A high baryon fraction in massive haloes at \( z \sim 3 \)

Gabriele Pezzulli\(^1\) and Sebastiano Cantalupo\(^1\)

\(^1\) Department of Physics, ETH Zurich, Wolfgang-Pauli-Strasse 27, 8093 Zurich, Switzerland

ABSTRACT

We investigate the baryon content of the circumgalactic medium (CGM) within the virial radius of \( M_h \sim 10^{12} \, M_\odot \) haloes at \( z \sim 3 \), by modelling the surface brightness profile of the giant Ly\( \alpha \) nebulae recently discovered by MUSE around bright QSOs at this redshift. We initially assume fluorescent emission from cold photo-ionized gas confined by the pressure of a hot halo. Acceptable CGM baryon fractions (equal or smaller than the cosmological value) require that the cold gas occupies \( \sim 1\% \) of the volume, but is about as massive as the hot gas. CGM baryon fractions as low as 30\% of the cosmic value, as predicted by some strongly ejective feedback models at this redshift, are not easy to reconcile with observations, under our assumptions, unless both the QSO-hosting haloes at \( z \sim 3 \) are more massive than recent BOSS estimates based on clustering and the photo-ionized gas is colder than expected in a standard QSO ionizing radiation field. We also consider the option that the emission is dominated by photons scattered from the QSO broad line region. In this scenario, a very stringent lower limit to the baryon fraction can be obtained under the extreme assumption of optically thin scattering. We infer in this case a baryon fraction of at least 70\% of the cosmic value, for fiducial parameters. Lower values require halo masses or gas temperatures different than expected, or that some mechanism keeps the cold gas systematically over-pressured with respect to the ambient medium.

Key words: galaxies: haloes – intergalactic medium – galaxies: formation – galaxies: evolution

1 INTRODUCTION

Modern galaxy formation models rely on means to keep outside of the star-forming body of galaxies most of the baryons theoretically associated to their haloes. Different models achieve this result in different ways. In particular, the so-called missing baryons can be stored either relatively close to galaxies (in the circumgalactic medium, or CGM) or very far away from them (in the intergalactic medium, or IGM).

As a consequence, observational estimates of the (baryonic) mass of the CGM, as a function of halo mass and redshift, are of great importance to further constrain existing models and guide advancement of galaxy formation theory.

Shock heating as a consequence of gravitational collapse was first identified as a possible means to bring most of the baryons falling into the potential wells of forming galaxies to the virial temperature, leading to the formation of a low density, hot and massive CGM (Rees & Ostriker 1977, White & Rees 1978). Whenever the dynamical time is long compared to the cooling time, however, the process is largely compensated by radiative cooling, which would lead to the collapse of most of the baryons towards the centre and to excessive star formation compared to observations (e.g. White & Frenk 1991). In some conditions, a virial shock does not even develop, or it is unstable, and filaments of cold gas can penetrate deep into the halo, possibly fuelling substantial star formation at the centre (e.g. Birnboim & Dekel 2003, Binney 2004, Keres et al. 2005). The conditions for efficient radiative cooling are far from trivial, as they crucially depend on metallicity (e.g. Sutherland & Dopita 1993), on the intensity and shape of a meta-galactic ionizing background (e.g. Efstathiou 1992, Wiersma, Schaye & Smith 2009), on the time-scales needed to ensure collisional or photo-ionization equilibrium (Cnat 2017) and on whether local ionizing sources are consistently taken into account (Cantalupo 2010). Some of these effects have been shown to be important especially at low masses (e.g. Kannan et al. 2016), while, at least at high masses, it is most of the times assumed that catastrophic cooling must be compensated by fresh injection of energy, which either provides additional heating to the CGM, or mechanically removes a substantial
fraction of the CGM from galaxy haloes. The second option can be referred to as **ejective feedback**.

The most commonly advocated energy sources are supernovae (SNe). First proposed by Dekel & Silk (1986), SN-driven *galactic winds* can, if certain conditions are met, eject a large fraction of the baryons out of the potential wells of galaxies and thus solve the overcooling problem. For this reason, they have now become a key ingredient of most semi-analytical galaxy formation models (e.g. Henriques et al. 2013 and references therein), as well as hydrodynamical cosmological simulations (e.g. Springel & Hernquist 2003; Oppenheimer & Davé 2008; Dalla Vecchia & Schaye 2012). Ejection of baryons due to SN feedback is expected to be particularly effective for low mass galaxies, due to their low escape speed. Crucially, however, the process will impact the baryon content of high mass haloes as well. Baryons ejected from the shallow potential wells of dwarf galaxies at very high redshift can reach considerably large distances. As a consequence, more massive haloes, formed later from the merging of the low mass ones, would find themselves relatively deficient in baryons already at virialization. The baryon fraction in these haloes would then slowly increase towards the cosmological value, as the gas ejected at early times is gradually accreted back (‘re-incorporated’). With some dependence on implementation details, this scenario provides predictions for the baryon fraction within the virial radius as a function of halo mass and redshift (e.g. Davé 2009; Muratov et al. 2015; Liang et al. 2016; Mitchell et al. 2018), thus indicating a way to test or refine the paradigm observationally.

There are reasons to expect that reality may be more complicated than the simple scenario above. Models based on the described paradigm usually require that the energy produced by SN explosions (as well as by the stellar winds of young stars) must be transferred to the surrounding medium with an efficiency close to unity. Solutions must then be envisaged to stem the radiative losses that are expected on theoretical grounds (e.g. Weaver et al. 1977; Gatto et al. 2017; Naab & Ostriker 2017 and references therein). Even at maximum efficiency, it is widely believed that SN feedback is insufficient to explain the high-mass end of the galaxy stellar mass function, so that additional mechanisms are often advocated to regulate the evolution of galaxies in that regime, the most common feedback being from active galactic nuclei (AGN, e.g. Ciotti & Ostriker 2001; Brightman & Mathews 2006; Beckmann et al. 2017). AGN feedback has been modelled in a variety of fashions: thermal, kinetic, radiative, continuous or intermittent, isotropic or anisotropic, with a variety of assumed efficiencies and a different impact on the baryon content of the CGM (e.g. Sijacki et al. 2007; Steinborn et al. 2015; Choi et al. 2019; Weinberger et al. 2017). In addition to the above, other mechanisms are being investigated which can influence accretion onto galaxies and star formation, while having a variety of impacts on the baryon content of haloes. These include, among others, various kinds of radiation pressure (e.g. Roskar et al. 2014; Kimm et al. 2018), cosmic rays (e.g. Samui, Subramanian & Srianand 2018; Jacob et al. 2018), thermal conduction (Nipoti & Binney 2007) and photo-ionization by local sources (Cantalupo 2010; Kannan et al. 2016).

Observational determinations of the baryonic mass in the CGM around galaxies, as a function of halo mass and redshift, are of great help to confirm galaxy formation models or direct theoretical investigations further. At low redshift, a wealth of information comes from the X-ray bremsstrahlung of hot (close to virial temperature) gas in groups and clusters, as well as around galaxies (e.g. Lagana et al. 2013; Bogdán et al. 2017; Li et al. 2018). Some of these observations (e.g. Lovisari, Reiprich & Schellenberger 2015; Eckert et al. 2016) suggest that at $z \approx 0$ the fraction of baryons (at least of those in the hot phase) found within haloes increases with halo mass, an input which has already proven useful to calibrate feedback models in some cosmological simulations (e.g. Le Brun et al. 2014; Pillepich et al. 2018). In addition to the hot gas, an increasing amount of data are becoming available about the colder phase ($T \approx 10^4$ K), entrained within hot haloes and mostly observable in absorption along the sightlines of background QSOs. Estimating the mass of this component is more difficult, because of the intrinsically serendipitous nature of absorption studies, as well as the need to resort to photo-ionization modelling to convert the observed column densities of the absorbing species into a total gas mass. It has however been suggested that the mass in the cold phase can in some cases also be significant (e.g. Werk et al. 2014; Stern et al. 2016; Zahedy et al. 2019).

The baryon content of the CGM of high $z$ galaxies is of special value to galaxy formation theory, as its determination would provide the most direct information on the nature of feedback mechanisms – and in general on the processes determining galaxy formation – in the earliest stages of their development. In particular, at early epochs, the amount of retained baryons is expected to change, with halo mass and redshift, in a way that is mostly determined by the efficiency of SN feedback (e.g. Henriques et al. 2018). Therefore, measurements of the CGM baryon fraction at high $z$ can, at least in principle, add useful constraints to the choice of relevant parameters, such as the fraction of SN energy that is converted into gas kinetic energy, eventually modulating the ejection of baryons from galaxy haloes. Unfortunately, secure measurements of bremsstrahlung emission from hot haloes at high $z$ require very long integration times with the current instrumentation. Progress is expected from complementary probes, such as the Sunyaev-Zeldovich effect. This is unaffected by cosmological dimming and has already provided interesting indications at intermediate redshifts at the mass scale of galaxy clusters (e.g. Chiu et al. 2018 and references therein). The mass of the hot gas can also be estimated indirectly from that of the cold gas, when observations of the latter are available, based on the assumption of pressure balance between the two phases. This idea was initially proposed by Spitzer (1956), who successfully applied it to the halo of the Milky Way, and was proven applicable up to the scale of galaxy clusters (e.g. Churazov et al. 2001; Zhuravleva et al. 2016). Absorption data provide some constraints on the cold gas around star forming galaxies (Rudie et al. 2012) and quasars (Prochaska, Hennawi & Simcoe 2013) out to $z \sim 2$ and suggest the presence of significant amounts of cold gas at projected distances comparable to virial radius, though a quantitative determination on the total gas mass is not straightforward for the reasons mentioned above. On the other hand, the cold phase of the CGM at $z \gtrsim 2$ is sufficiently bright in fluorescent Lyα emission to be detected with current facilities, as long as it is illumin-
ated by an intense source of ionizing photons, such as a bright QSO. Indeed, numerous extended Lyα nebulae associated to AGN have been discovered around radio galaxies (e.g. Villar-Martín et al. 2007 and references therein) and more recently around radio-quiet QSOs (e.g. Cantalupo et al. 2014, Hennawi et al. 2015).

Thanks to the Multi Unit Spectroscopic Explorer (MUSE, Bacon et al. 2010), Borisova et al. (2016) found 100% detection rate of giant Lyα nebulae around a sample of 17 radio-quiet and 2 radio-loud very luminous (L \sim 10^{44} \text{ erg s}^{-1}) QSOs at z \geq 3. A similar result was also found by the MUSEUM survey (Arrigoni Battaia et al. 2019). Giant Lyα nebulae around luminous radio-quiet QSOs are the ideal laboratory where to investigate the CGM baryon fraction, for at least two reasons. First, bright QSOs are able to ionize the illuminated part of their own CGM almost completely, so that in this region contributions to Lyα emission from collisional excitation and scattering either are drastically reduced or their treatment is much simpler than in more ordinary environments. Second, both the typical QSO life time and the CGM light crossing time are much shorter than the CGM sound crossing time. This gives us some confidence that fluorescent nebulae around radio-quiet QSOs provide a genuine and unbiased snapshot of the state of CGM, before it might be significantly displaced by possible mechanical feedback from the current QSO activity event.

The aim of the present paper is to build upon the observations of Borisova et al. (2016) and provide an estimate of the CGM baryon fraction in relatively massive (M_h \sim 10^{12} M_\odot) haloes at z \sim 3. We find that a Lyα surface brightness as high as the observed one is difficult to explain with baryon fractions significantly smaller than the cosmological average. Lower values, typical of strong ejective feedback models, can be accounted for under specific conditions, which we discuss in detail and which can be tested by independent observations. Much of our analysis relies on the hypotheses that i) Lyα emission is dominated by recombination radiation and ii) that the cold and the hot phase are in pressure equilibrium with each other. We also discuss, however, the conditions and consequences of relaxing these assumptions.

In Section 2 we use the observations of Borisova et al. (2016) to infer the density profile and the total mass of the cold gas in massive haloes at z \sim 3. In Section 4 we estimate the mass of the hot gas based on the assumption of pressure equilibrium. In Section 5 we discuss the total CGM mass, its dependence on model parameters and the conditions to reconcile it with the strong ejective feedback paradigm. In Section 6 we check the self-consistency of our model in terms of the ionization structure of the CGM. Section 7 addresses deviations from a pure-recombination scenario and in particular the impact of Lyα scattering. In Section 8 we discuss the possible origin and physical properties of the cold gas, including possible deviations from pressure equilibrium. A summary is given in Section 9.

In this paper we consistently adopt a flat cosmology with parameters H_0 = 67.74 \text{ km s}^{-1} \text{ Mpc}^{-1}, \Omega_m = 0.3089 and a cosmic baryon fraction \Omega_b/\Omega_m = 0.1573 (Planck Collaboration 2016). Using cosmological parameters as in Borisova et al. (2016) has negligible impact on our results.

2 THE COLD CGM IN MUSE QSO LYα NEBULAE

2.1 Recombination radiation from the CGM

2.1.1 Lyα recombination surface brightness

If a galaxy hosts a sufficiently bright QSO, large portions of its CGM can be almost completely photo-ionized (see Section 4 below for more details). Recombinations in the medium would produce Lyα radiation, with a surface brightness given by the integral along the line of sight of the recombination Lyα emissivity

\[ \Sigma = \int f_v B(T)n^2, \]

(1)

where \( f_v \) and \( n \) are, respectively, the volume filling factor and the hydrogen number density of the Lyα emitting gas, while the coefficient \( B \) is given by

\[ B(T) = \frac{E_{\text{Ly}\alpha}}{4\pi} \left(1 + 2 \frac{n_{\text{He}}}{n_H} \right) \alpha_{\text{Ly}\alpha}(T), \]

(2)

where \( E_{\text{Ly}\alpha} \) is the energy of the Lyα line, \( n_{\text{He}}/n_H = 0.25Y/(1-Y) = 1/12 \) is the relative proportion of helium-to-hydrogen nuclei for a helium abundance by mass \( Y = 0.25 \) and finally (posing \( T_4 = T/10^4 \text{ K} \))

\[ \alpha_{\text{Ly}\alpha} = 1.6 \times 10^{-13} T_4^{-0.96} \text{ cm}^3 \text{ s}^{-1}, \]

(3)

is the effective Lyα recombination coefficient, as derived by fitting the calculations by Pengelly (1964) in the temperature range \( 5 \times 10^3 \text{ K} < T < 8 \times 10^4 \text{ K} \). We adopt here case A coefficients. Note however that case A and B effective Lyα recombination coefficients differ by only 5% (Pengelly 1964). They must not be confused with the total recombination coefficients, which differ in the two cases by a factor \( \sim 1.6 \) (e.g. Osterbrock & Ferland 2006).

