



**HAL**  
open science

## Predicting LyC emission of galaxies using their physical and Ly $\alpha$ emission properties

Moupiya Maji, Anne Verhamme, Joakim A Rosdahl, Thibault Garel, Jérémy Blaizot, Valentin Mauerhofer, Marta Pittavino, Maria-Pia Victoria Feser, Mathieu Chuniaud, Taysun Kimm, et al.

### ► To cite this version:

Moupiya Maji, Anne Verhamme, Joakim A Rosdahl, Thibault Garel, Jérémy Blaizot, et al.. Predicting LyC emission of galaxies using their physical and Ly $\alpha$  emission properties. *Astronomy and Astrophysics - A&A*, 2022, 663, pp.A66. 10.1051/0004-6361/202142740 . insu-03777343v1

**HAL Id: insu-03777343**

**<https://hal.science/insu-03777343v1>**

Submitted on 17 Oct 2022 (v1), last revised 15 Sep 2022 (v2)

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Predicting LyC emission of galaxies using their physical and Ly $\alpha$ emission properties

Moupiya Maji<sup>1</sup>, Anne Verhamme<sup>1</sup>, Joakim Rosdahl<sup>2</sup>, Thibault Garel<sup>1</sup>, Jérémy Blaizot<sup>2</sup>, Valentin Mauerhofer<sup>1</sup>, Marta Pittavino<sup>3</sup>, Maria-Pia Victoria Feser<sup>3</sup>, Mathieu Chuniaud<sup>2</sup>, Taysun Kimm<sup>4</sup>, Harley Katz<sup>5</sup>, and Martin Haehnelt<sup>6</sup>

<sup>1</sup> Observatoire de Genève, Université de Genève, Chemin Pegasi 51, 1290 Versoix, Switzerland

<sup>2</sup> Univ Lyon, Univ Lyon1, Ens de Lyon, CNRS, Centre de Recherche Astrophysique de Lyon, Saint-Genis-Laval, France

<sup>3</sup> Research Center for Statistics, Université de Genève, 24 rue du Général-Dufour, 1211 Genève 4, Switzerland

<sup>4</sup> Yonsei University, 625 Science Hall, 50 Yonsei-ro, Seodaemun-gu, Seoul, 03722, South Korea

<sup>5</sup> University of Oxford, Clarendon Laboratory, Parks Road, Oxford, UK

<sup>6</sup> University of Cambridge, Madingley Road, Cambridge, UK

April 7, 2022

## ABSTRACT

**Aims.** The primary difficulty in understanding the sources and processes that powered cosmic reionization is that it is not possible to directly probe the ionizing Lyman Continuum (LyC) radiation at that epoch as those photons have been absorbed by the intervening neutral hydrogen. It is therefore imperative to build a model to accurately predict LyC emission using other properties of galaxies in the reionization era.

**Methods.** In recent years, studies have shown that the LyC emission from galaxies may be correlated to their Lyman-alpha (Ly $\alpha$ ) emission. In this paper we study this correlation by analyzing thousands of simulated galaxies at high redshift in the SPHINX cosmological simulation. We post-process these galaxies with the Ly $\alpha$  radiative transfer code RASCAS and analyze the Ly $\alpha$ -LyC connection.

**Results.** We find that the Ly $\alpha$  and LyC luminosities are strongly correlated with each other, although with dispersion. There is a positive correlation between the escape fractions of Ly $\alpha$  and LyC radiations in the brightest Ly $\alpha$  emitters (escaping Ly $\alpha$  luminosity  $L_{\text{esc}}^{\text{Ly}\alpha} > 10^{41}$  erg/s), similar to that reported by recent observational studies. However, when we also include fainter Ly $\alpha$  emitters (LAEs), the correlation disappears, which suggests that the observed relation may be driven by selection effects. We also find that the brighter LAEs are dominant contributors to reionization, with  $L_{\text{esc}}^{\text{Ly}\alpha} > 10^{40}$  erg/s galaxies accounting for  $> 90\%$  of the total amount of LyC radiation escaping into the inter-galactic medium in the simulation. Finally, we build predictive models using multivariate linear regression where we use the physical and the Ly $\alpha$  properties of simulated reionization era galaxies to predict their LyC emission. We build a set of models using different sets of galaxy properties as input parameters and predict their intrinsic and escaping LyC luminosity with a high degree of accuracy (adjusted  $R^2$  of these predictions in our fiducial model are 0.89 and 0.85 respectively, where  $R^2$  is a measure of how much of the response variance is explained by the model). We find that the most important galaxy properties to predict the escaping LyC luminosity of a galaxy are its  $L_{\text{esc}}^{\text{Ly}\alpha}$ , gas mass, gas metallicity and star formation rate.

