to 3σ for normal statistics). We call this sample: $\mathcal{S}^P_{c2}$. One advantage of this selection is the possibility to explore candidates where one of the two parameters is peculiar for the given value of the other parameter. In particular, some systems may have strong H_2 lines (i.e. low FR), visually recognisable despite a low significance of correlation due to noisy data etc. There is a natural overlap between the two selections, with 78 candidates in common out of a total of 280 (coloured and black points in Fig. 2).

We visually inspected all these 280 candidates. During the visual inspection, not only did we pay attention to the region covering the position of the expected H_2 lines, but also to the overall

SDSS spectrum, looking for the presence of other signatures of absorption systems, such as metal, H I lines and dust features. Our visual inspection led to the confirmation of 50 strong proximate H_2 systems, coloured green in Fig. 2. For another 8 candidates (filled yellow), H_2 lines are likely present but it remains difficult to disregard the possibility that the lines are coincidence from the Ly-α forest. We assign a visual grade "A" for the former 50 and "B" for the latter 8 in Table 1. The remaining candidates are either clearly false positive systems or systems for which the data are inconclusive. The spectral regions covering the H_2 and H I lines are shown in the Appendix A.

We finally note that the visual inspection remains somewhat subjective by nature and it is still possible that systems graded A or B are spurious or that we missed H_2 systems among the selected (hence inspected) candidates. While we believe these fractions to be very small, follow-up data with higher S/N and resolution are required to firmly establish the quality of our visual inspection.

2.4. Additional proximate H_2 systems

In spite of effective selection criteria, during the code testing, we came across several candidates that were quite evident by visual inspection but remain inside the main locus of the parameter space (i.e. less significant than the 99.73% confidence level imposed above). Some proximate[1] H_2 systems may also be located outside the redshift window used to build our statistical sample. This can happen when the quasar redshifts provided by the DR14Q catalogue are wrong or when the absorbers are very significantly redshifted (i.e. more than our limit of 2000 km s^{-1}).

In order to explore differently or further inside the main locus or even systems not considered in the previous search, we performed a second, independent search using a method similar to that presented by Balashev et al. (2014). This independent method proved to be an efficient way to identify strong intervening H_2-bearing DLAs in the SDSS. We slightly modified the method, adjusting the numerical values that specify the criteria used to search for H_2-bearing DLAs. We again searched all $z > 2.5$ quasars, but used a 3000 km s^{-1} search window around the best redshift value reported by Pâris et al. (2018). The identification of probable H_2 systems is based on a "χ^2-like" selection function and the probabilities of false detection for the candidates were estimated using Monte Carlo simulation, as described by Balashev et al. (2014).

As before, we then visually inspected all 23 additional systems, i.e. new systems found by this second procedure, systems found by our main code but outside the selected statistical sample as well as serendipitous systems. Unsurprisingly, it is also generally more difficult to judge the reality of these additional systems, so that we ended up having a high fraction of grade B (11 out of 23) compared to our main selection. We also include these in Table 1; However, they are not considered for the statistical analysis of the incidence rate. The systems, for which we have measurements of the parameters FR and P at the same redshift but where FR and P fall within the rejection contour, are over-plotted in Fig. 2 as red and orange dots corresponding to visual grade A and B, respectively.

[1] Any system with $z_{abs} > z_{em}$ is naturally considered as proximate, independent of the exact velocity shift.

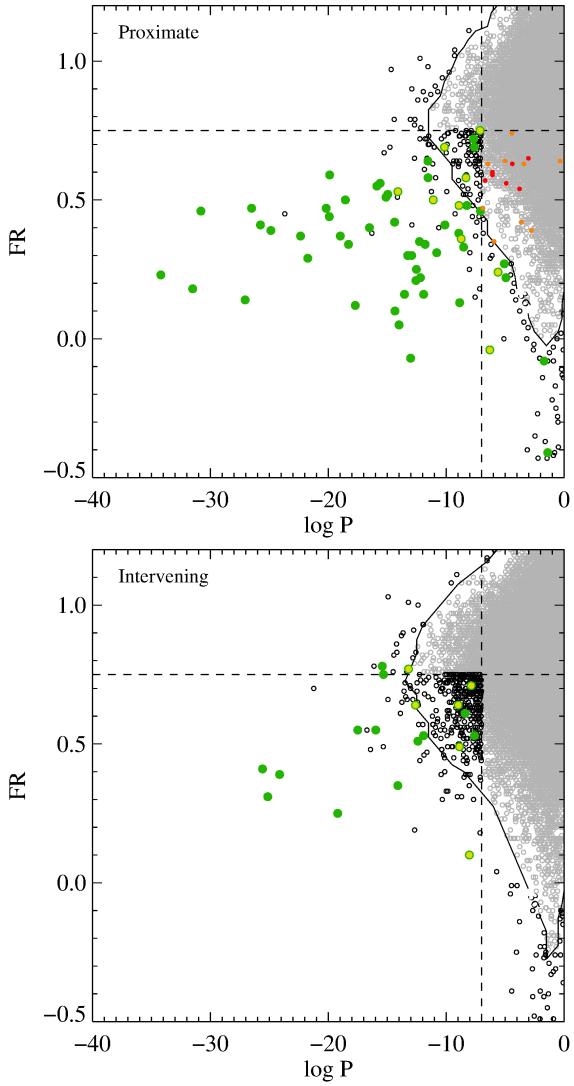

Fig. 2. Core-to-continuum median flux ratio versus significance of the Spearman correlation for all quasar spectra searched in a proximate velocity window (*top*, within a velocity window encompassing the pipeline and visual redshifts estimates, extended 2000 km s^{-1} on each side) and an intervening window with the exact same width for each spectrum, but shifted by 5000 km s^{-1} bluewards (*bottom*). The vertical and horizontal dotted lines show our cuts defining the samples $\mathcal{S}^P_{c1}$ (*top*) and $\mathcal{S}^I_{c1}$ (*bottom*). Points located outside the solid contour (containing 99.73% of the points) define, respectively, $\mathcal{S}^P_{c2}$ (*top*) and $\mathcal{S}^I_{c2}$ (*bottom*). Candidates belonging to either one or both of these selections (black points) were visually checked and coloured green when strong H_2 is confirmed (grade A) or yellow when considered tentative only (grade B). Red and orange points correspond to additional systems described in Sect. 2.4 with, respectively, grade A and B.

2.5. Note on sample completeness

The detection of additional H_2 systems inside our rejection contour indicates that the detection of strong H_2 in the overall parent sample of quasar is not complete. Indeed, the actual completeness of our statistical sample is expected to be a complex function of the quasar redshift and the S/N over the wavelength range where the H_2 bands are located. Furthermore, it depends on the column density of the H_2 system, the strength and exact location of Ly-α forest lines, and the presence of other absorption systems. In principle, this prevents us from deriving the *absolute* incidence of strong H_2 systems but should have little impact

Table 1. Sample of strong proximate H_2 absorbers in SDSS.

RA (J2000)	Dec (J2000)	MJD-plate-fiber [a]	z_{H2}	$\log N(\mathrm{H\,I})$ ($\mathrm{cm^{-2}}$)	$\log N(H_2)$ [b] ($\mathrm{cm^{-2}}$)	A_V [c] (mag)	$F_{\mathrm{Ly}-\alpha}$ (10^{-17} erg $\mathrm{s^{-1}\,cm^{-2}}$)	f_{leak} [d]	Flag [e]
00:15:14.82	18:42:12.34	56270-06111-0908	2.628	20.85	19.8	0.09	49.7 ± 1.9	0.19	3 A
00:19:30.55	−01:37:08.46	55536-04366-0874	2.529	21.00	20.0	0.03	4.5 ± 1.9	0.02	3 A
00:46:05.89	00:43:27.81	55444-04222-0981	2.940	20.00	19.3	0.10	-0.4 ± 0.8	0.00	0 A
00:59:17.64	11:24:07.70	56165-05706-0118	3.034	21.75	19.0	0.07	12.4 ± 2.4	0.04	1 A
01:02:11.89	02:52:07.18	57281-07858-0826	2.657	20.80	19.5	−0.00	-0.5 ± 1.5	0.00	0 A
01:25:55.11	−01:29:25.00	56898-07877-0966	2.665	21.75	20.2	0.17	70.0 ± 5.0	0.27	2 A
01:26:54.45	11:38:23.29	55831-04669-0080	2.603	20.50	19.2	0.04	21.3 ± 1.3	0.07	3 A
01:36:44.02	04:40:39.10	55508-04274-0691	2.779	20.75	19.5	−0.04	1.6 ± 1.3	0.01	3 A
02:16:02.33	04:13:57.35	55486-04266-0012	2.661	20.55	19.7	−0.34	16.3 ± 1.4	0.12	3 A
02:19:26.55	−01:10:57.30	55478-04237-0364	2.812	20.00	19.0	0.42	7.9 ± 0.6	0.23	0 B
02:23:16.90	−03:07:21.42	56904-07832-0378	2.583	22.00	20.0	0.01	23.9 ± 3.6	0.08	3 A
07:54:02.68	43:59:28.25	57067-08276-0092	2.948	21.40	19.5	0.22	71.6 ± 3.0	0.36	3 A
07:56:34.69	11:23:30.35	55602-04511-0874	3.315	21.00	20.1	−0.15	1.9 ± 1.7	0.03	3 A
07:59:01.28	28:47:03.43	55535-04453-0850	2.822	21.30	19.5	−0.03	15.6 ± 1.6	0.04	3 A
08:03:51.64	50:03:17.65	56992-07317-0693	2.977	20.65	19.0	0.62	3.6 ± 1.0	0.03	0 A
08:07:57.45	51:52:34.24	56741-07377-0768	3.124	21.00	19.3	0.11	8.8 ± 1.4	0.10	0 A
08:16:14.21	03:58:19.77	55869-04763-0096	3.682	20.00	19.0	−0.00	5.9 ± 1.4	0.08	3 B
08:21:26.13	36:26:06.10	57449-08857-0340	2.597	21.30	19.3	−0.01	-4.0 ± 2.3	0.00	3 A
08:55:02.19	42:09:37.16	57375-08296-0646	2.719	22.30	19.8	−0.16	-3.6 ± 3.8	0.00	3 A
08:58:59.67	17:49:25.19	55913-05297-0566	2.625	20.40	19.8	0.05	-0.5 ± 1.5	0.00	3 A
09:11:46.70	41:10:48.00	55999-04603-0104	2.840	21.15	20.0	−0.06	38.8 ± 2.2	0.12	3 A
09:18:50.50	52:20:03.56	57039-07289-0121	3.230	21.35	20.0	0.25	1.6 ± 1.4	0.02	3 A
09:47:38.40	06:38:00.16	55926-04873-0354	2.954	20.80	19.2	−0.12	13.8 ± 1.4	0.23	0 A
09:48:27.31	07:22:36.97	55926-04873-0684	2.671	21.40	20.0	−0.01	10.8 ± 2.0	0.05	3 A
09:55:49.68	60:57:41.34	56014-05719-0786	2.894	20.10	18.8	0.08	4.3 ± 1.2	0.05	1 B
10:06:40.78	36:02:34.14	57453-08856-0500	3.210	21.20	19.1	0.25	-2.5 ± 2.1	−0.01	0 A
10:08:52.61	47:55:52.51	56338-06663-0946	2.852	21.55	20.1	−0.13	-2.3 ± 2.7	0.00	3 A
10:11:04.27	32:39:11.00	56329-06461-0750	2.540	21.35	19.8	0.08	0.6 ± 1.8	0.01	2 A
10:14:04.96	27:23:11.48	56334-06470-0814	3.025	21.75	19.9	0.24	7.6 ± 2.5	0.07	2 B
10:31:43.87	22:40:15.19	56298-06425-0304	2.516	21.50	20.1	0.09	10.3 ± 1.6	0.02	3 A
10:45:02.97	15:02:11.36	56009-05350-0802	3.657	21.60	20.6	0.07	2.3 ± 1.8	0.03	3 A
10:47:21.79	29:15:47.78	56356-06449-0618	2.978	20.65	19.5	0.14	0.6 ± 1.5	0.01	3 A
10:52:51.23	18:09:15.66	56035-05887-0388	2.866	21.50	20.8	0.36	1.0 ± 2.1	0.01	3 A
11:10:41.05	67:15:50.46	56741-07111-0732	3.343	20.60	19.2	−0.03	3.4 ± 1.1	0.04	1 A
11:11:03.78	33:16:14.09	56370-06440-0770	2.960	20.00	19.0	0.10	10.4 ± 1.0	0.20	0 B
11:11:24.18	14:47:56.09	56017-05362-0176	3.049	21.70	19.1	0.05	-1.8 ± 2.7	0.00	3 A
11:22:14.56	42:19:58.69	56013-04686-0206	2.620	20.60	19.9	−0.02	15.6 ± 1.3	0.12	3 A
11:23:40.09	15:53:53.66	55986-05367-0926	3.392	21.20	19.6	0.27	-0.2 ± 1.5	0.00	0 B
11:31:55.38	08:12:39.18	55947-05374-0724	2.716	21.20	19.2	0.22	36.7 ± 1.2	0.44	0 B
11:35:17.10	29:57:22.08	56342-06405-0832	2.694	21.40	19.3	0.08	4.5 ± 2.0	0.02	0 B
11:44:10.88	47:34:51.98	56401-06678-0237	2.588	21.00	19.7	0.32	11.7 ± 2.6	0.08	2 A
11:44:52.49	20:36:44.15	56309-06432-0648	3.118	21.05	19.0	0.09	0.9 ± 1.3	0.01	0 B
11:45:52.44	49:51:14.26	56412-06684-0118	3.043	20.80	19.0	0.13	5.6 ± 1.2	0.04	0 A
11:53:14.86	58:14:40.20	56658-07091-0724	2.663	21.00	20.0	0.77	0.0 ± 1.4	0.01	2 A
12:00:51.84	14:08:31.46	56009-05387-0256	3.034	19.35	19.3	0.19	3.9 ± 0.5	0.16	3 B
12:28:56.08	31:47:53.86	56364-06479-0390	2.544	20.55	19.5	0.02	2.7 ± 1.2	0.04	0 B
12:36:02.11	00:10:24.54	55647-03848-0266	3.033	20.55	19.5	−0.09	-0.1 ± 1.4	0.00	3 A
12:41:10.70	38:35:36.46	57424-08836-0360	2.950	20.80	19.3	0.14	-2.2 ± 1.5	−0.01	3 B
12:42:01.74	20:06:25.54	56088-05855-0612	3.387	20.95	19.7	0.12	7.5 ± 1.4	0.06	3 A
12:42:34.82	44:48:13.10	56365-06617-0748	2.872	21.40	18.9	0.21	36.9 ± 1.6	0.28	0 A
12:45:27.17	32:21:36.08	56358-06482-0070	3.206	20.35	18.7	0.10	2.4 ± 0.9	0.04	0 B
12:48:29.51	06:39:35.62	56016-04835-0844	2.530	20.55	19.8	−0.07	-1.2 ± 1.3	0.00	3 A
12:59:17.31	03:09:22.51	55325-04005-0624	3.246	21.40	19.1	−0.09	24.7 ± 2.1	0.03	3 A
13:07:09.97	30:28:18.71	56362-06487-0380	3.594	21.35	19.8	0.13	22.9 ± 3.0	0.14	3 A
13:11:29.11	22:25:52.58	56066-05992-0136	3.092	20.80	19.4	0.09	-3.0 ± 1.5	0.00	3 A
13:19:04.34	13:56:49.57	56003-05425-0142	3.253	20.65	18.9	−0.01	4.2 ± 0.9	0.05	1 B