2.1.2 Filling factor

In a multiphase CGM, hot and diffuse gas at \( T \sim 10^6-10^7 \text{ K} \) occupies the majority of the volume and compresses the relatively cold phase \( (T \sim 1-5 \times 10^4 \text{ K}) \) to large densities and a small fraction of the volume. Given its dependence on density and temperature (equations 1 and 2), recombination radiation will be dominated by the cold and dense gas for any filling factor larger than \( \sim 10^{-6} \). This justifies the approximation that Lyα emission is produced by the cold phase only and explains the importance of including the filling factor \( f_v \) in equation 1.

We do not make assumptions here about the geometry or topology of the cold Lyα emitting regions (for instance, whether they are disconnected clouds or clumps, or a web of sheets and filaments), nor on their typical size, and we will loosely refer to them as cold gas structures. The possible nature and origin of the cold gas structure, as well as their typical sizes, are discussed further in Section 8.

\[ \odot 201X \text{ RAS, MNRAS 000. } 122 \]
If not stated otherwise, we interchangeably indicate with \( n_{\text{cold}} \) or simply with \( n \) the number density of hydrogen nuclei within the cold gas structures. This should not be confused with the \textit{volume averaged} density of cold gas, which is given by

\[
\langle n \rangle = f_V n, \tag{4}
\]

where the average is performed on volumes small with respect to the nebula but large with respect to the size of the emitting structures. Note that the term \( f_V n \) in equation (1) could equivalently be replaced by \( C \langle n \rangle^2 \), where

\[
C = f_V^{-1} \tag{5}
\]

is the \textit{clumping factor} of the emitting gas, as due to the fact that the latter only occupies a fraction of the total volume.

In this study, we neglect \textit{small-scale} variations of the density \textit{within} the structures themselves, i.e. their \textit{internal clumpiness}, which would appear as an additional multiplicative factor \( C_{\text{int}} \) at the right hand side of equation (5) (though see a discussion in Section 7.3 below and see also Cantalupo et al. 2019). We do, however, allow for \textit{large-scale} variations of the density \( n \) with the physical distance \( r \) from the centre of the nebula. This is both necessary – to explain the observed radially declining surface brightness profile (Section 2.2) – and expected, as the pressure of the confining hot gas is likely to decline with radius due to the gravitational pull of the halo (Section 3). The filling factor \( f_V \) and the temperature of the cold gas are also in principle a function of radius, though in the following we assume them to be constant, for simplicity.

2.1.3 Main caveats

In the rest of this section, we apply equation (1) to the observed \textit{Ly} \( \alpha \) surface brightness profile of the MUSE QSO Giant Ly \( \alpha \) nebulae, to infer the density and mass of the cold phase of the CGM around bright QSOs at \( z \sim 3 \). When doing so, one must be aware of a number of effects which could bias the result in one direction or the other. Dust absorption can attenuate the \textit{Ly} \( \alpha \) flux. We neglect dust, as we do not expect it to be present in the CGM or IGM in large amounts. Obviously, however, if the intrinsic \textit{Ly} \( \alpha \) surface brightness is larger than what we measure, our derived densities and masses should be interpreted as lower limits, strengthening our main finding of a large CGM baryon fraction. Similarly, if only a small fraction of the CGM is highly ionized (which may happen, for instance, because of a narrow aperture of the QSO ionization cone), then a significant portion of the existing cold CGM would not be visible in recombination radiation. This would also imply that our mass estimates are lower limits. On the other hand, mechanisms other than recombination, such as collisional excitation and scattering (or ‘photon pumping’) can contribute to the observed surface brightness. Scattering within the CGM can also mix photons belonging to different lines of sight, complicating the analysis. The most important of these points are discussed in greater detail in Sections 5 and 6. We argue there that they either do not affect our results, or can be taken into account by relatively simple modelling.

2.1.4 A note on ionization by stars

For some aspects of our problem – in particular, to estimate the temperature of the emitting gas, which determines the value of the recombination coefficient (see equation 3 and Section 4.3 below) – it is important to determine what the main source of ionizing photons is. In the close vicinity of a QSO, the QSO itself is obviously the most likely dominant source of ionizing photons. However, other sources, such as young stars, could in principle provide ionizing photons as well. For young stars to dominate the photoionization budget of the MUSE Giant Ly \( \alpha \) nebulae, \textit{both} the following two conditions must be met: i) that the ionization cone of the QSO is sufficiently narrow that most of the CGM is not illuminated by the QSO and ii) that there is sufficient star formation (either in the central galaxy or in the CGM itself) to keep the CGM completely ionized. The first condition cannot be discarded, considering the large observational uncertainties (see e.g. Ichikawa et al. 2019 and references therein), although recent work suggests that the most luminous QSOs may have rather wide opening angles (Ricci et al. 2017). The second necessary condition, on the other hand, is disfavoured based on energetic grounds and observed scaling relations. To power the observed median luminosity \( L_{\text{Ly} \alpha} = 10^{44} \text{ erg s}^{-1} \), one would need a star formation rate of 100 \( M_\odot \) yr\(^{-1}\) (e.g. Dijkstra & Westra 2010), assuming an implausibly large escape fraction of ionizing photons \( f_{\text{esc}} = 1 \), or \( 4 \times 10^3 \ M_\odot \) yr\(^{-1}\) assuming a more realistic value \( f_{\text{esc}} \approx 2.5\% \) (e.g. Haardt & Madau 2012 and references therein). For comparison, the observed stellar-to-halo mass relation (e.g. Behroozi et al. 2013; Moster, Naab & White 2013) and main sequence of star forming galaxies (e.g. Speagle et al. 2014 and references therein) imply that a halo with mass \( M_{\text{halo}} \approx 10^{12} \ M_\odot \) (close to the peak of star formation efficiency) at \( z \sim 3 \) would typically host a central galaxy with stellar mass \( M_\star = 10^{10.4} \ M_\odot \) and a star formation rate of 80 \( M_\odot \) yr\(^{-1}\), i.e. 50 times smaller than required. Similarly, diffuse star formation within the CGM itself is an unlikely explanation for the ionization state of MUSE nebulae, as it would require that the CGM forms stars at a rate equal to, or larger than, that in the central galaxy.

2.2 De-projected cold gas density profile

The MUSE QSO Giant Ly \( \alpha \) nebulae (Borisova et al. 2016) have a median intrinsic (redshift-dimming corrected) surface brightness profile which is well described by a power-law function of the projected radius \( R \):

\[
\Sigma(R) = 3.9 \times 10^{-4} \left( \frac{R}{10 \text{ kpc}} \right)^{-1.5} \text{ erg s}^{-1} \text{ cm}^{-2} \text{ sr}^{-1}. \tag{6}
\]

Note that the sample of Borisova et al. (2016) contains 17 radio quiet at 2 radio loud QSOs. We used here the entire sample; excluding the two radio-loud QSOs has negligible impact on our results. The slope in equation (6) is different from the slope reported in Borisova et al. (2016). Our slope was obtained by fitting to the median of the individually dimming-corrected surface brightness profiles, improving on the rough first estimate in Borisova et al. (2016). Our results are not sensitive to small changes in the slope of the surface brightness profile (see also Section 5.2 below). We did not attempt to model individual nebulae, mainly because a
measurement of the baryon fraction requires an estimate of the halo mass and this is known only in a statistical sense (see Section 4.2).

Assuming spherical symmetry (or as long as we focus on spherically averaged properties), we can deproject equation (6) by means of equation (1), to obtain the intrinsic radial profile of the density of the cold (Ly$\alpha$-emitting) gas:

$$n(r) = 0.15 f_V^{-1/2} T_4^{-1/2} \left( \frac{r}{10 \text{ kpc}} \right)^{-1.25} \text{ cm}^{-3},$$  \hfill (7)

where $r$ is the spherical radius and $t = 0.96$ is the exponent in the relation (9). More in detail, the normalization and slope $(n_0, \gamma)$ of the de-projected density profile (equation 7) are obtained from the normalization and slope $(\Sigma_0, \beta)$ of the projected brightness profile (equation 6) by means of the relations

$$n_0 = f_V^{-1/2} \left( \frac{\Sigma_0}{2B(T)\chi(1+\beta)R_0} \right)^{1/2},$$  \hfill (8)

$$\gamma = \frac{1+\beta}{2},$$

where $R_0 = 10 \text{ kpc}$ and the dimensionless projection factor $\chi$ is defined (see also Pezzulli, Fraternali & Binney 2017) as:

$$\chi(a) = \int_{-\infty}^{+\infty} \left( 1 + x^2 \right)^{-a/2} \, dx.$$  \hfill (9)

The derivation above, which is equivalent to a classic Abel de-projection, formally relies on the assumption that the surface brightness profile (equation 6) and the density profile (equation 7) extend to infinity. As this is not realistic, we show in Figure 1 the effect, on the projected Ly$\alpha$ surface brightness profile, of introducing a truncation radius $r_{\text{max}}$ in the density profile (7). Predictions are shown for some values of $r_{\text{max}}$, equal or larger than the virial radius of a $10^{12} \, M_\odot$ halo at $z = 3$ ($r_{\text{vir}} \approx 74 \, \text{kpc}$, see equation (11) below). Truncation radii equal or larger than $r_{\text{vir}}$ are all consistent with the current data. Recent results from the MUSEUM survey

(Arrigoni Battaia et al. 2019) possibly support a truncation radius not much larger than the virial radius. On the other hand (as already noted by Borisova et al. 2016), there is at least one example of a Giant Ly$\alpha$ nebula whose surface brightness profile is identical to those of the MUSE nebulae and extends to considerably larger physical radii (the Slug nebula, Cantalupo et al. 2014). We note, however, that azimuthal symmetry is a poor description of the Slug nebula, as well as of the outer regions of some of those MUSE nebulae that can be traced at radii larger than $r_{\text{vir}}$, so that the interpretation of profiles extending beyond the virial radius is not obvious. In the following, we will conservatively assume that the profile (7) holds within the virial radius, which is what we need to estimate the CGM baryon fraction there.

In any case, we emphasize that our analysis is not meant to be applied far out in the IGM, where both strong asymmetries and lack of pressure confinement from a hot atmosphere are expected.

The dependencies in equation (7) on $f_V$ and $T_4$ are readily understood: if a smaller fraction of the volume is occupied by Ly$\alpha$ emitting gas, individual structures must be denser in order to explain the observed emission; at the same time, a larger temperature implies a smaller effective recombination coefficient (equation 5) and therefore again a larger density to give account of a fixed Ly$\alpha$ surface brightness. Note that the relation between inferred density and filling factor in equation (7) is equivalent to the familiar relation between inferred average density and clumping factor $\langle n \rangle \propto C^{-1/2}$, as it follows from equations (4) and (5).

2.3 The baryon fraction of the cold CGM

The density profile (7) can be integrated over volume to obtain the total mass of cold gas within the virial radius:

$$M_{\text{cold}} = \int_0^{r_{\text{vir}}} 4\pi f_V m_p \Sigma(r) r^2 n(r) \, dr,$$  \hfill (10)

where $m_p$ is the mass of the proton and $X = 0.75$ is the cosmic hydrogen abundance by mass, while, as customary, we approximate the virial radius with the radius $r_{200}$ enclosing a 200-fold matter overdensity:

$$r_{\text{vir}} = \frac{1}{1+z} \left( \frac{GM_{\text{halo}}}{100 \, \Omega_m H_0^2} \right)^{1/3} = 74 \, M_{12}^{1/3} \, \text{kpc},$$  \hfill (11)

where $M_{12} = M_{\text{halo}}/10^{12} \, M_\odot$ is the halo mass in units of $10^{12} \, M_\odot$ and $z = 3.2$ is the median redshift of the MUSE QSO Ly$\alpha$ nebulae. Finally, dividing $M_{\text{cold}}$ by the cosmological baryon fraction gives the fraction, among all the baryons nominally associated to the halo, that are stored in the cold phase of the CGM:

$$f_{\text{CGM,cold}} = \frac{M_{\text{cold}}(r_{\text{vir}})}{(\Omega_b/\Omega_m) M_{\text{halo}}} = f_1 f_V^{-1/2} T_4^{-1/2} M_{12}^{-\gamma/3},$$  \hfill (12)

where $f_1 = 7.3$, $t = 0.96$ and $\gamma = 1.25$.

As $f_{\text{CGM,cold}}$ cannot exceed unity, equation (12) implies that the filling factor of the cold gas $f_V$ must be smaller than 1 (or, equivalently that the clumping factor must be larger than 1, cfr. equation 5), in agreement with previous findings (e.g. Cantalupo et al. 2014; Arrigoni Battaia et al. 2015).

We now consider the option that the physical reason why
the cold gas has $f_V < 1$ is pressure confinement by a hot corona, which is likely present in the halo, although invisible in Lyα emission.

3 THE MASS OF THE HOT CGM

Here we calculate the mass of the hot phase of the CGM that is needed to confine the cold phase to the densities derived in Section 2. Our treatment is based on the assumption of pressure equilibrium between the two phases and is not dissimilar, in its principles, to other works on the subject (e.g. Haiman & Rees 2001; Dijkstra & Loeb 2009; Vedantham & Phinney 2019). The possible presence and impact of deviations from pressure equilibrium are briefly discussed at the end of this Section and in more detail in Section 7.

Pressure equilibrium dictates that the ratio $\varepsilon$ of the densities of the two phases (i.e. the density contrast) equals the inverse ratio of their temperatures:

$$\varepsilon = \frac{n_{\text{hot}}}{n_{\text{cold}}} = \frac{T_{\text{cold}}}{T_{\text{hot}}}, \quad (13)$$

where $n_{\text{hot}}$ and $T_{\text{hot}}$ are the hydrogen number density and the temperature of the hot gas, respectively. Note that, in equation (13), we do not need to account for different ionized fractions (and therefore a different mean molecular weight) for the cold and the hot gas, as both phases are assumed here to be virtually completely ionized (although by entirely different processes: collisional ionization and photo-ionization for the hot and cold phase, respectively).

If one assumes that both $T_{\text{cold}}$ and $T_{\text{hot}}$ are independent on radius, then the density contrast $\varepsilon$ is also a constant, with the consequence that the inferred density profile of the hot gas follows a similar power law as the cold gas (equation 2), with a downward normalization shift, due to the fact that the hot gas is more diffuse:

$$n_{\text{hot}}(R) = \varepsilon n_0 \left( \frac{R}{R_0} \right)^{-\gamma}. \quad (14)$$

Interestingly, a power-law profile like that in equation (14) is indeed consistent with isothermal hydrostatic equilibrium in a dark matter halo gravitational potential, for a gas with close-to-virial temperature

$$T_{\text{hot}} = \frac{1}{\gamma} \frac{\mu m_p}{k_B} \frac{G M_{\text{halo}}}{r_{\text{vir}}}, \quad (15)$$

where $m_p$ is the mass of the proton, $k_B$ the Boltzmann constant and $\mu = 0.593$ the mean molecular weight. This suggests that temperature gradients, if present, are not dramatic.

By combining equations (11), (13) and (15), the density contrast can be written

$$\varepsilon = \frac{\gamma k_B T_{\text{cold}}}{\mu m_p} (1 + z)^{-1} (100 \Omega_m H_0^2)^{-1/3} (G M_{\text{halo}})^{-2/3} \left( \frac{r}{r_{\text{vir}}} \right)^{-\gamma}$$

$$= \varepsilon_1 T_4 M_{12}^{-2/3}, \quad (16)$$

with

$$\varepsilon_1 = 1.0 \times 10^{-2} \frac{\gamma}{1+z} = 3.0 \times 10^{-3} \quad (17)$$

for $z = 3.2$ and $\gamma = 1.25$.

The total mass of the hot gas within the virial radius is readily obtained by integrating equation (14). Equivalently, the ratio of hot to cold gas mass is given by the ratio of densities times the ratio of the volumes occupied by the two phases

$$\frac{M_{\text{hot}}}{M_{\text{cold}}} = \varepsilon \frac{1 + f_V}{f_V} = \varepsilon_1 f_V^{-1} T_4 M_{12}^{-2/3}, \quad (18)$$

where we have used equation (19) and taken into account that $f_V \ll 1$ for every viable model (cfr. Section 2.3).

The quantitative implications of equations (12) and (18) are presented in Section 4. Before doing so, we briefly discuss here whether any obvious large errors on the mass of the hot gas are expected to be introduced by our simplifying assumptions.

It is easy to verify that the simple inference above is not affected by the detailed shape of the gravitational potential. In a realistic potential (e.g. Navarro, Frenk & White 1996), the hydrostatic equilibrium density profile of an isothermal hot gas would not be exactly described by a power-law, but rather require an additional radial-dependent factor equal to $(V_{\text{circ}}(r)/V_{\text{vir}})^2$, where $V_{\text{circ}}(r)$ is the circular speed at radius $r$ and $V_{\text{vir}} = V_{\text{circ}}(r_{\text{vir}})$. In the situation of interest, this correction is remarkably small (less than 10% for $0.2 < r/r_{\text{vir}} < 1$), mostly due to the relatively low concentration of dark matter haloes at this redshift ($c \sim 3.5$, e.g. Dutton & Macciò 2014). The impact on the total mass is even smaller (less than 1%), mainly because most of the mass is at large radii, where the approximation $V_{\text{circ}} \approx V_{\text{vir}}$ is the most accurate. Coherent rotation, which must be present as the CGM necessarily has non-negligible angular momentum, also affects the density profile by providing centrifugal support against gravity, but also in this case the impact is mostly limited to the very innermost regions, which contain a very small fraction of the mass (Pezzulli, Fraternali & Binney 2017; Oppenheimer 2018).

More than the detailed shape of the potential or the presence of coherent rotation, deviations can arise due to large scale turbulence, or other sources of non-thermal pressure. Crucially, however, such deviations would contribute both to enhance the compression of the cold phase and to provide additional support to the hot gas against gravity, with the result that the density contrast $\varepsilon$ (which is essentially estimated by taking the ratio of the two effects) would remain largely unaffected. A formal way to see this is to reformulate equation (13) in terms of a ratio of sound speeds squared $\varepsilon = c_{\text{hot}}^2/c_{\text{cold}}^2$. If large scale turbulence is in place, the cold gas would be subject to (and compressed by) additional ram pressure, which would be, at least to the first order, encapsulated by the replacement $c_{\text{hot}}^2 \rightarrow c_{\text{hot}}^2 + c_{\text{turb}}^2$. An analogous reformulation and replacement would be needed in equation (15) as well, as the very same turbulent motions would also provide support to the hot gas against gravity.

Although we cannot claim this result to be exact, physical insight leads us

3 viz, the equation $c_{\text{hot}}^2 = GM_{\text{halo}}/\gamma r_{\text{vir}}$ (equivalent to equation 15) should be replaced by $(c_{\text{hot}}^2 + c_{\text{turb}}^2) = GM_{\text{halo}}/\gamma r_{\text{vir}}$. Note that we are using here the ‘isothermal sound speed’ $c_s \equiv P/\rho$ to simplify the notation and without loss of generality.

4 The same cannot be said for the temperature of the hot gas and...
to believe that additional corrections, if present, should be second-order and sub-dominant. Similar arguments should hold essentially for any large scale source of non-thermal pressure. Considerations about other kinds of pressure imbalance are discussed in Section 7.

4 THE TOTAL CGM BARYON FRACTION

By combining equations (12) and (18), we finally obtain the total (cold plus hot) CGM baryon fraction within the virial radius:

\[ f_{\text{CGM, tot}} = \frac{M_{\text{cold+hot}}(r_{\text{vir}})}{(\Omega_b/\Omega_m) M_{\text{halo}}} = f_\gamma (1 + \varepsilon_1 f_V^{-1} T_4 M_{12}^{2/3}) f_V^{1/2} T_4^{1/2} M_{12}^{-\gamma/3}, \]  

(19)

where we recall that \( f_\gamma = 7.3, \varepsilon_1 = 3.0 \times 10^{-3}, t = 0.96 \) and \( \gamma = 1.25 \) (equations [3] [8] [12] and [17]).

The total CGM baryon fraction thus depends on three parameters: \( f_V, M_{\text{halo}} \) and \( T_{\text{cold}} \). We discuss these dependencies below, one at a time.

4.1 Dependence on the filling factor

In Figure 2 we show how the total CGM baryon fraction, inferred from MUSE observations and from our model, depends on the filling factor of the cold gas \( f_V \). We adopt here a halo mass \( M_{\text{halo}} = 10^{12.3} M_\odot \) and a temperature of the cold gas \( T_{\text{cold}} = 10^4 K \), but we stress that these values are fixed here for illustrative purposes only. The effect of varying \( M_{\text{halo}} \) and \( T_{\text{cold}} \) is discussed in detail in Sections 4.2 and 4.3 respectively. In Figure 2 we also show the individual contributions of the cold gas and the hot gas to the total baryon fraction of the CGM.

As it is clear from the figure and from equation (12), the cold and the hot gas have opposite dependencies on \( f_V \). The estimated mass of the cold CGM increases with increasing \( f_V \); this is a consequence of the dependence of the recombination emissivity on the density squared (see Section 2). On the other hand, the estimated mass of the hot gas increases with decreasing \( f_V \); this is mostly because a larger amount of hot gas is needed to compress the cold gas to the larger densities that are associated to smaller filling factors (Section 3).

These opposite dependencies have two obvious consequences. First, there is at most a finite range of filling factors that are consistent with a physically meaningful baryon fraction (equal or smaller than the cosmological value, shown as a horizontal black dashed line). Second, for any given choice of the other two parameters (\( M_{\text{halo}} \) and \( T_{\text{cold}} \)), there is a minimum value of the total CGM baryon, below which no model based on recombination radiation and pressure balance would able to explain the observed Lyα surface brightness of the MUSE nebulae. We will use this minimum in Section 4.4 to put a lower limit to the total CGM baryon fraction, as inferred from the MUSE observations within our assumptions. We emphasize that the minimum CGM baryon fraction is obtained when the two phases contribute equally to the total mass: if either of the two phases dominates in mass, the total baryon fraction will be larger than this lower limit.

4.2 Dependence on halo mass

Figure 3 shows the dependence of the inferred CGM baryon fraction on the assumed mass of the halo. Different lines correspond to the total (cold plus hot) CGM baryon fraction, for different values of the assumed halo mass and for one fixed value of the temperature of the cold gas \( T_{\text{cold}} = 10^4 K \).

The inferred CGM baryon fraction decreases with increasing assumed halo mass. This is partly due to the obvious fact that increasing the halo mass also increases the amount of baryons theoretically associated to the halo, so that a lower fraction of those is needed to explain the observed Lyα surface brightness. Note however that the virial radius and the virial temperature are also changing accordingly, which explains why the dependence on halo mass is not trivially linear (equation 19).

With decreasing halo mass, it becomes increasingly difficult to accommodate a total CGM baryon fraction significantly smaller than the cosmological value, until eventually no physically acceptable solution is found. This makes the knowledge of the typical mass of QSO-hosting haloes at \( z \sim 3 \) crucial to draw conclusions from our results.

Probably the most reliable estimates of QSO-hosting halo masses are those based on large scale QSO clustering, or on QSO-galaxy cross-correlation. Several studies of
Figure 3. Dependence of the total CGM baryon fraction on the filling factor of the cold gas, for different values of the assumed halo mass and a fixed temperature of the cold CGM $T_{\text{cold}} = 10^4$ K. Results are shown for halo masses within the range inferred from galaxy-QSO cross-correlation (Trainor & Steidel 2012). Relatively low halo masses ($M_{\text{halo}} \sim 10^{12}$ M$_\odot$), as recently suggested by BOSS QSO clustering (Eftekharzadeh et al. 2015) would imply a CGM baryon fraction close to the cosmological value.

This kind have been attempted at a redshift $z \sim 3$ relevant to our application, finding a broad range of results, from $M_{\text{halo}} \sim 10^{12}$ M$_\odot$ to $M_{\text{halo}} \sim 10^{13}$ M$_\odot$ (see e.g. Shen et al. 2007 and references therein). Big uncertainties are to a large extent due to low-number statistics, inherent to studies of rare objects such as QSOs, but possibly also to differences in selection techniques and especially in the methods used to determine the redshift of the sources, which is crucial for clustering measurements. Taking these aspects into account, we chose to use, as a reference, the study by Eftekharzadeh et al. (2015). Their sample in fact is — by far — the one with the cleanest selection and the better statistics to our knowledge at the redshift of interest, with more than 20 000 QSOs with spectroscopic redshifts in the range $2.64 < z < 3.4$, as determined from the final sample of the Baryon Oscillation Spectroscopic Survey (BOSS, Dawson et al. 2013). These authors find a typical halo mass $M_{\text{halo}} = 10^{12}$ M$_\odot$, which is on the low side of the range of masses mentioned above. We refer to Eftekharzadeh et al. (2015) for a discussion of why larger values had been found by the previous work of Shen et al. (2007) on a smaller sample.

Despite the lack of new studies with sample sizes comparable to that of Eftekharzadeh et al. (2015), the debate on the mass of QSO-hosting haloes at $z \sim 3$ is still ongoing in the recent literature. For instance, Allevato et al. (2016) inferred a large mass $M_{\text{halo}} \sim 10^{12.9}$ M$_\odot$ based on a sample of ~ 400 X-ray selected AGN in a range $2.9 < z < 5.5$, ~ 100 of which with spectroscopic redshift. Also, Timlin et al. (2018) determined halo masses with a large uncertainty $M_{\text{halo}} = 1.70 - 9.83 \times 10^{12}$ h$^{-1}$ M$_\odot$ based on a sample of 1378 objects, identified as QSO candidates and assigned a photometric redshift in the range $2.9 < z < 5.1$ by a machine learning algorithm. The results by Allevato et al. (2016) and Timlin et al. (2018) emphasize the importance of further investigations to attenuate uncertainties in this field. However, given their much smaller samples with respect to Eftekharzadeh et al. (2015), as well as the less robust selection techniques (especially concerning the determination of redshift) we still think that the BOSS result should, for the time being, be considered as the reference.

Even if the BOSS result is (at least formally) the best measurement available to date, it is still not obvious that it can be safely be applied to our study. The sample of Eftekharzadeh et al. (2015) in the relevant redshift range contains QSOs as luminous as $\log(L_{\text{bol}}/\text{erg s}^{-1}) = 47.72$. This overlaps with the luminosity range of the QSOs in the sample of Borisova et al. (2016) (47.29 < $\log(L_{\text{bol}}/\text{erg s}^{-1})$ < 48.14, using equation 1 in Shen et al. 2009), but the latter also contains objects up to 2.6 times more luminous than the maximum luminosity probed by Eftekharzadeh et al. (2015). If, as some theoretical models predict, the halo mass correlates strongly with QSO luminosities (see e.g. Hopkins et al. 2007; Conroy & White 2013), it is possible that the QSOs of Borisova et al. (2016) live in haloes with a larger mass than the QSOs in the BOSS sample. On the other hand, a number of studies have recently shown that statistical measurements of QSO host halo masses correlate surprisingly weakly with QSO luminosities (e.g. Shen et al. 2009; Padmanabhan et al. 2009; Eftekharzadeh et al. 2015; Uchiyama et al. 2018; He et al. 2018), so that this option is not very clearly supported by the current data.

Furthermore, we notice that the MUSEUM survey (Argigoni Battaina et al. 2019) has investigated Giant Ly$\alpha$ nebulae around a sample of QSOs less luminous than those of Borisova et al. (2016), reaching a maximum luminosity of $\log(L_{\text{bol}}/\text{erg s}^{-1}) = 47.87$. This is very similar to the maximum luminosity probed by Eftekharzadeh et al. (2015), possibly making comparisons less prone to potential bias. The Giant Ly$\alpha$ nebulae in the MUSEUM sample have a surface brightness profile that is undistinguishable from those of Borisova et al. (2016) in the radial range of interest for this study, possibly suggesting that a bias in QSO luminosity is not the main explanation of the large surface brightnesses observed in all these objects, although studying Ly$\alpha$ nebulae around QSOs with luminosities even lower than those in the MUSEUM survey is needed to achieve a more complete picture on this matter.

Based on the discussion above, in the following, we will sometimes use as fiducial the halo mass $M_{\text{halo}} = 10^{12}$ M$_\odot$ found by Eftekharzadeh et al. (2015). However, we will also show and discuss results obtained in a broad range of masses $M_{\text{halo}} = 10^{12.3 \pm 0.5}$ M$_\odot$ (Trainor & Steidel 2012), which we chose as representative of the range of uncertainty in the value of the halo mass.

If a low value $M_{\text{halo}} \sim 10^{12}$ M$_\odot$ is confirmed by future studies, it would imply that it is not easy to explain the surface brightness of MUSE Giant Ly$\alpha$ nebulae with a baryon fraction significantly smaller than cosmological, at least within the assumptions discussed so far.

On the other hand, a significantly larger value for the halo mass would not necessarily imply that the observed Ly$\alpha$ surface brightness is trivial to reproduce. Most models in fact agree that at sufficiently large halo masses, most of the gas is expected to be in the hot phase. This would violate the first condition to maximize the predicted Ly$\alpha$
In Figure 4 we show the dependence of the inferred total gas. A baryon fraction smaller than cosmological requires rather low temperatures, which are, however, difficult to achieve in the photo-ionization field of a QSO (see Figure 5 and text for details).