Notes. [a]Modified Julian Date, plate number and fibre number corresponding to the SDSS spectra used in this work. [b]H_2 column densities values correspond to those used in the figures shown in the Appendix A. These should be considered as indicative only and require confirmation at higher spectral resolution (see Sect. 4.1). [c]Typical uncertainty of 0.12 mag from intrinsic quasar power-law variations (see Sect. 4.4). [d]Conservative uncertainty of about 0.3 dex (see Sect. 4.2). [e]The flag number corresponds to the selection (0: non-statistical systems described in Sect. 2.4), 1: systems belonging to $\mathcal{S}^P_{c1}$; 2: systems belonging to $\mathcal{S}^P_{c2}$; 3=1+2. The flag letter corresponds to the visual classification (grade A or B).

Table 1. continued.

RA (J2000)	Dec (J2000)	MJD-plate-fiber [a]	z_{H2}	$\log N(\mathrm{H\,I})$ ($\mathrm{cm^{-2}}$)	$\log N(\mathrm{H_2})$ [b] ($\mathrm{cm^{-2}}$)	A_V [c] (mag)	$F_{\mathrm{Ly}-\alpha}$ (10^{-17} erg $\mathrm{s^{-1}\,cm^{-2}}$)	f_{leak} [d]	Flag [e]
13:26:18.07	25:58:51.56	56096-05996-0912	2.887	21.60	20.2	0.31	-1.6 ± 2.4	0.00	3 A
13:31:11.41	02:06:09.06	55622-04045-0836	2.922	21.30	19.6	−0.01	34.9 ± 2.3	0.15	3 A
13:36:50.63	25:13:28.32	56089-05999-0079	2.810	20.25	19.5	0.03	0.1 ± 0.8	0.01	0 B
13:58:08.94	14:10:53.29	56014-05446-0139	2.892	21.20	19.4	0.02	4.1 ± 1.8	0.02	1 A
13:59:35.08	09:00:51.36	55958-05447-0810	2.766	20.30	19.5	0.08	0.2 ± 1.7	0.01	0 A
14:11:18.14	02:15:12.03	55634-04030-0564	2.969	20.10	19.2	−0.02	1.0 ± 1.0	0.03	0 A
14:17:45.90	36:21:27.46	55246-03859-0592	3.339	22.45	19.9	0.12	3.6 ± 2.1	0.02	3 A
14:21:41.26	52:45:51.69	56799-07339-0068	2.653	21.45	19.6	0.40	5.8 ± 1.5	0.04	0 A
14:30:42.90	05:49:07.95	55691-04860-0172	2.837	20.80	20.0	0.02	3.4 ± 1.9	0.05	3 A
14:49:24.08	24:41:23.11	56067-06021-0794	3.293	20.80	19.0	0.17	1.0 ± 0.9	0.02	3 A
15:04:50.99	30:22:45.32	55245-03876-0340	2.781	21.60	20.2	0.04	7.7 ± 1.7	0.06	3 A
15:12:22.27	38:21:07.40	56066-05167-0082	2.977	21.65	19.3	−0.09	4.9 ± 2.7	0.03	3 A
15:15:04.52	14:48:23.19	56030-05486-0112	2.782	20.70	20.0	0.77	3.7 ± 1.4	0.08	2 B
15:44:43.05	18:29:45.89	55352-03937-0212	3.166	21.45	19.4	−0.18	15.5 ± 2.1	0.08	3 B
15:46:20.43	42:44:51.62	56101-06042-0364	3.324	20.90	19.5	−0.11	1.2 ± 1.4	0.02	3 A
16:04:32.33	57:57:22.03	56448-06786-0724	2.867	20.65	19.0	−0.15	0.0 ± 1.3	0.00	1 A
16:06:38.54	33:34:32.98	55721-04965-0091	3.085	20.80	19.0	0.05	-0.9 ± 1.2	0.00	3 A
16:38:31.22	48:57:26.04	57186-08056-0154	2.757	20.30	19.8	0.22	17.0 ± 1.2	0.13	0 A
16:41:04.00	43:33:56.06	56091-06031-0436	3.024	20.20	19.0	0.45	-1.2 ± 1.0	−0.03	0 B
17:06:15.68	37:56:13.71	55836-04983-0122	2.831	21.65	20.1	0.19	16.3 ± 2.6	0.05	3 A
21:58:28.89	24:04:35.77	57311-07640-0252	2.688	21.30	20.0	0.28	1.2 ± 1.8	0.00	3 A
22:28:07.36	−02:21:17.17	55857-04380-0296	2.770	20.85	19.5	−0.08	8.8 ± 1.2	0.01	3 A
22:55:06.64	19:49:10.14	56959-07609-0172	2.914	20.50	19.2	0.05	-0.7 ± 1.1	−0.01	0 B
23:25:06.62	15:39:29.31	56267-06143-0586	2.615	21.70	20.0	0.02	3.8 ± 2.6	0.01	3 A
23:36:07.16	22:53:25.83	56566-06519-0141	2.701	21.00	19.7	0.04	11.3 ± 1.6	0.08	3 A

on the *relative* incidence between proximate and intervening systems discussed in the next section. We can still roughly estimate the overall H_2 detection rate in PDLAs using the statistical sample (i.e. $\log N(\mathrm{H\,I}) > 21.1$) of metal-selected PDLAs from Fathivavsari et al. (2018a). This sample contains 201 systems with $z > 2.5$ searched by our code, among which we found 20 H_2-bearing systems (18 grade A and 2 grade B) within our statistical selection, plus another 5 in our list of additional systems. This implies a H_2 covering fraction higher than 10% in strong ($\log N(\mathrm{H\,I}) > 21.1$) metal-selected PDLAs. This appears to be in qualitative agreement with the H_2 covering fraction for intervening systems. For example, Balashev & Noterdaeme (2018) found 4% (DLAs/sub-DLAs with $\log N(\mathrm{H\,I}) > 20$), 8% (DLAs with prominent metal lines) and 37% (extremely strong DLAs with $\log N(\mathrm{H\,I}) > 21.7$).

3. The excess of strong proximate H_2 absorbers

In this section, we investigate whether or not there is an excess of strong proximate H_2 systems compared to what is expected from intervening systems. In other words, we wish to quantify whether or not there is a higher probability for a H_2 cloud to be located close to the quasar in velocity space. To do this, we apply the exact same procedure, selection and visual inspection as for our statistical sample of proximate systems, with the only difference that we shifted the search window for each individual spectrum by 5000 $\mathrm{km\,s^{-1}}$ to the blue. This velocity shift corresponds to what is typically considered a safe limit to treat the systems as intervening. At the same time, the velocity shift is large enough to avoid overlap of the search window with that used for proximate H_2 systems while being small enough so that the probed spectral regions and the redshifts remain very similar. In spite of this, a slight shift is observed for the main locus in the (P, FR) parameter space as compared to proximate systems. This results in a larger number of candidates following selection 1 ($\mathcal{S}^I_{c1}$ with 396 candidates). However, these are mostly seen close to the chosen limits and the bottom-left corner of the plot (with a high probability of a given system to be real) is clearly much less populated than for proximate candidates. This alone already tells us that the incidence of strong intervening systems per velocity bin is much lower than for the proximate systems. Applying our outlier selection (#2), we obtain a total of 174 candidates ($\mathcal{S}^I_{c2}$). From visual inspection of all 525 candidates (45 are in common between the two selections), only 13 are graded A ($\mathcal{S}^I_A$) and 6 are graded B ($\mathcal{S}^I_B$).

In Fig. 3 we present the distribution of velocity offsets,

$$\Delta v \equiv c\frac{R^2-1}{R^2+1}, \tag{1}$$

where $R \equiv (1+z_{\mathrm{abs}})/(1+z_{\mathrm{em}})$ for the strong H_2 systems detected in both search windows (i.e. centred on z_{em} and shifted bluewards by 5000 $\mathrm{km\,s^{-1}}$). For a fair comparison of the two distributions, we used only those systems satisfying selection 2, but note that the results do not change significantly when using selection 1 or the union or intersection of both selections. The shaded regions in Fig. 3 show the minimal 4000 $\mathrm{km\,s^{-1}}$-wide search windows. Both the intervening and the proximate distribution slightly extend beyond these boundaries as the search windows for each spectrum were defined to take into account the uncertainties on the quasar redshift. The statistical results discussed below are however strictly restricted to systems falling in the respective 4000 $\mathrm{km\,s^{-1}}$ windows. The intervening systems are uniformly distributed over the velocity interval, which is expected for systems randomly intercepted by a quasar line of sight. On the other hand, proximate systems are on average 5 times more numerous (4.2 if including grade B systems as well) at $\Delta v = 0 \pm 2000\,\mathrm{km\,s^{-1}}$ than at $\Delta v = -5000 \pm 2000\,\mathrm{km\,s^{-1}}$ (shaded areas in Fig. 3). These are conservative lower limits

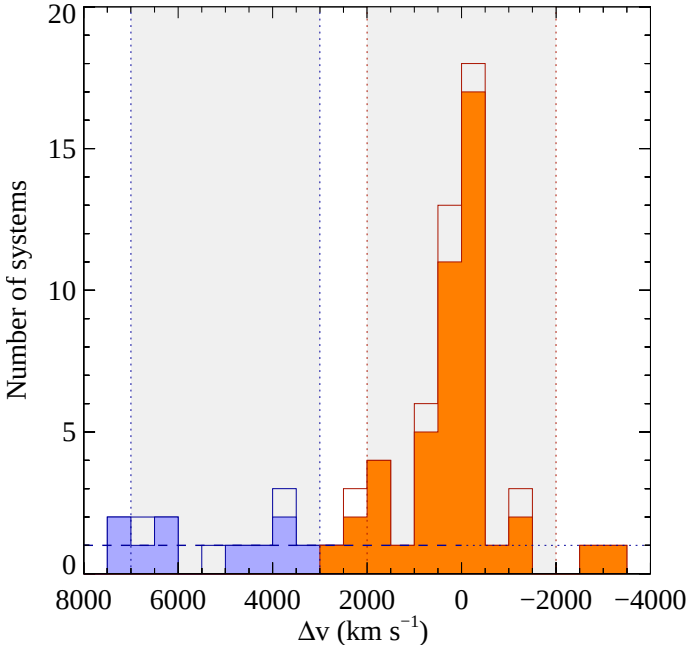

Fig. 3. Distribution of relative velocities with respect to the quasar redshift for our sample of strong proximate H_2 systems (orange histograms) compared to those found in a region shifted by 5000 km s^{-1} (blue). We here used the "zbest" provided by the DR14Q catalogue as the quasar redshift and the z_{abs} measurement directly from our search algorithm. Negative velocities indicate $z_{abs} > z_{em}$. Note that the x-axis goes from positive velocities (blueshifted compared to the quasar) on the left to negative velocities (redshifted) to the right. Both distributions are restricted to visually-checked systems (unfilled histograms: grade A or B, filled histograms: grade A only) isolated using the outlier selection (# 2). The grey regions show the corresponding minimal search windows. Systems falling outside these regions are not considered when comparing incidence rates. The horizontal dashed line shows the mean number of intervening strong H_2 systems per velocity bin (∼1 per 500 km s^{-1} bin). A significant excess of H_2 systems at the quasar redshift is observed and cannot be explained by intervening statistics.

since the number of intervening systems at significantly negative velocities (i.e. $z_{abs} > z_{em}$) should be close to zero, as we expect little peculiar velocities of intervening gas to shift systems in that region. The distribution of proximate systems is also clearly peaked around the quasar redshift. The excess of proximate systems is about a factor of 2.5 in the velocity range from 1000 to 2000 km s^{-1} compared to what is expected from the statistics of purely intervening systems (dashed blue horizontal line). In the central 1000 km s^{-1}, however, 28 strong H_2 systems are seen when ∼2 are expected from intervening statistics. We note that the uncertainty on the quasar emission redshifts as provided by the SDSS quasar catalogue is of the order of 500−1000 km s^{-1}. Hence, the observed distribution of proximate H_2 absorbers may well appear wider than it is intrinsically.