In Figure 4, similar to Figure 3 but for a fixed halo mass $M_{\text{halo}} = 10^{12.3} \, M_\odot$ and different values of the temperature of the cold gas. A baryon fraction smaller than cosmological requires rather low temperatures, which are, however, difficult to achieve in the photo-ionization field of a QSO (see Figure 5 and text for details).

4.3 Dependence on the temperature of the cold gas

In Figure 4 we show the dependence of the inferred total CGM baryon fraction on the assumed temperature of the cold photo-ionized gas $T_{\text{cold}}$, for a fixed halo mass $M_{\text{halo}} = 10^{12.3} \, M_\odot$. The inferred CGM baryon fraction increases with increasing assumed $T_{\text{cold}}$. The reason is twofold. First, as the recombination coefficient is a declining function of temperature (equation 3), a larger $T_{\text{cold}}$ requires a larger amount of cold gas to explain the same observed Lyα emission. Second, a larger $T_{\text{cold}}$ also implies a larger pressure of the cold gas and therefore a larger amount of hot gas needed to keep it confined. The magnitude of the effect is significant and it is therefore important to make physically motivated assumptions concerning the value of $T_{\text{cold}}$.

In an almost fully photo-ionized gas, the temperature is mainly determined by the equilibrium between photo-electric heating and recombination cooling and therefore, ultimately, by the shape of the ionizing spectrum (e.g. Osterbrock & Ferland 2006). In our case, the ionizing spectrum is completely dominated by QSO (we will see in Section 5.1 below) and the UVB can be neglected by orders of magnitude. In a pure hydrogen nebula, the resultant equilibrium temperature would be

$$T_{\text{eq}} = \frac{2h\nu_{\text{LL}}}{3(p+2)k_B} = 2.8 \times 10^4 \left(\frac{p+2}{3.7}\right)^{-1} \, K,$$

(20)

where $h$ is the Planck’s constant, $\nu_{\text{LL}}$ is the Lyman-limit frequency and $p$ is the power-law slope of the hydrogen ionizing continuum, for which we take $p = 1.7$ as a fiducial value (Lusso et al. 2015 see also Section 5 below).

In a QSO ionizing field, however, the photo-electric effect due to the second ionization of helium is also a very important heating channel. Taking this into account and assuming that a single power-law spectrum can be extrapolated to $E > 4$ Ry, the equilibrium temperature increases by a factor $(1 + 4A(T))/(1 + A(T))$, where

$$A(T) = \frac{n_{\text{He}} \alpha_{\text{He}}(T)}{n_{\text{H}} \alpha_{\text{H}}(T)}.$$  

(21)

By using the temperature dependent recombination coefficients as in Meiksin (2009), we then find an equilibrium temperature

$$T_{\text{eq}} = 5.6 \times 10^4 \, K.$$  

(22)

This is large enough that no physically acceptable solution can be found for a halo mass $M_{\text{halo}} = 10^{12.3} \, M_\odot$, as it is clear for instance from Figure 4. Note that the problem would only be exacerbated for $M_{\text{halo}} = 10^{12} \, M_\odot$, as suggested by clustering measurements (Section 4.2).

Deviations from the simple estimate above would mostly come from i) line cooling from residual non completely ionized hydrogen and helium and ii) traces of metals, which would constitute additional line cooling channels, but also additional targets to photo-electric heating. Other potentially relevant effects are those whereby the energy of a photon can be distributed to more than one electron and primarily i) ionization of hydrogen by high-energy HeII recombination photons and ii) secondary ionization (of hydrogen or other species) by high-energy electrons produced in a previous ionization event (Osterbrock & Ferland 2006).

To assess the role of all these effects, we performed more detailed calculations of the equilibrium temperature using the photo-ionization code Cloudy (version 13.04, last described by Ferland et al. 2013). We modelled the hydrogen and helium ionizing spectrum as a single power law with slope $p = 1.7$. We also included a hardening at X-ray frequencies (constant $\nu_{\text{LL}}$ in the energy range $2 \, \text{keV} < E < 100 \, \text{keV}$) and a cut-off for $E > 100 \, \text{keV}$, though we found the details of the X-ray spectrum to be unimportant. We adopted a specific luminosity at the Lyman limit $L_{\text{H} \alpha} = 5.2 \times 10^{31} \, \text{erg s}^{-1} \, \text{Hz}^{-1}$ (see Section 5.1 below) and a distance from the source $d = 50 \, \text{kpc}$. Results are shown in Figure 5 for different values of gas density and metallicity.

We see that the results are in broad agreement with the simple analytical estimate above (equation 22), especially for intermediate densities and metallicities. As also expected, the equilibrium temperature decreases with increasing density, due to collisionally excited line cooling, the effect being larger for larger metallicity. Note that at low densities the effect of non-primordial metallicity is reversed, as photo-electric heating on highly ionized metal species (as well as a small contribution of Compton heating) dominates over additional line cooling, so that the equilibrium temperature is even larger than the analytic estimate.

An equilibrium temperature equal or smaller than $T \sim 2 \times 10^4 \, K$, as desirable from inspection of Figure 4, can be reached only for a rather high density $n = 10 \, \text{cm}^{-3}$ and solar metallicity. Note that at a distance $d = 50 \, \text{kpc}$...
and a temperature $T = 2 \times 10^4 \text{ K}$, a density $n = 10 \text{ cm}^{-3}$ corresponds to a filling factor $f_{\nu} = 7.3 \times 10^{-6}$ (equation 7), which would give an absurdly large baryon fraction even at low temperature (Figure 4). Furthermore, it is not obvious that a metallicity as high as solar can be easily reached in the CGM at $z = 3$. Doing the calculation at smaller distances $d < 50 \text{ kpc}$ would not help in obtaining lower temperatures, as the ionization parameter increases with decreasing radius for our density slope $\gamma = 1.25 < 2$.

As a sanity check, we also repeated our calculation including only the (QSO-dominated) ionizing background at $z = 3.2$, a metallicity $Z = 0.1 \ Z_{\odot}$ and a density equal to ten times the critical density at $z = 3.2$, in order to mimic the conditions in the IGM far away from an individual QSO. We found in this case an equilibrium temperature $T_{\text{eq}} = 3 \times 10^4 \text{ K}$, in agreement with models of HeII re-ionization (e.g. McQuinn et al. 2009, Upton Sanderbeck, D’Aloisio & McQuinn 2016). The relatively low equilibrium temperature in this case is due to the relatively low ionization parameter at large distances from a single bright QSO. This emphasizes that temperatures as high as those shown in Figure 3 are expected, for a gas photo-ionized by a QSO, only in the very close vicinity to the source, such as in the CGM of the host galaxy, which is the main focus of this work.

Another possibility to achieve lower equilibrium temperatures at the small distances of interest could be to change the shape of the ionizing spectrum, especially at energies $E > 4 \text{ Ry}$, which contribute most of the heating and on which observational constraints are very scant. We therefore also considered a modified spectrum, identical to the one above but for being double as steep at $E > 4 \text{ Ry}$ than in the range $1 \text{ Ry} < E < 4 \text{ Ry}$. However, both analytic estimates and Cloudy modelling showed in this case only a modest $\sim 25\%$ decrease in $T_{\text{eq}}$. Virtually, only an exponential cut-off in the EUV spectrum would be effective in heating the temperature down as required. We note that this is in fact the theoretical expectation for the emission spectrum of a classical accretion disc around a supermassive black hole (see e.g. Shakura & Sunyaev 1973, Laor & Davis 2011). This is shown in Figure 6, where we assumed a spectrum of a classical (geometrically thin, optically thick) accretion disc (Shakura & Sunyaev 1973) with Eddington accretion around a supermassive black hole with mass $M_{\text{BH}} = 10^8 \ M_{\odot}$, ignoring in particular any comptonization effect. As expected, this kind of model easily predicts an equilibrium temperature $T_{\text{eq}} \approx 2 \times 10^4 \text{ K}$ for a metallicity $Z = 0.1 \ Z_{\odot}$ and a broad range of densities and even lower temperatures ($T_{\text{eq}} \approx 10^3 \text{ K}$) for solar metallicity. This kind of solution is formally appealing and it emphasizes the crucial role of assumptions on the shape of the ionizing spectrum of QSOs. Nonetheless, we refrain from considering this option as our fiducial case, as it would raise significant issues with a large body of independent observations including i) wide-spread evidence of comptonization in AGN (e.g. Done et al. 2012 and references therein) and ii) indirect evidence based on HeII re-ionization (e.g. Worseck et al. 2016 and references therein).

Motivated by the discussion above, in the following we will sometimes adopt as fiducial the value $T_{\text{cold}} = 5.6 \times 10^4 \text{ K}$ (cfr. equation 22 and Figure 3). We will nonetheless also present our results for a large range of temperatures to allow for more generality.

### 4.4 The minimum CGM baryon fraction

A useful way to summarise the results of the previous sections is to compute, for each value of the parameters $M_{\text{halo}}$ and $T_{\text{cold}}$, the minimum total CGM baryon fraction. As it is readily found from equation (19), this is equal to

$$f_{\text{min}} = 2 f_{\epsilon_1} T_4^{(1+\gamma)/2} M_{12}^{-(1+\gamma)/3}$$

© 201X RAS, MNRAS 000, 1
and is achieved for a filling factor $f_{V, eq}$ such that cold and hot gas contribute equally to the total mass:

$$f_{V, eq} = \varepsilon_1 T_a M_{12}^{-2/3}. \quad (24)$$

Equation (23) gives the lower limit to the total CGM baryon fraction needed to explain the observations within our assumptions. It is shown in Figure [as a function of the two parameters $M_{halo}$ and $T_{cold}$. The solid black line indicates the cosmological baryon fraction. The blank region at low masses and high temperatures corresponds to unphysical solutions (baryon fraction larger than cosmological), while progressively smaller baryon fractions are inferred at large halo masses and low temperatures.

The couple of fiducial values $M_{halo} \approx 10^{12} M_\odot$ and $T_{cold} \approx 5.6 \times 10^4$ K (determined in Section 4.2 and 4.3 respectively) fall in the unphysical region. The simplest explanation for this fact is that at least one of these values is incorrect. We do not have strong reasons to prefer a revision of either $M_{halo}$ or $T_{cold}$. We stress, however, that both parameters can be constrained by independent observations and in particular by further studies of clustering and spectral properties of luminous QSOs (Sections 4.2 and 4.3). Another option is that both our fiducial values are correct, but that some of our other assumptions (recombination dominated emission and pressure equilibrium) is wrong. These possibilities are discussed in Sections 6 and 7 but we anticipate that they are not trivially achieved either.

Figure [can also be compared to cosmological hydrodynamical simulations or to semi-analytic models, as long as these include predictions for the CGM baryon fraction in massive haloes at high redshift.

Cosmological hydrodynamical simulations with strong ejective feedback (see Section 4.1) typically predict a CGM baryon fraction equal to $\sim 30\% - 40\%$ of the cosmic average at the relevant halo mass and redshift (e.g. [David 2009]; Liang et al. 2016; Kulier et al. 2018; Mitchell priv. comm.). Unfortunately, not all simulation projects have published suitable predictions, so we cannot make an exhaustive comparison here. On the other hand, we hope that our results can encourage more of these predictions in the future. Based on the references above and on Figure 4, we expect that simulations with strong ejective feedback should be able to reproduce the observed Ly$\alpha$ surface brightness, under the assumption of fluorescent emission (recombination), only if the analysis is performed on rather massive haloes ($M_{halo} \gtrsim 10^{12.5} M_\odot$) and the cold CGM temperature is assumed to be sufficiently low ($T_{cold} \lesssim 2 \times 10^4$ K), despite photo-heating by the central QSO. Furthermore, it is required that the CGM mass is split almost equally among the cold and the hot phase, so that the former is both massive and dense enough to give rise to sufficiently bright recombination radiation. In Section 4.5 we discuss in some detail the only experiment published so far that can be directly compared to our expectations.

Among semi-analytic models with strong ejective feedback, the one of Henriques et al. (2018) predicts that a typical halo with $M_{halo} = 10^{12} M_\odot$ at $z = 3$ would have a median of 53% of its baryons in the CGM. This is slightly larger than what found by the simulations mentioned above and therefore potentially closer to what needed to explain the observations. Note however that the CGM in the model of Henriques et al. (2018) is all hot and therefore invisible in Ly$\alpha$, unless a large fraction of it is somehow assumed to be converted into cold CGM.

4.5 Comparison with Gronke & Bird (2017)

A special mention is needed to the study of Gronke & Bird (2017). These authors analysed massive haloes at $z = 2$ in the Illustris simulation (Vogelsberger et al. 2014) and calculated the expected recombination Ly$\alpha$ surface brightness, under the assumption that the CGM is highly ionized. They found agreement with the observed surface brightness profile of the Slug nebula and of MUSE Giant Ly$\alpha$ nebulae (note that the Borisova et al. 2016 profile in their Figure 9 should be shifted upwards by 0.58 dex due to differential redshift-dimming between the observed $z = 3.2$ and the simulated $z = 2$). As they assumed recombination radiation from a highly ionized medium, their experiment is directly comparable to our analytic calculations. As discussed in Section 4.4 we expect simulations with strong ejective feedback to be able to reproduce a Ly$\alpha$ surface brightness as high as observed, as long as the assumed halo mass and gas temperature are sufficiently large and small, respectively. Indeed, Gronke & Bird (2017) adopted a range of halo masses centered on $M_{halo} = 10^{12} M_\odot$, which is on the high side of those discussed in Section 4.3. Furthermore, their post-processing analysis neglected photo-heating from the QSO, with the result that they most likely underestimated the temperature of the photo-ionized CGM. Note that Illustris does include a recipe for radiative feedback from QSOs (Vogelsberger et al. 2013; Gronke & Bird 2017), however, did not select their haloes on the requirement that the central super massive black hole be in an active phase and therefore a QSO would be found therein with a probability that can be approximated by the QSO duty cycle $f_{duty} \approx 1$ (e.g. Haiman, Ciotti & Ostriker 2004; Eftekhارzadeh et al. 2015).

© 201X RAS, MNRAS 000 [122]
Note also that any imprint of past QSO activity on photo-ionized gas temperatures would die out on roughly one recombination time-scale, or less than 1 Myr in our case.

We therefore think that the formal success found by this work is not surprising and actually in line with our analytic expectations, but it may be driven by a particularly favourable choice of the parameters. In particular, our analysis shows that neglecting in post-processing photo-heating of the cold CGM by the central QSO can significantly bias the recombination rate and therefore the predicted Lyα surface brightness in the direction needed to make the comparison with observations easier.

Gronke & Bird (2017) also considered the option that most of the CGM is ionized by star formation rather than by the central QSO, in which case the temperature of the photo-ionized gas can be lower than our estimates. This option, however, is strongly disfavoured for the MUSE Giant Lyα nebulae, as we have argued already in Section 2.1.4.

5 IONIZATION STRUCTURE

Here we verify the self-consistency of our model. We start from the key requirement that the CGM can be almost completely photo-ionized by the central QSO (Section 5.2), before discussing the ionization of helium (Section 5.3) and the possible contribution of collisional excitation to the Lyα emission (Section 5.4). We obviously refer here only to the fraction of the solid angle that is illuminated by the QSO. We recall that considering any substantial QSO obscuration (by a dusty torus or any other intervening material) would only go in the direction of reinforcing our conclusions (see Section 2.1.3).

5.1 Hydrogen

The QSOs in the sample of Borisova et al. (2016) have a median specific luminosity at the Lyman limit frequency $\nu_{LL}$ of $L_{\nu_{LL}} = 5.2 \times 10^{41} \text{erg s}^{-1} \text{Hz}^{-1}$, which we estimated following Arrigoni Battaia et al. (2015) and using the absolute i-band magnitudes of the QSOs in the Borisova et al. (2016) sample (see their Table 2). The corresponding hydrogen ionization rate at distance $r$ from the QSO is

$$\Gamma_{\text{ion}}(r) = \frac{1}{4\pi r^2} \int_{\nu_{LL}}^{\infty} \frac{L_{\nu}}{\nu} \sigma_{\text{ion}}(\nu) d\nu$$

$$= \frac{1}{p + 3} \frac{L_{\nu_{LL}} \sigma_{\text{ion}}^{\text{local}}}{{4\pi hr^2}}$$

$$= 8.8 \times 10^{-7} \xi \left(\frac{r}{10 \text{ kpc}}\right)^{-2} \text{s}^{-1},$$

where $\sigma_{\text{ion}}(\nu) \simeq \sigma_{\text{ion}}^{\text{local}} (\nu/\nu_{LL})^{-3}$, with $\sigma_{\text{ion}}^{\text{local}} = 6.3 \times 10^{-18} \text{cm}^2$, is the hydrogen photo-ionization cross section and $L_{\nu} \simeq L_{\nu_{LL}}(\nu/\nu_{LL})^{-p}$ is a power-law approximation of the ionizing continuum of the QSO, with a typical slope $p = 1.7$ (e.g. Lusso et al. 2015), while the factor

$$\xi \equiv \left(5.2 \times 10^{31} \text{erg s}^{-1} \text{Hz}^{-1}\right) \left(\frac{p + 3}{4.7}\right)^{-1},$$

incorporates possible deviations of the normalization and slope of the QSO spectrum from our fiducial values. Note that adopting a more precise double power-law approximation to the frequency dependence of the hydrogen ionization cross section (see e.g. Meiksin 2009) is equivalent to replacing in the expressions above $\frac{L_{\nu}}{\nu} \rightarrow \frac{L_{\nu}}{\nu} \Gamma_{\text{rec}}(p + 3) - p$, with $s = 2.99$ and $p = 1.34$. This gives a correction of less than 7% for any $p > 1$ and is negligible to our purposes.

The photo-ionization rate given in equation (25) should be compared to the total ionization rate (per proton) for a fully ionized gas:

$$\Gamma_{\text{rec}} = \alpha_{\text{rec}}(T) \left(1 + 2 \frac{n_{\text{He}}}{n_{\text{H}}}\right)n$$

$$= 4.8 \times 10^{-13} T_4^{-0.75} \left(\frac{n}{\text{cm}^{-3}}\right) \text{s}^{-1}$$

where again we adopted a power-law approximation to a more complex temperature dependence (Meiksin 2009) over the temperature range of interest and we accounted for electrons from doubly-ionized helium. It is readily seen that $\Gamma_{\text{rec}} \propto \Gamma_{\text{ion}}$ for any plausible density for CGM conditions (the equality would require $n \simeq 2 \times 10^{14} \text{cm}^{-3}$ at $r = 100$ kpc and even larger densities at smaller radii). Note also that $\Gamma_{\text{rec}}$ exceeds the photo-ionization rate due to the cosmic UV background (e.g. Haardt & Madau 2012) by several orders of magnitude at any distance plausibly associated to the CGM.

Again following Meiksin (2009), the equilibrium hydrogen neutral fraction under these conditions is achieved on very short time-scale $t_{\text{ion}} = \frac{1}{\Gamma_{\text{ion}}}$ (everywhere smaller than a few years in our case) and very well approximated by $x_{\text{HI}} = \frac{\Gamma_{\text{rec}}}{\Gamma_{\text{ion}}}$, which gives

$$x_{\text{HI}} = 7.9 \times 10^{-8} f_{\nu}^{-1/2} \xi^{-1} T_4^{-0.27} \left(\frac{r}{10 \text{ kpc}}\right)^{0.75}. \tag{28}$$

Combining equations (7) and (28) we obtain the neutral hydrogen number density $n_{\text{HI}} = x_{\text{HI}} n_{\text{H}}$, which reads

$$n_{\text{HI}} = 1.1 \times 10^{-8} f_{\nu}^{-1/2} \xi^{-1} T_4^{-0.21} \left(\frac{r}{10 \text{ kpc}}\right)^{-0.5} \text{cm}^{-3}, \tag{29}$$

and therefore the volume averaged neutral hydrogen number density

$$\langle n_{\text{HI}} \rangle = f_{\nu} n = 1.1 \times 10^{-8} \xi^{-1} T_4^{-0.21} \left(\frac{r}{10 \text{ kpc}}\right)^{-0.5} \text{cm}^{-3}, \tag{30}$$

which is independent on the filling factor. This is not surprising, if one considers that both recombination radiation (on which our inferred density profile of equation 24 is based) and the neutral fraction are determined by the same physical process (obviously, recombination). It is similarly easy to show that the the inferred volume averaged neutral hydrogen number density $\langle n_{\text{HI}} \rangle$ is actually independent also on assumptions on the ionization state of helium (which enters in both equations 7 and 27) and on possible internal clumpiness $C_{\text{int}}$ within the cold gas structures (cfr. Section 2.1.3), as all these factors equally contribute to the recombination emissivity and the recombination rate. The small dependence on temperature is due to a non constant efficiency of the ionizing continuum from our fiducial values (see e.g. Haardt & Madau 2012) by several orders of magnitude at any distance plausibly associated to the CGM.
If radially integrated, equation (29) gives the neutral hydrogen column density between the QSO and a point in the nebula at distance $r$ from the centre:

$$N_{\text{HI}}(r) = 7.1 \times 10^{14} \xi^{-1} T_4^{0.21} \left( \frac{r}{10 \text{ kpc}} \right)^{0.5} \text{cm}^{-2},$$ (31)

which is everywhere much smaller than the Lyman limit column density $N_{\text{HI,LL}} = \sigma_{\text{ion}}(\nu_{\text{LL}})^{-1} = 10^{17.2} \text{ cm}^{-2}$, for any plausible value of $\xi$ and $T_4$. This confirms that our main assumption that the entire nebula is almost completely ionised and transparent to hydrogen ionizing radiation is self-consistent. The implications for Lyα emission, or that the majority of the emission comes from regions that are dense enough that the helium is only partially ionized despite being fully illuminated (e.g. Arrigoni Battaia et al. 2015). This in our case would imply a very low filling factor $f_V < 10^{-4}$, or additional internal clumpiness within the cold gas structures (see Section 7.3 below, see also Cantalupo et al. 2019). It is of course also possible that the explanation is a combination of those mentioned above and it may be different for different nebulae. We plan to explore these aspects in more detail in future work.

5.2 Helium

Similarly, the column density of HeII is

$$N_{\text{HeII}} = 1.3 \times 10^{16} \xi_{\text{He}}^{-1} T_4^{0.27} \left( \frac{r}{10 \text{ kpc}} \right)^{0.5} \text{cm}^{-2},$$ (32)

with

$$\xi_{\text{He}} = 4^{-(p_1 - 1.7)} \frac{p_1 + 3}{p_2 + 3} \xi,$$ (33)

where we used again atomic coefficients as in Meiksin (2000) and we further assumed that the filling factor $f_V > 10^{-4}$, while $p_1$ and $p_2$ are the slopes of the QSO ionizing continuum in the ranges $1 \text{ Ry} < E < 4 \text{ Ry}$ and $E > 4 \text{ Ry}$, respectively. If for simplicity we assume $p_2 = p_1 = 1.7$ (that is, if the spectrum by Lusso et al. 2015 can be extrapolated above 4 Ry), we see that the column density (32) is still safely below the limit for self-shielding $N_{\text{HeII,crit}} = 6.3 \times 10^{17} \text{ cm}^{-2}$, implying that the entire nebula should be transparent to the helium ionizing radiation as well.

Note, however, that Borisova et al. (2016) estimated very low HeII/Lyα line ratios, inconsistent with recombination radiation from fully ionized hydrogen and helium. Though the estimate above suffers from several approximations and extrapolations (most notably, the shape of the ionizing spectrum and the extrapolation of the modelled densities and filling factors to radii smaller than 10 kpc), we found that it is not easy to accommodate HeII column density that are high enough to guarantee self-shielding. If this is true, and if the very low limits on the line ratios are confirmed, this could suggest that either some of the Lyα emission is contributed by processes other than pure recombination (see Sections 5.3 and 6 below, see also Humphrey et al. 2019), or that the majority of the emission comes from regions that are dense enough that the helium is only partially ionized despite being fully illuminated (e.g. Arrigoni Battaia et al. 2015). This in our case would imply a very low filling factor $f_V < 10^{-4}$, or additional internal clumpiness within the cold gas structures (see Section 7.3 below, see also Cantalupo et al. 2019). It is of course also possible that the explanation is a combination of those mentioned above and it may be different for different nebulae. We plan to explore these aspects in more detail in future work.

5.3 Collisional excitation

We finally verify that the contribution of collisional excitation to the Lyα emission can be consistently neglected in our model.

The local Lyα emissivity due to collisional excitation and recombination are, respectively:

$$j_{\text{coll}} = \frac{E_{\text{Lyα}}}{4\pi} \alpha_{\text{coll}} n_{\text{HI}} n_e$$ (34)

and

$$j_{\text{rec}} = \frac{E_{\text{Lyα}}}{4\pi} \alpha_{\text{rec}} n_{\text{HI}} n_e$$ (35)

we omitted ‘Lyα’ super-scripts or sub-scripts $j$’s and $\alpha$’s for clarity) and their ratio is

$$\frac{j_{\text{coll}}}{j_{\text{rec}}} = \frac{\alpha_{\text{coll}}}{\alpha_{\text{rec}}} \frac{x_{\text{HI}}}{1 - x_{\text{HI}}},$$ (36)

where we take $x_{\text{HI}}$ from equation (28) and the ratio of recombination to collisional excitation coefficients as in Cantalupo, Porciani & Lilly (2008). The ratio $\alpha_{\text{coll}}/\alpha_{\text{rec}}$ increases rapidly with temperature (much more rapidly than the decrease with temperature of $x_{\text{HI}}$, see equation 28). To obtain a conservative upper limit to the contribution of collisional excitation we therefore consider a fairly high temperature $T = 5 \times 10^4 \text{ K}$ and also a fairly large distance from the centre $r = 100 \text{ kpc}$ (to maximize the neutral fraction). Note that choosing even higher temperatures would not help, as collisional ionization would start to dominate over photoionization driving an additional drop of the neutral fraction and therefore of the collisional excitation rate. In these conservative conditions we obtain an upper limit:

$$\frac{j_{\text{coll}}}{j_{\text{rec}}} < 1.4 \times 10^{-2} \xi^{-1} f_V^{-1/2}$$ (37)

so that we can safely ignore collisional excitation as a powering mechanism for filling factors larger than $10^{-4}$ (a limit which goes further down for smaller distances and temperatures).

6 THE IMPACT OF SCATTERING

Here we discuss the possible impact on our results of Lyα scattering. We distinguish between the addition of Lyα photons, which may be produced by the central QSO and scattered into the line of sight (Section 6.1), from the redistribution of Lyα photons due to scattering within the CGM (Sections 6.2 and 6.3).

6.1 Scattering from the broad line region

Even if the CGM is almost completely ionized, some of the very rare neutral hydrogen atoms can still intercept some of the Lyα photons coming from the broad line region (BLR) of the central QSO and scatter them in to the line of sight. These photons would contribute to the observed Lyα surface brightness of the nebulae, in addition to the recombination radiation we have focused on so far.

To quantify this effect, we must convert the neutral hydrogen column density as seen by the QSO (equation 31).
into a line center optical depth to Ly$\alpha$ scattering (following e.g. Meiksin 2009). We find

$$
\tau_{\text{Ly} \alpha, \text{QSO}} (r) = 0.99 \xi^{-1} T_4^{0.21} \sigma_{300}^{-1} \left(\frac{r}{10 \text{kpc}}\right)^{0.5},
$$

where $\sigma_{300}$ is the 1D velocity dispersion of the cold CGM, as seen by the QSO, in units of 300 km/s.

We have therefore found a model-dependent optical depth of order unity. This is the most difficult regime for analytic treatment, as both the single scattering and the diffusion limit are prone to be inadequate. One the other hand, a brute-force numerical approach would not give a clear-cut solution either, as the limiting factor here is not the calculation itself, but rather that the result is bound to be strongly dependent on the assumptions on the intrinsic kinematics of the cold gas, which is largely unknown.

The reference value for $\sigma$ that we used in equation (38) corresponds to the typical width of the Ly$\alpha$ line as observed in emission (Borisova et al. 2016). Note however that this can be a biased estimate of the true value of $\sigma$ for at least two reasons. First: if radiative transfer effects were indeed important (which is what we would like to establish here), the observed line width would likely be larger than the intrinsic one, with the result that the fiducial $\sigma$ above would be overestimated and the optical depth underestimated. Second, the velocity dispersion as seen by the QSO will be different from that along the line of sight unless the kinematics of the cold CGM is isotropic. For instance, if the cold CGM were dominated by coherent radial flows (such as inflows or outflows) the dispersion seen by the QSO (as determined by the acceleration or deceleration of the cold gas flows) would again be smaller than that along the line of sight (which would instead be dominated by projection effects). Both effects would go in the direction of favouring optical depths larger than unity. As we show below, a large optical depth would imply a small contribution of BLR photons to the observed emission and therefore that the model considered so far is self-consistent in this respect (although not necessarily unique).

Our ignorance on the intrinsic kinematics of the CGM prevents us from calculating the true contribution of scattered BLR photons to the observed emission. We can none the less easily put a very conservative upper limit to this contribution by calculating the maximum possible number of BLR Ly$\alpha$ scattered photons per fluorescent Ly$\alpha$ photon produced in a recombination event, based on fundamental atomic physics, as described below (we have also verified the validity of our reasoning with Cloudy simulations).