In summary, we observe more than an order of magnitude excess of H_2 absorbers close to the quasar compared to what is expected from chance alignment with the quasar. This means that most of the proximate H_2 systems presented in this work must be related to the quasar environment and not to intervening galaxies in the Hubble flow. The question now becomes whether these systems are directly associated to the quasar, its host galaxy, or arise from galaxies in the quasar group environment. In the absence of detailed understanding of the physical conditions in the clouds, this is a difficult question to answer. In the following sections, we shed light on this from the observed properties of the proximate H_2 systems as seen in the SDSS data.

4. Properties of the proximate H_2 systems

In this section, we derive some of the main properties of the proximate H_2 systems from the SDSS data alone. These are the atomic and molecular hydrogen column densities, Ly-α emission, metal content, and dust properties.

4.1. H I and H_2 column densities

We fitted a Voigt profile to the damped Ly-α line keeping the redshift fixed to that obtained from H_2 and metal lines. We also simultaneously fitted the other lines from the Lyman series and estimated the quasar continuum using a spline function. Since the latter task is complicated by the quasar blended Ly-α and N v emission lines, we guided the placement of the spline knots using the quasar composite spectrum from Vanden Berk et al. (2001) matched to the spectrum redwards of the quasar Ly-α emission. When necessary we adjusted the strength of the Ly-α+N v emission line by considering those of other emission lines in the spectrum. However, the derivation of the exact unabsorbed continuum will inevitably partly rely on implicit assumptions about the shape and strength of the Ly-α+N v emission line, which are hard to quantify. We therefore paid particular attention to the width of the profile close to the bottom, which is little influenced by the exact continuum placement but note that in some cases, it can still be affected by the presence of strong leaking Ly-α emission. The measurement of the H I column density was also helped by the presence of Ly-β and other Lyman series lines for which the emission-line-to-continuum ratio is different. The obtained N(H I) then sets the strength of the DLA (or sub-DLA) wings, and the continuum is then re-adjusted if necessary until we obtain a satisfactory fit. During this process, we remarked that the obtained H I column densities typically varied by no more than 0.2 dex. Our final N(H I) measurements are given in Table 1 and the corresponding figures in the Appendix A. We note that automatically determined N(H I)-measurements of intervening DLAs based on Ly-α absorption only have typical uncertainties of 0.2 dex in SDSS (Noterdaeme et al. 2009). In the case of proximate DLAs, follow-up observations by Ellison et al. (2010) are actually in very good agreement (∼0.05 dex) with those obtained by Prochaska et al. (2008a) from SDSS data using a manual fitting scheme very similar to the one used here. Nevertheless, since we here discuss the overall population, the N(H I) uncertainty for individual systems does not affect the main results and conclusion of the paper.

We also obtain rough estimates of the H_2 column densities by manually adjusting the total column density of a H_2 template with a fixed excitation temperature and fixed Doppler parameter. We derive typical column densities of $\log N(H_2) \sim 19.5$ but caution that individual values are very uncertain in the absence of high-S/N, medium/high-resolution spectroscopy. The values of $N(H_2)$ provided in Table 1 should then be considered as indicative only. We remark that we already have medium or high resolution data for several H_2-bearing DLAs (including four from this sample: J1311+2225, Noterdaeme et al. 2018, J0136+0440, J0858+1749, J1236+0010, Balashev in prep.), in which we found that SDSS-based values typically underestimate the H_2 column density by up to 0.3 dex. In one outlier, however, the H_2 column density differs by about 0.8 dex compared to the SDSS-based estimate. Therefore, while the H_2 lines are intrinsically in

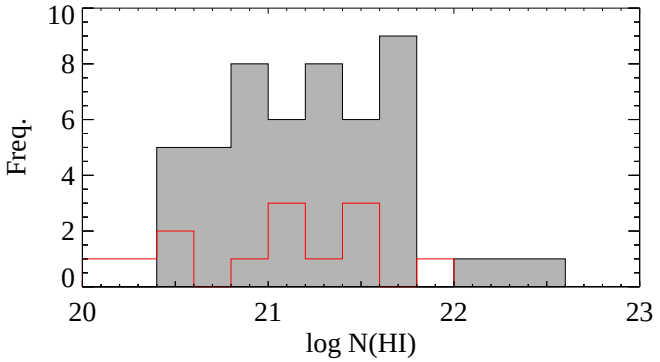

Fig. 4. Distribution of H I column densities in our statistical samples with visual grade A (proximate: filled, intervening: red).

the saturated regime, we do not use the column density estimates in the following.

The distribution of H I column densities for intervening and proximate DLAs is shown in Fig. 4. The observed distribution for H_2-bearing PDLAs is slightly shifted towards higher H I column densities (by about 0.3 dex) compared to intervening H_2-bearing DLAs. This may be due to the fact that higher H I column densities are necessary for the H_2/H I transition closer to a strong UV source, as expected from transition theories (e.g. Krumholz et al. 2008; Sternberg et al. 2014), if the other parameters are kept unchanged. It is also possible that part of the observed H I is unrelated to the H_2 gas, and the excess column density is only due to a more gas-rich environment close to the quasar.

Our H_2-selection of PDLAs can also provide an independent estimate of PDLA clustering close to the quasar. Indeed, if the conditions for the formation of H_2 are not very different, then the observed factor of 5 excess of proximate H_2 over intervening H_2 systems corresponds to the excess of proximate DLAs over intervening DLAs. This is well above the factor of two excess found by Prochaska et al. (2008a) in the SDSS-II. If, in turn, H_2 is more difficult to form in the quasar environment (as we could naively expect from the strong UV field), then the discrepancy is even larger. We note however that the PDLA detection algorithm from Prochaska et al. (2008a) was based on the zero-flux in the core of the DLA and hence likely missed most of the systems with leaking Ly-α emission. The clustering of neutral gas around the quasar could also depend on the column density, being stronger at high N(H I) (as observed here) than for the overall population of DLAs. Finally, it remains possible that H_2 is instead formed more efficiently in the quasar environment (i.e. a positive AGN feedback) owing to higher metallicities, larger total surface of dust grains or gas compression.

4.2. Leaking Ly-α emission

Significant (>3σ) residual flux in the core of the DLA absorption is the most evident peculiar feature of our systems and observed in about half of our sample. We measured the total Ly-α flux ($F_{\mathrm{Ly}-\alpha}$) for each system by integrating the observed flux spectrum over the DLA trough. The associated uncertainty is obtained from the error spectrum. These values are robust since they do not depend on the assumed unabsorbed quasar emission and are provided in Table 1 for reference. However, since the Ly-α emission can be strong and is generally broad, it most likely corresponds to leaking Ly-α photons from the background quasars' emission line regions rather than arising solely

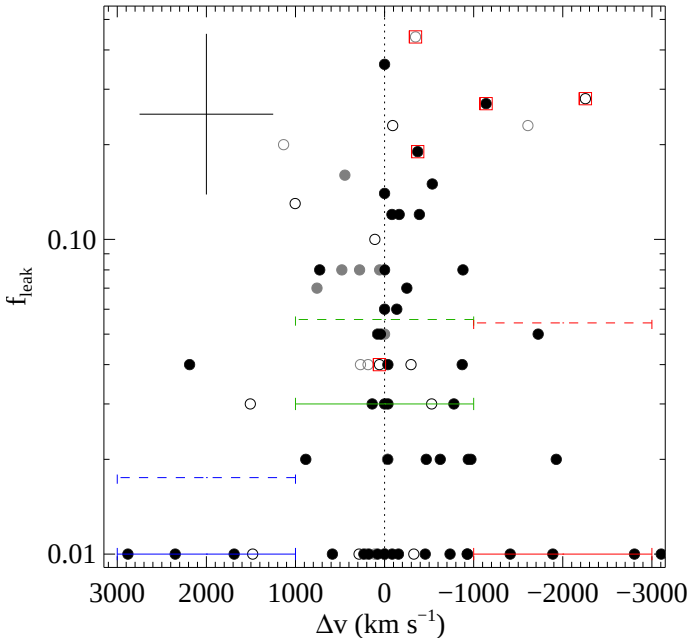

Fig. 5. Fraction of leaking Ly-α photons at the core of the DLAs as a function of its relative velocity to the quasar redshift. Filled points correspond to systems from the two statistical selections described in Sect. 2 (flag $\neq$ 0 in Table 1). Unfilled symbols correspond to the additional systems described in Sect. 2.4 (flag = 0). The colour indicates the visual classification (black:A, grey:B). Finally, red squares are overplotted on top of systems with clear Si II* absorption. The solid (resp. dashed) segments correspond to the median (resp. mean) values in different velocity bins, using only statistical rank A systems. Values measured to be less than 0.01 are set to 0.01 for plotting convenience. The cross at the top-left corner shows typical (albeit conservative) uncertainties along both axes.

from local star-formation activity in the quasar host. Therefore, the most interesting quantity to consider is actually the *fraction* of leaking photons at the DLA wavelength rather than the actual luminosity of this residual. Thus, we define f_{leak} as the ratio of the observed flux integrated in the DLA core over the unabsorbed flux integrated over the same region. In spite of our efforts to reconstruct the unabsorbed quasar continuum (see previous section), the fraction f_{leak} is highly uncertain. However, it remains a convenient way to distinguish between systems that allow a significant fraction of photons to leak at the DLA wavelength, and those systems that do not support such a leakage[2]. We assign a conservative estimate of the uncertainty of a factor of two to take into account the observed dispersion of Ly-α-emission-to-continuum ratio seen between different quasars (e.g. Selsing et al. 2016).

Splitting the sample into two sub-samples with f_{leak} above or below the median value (0.02), we then found that the systems with high f_{leak} are located on average twice closer in velocity space than those with low f_{leak} ($|\Delta v| \sim 500$ vs 1000 km s^{-1}). Figure 5 illustrates this further with f_{leak} plotted as a function of the relative velocity with respect to the quasar redshift[3].

[2] We note that f_{leak} represents the total leaking fraction of photons at the DLA wavelength. In other words, this includes not only Ly-α photons but also photons from the continuum. The actual fraction of escaping Ly-α photons at the DLA wavelength should then be slightly higher than the f_{leak} values.

[3] Whenever necessary, we corrected the quasar redshift provided in the DR14Q catalogue through a careful reassesment using the reddened composite.

Systems without significant emission span the full range of velocities, while systems with high $f_{\rm leak}$ tend to concentrate closer to zero velocities. Separating the systems according to their velocity shift to the quasar, we can indeed see that the mean and median $f_{\rm leak}$ values are higher at small velocity separation than at high velocity separation. Interestingly enough, we note that the leaking fraction seems to be higher for systems with $\Delta v < -1000\,{\rm km\,s^{-1}}$ than for those with $\Delta v > 1000\,{\rm km\,s^{-1}}$. To summarise, it appears that DLAs with absorption redshift very close to that of the quasar emission cover less of the corresponding Ly-α photons than those with significant velocity shifts. Among the latter, those redshifted compared to the quasar (i.e. moving towards the quasar) tend to cover less than those moving away from the quasar.

The observed dependence of $f_{\rm leak}$ on the relative absorber to quasar velocity can in principle be explained as a purely observational effect. DLAs redshifted onto either wing of the quasar Ly-α emission will absorb Ly-α photons with wavelengths shifted relatively far away from resonance ($1215.67\,{\rm Å} \times (1+z_{\rm QSO})$) and hence arising mostly from the BLR. Conversely, DLAs located exactly at the quasar redshift correspond to Ly-α photons arising both from the BLR and from narrower Ly-α emission arising from regions further away from the central engine, up to the very outskirts of the quasar host (see e.g. Fathivavsari et al. 2015). This "narrow" and likely more extended component can therefore more easily leak through the absorbers. If this is the case, then we can expect that intervening H_2-bearing clouds also have projected sizes smaller than emission region of the quasar at the peak Ly-α wavelength. This potentially could be detected as a partial coverage effect in the metal absorption lines. Indeed, Balashev et al. (2017) have recently observed an unambiguous partial coverage of the Ly-α emission by the S II absorbing gas (see their Fig. 12) associated to an intervening DLA ($z_{\rm abs} = 2.786$, $z_{\rm QSO} = 2.92$) with damped H_2 lines. A systematic study of the partial coverage of Ly-α emission by different absorbing clouds, and as a function of wavelength shift compared to systemic redshift, would provide clues on the origin and extent of the different Ly-α emission components.

However, there may also be a physical reason for the clouds at small velocity separation covering statistically less of the Ly-α emission than those at large velocity separation. Indeed, neutral gas clouds close to the UV source may typically have higher density and hence be smaller (for a given column density) than those located farther away, as proposed by Fathivavsari et al. (2017). This would also explain the observations if systems close in velocity space are also statistically closer in distance. This is a valid possibility as clouds rotating with the quasar galaxy host should have little velocity along the line of sight while those located in other galaxies of the group could have larger $|\Delta v|$. Gas flows (either winds or infall) can however complicate the picture, being located relatively close to the source but still possibly having large relative velocities. Interestingly, there is a trend for systems with positive velocities (possibly due to infalling gas) to have larger leaking fraction and also featuring at the same time excited levels of silicon (red squares on Fig. 5). Both collisions (denser cloud) and enhanced UV field (closer to quasar) would help populating the fine-structure levels.

All this means that the presence of leaking Ly-α alone is probably not enough to differentiate between wavelength dependence of the emission size or distance dependence of the size of the absorbing cloud. However, metal lines (in particular in excited states) as well as molecular lines may provide further information in order to distinguishing between the different scenarios.

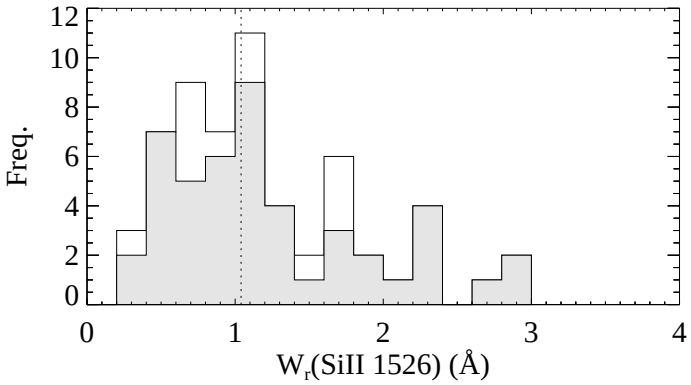

Fig. 6. Distribution of the rest-frame equivalent widths of Si II λ1526 for grade A systems, including (unfilled) or not (filled) the additional systems described in Sect. 2.4. The vertical line shows the median value (identical for the two samples).

Finally, we note that it is very likely that other clouds, similar to that giving rise to the DLA, are located in the same galaxy (e.g., the quasar host or a group member) yet spatially offset from the line of sight to the quasar central engine. While these clouds do not intercept the line of sight to the compact continuum source they may still contribute to the absorption of the spatially extended Ly-α emission. Absorption signatures of such clouds would however be very difficult to identify. Only detailed measurements of absorption lines falling on top of emission lines, arising from the spatially extended emission region, would reveal the presence of such complex absorption geometries. In order to carry out such detailed analyses of the absorption and emission geometry, higher resolution spectroscopy with better S/N is required.

We also caution that the uncertainties on Δv are large and dominated by the uncertainty on the quasar redshift. Measuring accurately the quasar systemic redshift through follow-up observations of the narrow forbidden emission lines in the near infrared would be imperative to confirm or reject the above discussed trends.

4.3. Metal lines

Metal absorption lines are systematically seen associated to the H_2 systems. However, at the typical S/N and given the low resolution of the SDSS spectra, the only information we can obtain is the equivalent width of strong lines, which are very likely intrinsically saturated. The equivalent width of such lines is therefore mostly determined by the velocity spread of the profile. Observationally, high resolution studies of DLAs indicate that the velocity extent of metal lines correlates well with the metallicity (Ledoux et al. 2006). This means that we can in principle use the observed equivalent width to get an idea of the metallicity. We measured the Si II λ1526 equivalent widths using an automated procedure and obtain the distributions shown on Fig. 6. The median equivalent width in our statistical sample is about 1 Å, i.e. similar to that observed by Balashev et al. (2014) for the population of strong intervening H_2 systems. Using the empirical relation $[{\rm X/H}] = -0.92 + 1.41\log(W_r\lambda1526)$ from Prochaska et al. (2008b), the median equivalent width corresponds to a metallicity of about one tenth Solar. However, we caution that this empirical equivalent-width metallicity relation has been obtained using intervening systems and thus may not actually apply here.

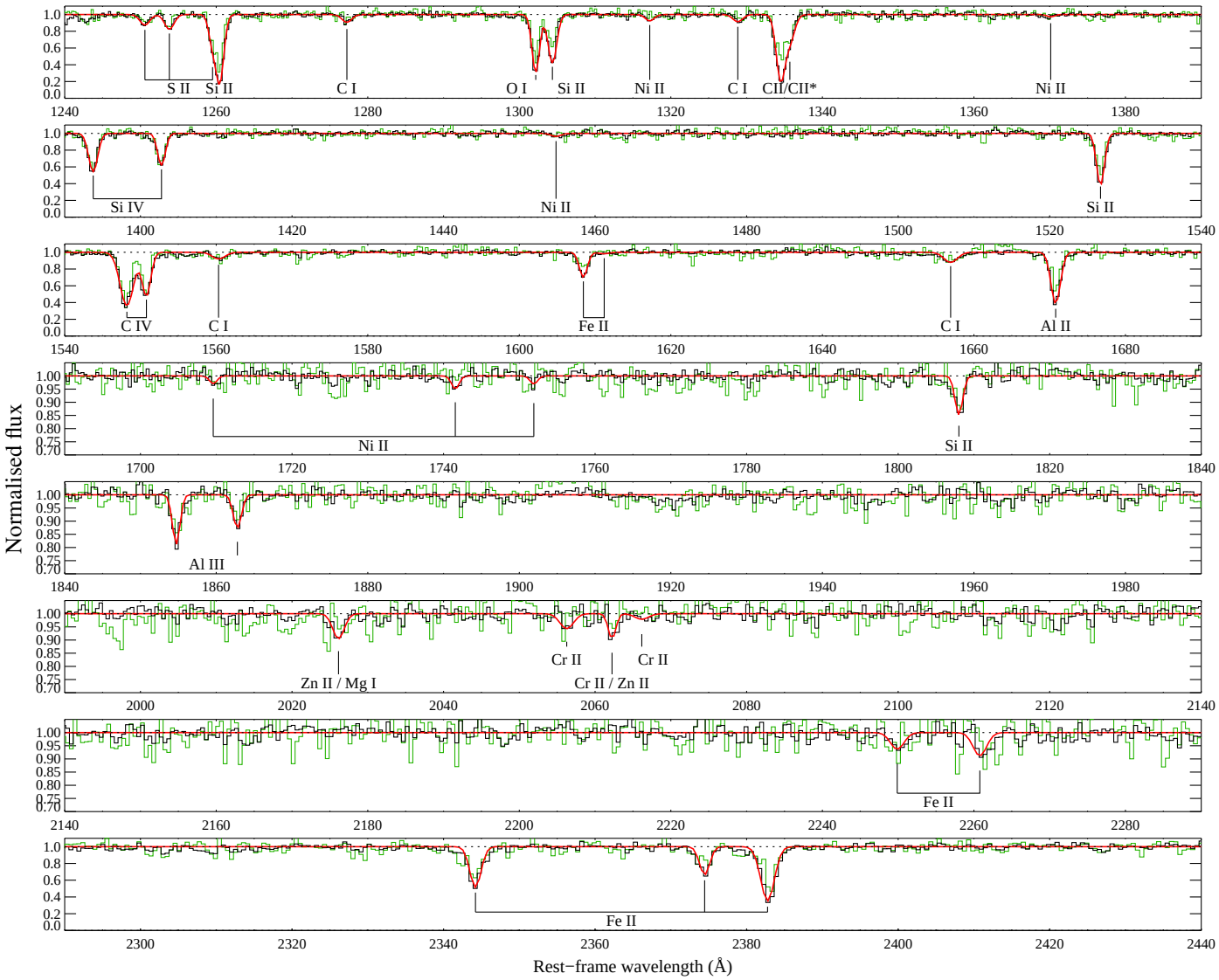

Fig. 7. Composite spectrum obtained by median-averaging all grade-A systems (black, with a Gaussian fit over-plotted in red) and a sub-set with significant Ly-α leakage ($f_{\rm leak} > 0.5$, green). The vertical scale is adapted for each panel to maximise the visibility of the lines.

Therefore, we further test this result using a stacked spectrum built by median averaging all systems visually classified A. The obtained composite spectrum, shown in Fig. 7 has a S/N of about 50, allowing us to detect weak absorption lines that are otherwise undetectable in individual spectra and whose equivalent width will then depend rather on the column density than the velocity extent. The typical species seen in the overall population of DLAs are detected but we also detect significant C I lines, that are otherwise much less frequent in DLAs (Ledoux et al. 2015). This is consistent with our H_2 selection since C I is known to be a good tracer of molecular gas (Noterdaeme et al. 2018).

Using the unblended and undepleted S II λ1253 line, and assuming optically thin regime, we obtain a metallicity of about [S/H] ~ -0.9 using the median $\log N({\rm H\,I}) = 21$. Similarly, we obtain [Zn/H] ~ -0.8 from Zn II λ2026 and [Si/H] ~ -1 from Si II λ1808. This exercise shows us that the average metallicity of our sample should be roughly 1/10th of the Solar value. This is higher than the typical value seen in DLAs, albeit lower than purely C I-selected systems, that have Solar metallicity (Zou et al. 2018, Ledoux et al., in prep.). Nonetheless, it is important to keep in mind that the metal equivalent widths in our proximate molecular systems spread over a wide range, so that the metallicities are also likely to differ significantly from one system to another. Still, we attempt to identify some global trends in the following.

We then compare the composite spectrum with that obtained for a subset with significant leaking Ly-α emission. Overall, there is no striking difference between the strength of the main metal lines. However, it appears that the equivalent width of the weak Si II λ1808 line remains almost unchanged while other Si II lines (λ1260, 1304, 1526) are weaker for systems with leaking Ly-α. This suggests that the column densities (and the metallicities, since the median $\log N({\rm H\,I})$ is unchanged) in the Ly-α-leaking sub-sample are similar to the overall average, but that systems with leaking emission may have smaller velocity spreads than the average. This could also explain the narrower C IV seen in the "leaking" sub-sample.

A more significant difference is seen for the C II line. While the overall median composite spectrum already shows clear evidence of C II* absorption in the wing of the C II λ1334 line, the composite spectrum corresponding to the Ly-α-leaking sub-sample apparently has a much higher

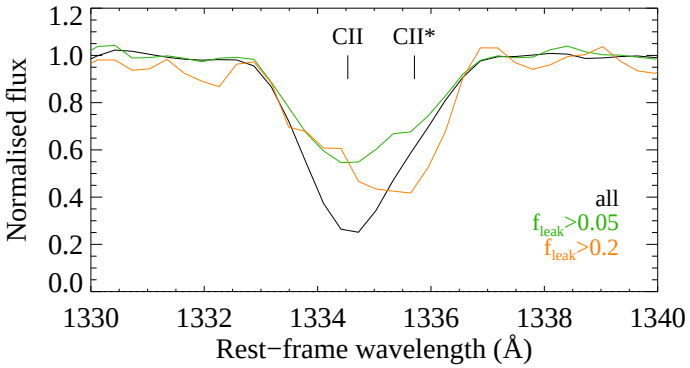

Fig. 8. Median spectra around the C II λ 1334 line for all grade-A systems (black) compared with sub-samples with $f_{\rm leak} > 0.05$ (green) and $f_{\rm leak} > 0.2$ (orange). The spectra are boxcar smoothed by 3 pixels for presentation purposes.

C II*/C II ratio ($\langle Wr(\mathrm{C\,II^*})/Wr(\mathrm{C\,II})\rangle \sim 0.4$ overall versus $\langle Wr(\mathrm{C\,II^*})/Wr(\mathrm{C\,II})\rangle \sim 0.8$ for Ly-α-leaking systems). A zoomed version of Fig. 7 is shown in Fig. 8, along with the composite spectrum built for systems with even stronger Ly-α leaking fraction ($f_{\rm leak} > 0.2$). In the last composite spectrum, albeit noisier given only four grade A systems contributing to the stack, the C II* line appears even stronger than C II. All this indicates an increasing excitation of C II with increasing leakage of Ly-α consistent with the findings of Fathivavsari et al. (2018a). Since the excited level of ionised carbon is mostly excited by collisions (Silva & Viegas 2002; Goldsmith et al. 2012), this would favour a dependence of Ly-α leaking fraction on the compactness of the cloud. However, detailed investigation through follow-up observations and numerical modelling is needed to confirm the higher C II* excitation and to understand its origin.

4.4. Dust reddening

In order to obtain a measure of the reddening induced by dust, we fitted the individual spectra using the quasar template by Selsing et al. (2016) assuming either the extinction law of the Small Magellanic Cloud (SMC) or that of the giant shell in the Large Magellanic Cloud (LMC2) as parameterised by Gordon et al. (2003). However, due to the limited wavelength coverage of the spectra, we were not able to significantly distinguish the two extinction laws. In what follows, all measurements of dust reddening are therefore reported assuming the SMC extinction curve. Since the broad emission lines may vary significantly from one quasar to another, we masked out the corresponding parts of the spectra. This was done by defining "bona fide" continuum regions in the quasar rest-frame which were used to constrain the fit. These regions were defined as: 1314−1351, 1430−1490, 1585−1600, 1700−1830, and 2000−2225 Å.

The best-fit values of A_V are given in Table 1. Due to the intrinsic variations of the spectral power-law index of quasars, we report negative reddening for some targets. This does not necessarily mean that there is no dust reddening, but it is not possible to break the degeneracy without spectroscopic data covering the full rest-frame optical range of the quasar spectral energy distribution.

We can quantify the significance of the A_V measurements by calculating the expected dispersion in A_V introduced by variations in the power-law index. Based on the measured intrinsic dispersion of the quasar power-law index of $\sigma_\beta = 0.186$ (Krawczyk et al. 2015), we calculate an expected 1-σ dispersion in A_V of $\sigma_{A_V} = 0.12$ mag. We can therefore state that any target with $A_V > 2\,\sigma_{A_V}$ is significant at 95% confidence level, and any value below this threshold should be considered an upper limit, i.e., $A_V < 0.24$ mag. In spite of a few exceptions, most of the quasars present no significant reddening (see Fig. 9), with a median A_V of only 0.04 mag, which is consistent with the value measured for the sample of intervening H_2-bearing DLAs selected in SDSS (Balashev et al. 2014). The typical dust-to-gas ratio in our sample is then roughly $A_V/N(\mathrm{H}) \sim (1-2)\times10^{-23}$ mag cm^2, which is similar or less than the typical value for intervening DLAs ($\sim(2-4)\times10^{-23}$ mag cm^2, Vladilo et al. 2008) and much lower than values measured in the local ISM (where the dust-to-gas ratio is about 30 times higher, e.g. Watson 2011) and in C I-selected molecular-rich intervening systems (Ledoux et al. 2015; Zou et al. 2018) that also typically have Solar metallicities and low N(H I). Our current sample may be biased against systems with high reddening, not only because the colour selection may preclude their presence in the SDSS-III spectroscopic database, but also because of the decreased S/N in the blue, impeding the detection (and visual confirmation) of the H_2 lines. Indeed, including the additional (non-statistical) systems, the median A_V/N(H) increases by a factor of two, owing to the inclusion of several significantly reddened systems with lower N(H I) values. Given the low dust-to-gas ratios, the presence of H_2 might then rather be due to higher densities than those typically derived in intervening H_2-bearing DLAs (50−100 cm^{-3}, see e.g. Srianand et al. 2005; Noterdaeme et al. 2017), with the notable exception of the extremely strong H_2 system towards SDSS J0843+0221 (Balashev et al. 2017), which has a low metallicity ([Zn/H] ∼ −1.5) and high density, $n_{\rm H} \sim 300$ cm^{-3}.