Recombination Ly$\alpha$ photons are produced, with some probability $\eta$, every time an electron recombines with a proton to form an atom of neutral hydrogen. Statistically, the electron will be freed again (and therefore become again eligible to recombination and Ly$\alpha$ emission) in a time equal to the inverse of the photo-ionization rate $\tau_{\text{ion}} = \Gamma_{\text{ion}}^{-1}$. During this time, the neutral atom is available to undergo a number of scattering events, statistically equal to $\Gamma_{\text{scatt}} \tau_{\text{ion}} = \Gamma_{\text{scatt}} / \Gamma_{\text{ion}}$, where $\Gamma_{\text{scatt}}$ is the scattering probability per unit time

$$
\Gamma_{\text{scatt}} = \int_0^{+\infty} \frac{F_\nu (\nu)}{\hbar \nu} \sigma_{\text{scatt}} (\nu) d\nu = \frac{F_{\nu, \text{Ly} \alpha}}{\hbar \nu_{\text{Ly} \alpha}} \bar{\sigma},
$$

where $\bar{\sigma} = \int_0^{+\infty} \sigma_{\text{scatt}} (\nu) d\nu = 1.1 \times 10^{-2}$ cm$^2$ Hz is the frequency-integrated cross-section to Ly$\alpha$ scattering (e.g. Rybicki & Lightman 1979; Meiksin 2009) and

$$
F_{\nu, \text{Ly} \alpha} = \frac{L_{\nu, \text{Ly} \alpha}}{4 \pi r^2} e^{-\tau_{\text{Ly} \alpha, \text{QSO}} (r)}
$$

is the specific flux density of Ly$\alpha$ photons propagating directly from the BLR to distance $r$ from the QSO. The cumulative Ly$\alpha$ optical depth $\tau_{\text{Ly} \alpha, \text{QSO}} (r)$ describes attenuation by scattering at smaller radii. Note that we do not need to introduce an attenuation factor for the Lyman limit photons, as the optical depth to photo-ionization is negligible irrespective of assumptions on kinematics at the column densities of interest (Section 5.1).

The local ratio of the scattering to recombination Ly$\alpha$ emissivity is given by:

$$
\frac{J_{\text{scatt}}}{J_{\text{rec}}} = \frac{\Gamma_{\text{scatt}}}{\eta \Gamma_{\text{ion}}} = \frac{p + 3}{\eta} \bar{\sigma} \frac{F_{\nu, \text{Ly} \alpha}}{\eta F_{\nu, \text{Ly} \alpha}} = 0.71 p + 3 \frac{F_{\nu, \text{Ly} \alpha}}{F_{\nu, \text{Ly} \alpha}},
$$

where the case $\Lambda$ Ly$\alpha$ production efficiency per recombination is well approximated by

$$
\eta (T) = 0.41 T_4^{-0.23}.
$$

It is clear from equations (40) and (41) that a very conservative upper limit for the ratio of scattering to recombination emissivity is found in the Ly$\alpha$ optically thin limit $\tau_{\text{Ly} \alpha, \text{QSO}} \ll 1$. Furthermore, in this limit, the ratio $J_{\text{scatt}} / J_{\text{rec}}$ is just a constant, determined by the intrinsic spectrum of the QSO. This is not surprising if one considers that, in an almost completely ionized gas, one single physical mechanism (recombination) is responsible both for Ly$\alpha$ fluorescence and for the creation of scattering targets (neutral atoms) and that in both cases the energy source is the same (the QSO). Assuming the typical QSO spectrum from Lusso et al. (2015) and a UV slope $p = 1.7$ we then find

$$
R = \frac{J_{\text{scatt}}}{J_{\text{rec}}} \leq R_{\text{max}} (T) = 26 T_4^{-23}.
$$

We stress that the value above is a very strict upper limit on the true ratio of the scattering-to-recombination emissivity, as it assumes that the whole CGM around the QSO sees the Ly$\alpha$ from the BLR entirely and unattenuated. Having clarified this, in such a limit one would have that the intrinsic recombination radiation surface brightness is $(1 + R_{\text{max}} (T))$ smaller than the observed one. As a consequence, all the densities and the masses discussed in the previous sections would need to be rescaled downwards by a factor $\sqrt{1 + R_{\text{max}} (T)}$. For instance, for a distance $r = 50$ kpc, a temperature $T_{\text{cold}} = 5.6 \times 10^4$ K and a volume filling factor $f_v = 10^{-3}$, the inferred density of the cold gas decreases from $n = 1.4$ cm$^{-3}$, in the recombination dominated case, to $n = 0.23$ cm$^{-3}$, assuming maximum scattering. The neutral hydrogen column density and optical depth should also be rescaled, this time by a factor $1 + R_{\text{max}} (T)$, due to the extra dependence of the neutral fraction of density (cfr. Section 5.1).

5 Obviously attenuation here is not synonym to absorption. The fate of intercepted BLR photons after first scattering is addressed in Section 6.2.
The minimum CGM baryon fraction assuming maximum possible contribution from scattering is given in Figure 8 as a function of the assumed halo mass $M_{\text{halo}}$ and temperature of the cold gas $T_{\text{cold}}$. In this case, the fiducial values $M_{\text{halo}} = 10^{12} M_\odot$ and $T_{\text{cold}} = 5.6 \times 10^4 K$, see Sections 4.2 and 4.3, correspond in this case to a minimum CGM baryon fraction of 70% of the cosmological value. Note that this model provides a very strict lower limit on the CGM baryon fraction, as it neglects any (possibly severe) attenuation of BLR photons while traversing the CGM (see text for details).

We end this section by briefly discussing the self-consistency and the plausibility of a scattering dominated scenario. As mentioned above, the neutral hydrogen column density in this case needs to be rescaled, with respect to a recombination dominated scenario, by a factor $1 + R_{\text{max}}$. Equations (31) and (38) therefore become

$$N_{\text{HI}}(r) = 2.7 \times 10^{13} \xi^{-1} \left( \frac{r}{10 \text{kpc}} \right)^{0.5} \text{ cm}^{-2},$$

and

$$\tau_{\text{Ly}\alpha, \text{QSO}}(r) = 3.7 \times 10^{-2} \xi^{-1} \sigma_{\text{HI}0} \left( \frac{r}{10 \text{kpc}} \right)^{0.5},$$

respectively, where we omitted for clarity a negligible explicit dependence on temperature. Equation (15) in particular implies that a Ly$\alpha$ optically thin scenario is self-consistent out to a distance $r = 100$ kpc for an intrinsic velocity dispersion as low as $\sigma = 35$ km s$^{-1}$ (assuming $\xi = 1$).

We have therefore found that a recombination-dominated and scattering-dominated scenario are both formally self-consistent. As, however, they have distinctly different implications for the CGM baryon fraction (see Figures 7 and 8), it would be desirable to find additional means to discriminate between these two options.

One possible way to verify the viability of the scattering-dominated scenario could be to look for proximate absorption signatures in the QSO Ly$\alpha$ spectra. We have estimated that, in order to extract their luminosities from the BLR of the central QSOs, the MUSE Giant Ly$\alpha$ nebulae should offer to the central source, on average, a rest-frame equivalent width

$$EW_0 = \frac{L_{\text{Ly}\alpha} / L_{\lambda, \text{QSO}}}{\frac{\xi}{4\pi}} = 0.3 \text{ Å},$$

where we have used the median observed luminosity of the nebulae $L_{\text{Ly}\alpha} = 10^{44} \text{ erg s}^{-1}$ (Borisova et al. 2010), a median QSO Ly$\alpha$ specific luminosity $L_{\lambda, \text{QSO}} = 3.3 \times 10^{44} \text{ erg s}^{-1} \text{ Å}^{-1}$ (again assuming the typical QSO spectrum of Lusso et al. 2015) and we finally assumed that the nebulae surround their central QSOs completely, so that $\Omega = 4\pi$. We have looked for associated absorption signatures in the spectra of the radio-quiet QSO of Borisova et al. (2016) within the wavelength range defined by the FWHM of the nebular spectra and we have found an average rest-frame equivalent width $EW_0 = 0.27$ Å, fairly close to the expected value. We emphasize however that this kind of measurements should be taken with great caution, for a number of important reasons, including i) the inherent difficulty of modelling absorption features on top of an emission line (here we simply modelled the ‘continuum’ with a local linear interpolation), ii) we cannot securely distinguish absorption features arising in the CGM from those associated to the BLR itself, the ISM of the host galaxy or even the IGM, so that the measurement should probably be interpreted as an upper limit and iii) the down-the-barrel sightline offers a very limited and possibly biased view, which may not be representative of the average. We speculate however that studies of associated absorbers on larger samples and with a more detailed modelling of kinematics and radiative transfer could possibly help to distinguish between scenarios.

Finally, we want to emphasize that there is at least one example of a Giant Ly$\alpha$ nebula (the Slug nebula, Cantalupo et al. 2014) where a scattering dominated scenario can be excluded at very high confidence. If applied to the Slug, a calculation similar to that in equation (46) leads to a very different result than for the MUSE nebulae. This is due to i) the relatively high luminosity of the Slug ($L_{\text{Ly}\alpha} = 2 \times 10^{44} \text{ erg s}^{-1}$, Cantalupo et al. 2014) and

6 Note that the illuminating QSO is instead very similar to those
and most importantly ii) to the fact that the Slug nebula is strongly asymmetric. Most of the luminosity arises from a region with size $l \approx 80$ kpc at a projected distance $R \approx 80$ kpc (Cantalupo et al. 2014). The solid angle subtended by the nebula as seen by the QSO can therefore be estimated as $Ω \approx 0.79$ sr, or $Ω/4π = 0.06$ (assuming that the true distance and projected distance coincide) or even smaller (if the true distance is larger than the projected one, as argued for instance in Cantalupo et al. 2019) though note that we have little knowledge of the structure of this nebula along the line of sight). As a consequence, the rest-frame equivalent width necessary to power the nebula with scattering is exceptionally high: $EW_0 = 9.6$ Å, or possibly even larger that that. Note that to be consistent with the optically thin assumption this would require a velocity dispersion in excess of 1000 km s$^{-1}$ (e.g. Meiksin 2009). We can therefore conclude that scattering cannot be the dominant powering mechanism of the Slug nebula, as also found by the radiative transfer simulations of Cantalupo et al. (2014). This is also in agreement with the recent detection, in the same object, of the non-resonant Hα line, with a Lyα/Hα ratio consistent with recombination radiation (Leibler et al. 2018). Future Hα observations, possible with the James Webb Space Telescope (JWST), will help to determine whether this is also the case for the MUSE Giant Lyα nebulae. Finally, polarization studies of Giant Lyα nebulae would also help constraining the contribution of scattering from a central source to the total observed emission (e.g. Hayes, Scarlata & Siana 2011).

### 6.2 Scattering within the nebula

Some of the Lyα photons, coming either from recombinations or from the first scattering of a BLR photon, can be scattered (or further scattered) within the nebula itself, thus affecting the surface brightness profile, while keeping the luminosity constant. A difference between the intrinsic and observed surface brightness profiles would imply a difference between the intrinsic density distribution of the cold gas and the one derived neglecting internal scattering (Section 2.2). Although the total luminosity is conserved in the process, it is not obvious that the inferred total mass (which is our main concern here) would remain unaffected as well. Here we show that this is indeed not the case, but that the required correction is small enough to have negligible impact to our conclusions.

Suppose that the ‘intrinsic’ surface brightness distribution due to recombination is different from the observed one:

$$\Sigma_{\text{oss}}(R) = \Sigma_{0,\text{oss}} \left( \frac{R}{R_0} \right)^{-\beta_{\text{oss}}} \tag{47}$$

$$\Sigma_{\text{int}}(R) = \Sigma_{0,\text{int}} \left( \frac{R}{R_0} \right)^{-\beta_{\text{int}}} \tag{48}$$

with $\Sigma_{0,\text{int}} \neq \Sigma_{0,\text{oss}}$ and $\beta_{\text{int}} \neq \beta_{\text{oss}}$. Then, using the notation of Section 2.2, the intrinsic density profile would be described by some intrinsic normalization and slope ($\eta_{0,\text{int}}, \gamma_{\text{int}}$) that are different from those that we derived from the observed profile, which we can call here ($\eta_{0,\text{oss}}, \gamma_{\text{oss}}$). We can integrate equations (17) and (18) to obtain the total luminosity and the luminosity due to recombinations, respectively:

$$L_{\text{tot}} = \frac{8\pi^2 \Sigma_{0,\text{oss}} R_0^2}{2 - \beta_{\text{oss}}} \left( x^{2 - \beta_{\text{oss}}} - 1 \right), \tag{49}$$

$$L_{\text{rec}} = \frac{8\pi^2 \Sigma_{0,\text{int}} R_0^2}{2 - \beta_{\text{int}}} \left( x^{2 - \beta_{\text{int}}} - 1 \right), \tag{50}$$

where

$$x \equiv \frac{R_{\text{max}}}{R_0}. \tag{51}$$

Note that in the integration above we use, as a lower limit, $R_0 = 10$ kpc, which is a reasonable estimate for the CGM/ISM transition, while $R_{\text{max}}$ is left free, but can be assumed to be of the order of the virial radius (equation [11]).

To derive the necessary corrections to the density profile, we then impose that the total luminosity be

$$L_{\text{tot}} = (1 + f_{\text{BLR}}) L_{\text{rec}}, \tag{52}$$

where $f_{\text{BLR}}$ is the global average of the ratio defined in equation (41) and will therefore vary in the range $0 < f_{\text{BLR}} < R_{\text{max}}$ (see equation [43]). Combining equations (49), (50) and (52) with equations (4) we find

$$\frac{n_{0,\text{int}}}{n_{0,\text{oss}}} = \frac{1}{\sqrt{1 + f_{\text{BLR}}} \left( \frac{\chi(1 + \beta_{\text{oss}})}{\chi(1 + \beta_{\text{int}})} \right)^{1/2}} \frac{\beta_{\text{int}}}{\beta_{\text{oss}}} \left( \frac{x^{2 - \beta_{\text{oss}}} - 1}{x^{2 - \beta_{\text{int}}} - 1} \right)^{1/2}. \tag{53}$$

The total cold gas mass is then found integrating over radius (as in Section 2.2), thence

$$\frac{M_{\text{int}}}{M_{\text{oss}}} = \frac{3 - \gamma_{\text{oss}}}{3 - \gamma_{\text{int}}} \frac{n_{0,\text{int}}}{n_{0,\text{oss}}} x^{\gamma_{\text{oss}} - \gamma_{\text{int}}}, \tag{54}$$

where the $\gamma$’s are as in equation (8).