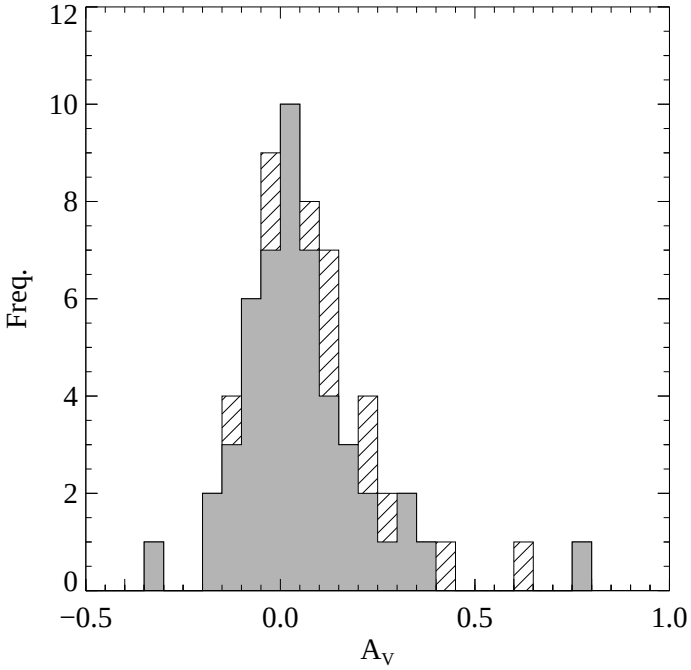

Fig. 9. Distribution of A_V measurements in all grade-A systems (hashed histogram) and in our statistical sub-sample (filled).

5. Discussion

By construction, we select only saturated H_2 systems (with $\log N(H_2) \sim 20$). At such large H_2 column densities, we expect

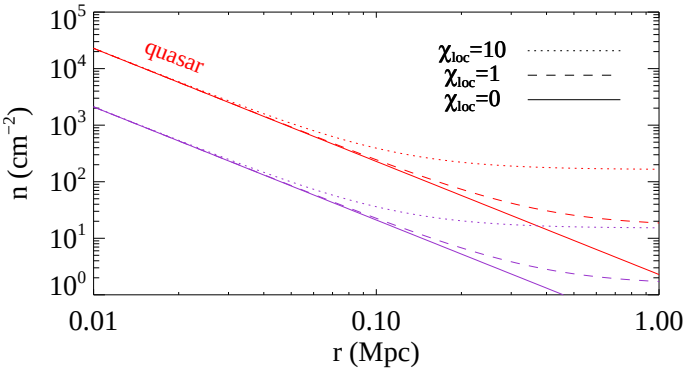

Fig. 10. Density required to produce H_2 as a function of the distance to the quasar. We assumed here a typical situation, with column density equating the median observed value ($\log N(\mathrm{H\,I}) = 21.3$), assuming $\tilde{\sigma_g}$ = 0.1 (red) and $\tilde{\sigma_g}$ = 0.5 (purple) and a quasar with the median luminosity observed at the Lyman-Werner wavelength range. The different curves are when including a local UV field, in units of Draine field ($\chi_{\rm loc}$ = 0, 1, 10).

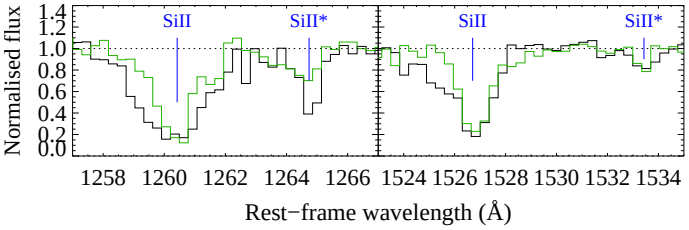

Fig. 11. Composite spectra obtained by median-averaging the spectra of five systems with clear Si II* detection (black) and another five where this detection is tentative (green).

that the H I-H_2 transition has already occurred. We can then use the theoretical description of the H I-H_2 transition by Sternberg et al. (2014, see also Bialy & Sternberg 2016) to constrain the physical properties of the cloud. Following their formalism, the surface density of H I at which the transition occurs is given by

$$\Sigma_{\mathrm{H\,I}} = \frac{3.35}{\tilde{\sigma_g}} \ln\left(\frac{\alpha G}{3.2} + 1\right) M_\odot\,\mathrm{pc}^{-2}, \quad (2)$$

where

$$\alpha G = 2.85 \times 10^{-8} F_0 \left(\frac{100\,\mathrm{cm}^{-3}}{n_{\rm H}}\right)\left(\frac{9.9}{1 + 8.9\tilde{\sigma_g}}\right)^{0.37}. \quad (3)$$

In these equations, $\tilde{\sigma_g} \equiv \sigma_g/(1.9 \times 10^{-21}\,\mathrm{cm}^2)$ is the dust grain Lyman-Werner (LW = 11.2–13.6 eV, 911.6 Å–1107 Å) photon absorption cross section per hydrogen nucleon normalised to the fiducial Galactic value. $n_{\rm H}$ is the hydrogen number density of the cloud and F_0 is the free-space LW photon flux ($\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$) irradiating the cloud (see Bialy et al. 2015, 2017). Note that the constant factor in Eq. (2) is a factor of two lower than that used in previous works (e.g., Ranjan et al. 2018) considering a slab of gas illuminated on both sides while we here consider one-sided illumination dominated by the quasar. Knowing the quasar luminosity at the LW band and H I column density in the cloud, we can then derive the number density of the H_2 cloud as a function of its distance to the quasar, for a given dust enrichment. In Fig. 10, we illustrate the relation between the cloud density and its distance to the quasar UV source for the typically observed quasar and cloud properties. More specifically, the relation is calculated for the median quasar luminosity at the LW band assuming a median H I column density of $\langle \log N(\mathrm{H\,I}) \rangle = 21.3$. We considered a typical value of $\tilde{\sigma_g} = 0.1$, corresponding to the median A_V and $N(\mathrm{H\,I})$ values of our sample ($\tilde{\sigma_g} = 4.8 \times 10^{20} A_{\rm LW}/N(\mathrm{H})$, but we also included a calculation for $\tilde{\sigma_g} = 0.5$. Finally, we considered two calculations: one with and one without a local source of UV photons, $\chi_{\rm loc}$, expressed in units of the interstellar radiation field as measured by Draine (1978).

We find that, farther than about 0.3 Mpc, atomic hydrogen can transition to H_2 in relatively low-metallicity clouds with density $n_{\rm H} \sim 100\,\mathrm{cm}^{-3}$, similar to what has been derived in intervening H_2-bearing DLAs observed so far (e.g. Srianand et al. 2005; Noterdaeme et al. 2017; Ranjan et al. 2018). In other words, at such distances, the conditions for the formation of H_2 become similar to those of intervening clouds, as seen from the inflexion point where the influence of a realistic local UV field becomes comparable to that of the quasar. Since such clouds would typically be of parsec scales, it is not surprising that Ly-α photons from the narrow line region of the quasar (and at fortiori from extended emission regions) can leak around the absorbing cloud.

Closer than 0.1 Mpc, the quasar UV flux likely dominates and the density must be higher ($n_{\rm H} \propto r^{-2}$) for H_2 to form efficiently. It is important to note, however, that this depends strongly on the $\tilde{\sigma_g} \times N(\mathrm{H\,I})$ product and hence on the total dust extinction, with $n_{\rm H} \propto \exp(A_V) - 1$ (when ignoring the slow dependence on $\tilde{\sigma_g}$ of the second factor in Eq. (3)). For example, while keeping the same $N(\mathrm{H\,I})$, a value of $\tilde{\sigma_g} \sim 0.5$ results in a decrease of the required density for H_2 formation by about an order of magnitude. This may be the case for the most reddened systems in our sample. As we get closer to the quasar, we expect that higher densities, together with a stronger UV field, will result in the excitation of fine-structure levels of species like Si II and O I. While we do not see any evidence for excited fine-structure levels in most of the systems, nor in the median stack, we do find clear evidence of Si II* in five systems (J0015+1842, J0125−0129, J1131+0812, J1242+4448 and J1421+5245) as well as tentative evidence in another five systems (J0756+1123, J0911+4110, J1135+2957, J1358+1410 and J1512+3821). Composite spectra of these systems around the main Si II* lines are shown in Fig. 11. Interestingly, these systems with Si II* tend to have stronger and wider leaking Ly-α emission than seen on average, while not necessarily being located at the exact quasar redshift[4].

Similarly, Fathivavsari et al. (2018a) show that excited levels of silicon and oxygen are systematically seen in proximate (metal-selected) DLAs with Ly-α emission in their trough. The authors find a sequence in which the equivalent width of the fine-structure lines increases with increasing leaking Ly-α emission. In the case of eclipsing DLAs, the fine-structure lines are weak whereas the lines are much stronger in the case of ghostly DLAs, which the authors interpret as an effect arising from clouds so compact that the BLR is not fully covered. However, in the absence of detailed investigation through follow-up studies, the number density remains degenerate with the strength of the UV flux since an increase of both these quantities increases the excitation of the fine-structure lines. The presence of H_2 should help break this degeneracy since an atomic-to-molecular transition

[4] One of them (J0125−0129) is particularly intriguing since the equivalent width of the Si II* λ1264 line is larger than that of the nearby Si II λ1260. This is confirmed by a significant Si II* λ1533 line, despite a 10 times lower oscillator strength than Si II* λ1264. This system also has very significant Ly-α emission in the DLA trough, with about 30% photon leakage at the corresponding wavelength.

requires the cloud to be denser when the UV field is stronger (or equivalently when the cloud is located closer to the quasar). Additionally, the excitation of high rotational levels of H_2 could also be efficiently used to discriminate between enhanced UV flux and increased number density, since these are predominantly populated via UV pumping.

The distance-density constraint can be converted into a constraint between cloud-size and distance, using $l = N(\mathrm{H})/n_\mathrm{H}$, where $N(\mathrm{H}) \sim N(\mathrm{H\,I})$. For example, at 10 kpc, the required density for a H I-H_2 transition ($n_\mathrm{H} \sim 2 \times 10^4\,\mathrm{cm^{-3}}$ for $\tilde{\sigma_\mathrm{g}} = 0.1$, $\log N(\mathrm{H\,I}) = 21.3$) would imply a cloud-size less than 0.1 pc. This is a strict upper limit since part of the observed column density may be unrelated to the H_2 cloud. Indeed, not only the numerator in the expression of l is decreased, but the denominator is also increased through Eqs. (2) and (3). On the other hand, we can estimate the size of the BLR using the relation between quasar luminosity and BLR size obtained from reverberation mapping. For the typical quasar luminosity $\lambda L_\lambda(1350\,\text{Å}) \approx 10^{46}\,\mathrm{erg\,s^{-1}}$ in our sample and using the relation from Kaspi et al. (2007), we obtain a C IV BLR size of about 0.1 pc. This is already comparable to the expected cloud size at 10 kpc derived above. Furthermore, the Ly-α BLR is likely to be more extended than the C IV BLR owing to scattering. In other words, the compression of neutral clouds required for an atomic-to-molecular transition to occur, if located closer than 10 kpc, could be such that the projected size of the cloud becomes comparable to that of the BLR. When the partial covering of the BLR gets significant, the system may be seen as a ghostly DLA. Since this is not the case for our systems, these are most likely located farther away, i.e. in other galaxies from the same group or in large-scale gas flows. Notwithstanding, H_2 may still form at distances of ~10 kpc from the quasar in more diffuse clouds (hence possibly covering fully the BLR, i.e. non-ghostly DLAs) provided their metallicity is high enough (e.g. purple line on Fig. 10).

Because Ly-α transfer complicates the apparent velocity and spatial extent of the emission compared to that of the gas producing it[5], it will be interesting to look for signatures of partial coverage of other emission lines by different species as done for intervening systems by e.g. Balashev et al. (2011) and Bergeron & Boissé (2017). C I is an interesting species since not only does it trace the same gas as that seen in H_2, but it has several transitions, one of which (at 1560 Å) falling on the wing of the C IV emission line, when other C I lines are located on the quasar continuum which arises from the extremely small accretion disc. The continuum, by selection, should be fully covered by the absorbing clouds.

Before summarising our results, we remark that the transition theories used in the discussion implicitly assume a steady-state regime. Accurate measurements of the density and dust content in the molecular phase would allow us to investigate whether the molecular formation has reached an equilibrium or not. This would provide additional insights into the understanding of H_2 in quasar environments.

6. Summary

We have developed a novel technique to directly detect strong H_2 absorbers in low-resolution spectra solely from their Lyman-Werner band absorption, without any prior on the associated H I or metal content. Applying our technique to the SDSS-DR14 database, we have assembled a significant sample of strong H_2 systems proximate to the quasar redshift, with $|\Delta v| \lesssim 2000\,\mathrm{km\,s^{-1}}$. We have studied the absorber statistics and investigated the basic characteristics that can be derived from the SDSS data. Our main findings are the following.

(1) We found that the incidence of proximate H_2 systems is about four to five times higher than that expected from the statistics of intervening systems. We further found that the excess of H_2 systems peaks at the quasar redshift, with an excess of more than an order of magnitude compared to intervening statistics. This shows that most of the proximate systems are actually associated to the quasar environment, arising either from galaxies in the same group, or to the quasar host itself. The observed velocities are hence not corresponding to the Hubble flow, but to the individual cloud velocities.

(2) Unsurprisingly, the proximate H_2 systems are also damped Ly-α systems. The column density distribution is however skewed to much higher values than the overall population of intervening DLAs, but only about a factor of two higher than our strong intervening H_2 systems selected the same way. The higher $N(\mathrm{H\,I})$ values could be expected in order to shield H_2 clouds closer to a strong UV source.

(3) We detected significant Ly-α emission in the core of the DLA profile for about half of our sample. We showed that the fraction of leaking Ly-α photons is higher when the DLA is located at small velocity separation from the quasar's systemic redshift. This indicates that the relative projected sizes of the absorbing cloud and the Ly-α emission region decreases with decreasing velocity separation. This effect can then be explained by Ly-α emission at the emission peak arising from both the broad line region and gas located farther out (narrow line region, or even kpc-scales), while photons in the wings of the Ly-α emission arise only from the compact broad line region, and hence are easily covered by the cloud. It is also possible that clouds with smaller velocity separation belong to the quasar host compared to those at high velocities which could be due to other galaxies in the group. In this case, clouds located closer to the UV source could be more compact, as suggested by Fathivavsari et al. (2018a), hence covering less the quasar emission.

(4) The equivalent width distribution as well as the average metal strength seen in a composite spectrum indicates that the proximate H_2 systems have metallicities around one tenth Solar, albeit with a wide dispersion between individual systems. We also identify several cases with signatures of high excitation, namely the presence of fine-structure lines of Si II and C II. These tend to be related to the fraction of leaking Ly-α photons, suggesting that the corresponding clouds are indeed more compact than typical DLA clouds.

(5) The measured high H_2 abundance allows us to bring further clues to the understanding of the clouds' origin. Following the H I-H_2 transition theory developed by Sternberg et al. (2014), we show that the number density required for a transition to occur depends strongly on the distance to the quasar, for a given metallicity and column density. Clouds located in galaxies from the group further than about 100 kpc from the quasar may have characteristics very similar to intervening clouds. In turn, clouds located within the quasar host or belonging to flows to or from the quasar would need $n_\mathrm{H} \sim 10^4$–$10^5\,\mathrm{cm^{-3}}$ to form H_2 and hence have very small dimensions. This could be the case for the systems with the highest excitation (dense gas, close to UV source) and large Ly-α leaking fraction (due to less coverage of the quasar emission line regions). On the other hand, it will be interesting to study the presence and excitation of H_2 in the

[5] For example, the velocity width of the Ly-α emission does not represent the bulk gas velocity since Ly-α photons escape more easily when scattering with atoms at the end of the velocity distribution.

overall population of proximate DLAs, in particular the ghostly DLAs, which are expected to be the sub-population located closest to the central engine (Fathivavsari et al. 2018b).

In conclusion, given the spread in absorber characteristics (metallicities, dust extinction, excitation of fine-structure lines, and the presence, strength and width of leaking Ly-α emission), it is likely that there is no single origin for such clouds. While a large fraction, even with leaking Ly-α emission, is likely to belong to other galaxies in the group, several systems in our sample may well be directly associated to the quasar host or flows to or from the quasar. Follow-up at higher spectral resolution is required to investigate the partial coverage of the emission line regions by the absorbing clouds, to measure the exact relative velocity between the quasar and the cloud, to estimate the chemical enrichment in individual systems, and finally to investigate the physical conditions in order to estimate the cloud's density and distance to the UV source. The excitation of fine structure levels of ionised silicon and carbon as well as neutral oxygen and carbon will bring important constraints, together with the presence and excitation of molecules.

Acknowledgements. We thank the referee, Sergei Levshakov, for a thorough reading of the paper and useful comments and suggestions. PN and JKK warmly thank the Ioffe institute in Saint Petersburg for hospitality where this work was initiated and the Russian-French collaborative programme (PRC) for supporting their visit. SB is supported by the Russian Science Foundation grant 18-72-00110. The research leading to these results received support from the French *Agence Nationale de la Recherche*, under grant ANR-17-CE31-0011-01 (Project "HIH2" – PI Noterdaeme). PN, RS and PPJ also acknowledge support from the Indo-French Centre for the Promotion of Advanced Research under contract 5504-B. HF thanks the Institut d'Astrophysique de Paris for hospitality and support from the ANR under grant ANR-16-CE31-0021 (Project "eBOSS" – PI Yèche). PN and JKK are also grateful to the ESO office for science for supporting a visit to the ESO headquarters in Santiago de Chile. We acknowledge the use of SDSS-III data. Funding for SDSS-III has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Science Foundation, and the US Department of Energy Office of Science. The SDSS-III web site is `http://www.sdss3.org/`. SDSS-III is managed by the Astrophysical Research Consortium for the Participating Institutions of the SDSS-III Collaboration including the University of Arizona, the Brazilian Participation Group, Brookhaven National Laboratory, Carnegie Mellon University, University of Florida, the French Participation Group, the German Participation Group, Harvard University, the Instituto de Astrofísica de Canarias, the Michigan State/Notre Dame/JINA Participation Group, Johns Hopkins University, Lawrence Berkeley National Laboratory, Max Planck Institute for Astrophysics, Max Planck Institute for Extraterrestrial Physics, New Mexico State University, New York University, Ohio State University, Pennsylvania State University, University of Portsmouth, Princeton University, the Spanish Participation Group, University of Tokyo, University of Utah, Vanderbilt University, University of Virginia, University of Washington, and Yale University.

References

Balashev, S. A., & Noterdaeme, P. 2018, MNRAS, 478, L7
Balashev, S. A., Petitjean, P., Ivanchik, A. V., et al. 2011, MNRAS, 418, 357
Balashev, S. A., Klimenko, V. V., Ivanchik, A. V., et al. 2014, MNRAS, 440, 225
Balashev, S. A., Noterdaeme, P., Rahmani, H., et al. 2017, MNRAS, 470, 2890
Bergeron, J., & Boissé, P. 2017, A&A, 604, A37
Bialy, S., & Sternberg, A. 2016, ApJ, 822, 83
Bialy, S., Sternberg, A., Lee, M.-Y., Le Petit, F., & Roueff, E. 2015, ApJ, 809, 122
Bialy, S., Bihr, S., Beuther, H., Henning, T., & Sternberg, A. 2017, ApJ, 835, 126
Borisova, E., Cantalupo, S., Lilly, S. J., et al. 2016, ApJ, 831, 39
Cantalupo, S., Arrigoni-Battaia, F., Prochaska, J. X., Hennawi, J. F., & Madau, P. 2014, Nature, 506, 63
Courbin, F., North, P., Eigenbrod, A., & Chelouche, D. 2008, A&A, 488, 91
De Cia, A., Ledoux, C., Petitjean, P., & Savaglio, S. 2018, A&A, 611, A76
Draine, B. T. 1978, ApJS, 36, 595
Ellison, S. L., Yan, L., Hook, I. M., et al. 2002, A&A, 383, 91
Ellison, S. L., Prochaska, J. X., Hennawi, J., et al. 2010, MNRAS, 406, 1435
Ellison, S. L., Prochaska, J. X., & Mendel, J. T. 2011, MNRAS, 412, 448
Fathivavsari, H., Petitjean, P., Noterdaeme, P., et al. 2015, MNRAS, 454, 876
Fathivavsari, H., Petitjean, P., Noterdaeme, P., et al. 2016, MNRAS, 461, 1816
Fathivavsari, H., Petitjean, P., Zou, S., et al. 2017, MNRAS, 466, L58
Fathivavsari, H., Petitjean, P., Jamialahmadi, N., et al. 2018a, MNRAS, 477, 5625
Fathivavsari, H., Petitjean, P., Jamialahmadi, N., et al. 2018b, ApJ, submitted
Finley, H., Petitjean, P., Pâris, I., et al. 2013, A&A, 558, A111
Foltz, C. B., Chaffee, Jr., F. H., & Black, J. H. 1988, ApJ, 324, 267
Fynbo, J. P. U., Jakobsson, P., Prochaska, J. X., et al. 2009, ApJS, 185, 526
Goldsmith, P. F., Langer, W. D., Pineda, J. L., & Velusamy, T. 2012, ApJS, 203, 13
Gordon, K. D., Clayton, G. C., Misselt, K. A., Landolt, A. U., & Wolff, M. J. 2003, ApJ, 594, 279
Jiang, P., Zhou, H., Pan, X., et al. 2016, ApJ, 821, 1
Jorgenson, R. A., Wolfe, A. M., & Prochaska, J. X. 2010, ApJ, 722, 460
Jorgenson, R. A., Murphy, M. T., Thompson, R., & Carswell, R. F. 2014, MNRAS, 443, 2783
Kaspi, S., Brandt, W. N., Maoz, D., et al. 2007, ApJ, 659, 997
Klimenko, V. V., Balashev, S. A., Ivanchik, A. V., et al. 2015, MNRAS, 448, 280
Krawczyk, C. M., Richards, G. T., Gallagher, S. C., et al. 2015, AJ, 149, 203
Krogager, J. K., Møller, P., Fynbo, J. P. U., & Noterdaeme, P. 2017, MNRAS, 469, 2959
Krumholz, M. R., McKee, C. F., & Tumlinson, J. 2008, ApJ, 689, 865
Ledoux, C., Petitjean, P., & Srianand, R. 2003, MNRAS, 346, 209
Ledoux, C., Petitjean, P., Fynbo, J. P. U., Møller, P., & Srianand, R. 2006, A&A, 457, 71
Ledoux, C., Noterdaeme, P., Petitjean, P., & Srianand, R. 2015, A&A, 580, A8
Levshakov, S. A., & Foltz, C. B. 1988, Sov. Astron. Lett., 14, 464
Levshakov, S. A., & Varshalovich, D. A. 1985, MNRAS, 212, 517
Neeleman, M., Kanekar, N., Prochaska, J. X., Rafelski, M. A., & Carilli, C. L. 2019, ApJ, 870, L19
North, P. L., Marino, R. A., Gorgoni, C., et al. 2017, A&A, 604, A23
Noterdaeme, P., Petitjean, P., Srianand, R., Ledoux, C., & Le Petit, F. 2007, A&A, 469, 425
Noterdaeme, P., Ledoux, C., Petitjean, P., & Srianand, R. 2008, A&A, 481, 327
Noterdaeme, P., Petitjean, P., Ledoux, C., & Srianand, R. 2009, A&A, 505, 1087
Noterdaeme, P., Petitjean, P., Carithers, W. C., et al. 2012, A&A, 547, L1
Noterdaeme, P., Krogager, J. K., Balashev, S., et al. 2017, A&A, 597, A82
Noterdaeme, P., Ledoux, C., Zou, S., et al. 2018, A&A, 612, A58
Pâris, I., Petitjean, P., Aubourg, É., et al. 2018, A&A, 613, A51
Péroux, C., McMahon, R. G., Storrie-Lombardi, L. J., & Irwin, M. J. 2003, MNRAS, 346, 1103
Prochaska, J. X., Herbert-Fort, S., & Wolfe, A. M. 2005, ApJ, 635, 123
Prochaska, J. X., Hennawi, J. F., & Herbert-Fort, S. 2008a, ApJ, 675, 1002
Prochaska, J. X., Chen, H.-W., Wolfe, A. M., Dessauges-Zavadsky, M., & Bloom, J. S. 2008b, ApJ, 672, 59
Rafelski, M., Wolfe, A. M., Prochaska, J. X., Neeleman, M., & Mendez, A. J. 2012, ApJ, 755, 89
Ranjan, A., Noterdaeme, P., Krogager, J. K., et al. 2018, A&A, 618, A184
Selsing, J., Fynbo, J. P. U., Christensen, L., & Krogager, J.-K. 2016, A&A, 585, A87
Sheffer, Y., Prochaska, J. X., Draine, B. T., Perley, D. A., & Bloom, J. S. 2009, ApJ, 701, L63
Silva, A. I., & Viegas, S. M. 2002, MNRAS, 329, 135
Srianand, R., Petitjean, P., Ledoux, C., Ferland, G., & Shaw, G. 2005, MNRAS, 362, 549
Sternberg, A., Le Petit, F., Roueff, E., & Le Bourlot, J. 2014, ApJ, 790, 10
Vanden Berk, D. E., Richards, G. T., Bauer, A., et al. 2001, AJ, 122, 549
Vladilo, G., Prochaska, J. X., & Wolfe, A. M. 2008, A&A, 478, 701
Vreeswijk, P. M., Ledoux, C., Smette, A., et al. 2007, A&A, 468, 83
Wakelam, V., Bron, E., Cazaux, S., et al. 2017, Mol. Astrophys., 9, 1
Watson, D. 2011, A&A, 533, A16
Wolfe, A. M., Gawiser, E., & Prochaska, J. X. 2005, Annu. Rev. Astron. Astrophys., 43, 861
Zou, S., Petitjean, P., Noterdaeme, P., et al. 2018, A&A, 616, A158
Zubovas, K., Nayakshin, S., King, A., & Wilkinson, M. 2013, MNRAS, 433, 3079