The final correction factor on the cold gas mass, which then propagates to a correction to the total (cold plus hot) inferred mass of the CGM, is readily obtained by combining equations (53) and (54). The resulting expression (which we do not write explicitly as it is as long as trivial) contains a factor $(1 + f_{\text{BLR}})^{-1/2} \propto (1 + R_{\text{max}})^{-1/2}$, which is just the same that was discussed already in Section 6.1. The remaining part of the expression gives any additional bias in the determination of the CGM mass due to internal scattering within the nebula. This is visualized in Figure 9 as a function of the intrinsic slope $\beta_{\text{int}}$, for various values of the radial range parameter $x$, chosen to comprise the range of virial radii plausible for our haloes. Though the internal scattering of recombination radiation would only act in making the profile shallower, we consider here both the options $\beta_{\text{int}} < \beta_{\text{oss}}$ and $\beta_{\text{int}} > \beta_{\text{oss}}$, as the first scattering from the BLR can instead result in a centrally concentrated contribution, especially in the high optical depth regime (large $\gamma_{\text{Ly}α,\text{QSO}}$, which however corresponds to small $R$ and $f_{\text{BLR}}$, as discussed in Section 6.1). It is clear from the figure that the correction is of the order of a few per cent at most and therefore has no significant impact on our results.
and therefore to weaken our conclusions. The idea can be  
alyzed indirectly by ionizing photons produced in recombi-
 the volume ionized by the QSO directly and the volume ion-
 high the observed Lyα by the intervening screen and therefore not contribute to
 the whole ‘far’ side of the nebula would be largely obscured
 ation enhancement would be counter-acted by the fact that
 normalization bias, as light is being added along some lines
 6.2 – but, in addition to that, it could also introduce a global
 ness profile – an effect which we addressed already in Section
 less ionized than others and offer higher neutral hydrogen
 from the line of sight, so that some regions of the CGM are
 if the central QSO is obscured along some directions away
 the line-of sight after interacting with a layer or region of
 material with a relatively high neutral fraction and neut-
 the line-of sight (see text for details). Corrections amounts to unity (by construction) for
 βintr = βos = 1.5. The bias due to internal scattering is in all cases smaller than 10% and has therefore negligible impact on our conclusions.

6.3 Scattering off a neutral screen

An additional redistribution effect, which can be partly treated with the same formalism of Section 6.2, would also arise if some of the Lyα photons originated in the CGM and travelling away from the observer are scattered back into the line-of sight after interacting with a layer or region of material with a relatively high neutral fraction and neutral hydrogen column density. This can happen, for instance, if the central QSO is obscured along some directions away from the line of sight, so that some regions of the CGM are less ionized than others and offer higher neutral hydrogen column density, as suggested, for instance, by absorption studies of the CGM ([Prochaska et al., 2013] around QSOs of similar redshift as (although about 1 dex less luminous than) those of [Borisova et al., 2016].

This process can change the shape of the surface brightness profile – an effect which we addressed already in Section 6.2 but, in addition to that, it could also introduce a global normalization bias, as light is being added along some lines of sight. At the same time, however, this potential normalization enhancement would be counter-acted by the fact that the whole ‘far’ side of the nebula would be largely obscured by the intervening screen and therefore not contribute to the observed Lyα flux. Insight suggests that the latter effect should in general be stronger that the former, so that the combination of the two is unlikely to bias the total flux high and therefore to weaken our conclusions. The idea can be intuitively illustrated as follows.

Suppose that the QSO is able to ionize only a fraction $f_{\text{ion}} < 1$ of the entire solid angle. We include in this both the volume ionized by the QSO directly and the volume ionized indirectly by ionizing photons produced in recombination events (e.g. [Osterbrock & Ferland, 2006]). Then only a fraction $0.5f_{\text{ion}}$ of the CGM volume would be bright in recombination radiation (or capable to directly see and scatter photons from the BLR) and also be directly visible by the observer. At the same time, some fraction $f_{\text{screen}}$ of the photons produced in this region would intercept the screen, where they can be scattered back (and possibly reach the observer) with some probability $f_{\text{back}}$, or find their way through the screen with a probability $1 - f_{\text{back}}$. Also, with a probability $f_{\text{screen}}(1 - f_{\text{back}})$, photons from the ‘far’ side of the CGM would intercept and traverse the screen and also contribute to the observed flux. The total observed Lyα flux would be different from that of an ideal case with perfect illu-
mination, by a factor

$$\frac{F_{\text{os}}}{F_{\text{ideal}}} = \frac{1}{2} f_{\text{on}}(1 + f_{\text{screen}}) + \frac{1}{2} f_{\text{on}} f_{\text{screen}}(1 - f_{\text{back}})$$

$$= \frac{1}{2} f_{\text{on}}(1 + f_{\text{screen}}) \leq 1 \tag{55}$$

Although highly idealized, the inequality above suggests that, if anything, the presence of a neutral screen is more likely to introduce a bias in the direction of strengthening our results, rather than vice versa. We plan to demonstrate this point more thoroughly with radiative transfer simulations in future work.

7 DISCUSSION

We discuss here the following inter-related questions: i) What is the typical size and the number of the cold gas structures which produce the emission in MUSE Giant Lyα nebulae? ii) What is the origin of the cold gas? iii) How well justified is our assumption of pressure equilibrium with the hot gas? When relevant, we also try to connect our discussion to resolution or other numerical effects which may be important for the interpretation of cosmological hydrodynamical simulations of the CGM.

7.1 The typical size of the cold gas structures

So far we have not addressed the question of what is the typical size and the number of the cold gas structures in MUSE Giant Lyα nebulae. This is because the main parameter in our calculations is the volume filling factor $f_v$ and any given volume filling factor can be obtained either with a large number of small clouds or with a smaller number of larger ones. One can however attempt to break the degeneracy, at least qualitatively, by considering the area covering factor $f_c$, which is more easily accessible to observations than $f_v$ and is given by:

$$f_c = \int f_c \frac{dV}{R_c}, \tag{56}$$

where the integral is along the line of sight and $R_c$ is the typical size of the individual cold gas structures along the same direction. Assuming that the intrinsic orientation of the structures is independent on our line of sight, we can take $R_c$ as the typical physical scale of the emitting structures, averaged over all directions; we stress that $R_c$ does not necessarily represent the radius of spherical clouds, as we did not make any assumptions on the geometry and not even on the topology of the emitting structures.
As an order of magnitude estimate, equation (56) provides

$$R_c \approx f_v r_{\ddot{v}} f_C^{-1} \lesssim 74 \left( \frac{f_v}{10^{-3}} \right) \text{pc},$$

(57)

where we assumed that the typical length of the line of sight through the CGM is of the order of the virial radius (equation (11)) and we also assumed $f_C \gtrsim 1$. While do not know what the true covering factor of the cold CGM is (due to the seeing-limited spatial resolution of MUSE), the continuity and smoothness of the emission suggests that it is not much smaller than unity. Arrigoni Battaia et al. (2015) has shown with synthetic observations that, even accounting for finite resolution, the observed morphology of extended Lyα emission requires an intrinsic covering factor $f_C > 0.5$. Large covering factors are also supported by absorption studies, which are more sparse but have virtually infinite angular resolution (e.g. Prochaska et al. 2013). On the other hand, $f_C > 1$ is possible if multiple structures overlap in projection along the line of sight, which explains why equation (57) is best interpreted as an upper limit. We stress that the one above is just an order of magnitude estimate and it is only useful as a guidance to discuss which scenarios for the nature and origin of the cold gas should be considered more plausible than others.

7.2 Most of the cold gas is not in gravitationally bound substructures

Using the first equality in (57) and the definition of volume filling factor, one can also estimate (again by order of magnitude) what number of cold structures with size $R_c$ can be found within the virial radius:

$$N \approx f_v^2 f_C^3 \gtrsim 10^4,$$

(58)

assuming $f_v \lesssim 10^{-2}$ and $f_C \gtrsim 1$. This simple estimate favours a scenario where that the cold gas structures are individually associated to dark matter sub-haloes (and therefore suggests that the Lyα emission cannot originate, for instance, from ISM bound to undetected satellite galaxies), as the latter are only a few hundreds at most (e.g. Moore et al. 1999, Rodríguez-Puebla et al. 2016). We recall that larger values of the filling factor $f_v$ would require an implausibly large mass of cold gas to explain the observed surface brightness (see Section 2.3); note also that this conclusion is independent on assumptions about pressure confinement.

The size estimate (57) also helps determining whether the cold gas structures are collapsed or on the verge of collapsing under their own self-gravity. This is an intriguing option, as it could imply that some of the Lyα emitting structures could be related to precursors of globular clusters. To this aim, we should compare $R_c$ to the Jeans length

$$\lambda_J = \frac{\sqrt{\pi^2 m}}{G \rho} = 16 T_4^{0.02} f_v^{1/4} \text{pc},$$

(59)

where again $T_4 = T/10^4$ K, $c_s$ is the sound speed of an ionized plasma at temperature $T$ and we used equation (7), evaluated at a reference radius of 50 kpc. The negligible dependence on temperature originates from a combination of the temperature dependence of the sound speed and the effective Lyα recombination coefficient. Combining equations 67 and (59) gives the ratio

$$r_J = \frac{R_c}{\lambda_J} \lesssim 4.6 T_4^{-0.02} f_v^{3/4}.$$  

(60)

We see that the condition $r_J < 1$ for stability is safely met for any filling factor of interest. We conclude that the cold gas structures are most likely stable against gravitational collapse. In the calculation above, we used our density estimate based on a pure recombination scenario (Section 2), but note that our conclusion would only be reinforced if scattering is non-negligible, as the estimated densities decrease in that case (Section 6.1), further enhancing stability against gravitational collapse.

7.3 Can the cold gas be over-pressured?

As the gravitational pull (by either dark matter sub-haloes or the gas itself) can be neglected (see Section 7.2), we are confident that the cold gas structures are most likely confined by the pressure of the surrounding medium, which justifies the main assumption of our Section 3. We also remind that such external pressure can include non-thermal contributions, such as those due to large-scale turbulence, and that the precise amount of thermal and non-thermal pressure has little impact on our results. With this specification in mind, any residual pressure imbalance should be smeared out roughly on a sound crossing time-scale:

$$t_{\text{sound}} = \frac{R_c}{c_s} = 2.0 \left( \frac{T_{\text{cold}}}{5.6 \times 10^4 \text{K}} \right)^{-1/2} \left( \frac{R_c}{74 \text{pc}} \right) \text{Myr},$$

(61)

where we scaled the result to the reference temperature and size estimated by means of equation (22) and (57), respectively.

One possibility to explain large Lyα surface brightness without resorting to large baryon fractions could therefore be that the majority of the cold gas structures were either created, or they acquired their current pressure, less than $\sim 2$ Myr ago. There is one simple way to achieve this. If the QSO lifetime is significantly smaller than $\sim 2$ Myr, the cold gas that we see bright in Lyα should be modelled as having been suddenly heated to the equilibrium temperature (see Section 4.3) and still find itself temporarily over-pressured with respect to the ambient medium, by an amount $T_{\text{cold}}/T_{\text{in}}$, where $T_{\text{in}}$ is its temperature before being illuminated by the QSO. Quantitatively, this sudden photo-heating scenario is equivalent to rescaling the factor $\varepsilon$ in equation (16) by $T_{\text{in}}/T_{\text{cold}}$. As a result, equation (23) for the minimum CGM baryon fraction reads in this case

$$f_{\text{min}} = 2 f_v \varepsilon_{1/2}^{1/2} T_{\text{in},4}^{3/2} T_4^{1/2} M_{h,12}^{-(1+\gamma)/3},$$

(62)

with $T_{\text{in,4}} \equiv T_{\text{in}}/10^4$ K and the other quantities as in Section 4.3. The minimum CGM baryon fraction for this model is shown in Figure 10 assuming an initial cold gas temperature $T_{\text{in}} = 10^4$ K and recombination dominated emission. Comparison to Figure 7 shows that sudden photo-heating facilitates the achievement of a physically acceptable solution, without requiring uncomfortable assumptions on the QSO spectrum or metallicity (Section 4.3), although relatively large halo masses $M_{\text{halo}} \gtrsim 10^{12.4}$ $M_\odot$ are still required. Short-lived overpressures could also originate, for instance, if some of the cold gas is ballistically ejected from
the ISM of a galaxy, without significantly interacting with the surrounding medium before reaching a considerable distance within the CGM. A discussion of the conditions under which this can happen would go beyond the scope of this paper. We note, however, that schemes corresponding to this description are sometimes implemented in some recipes for feedback in cosmological hydrodynamical simulations (e.g. Vogelsberger et al. 2013 and references therein). We therefore expect some amount of over-pressured gas to be found in such simulations, but we warn that it can be, at least to some extent, an unphysical consequence of the adopted hydrodynamic/feedback scheme and should therefore be considered with caution.

One should finally consider the option that the cold gas structures are *internally* turbulent. Super-sonic turbulence would induce a lognormal density distribution *within* the cold gas structures (e.g. Molina et al. 2012 and references therein), leading to an additional internal clumping factor (see Section 2.1.2), which we did not account for in our calculations, though there are some recent observational indications pointing in this direction (Cantalupo et al. 2019). Significant internal clumpiness would have an impact on the estimated CGM baryon fraction of the cold gas. For instance, following Nolan, Federrath & Sutherland (2015), internal supersonic turbulence with a Mach number $M = 2.5$ would induce in the cold gas a lognormal density distribution with dispersion $\sigma_{\text{ln}} = 0.7$. This would imply an internal clumping factor $C_{\text{int}} = e^{\sigma_{\text{ln}}^2} = 2$ and drive our mass estimates of the cold gas down by a factor 2. Two potential problems are worth being pointed out concerning this scenario. First, internal turbulence would imply a total (thermal plus turbulent) *internal* pressure which is about $(1 + \Gamma M^2)$ larger than the thermal component alone (with $\Gamma = 5/3$ for a monoatomic gas). This would require an equal enhancement of the *external* confining pressure and imply an increase of the estimated mass of the hot gas (e.g. by a factor 11 for $M = 2.5$), more than compensating the decrease in the estimated mass of the cold gas. The last conclusion does not hold, of course, if the phenomenon is short-lived and set to be smeared out on a turbulent crossing time. Note that for supersonic turbulence the turbulent crossing time would be even smaller than the sound crossing time in equation (61), by a factor $\sim M$, so that the emitting structures should have become turbulent, in this scenario, roughly 1 Myr ago. Also note, in contrast to the ‘sudden photo-heating’ scenario discussed above, that it is not clear why supersonic turbulence should have been set on so recently. Second, supersonic turbulence would quickly heat the cold gas, through repeated shocks, to high temperatures, where the Lyα emissivity drops, further lowering the survivability of this situation, by an amount which we did not try to quantify. We cannot exclude that a large internal clumpiness can be due to processes other than turbulence. We are not aware, however, of plausible alternatives. We recall that self-gravity (which would in principle be an appealing candidate) can be neglected here, as we have argued already in Section 7.2.