Appendix A: SDSS spectra of proximate H_2 systems

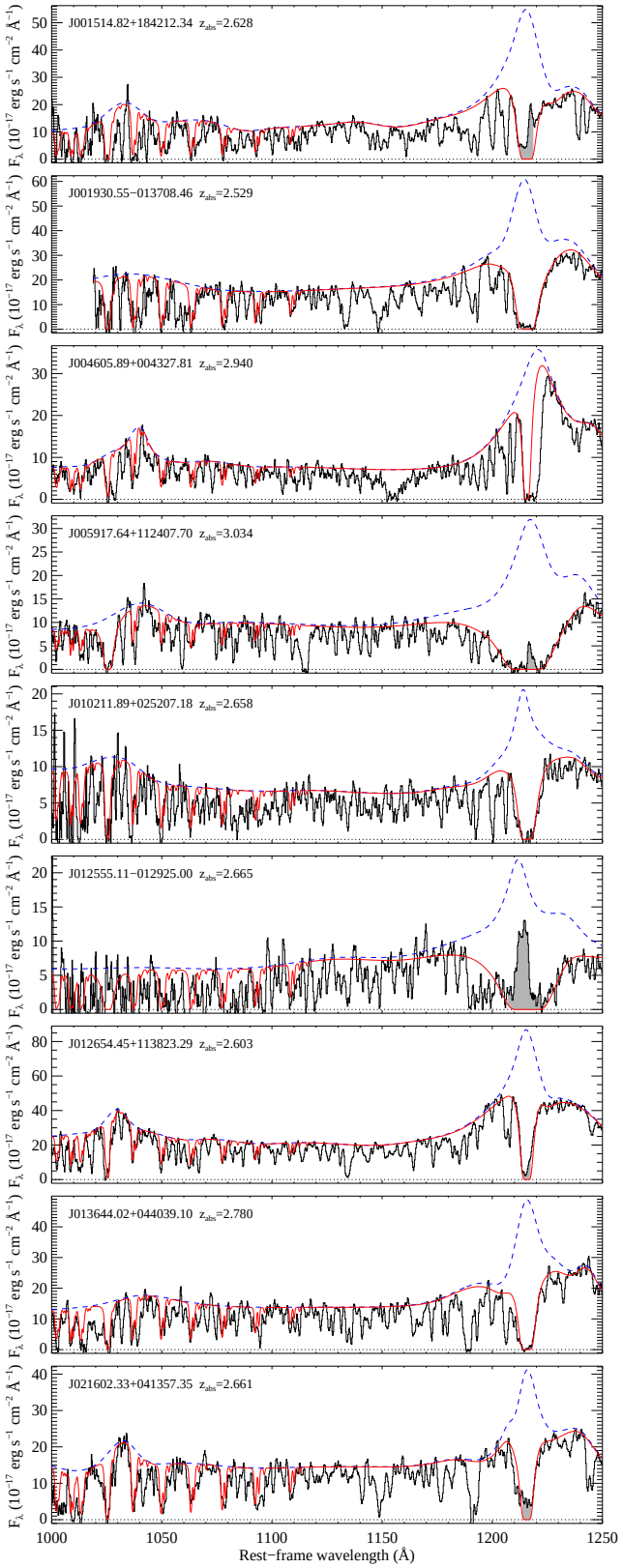

Fig. A.1. Proximate H_2 systems. The panels show a portion of the SDSS spectra (black), shifted at the quasar rest-frame. The estimated unabsorbed quasar spectrum is shown as dashed blue curve. The synthetic H I + H_2 profile is overplotted in red. The shaded area in the core of the PDLA highlights the leaking flux.

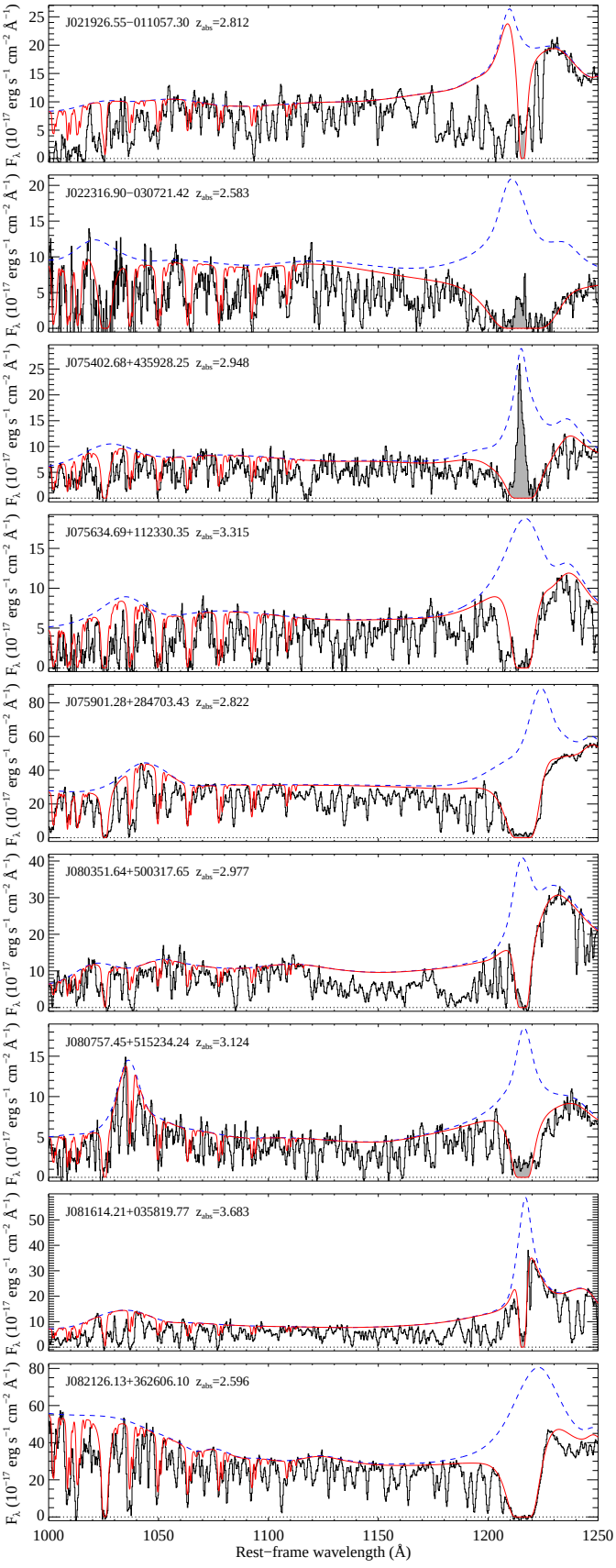

Fig. A.1. continued.

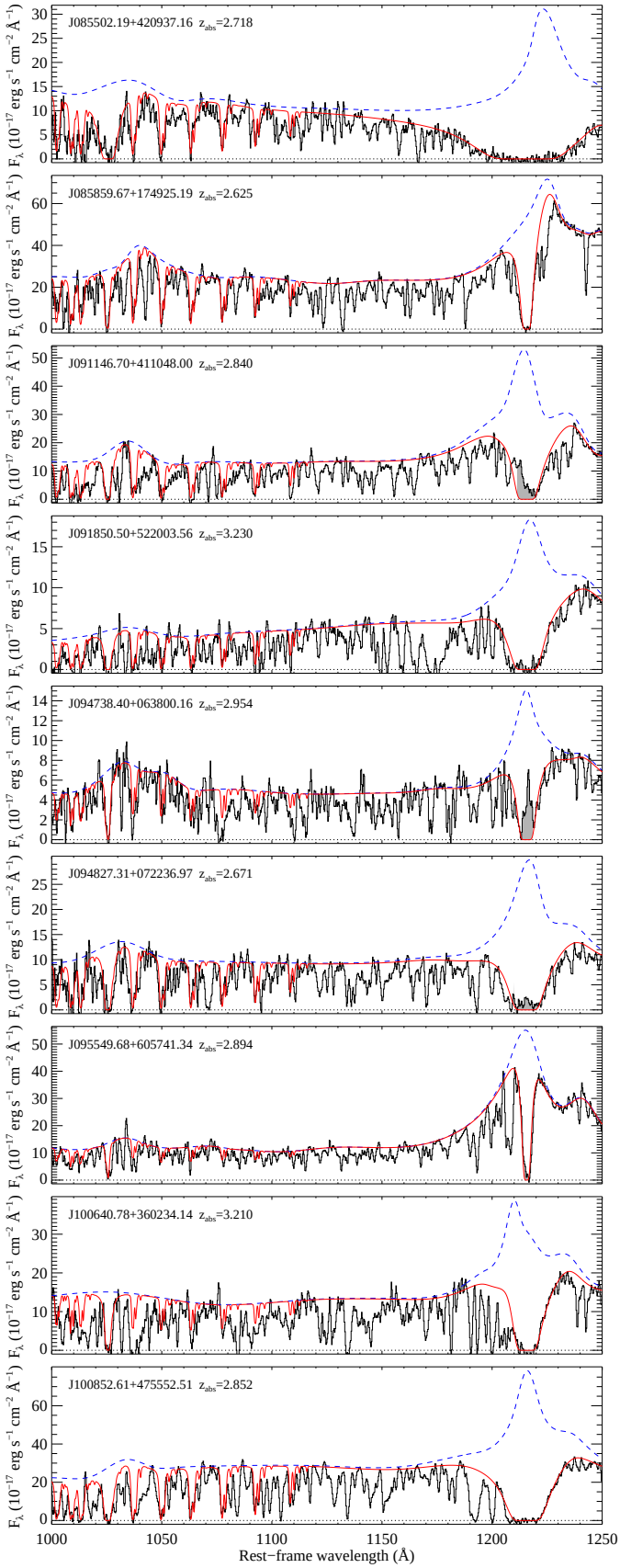

Fig. A.1. continued.

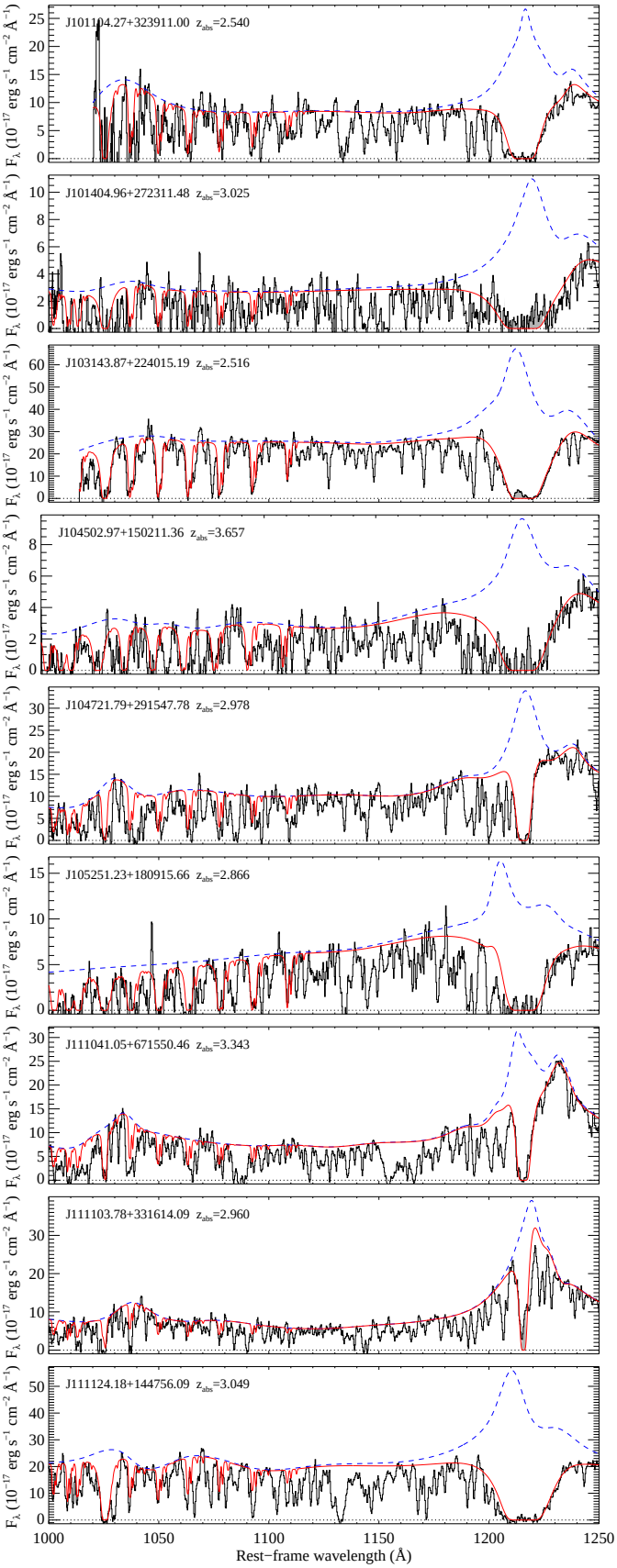

Fig. A.1. continued.

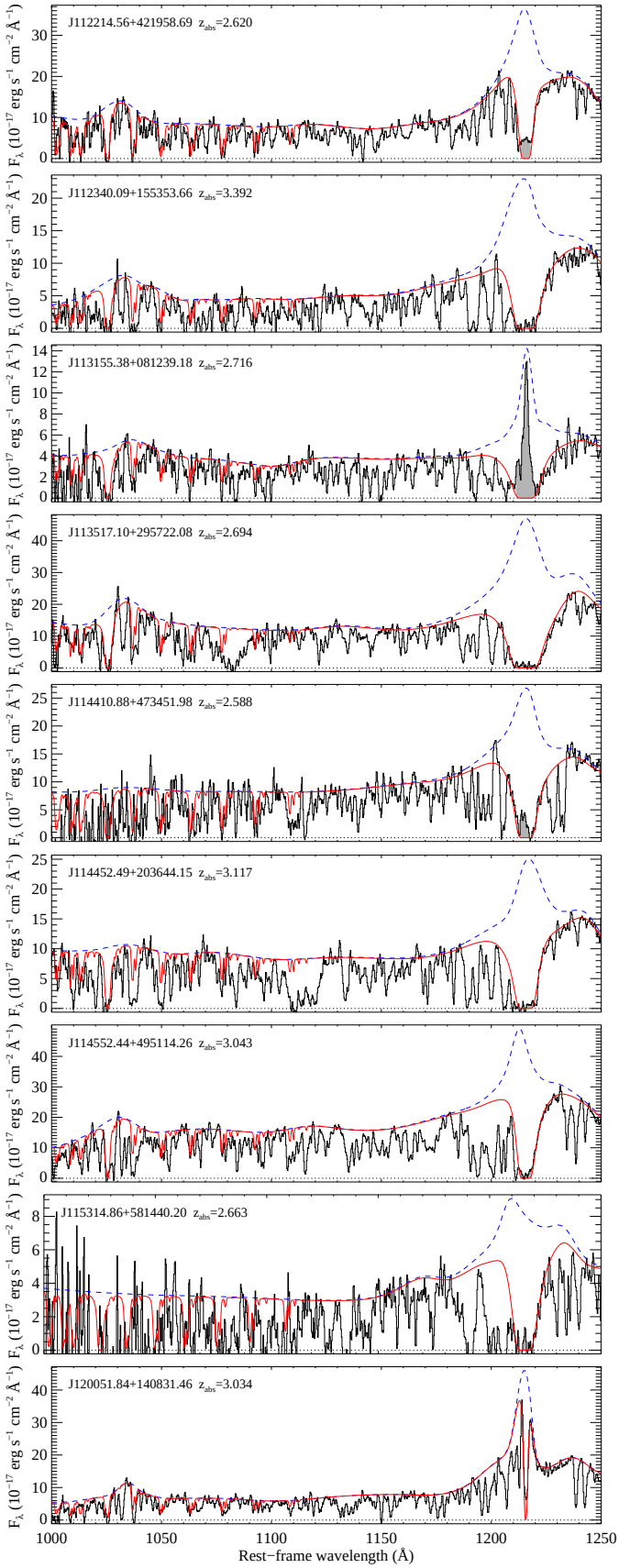

Fig. A.1. continued.