### 7.4 Progressive fragmentation and incomplete pressurisation

Though persistent over-pressures are challenging to explain, an under-pressured cold CGM can instead be achieved under somewhat easier circumstances, which we briefly discuss below.

Cold gas accreting onto a halo from IGM filaments can find itself, at the moment of entering the halo, at a lower pressure with respect to the hot gas already present there (see e.g. the phase diagrams in Nelson et al. 2016). The hot gas will then compress the incoming gas to higher densities, but the process is not necessarily well approximated as being instantaneous. It is therefore important to establish whether the incoming gas has the time to be pressurized before penetrating deep into the halo. One way to do this is to compare the sound-crossing time of a filament

$$t_{\text{sound, fil}} = \frac{R_{\text{fil}}}{c_s} = 640 \, T_{\text{fil}}^{1/2} \left( \frac{R_{\text{fil}}}{10 \, \text{kpc}} \right) \, \text{Myr}$$

and its halo crossing time

$$t_{\text{cross, fil}} = r_{\text{vir}} \left( \frac{r_{\text{vir}}}{V_{\text{vir}}} \right) \left( \frac{V_{\text{vir}}}{100 \, \text{km s}^{-1}} \right)^{-1} \, \text{Myr}$$

(64)

It is readily seen that sufficiently thick or fast filaments could in principle be able to travel significant distances within the CGM before being completely pressurized by the hot gas. Part of this gas would therefore be under-pressured and emit less Lyα than otherwise.

It is possible and even likely, however, that pressurisation proceeds much faster then the simple estimate above. Unless perfect symmetry conditions are met, in fact, an under-pressured structure would be progressively deformed, during compression, and possibly fragment in smaller units. Fragments (or highly deformed portions of the original structure) will have a smaller sound-crossing time and therefore be prone to even faster compression and further fragmentation. The process has a run-away nature and results in very rapid progressive fragmentation in structures with smaller and smaller size, until complete pressurisation is reached, or some other process (e.g. thermal conduction) takes over,
either halting further fragmentation or destroying the fragments altogether.\footnote{This process should not be confused with the interaction between the hot gas and an inflowing filament assumed to be in pressure equilibrium since the beginning and which can then be disrupted by the Kelvin-Helmholtz instability; this can result in moderate local deviations from pressure equilibrium (e.g. Mandelker et al. 2016; Padnos et al. 2018; Vossberg et al. in prep.)}

It is also possible that some of the cold gas originates from the condensation of the hot gas itself. Progressive fragmentation is indeed usually described in the context of thermal instability, where radiative cooling is the means to create local underpressures and initiate fragmentation. The final size of the cold gas structures in this case is roughly given by the acoustic length for which the cooling time equals the sound crossing time (e.g. Inoue & Omukai 2015, McCourt et al. 2018) recently described the phenomenon under the name of shattering and clarified that fragmentation would stop at the cooling length (the minimum acoustic length achieved during the cooling process), unless this is smaller than the Field length, below which thermal conduction would lead the cold gas to evaporate. We also notice that, if the radial profiles of heating and cooling meet certain conditions (specifically, if the heating rate declines with radius more steeply than the cooling rate), some cold gas can condensate in a shell of finite radius, possibly meeting the requirements for the onset of the Rayleigh-Taylor instability and eventually lead to the creation of filamentary structures within the CGM.\footnote{Thanks to A. Negri for bringing this option to our attention.}

Crucially, high resolution is required for the processes mentioned above to be captured in a hydrodynamical simulation. The acoustic length, in particular, can be as small as a fraction of a pc (McCourt et al. 2018). It is therefore possible that under-pressured gas may be present in some cosmological simulations, as a consequence of non-negligible nominal compression timescales (equation 63); but it is also possible that some of these features may partially be artefacts, due to the numerical resolution being insufficient to follow the shattering process and the resultant accelerated compression. Note that this last consideration should mostly apply to grid codes. SPH codes could suffer from different issues, including on one side spurious clumpiness (e.g. Torrey et al. 2012) and on the other side artificial under-pressures associated to the non-trivial nature of density evaluations in SPH codes (see e.g. Appendix D in Oppenheimer et al. 2018). We conclude by emphasizing that a systematically under-pressured cold CGM, if real, would imply that the hot halo is even more massive than what inferred based on pressure balance, and would therefore go in the direction of strengthening our results.

8 SUMMARY AND CONCLUSIONS

Galaxy formation models rely on mechanisms to keep most of the baryons outside of galaxies, but different scenarios make different predictions about how far away from galaxies the gas is expected to be found. Measuring what fraction of baryons is retained within the virial radius of haloes – especially at high redshift, when most of the relevant processes are expected to take place – is therefore crucial to inform and constrain theoretical models and especially different flavours of feedback.

In this work, we have tried to determine the total baryonic mass stored in the CGM of relatively massive \((M_{\text{halo}} \sim 10^{12} \, M_\odot)\) haloes at \(z \sim 3\), based on the observed surface brightness profile of Giant Ly\(\alpha\) nebulae discovered with MUSE around luminous QSOs at this redshift (Borissova et al. 2016). To this aim, we modelled the CGM as a multiphase medium, where a hot (collisionally ionized) plasma confines, due to its pressure, a colder (photo-ionized) phase into a small fraction of the volume and to large densities, with correspondingly large Ly\(\alpha\) emissivity. We have found that a CGM dominated in mass by any of the two phases would fall short in reproducing the observed surface brightness, either because the cold emitting gas is not enough, or because it has too low density (and therefore too low emissivity) due to insufficient compression from the ambient medium. On the other hand, by assuming that the two phases provide similar contributions to the total mass (i.e. by maximising the theoretical radiative output of the model), one can put a lower limit to the total mass of the CGM required to explain the observed surface brightness. Our CGM mass estimates are lower limits also because we have neglected a number of effects (dust extinction, IGM attenuation, partial illumination due to a narrow QSO ionization cone), which would diminish the predicted Ly\(\alpha\) luminosity and therefore increase the amount of gas needed to explain the observations.

We have thereby found that a large total (hot plus cold) baryonic mass of the CGM – close to the total amount of baryons nominally associated to these haloes – is required to give account for a Ly\(\alpha\) surface brightness as high as observed. The physical constraint that the baryon fraction be equal or smaller than the cosmological value requires at least one of the following four conditions to be met:

(i) The mass of QSO-hosting dark matter haloes at \(z \sim 3\) is larger than the current best estimates based on clustering \((M_{\text{halo}} \sim 10^{12} \, M_\odot)\), Section 4.2.

(ii) The equilibrium temperature of the photo-ionized gas is significantly smaller than what expected in a ‘standard’ QSO ionizing spectrum \((T_{\text{eq}} \sim 5 \times 10^4 \, \text{K}, \text{Section 4.3})\).

(iii) The intrinsic radial velocity dispersion of the CGM is large enough that the photo-ionized gas is optically thin to Ly\(\alpha\) scattering (Section 6.1).

(iv) Most of the cold gas is in a short-lived over-pressured phase, which we observe before it had the time to expand in the surrounding medium (Section 7.3).

All these options can potentially be constrained by independent observations. Particular help would come from further constraints on clustering and spectral properties of high-\(z\) luminous QSOs, as well as from follow-up studies on CGM non-resonant emission lines and/or polarimetry of the Ly\(\alpha\) emission. The last possibility (short-lived overpressures) is the most difficult to test, but we estimated that it could be fulfilled if the majority of bright QSOs at \(z = 3\) have a life-time shorter than \(\sim 2\) Myr.

At least two of the conditions above must be met, if the observed surface brightness is to be reconciled with relatively small CGM baryon fractions (30 – 40% of the cosmic
average) predicted at the relevant halo mass and redshift by the extremely strong ejective feedback sometimes adopted in galaxy formation models. More comparisons with models and simulations would be desirable to derive more robust conclusions. Regarding the comparison of QSO Giant Lyα nebulae with simulations, we stress in particular the importance of explicitly including photo-electric heating due to photo-ionization of HeII, as this significantly impacts the temperature, recombination coefficient and ultimately the emissivity of the gas.

In conclusion, our results seem to indicate a preference for a relatively high baryon fraction (close to cosmological) predicted at the relevant halo mass and redshift by the extremely strong ejective feedback sometimes adopted in galaxy formation models, such as the efficiency with which SN feedback can expel baryons from galactic haloes in the early Universe.

ACKNOWLEDGEMENTS

We thank Arif Babul, Andrea Ferrara, Max Gronke, Andrea Negri and Benny Trakhtenbrot for useful discussion, Raffaella Anna Marino for help and support and Peter Mitchell for sharing some of his analysis of the EAGLE simulations. This work was supported by the Swiss National Science Foundation grant PP00P2_163824.

References

Allevato V. et al., 2016, ApJ, 832, 70
Arrigoni Battaia F., Hennawi J. F., Prochaska J. X., Cantalupo S., 2015, ApJ, 809, 163
Arrigoni Battaia F., Hennawi J. F., Prochaska J. X., Oñorbe J., Farina E. P., Cantalupo S., Lusso E., 2019, MNRAS, 482, 3162
Bacon R. et al., 2010, in Proc. SPIE, Vol. 7735, Ground-based and Airborne Instrumentation for Astronomy III, p. 773508
Beckmann R. S. et al., 2017, MNRAS, 472, 949
Behroozi P. S., Marchesini D., Wechsler R. H., Muzzin A., Papovich C., Stefanon M., 2013, ApJL, 777, L10
Binney J., 2004, MNRAS, 347, 1093
Birnboim Y., Dekel A., 2003, MNRAS, 345, 349
Bogdán Á., Bourdin H., Forman W. R., Kraft R. P., Vogelsberger M., Hernquist L., Springel V., 2017, ApJ, 850, 98
Borisova E. et al., 2016, ApJ, 831, 39
Brighenti F., Mathews W. G., 2006, ApJ, 643, 120
Cantalupo S., 2010, MNRAS, 403, L16
Cantalupo S., Arrigoni-Battaia F., Prochaska J. X., Hennawi J. F., Madau P., 2014, Nature, 506, 63
Cantalupo S. et al., 2019, MNRAS, 483, 5188
Cantalupo S., Porciani C., Lilly S. J., 2008, ApJ, 672, 48
Chiu I. et al., 2018, MNRAS
Choi E., Ostriker J. P., Naab T., Oser L., Moster B. P., 2015, MNRAS, 449, 4105
Churazov E., Brüggen M., Kaiser C. R., Böhringer H., Forman W., 2001, ApJ, 554, 261
Ciotti L., Ostriker J. P., 2001, ApJ, 551, 131
Conroy C., White M., 2013, ApJ, 762, 70
Correa C. A., Schaye J., van de Voort F., Duffy A. R., Wyithe J. S. B., 2018, MNRAS, 478, 255
Dalla Vecchia C., Schaye J., 2012, MNRAS, 426, 140
Davé R., 2009, in Astronomical Society of the Pacific Conference Series, Vol. 419, Galaxy Evolution: Emerging Insights and Future Challenges, Jogee S., Marinova I., Hao L., Blanc G. A., eds., p. 347
Dawson K. S. et al., 2013, AJ, 145, 10
Dekel A., Silk J., 1986, ApJ, 303, 39
Dijkstra M., Loeb A., 2009, MNRAS, 400, 1109
Dijkstra M., Westra E., 2010, MNRAS, 401, 2343
Done C., Davis S. W., Jin C., Blaes O., Ward M., 2012, MNRAS, 420, 1848
Dutton A. A., Macciò A. V., 2014, MNRAS, 441, 3359
Eckert D. et al., 2016, A&A, 592, A12
Efstathiou G., 1992, MNRAS, 256, 43P
Efekekarzadeh S. et al., 2015, MNRAS, 453, 2779
Ferland G. J. et al., 2013, RMxAA, 49, 137
Gatto A. et al., 2017, MNRAS, 466, 1903
Gnat O., 2017, ApJS, 228, 11
Gronke M., Bird S., 2017, ApJ, 835, 207
Haardt F., Madau P., 2012, ApJ, 746, 125
Haiman Z., Ciotti L., Ostriker J. P., 2004, ApJ, 606, 763
Haiman Z., Rees M. J., 2001, ApJ, 556, 87
Hayes M., Scarlata C., Siana B., 2011, Nature, 476, 304
He W. et al., 2018, PASJ, 70, S33
Hennawi J. F., Prochaska J. X., Cantalupo S., Arrigoni-Battaia F., 2015, Science, 348, 779
Henriques B., White S., Lilly S., Bell E., Bluck A., Terrazas B., 2018, ArXiv e-prints
Henriques B. M. B., White S. D. M., Thomas P. A., Angulo R. E., Guo Q., Lemson G., Springel V., 2013, MNRAS, 431, 3373
Hopkins P. F., Lidz A., Hernquist L., Coil A. L., Myers A. D., Cox T. J., Spergel D. N., 2007, ApJ, 662, 110
Humphrey A., Villar-Martín M., Binette L., Raj R., 2019, A&A, 621, A10
Ichikawa K. et al., 2019, ApJ, 870, 31
Inoue T., Omukai K., 2015, ApJ, 805, 73
Jacob S., Pakmor R., Simpson C. M., Springel V., Pfrommer C., 2018, MNRAS, 475, 570
Kannan R., Vogelsberger M., Stinson G. S., Hennawi J. F., Marinacci F., Springel V., Macciò A. V., 2016, MNRAS, 458, 2516
Kereš D., Katz N., Weinberg D. H., Davé R., 2005, MNRAS, 363, 2
Kimm T., Haehnelt M., Blaizot J., Katz H., Michel-Dansac L., Garel T., Rosdahl J., Teyssier R., 2018, MNRAS, 475, 4617
Kulier A., Padilla N., Schaye J., Crain R. A., Schaller M., Bower R. G., Theuns T., Paillas E., 2018, ArXiv e-prints
Lagana A. T., Martinet N., Durret F., Lima Neto G. B., Maughan B., Zhang Y.-Y., 2013, A&A, 555, A66
Laor A., Davis S. W., 2011, MNRAS, 417, 681
Le Brun A. M. C., McCarthy I. G., Schaye J., Poron T. J., 2014, MNRAS, 441, 1270
Leibler C. N., Cantalupo S., Holden B. P., Madau P., 2018, MNRAS, 480, 2094
Li J.-T., Bregman J. N., Wang Q. D., Crain R. A., Anderson M. E., 2018, ApJL, 855, L24

© 201X RAS, MNRAS 000.