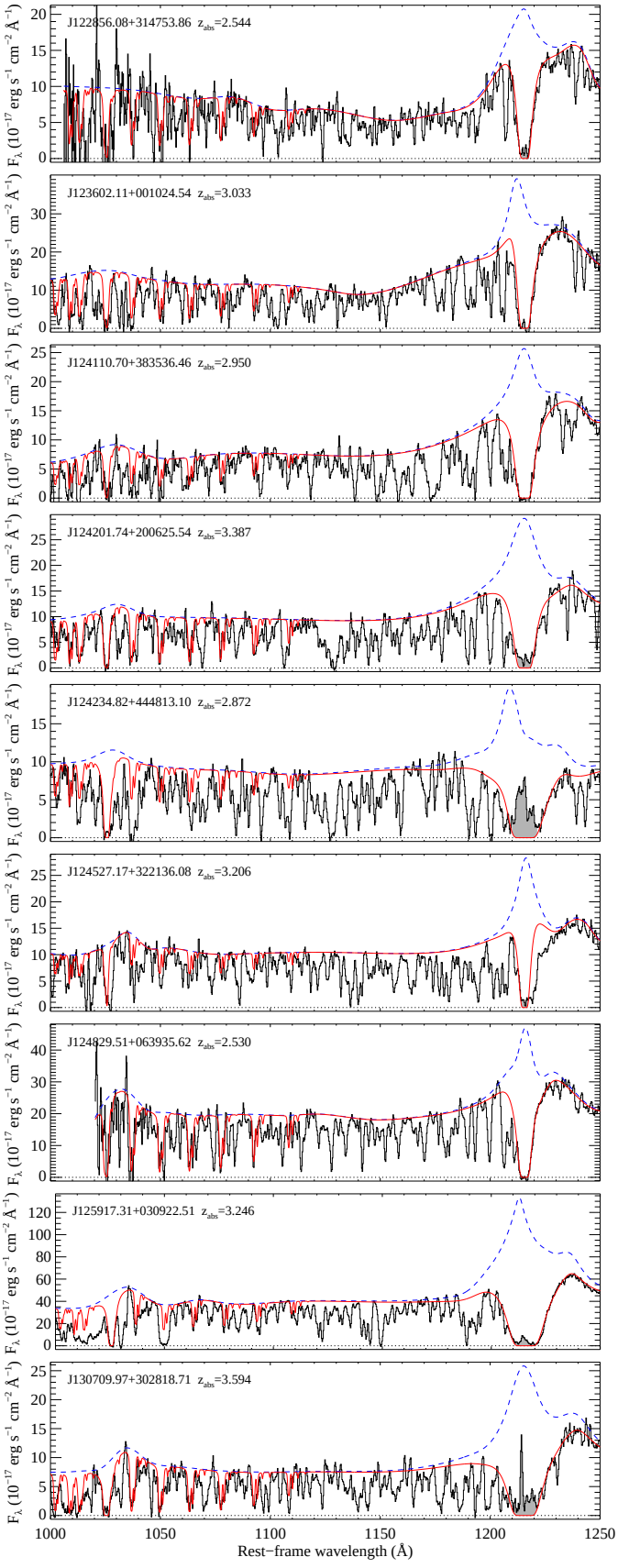

Fig. A.1. continued.

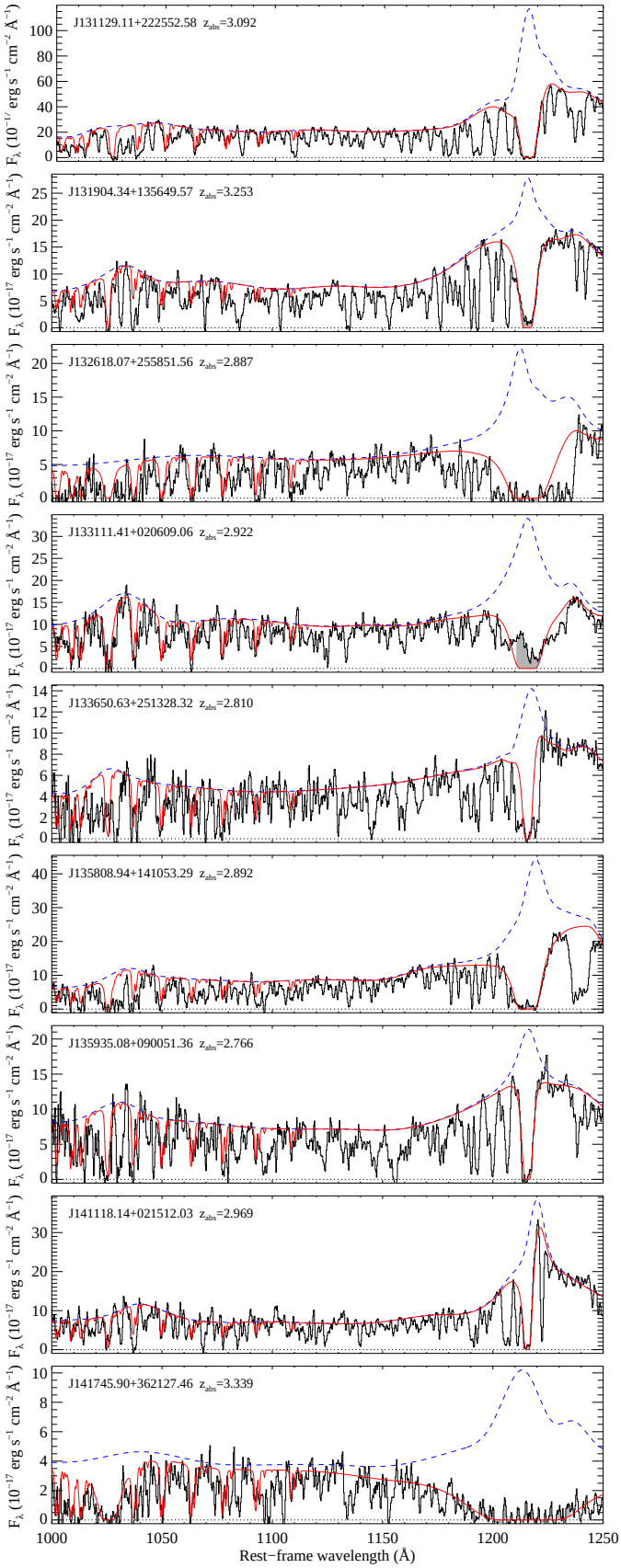

Fig. A.1. continued.

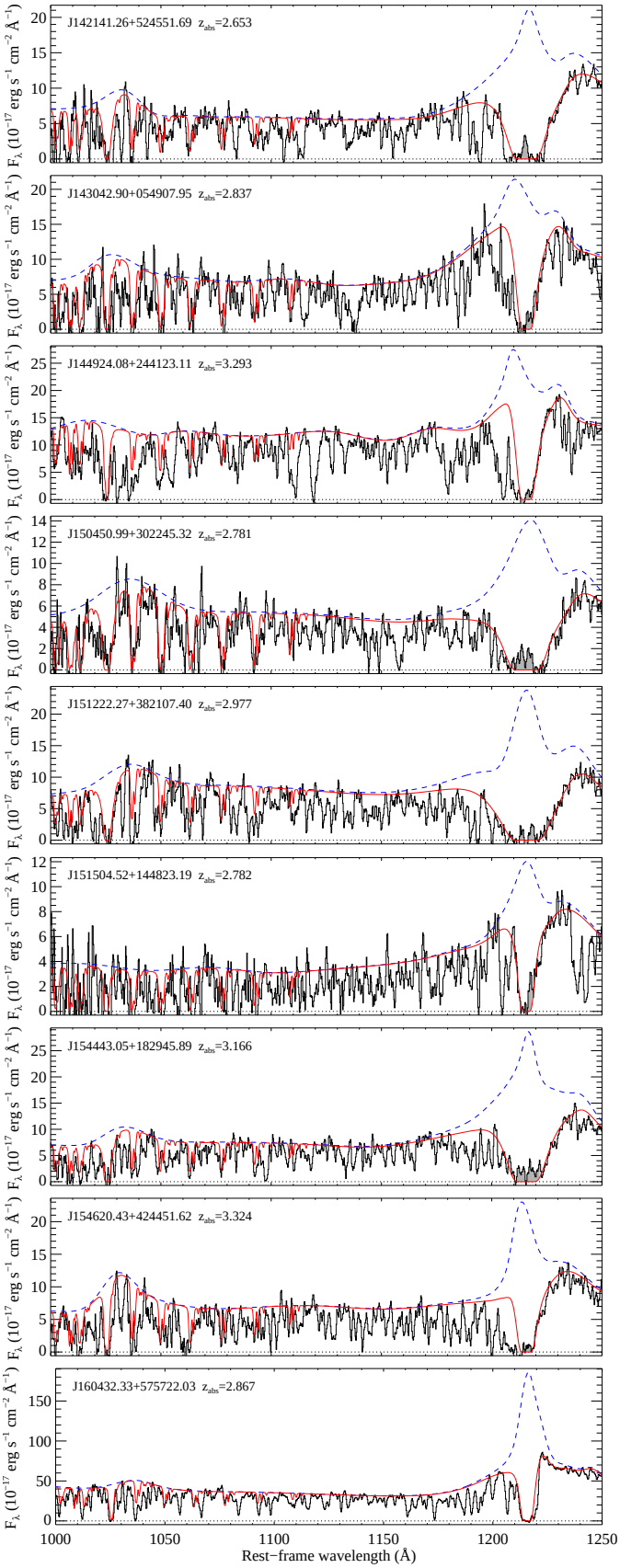

Fig. A.1. continued.

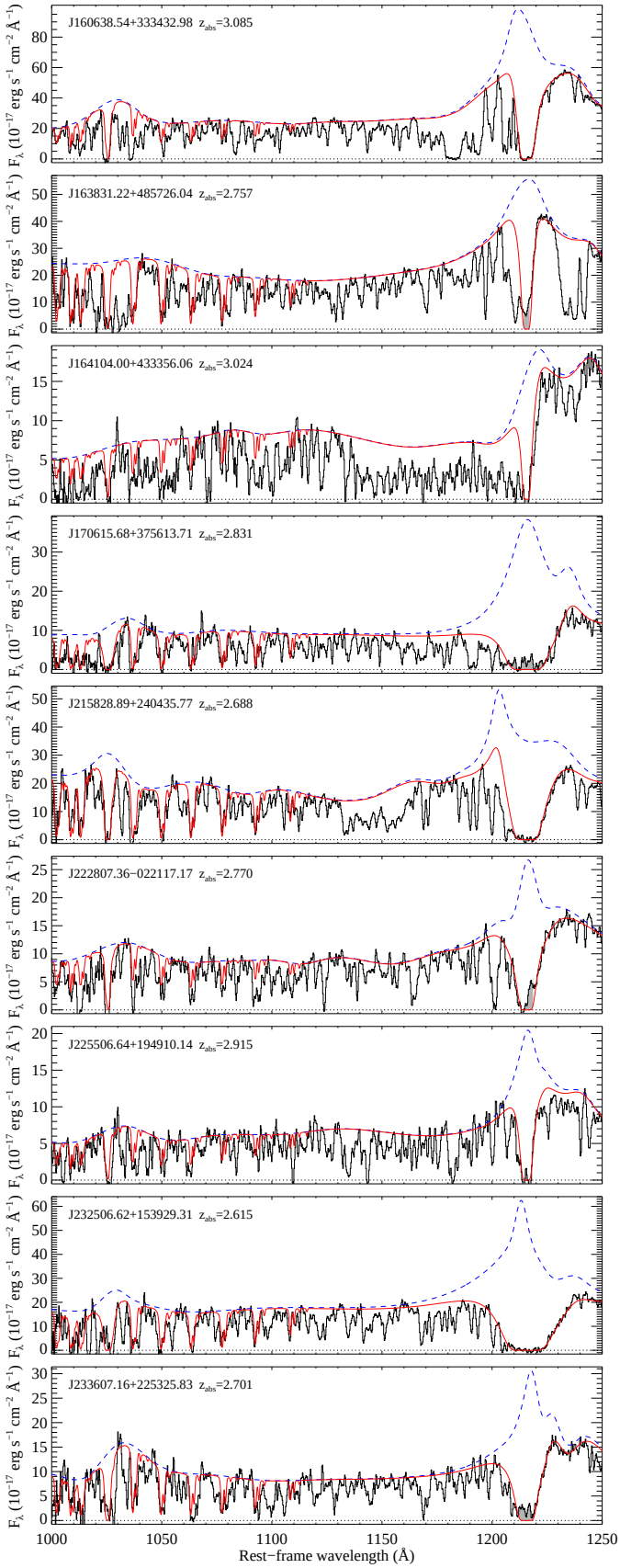

Fig. A.1. continued.