**Conclusions.** These results and the predictive models can be useful to predict the LyC emission from galaxies using their physical and Ly $\alpha$  properties and thus help us identify the sources of reionization.

## 1. Introduction

Cosmic reionization is an important period in the evolution of the Universe, when photons from energetic sources (i.e. first stars, galaxies or quasars) ionized the ubiquitous neutral hydrogen gas in the intergalactic medium. This milestone happened over the first billion years of the Universe, ending around  $z \sim 6$ , and it holds a key for understanding the formation and evolution of the first galaxies (Loeb & Barkana 2001; Stark 2016; Ocvirk et al. 2016; Rosdahl et al. 2018; Wise 2019). However, the Epoch of Reionization (EoR) is yet to be fully understood. One of the biggest outstanding question is determining the primary sources of the photons that ionize the Universe. The relative importance of two types of sources that are proposed in the literature, i.e., star-forming galaxies and quasars, is still somewhat debated but recent studies indicate that quasars were likely too rare at these redshifts to reionize the Universe (Cowie et al. 2009; Fontanot et al. 2012, 2014; Kulkarni et al. 2019; Faucher-Giguère 2020; Trebitsch, Maxime et al. 2021), and that photons from star formation are most probably the primary sources of reionization. Yet, it remains to be understood which types of galaxies are most profusely leaking ionizing radiation (photons with wavelength  $\lambda < 912\text{\AA}$ , also called Lyman continuum or LyC) and the properties and environments that can make a galaxy a LyC leaker.

The primary difficulty in understanding the processes and sources that powered cosmic reionization is that it is not possible to directly probe the ionizing radiation at that epoch as those photons are all absorbed by the intergalactic medium (IGM) on

their way to us (Madau 1995; Inoue et al. 2014). Due to this, it is imperative to find indirect tracers for LyC emission to identify the sources of reionization.

In recent years, several methods have been proposed in the literature to indirectly measure LyC emission from galaxies: weak ISM absorption lines (Heckman et al. 2011; Erb 2015; Chisholm et al. 2017, but see Mauerhofer et al. 2021), a high [OIII]/[OII] ratio (Jaskot & Oey 2013; Nakajima & Ouchi 2014, but also see Bassett et al. 2019; Katz et al. 2020), and the Ly $\alpha$  line of Hydrogen (Dijkstra 2014a; Verhamme et al. 2015; Dijkstra et al. 2016; Verhamme et al. 2017; Izotov et al. 2018a). Among these, the Ly $\alpha$  line is particularly interesting. Since it is a UV line, Ly $\alpha$  is observable over a wide range of redshifts, allowing one to probe galaxy formation with the same tool over several Gyrs of evolution. Indeed, over the last 20 years, a great amount of Ly $\alpha$  emitting galaxies have been observed: from the low redshift Universe, using space-based facilities (Lyman Alpha Reference Sample (LARS) and Extended LARS survey (eLARS), studying 14 and 28 LAEs, respectively, at  $0.03 < z < 0.18$  (Hayes et al. 2013; Östlin et al. 2014) and the Green Pea sample of 43 LAEs at  $z = 0.2$ ; Henry et al. 2015; Schaerer et al. 2016; Yang et al. 2017); from the ground, in optical from  $z \sim 2$  to  $z \sim 6$  (several thousands of spectroscopically confirmed LAEs, Erb et al. 2011; Bacon et al. 2015; Trainor et al. 2015; Urrutia et al. 2019) and in IR at the highest redshifts (e.g. SILVERRUSH survey using the Hyper Suprime-Cam (HSC) recently observed a large sample of 2230 Lyman Alpha emitters (LAEs) at  $z = 5.7$

- 6.6 with narrow band imaging data; Ouchi et al. 2018; Shibuya et al. 2018). At even higher redshift, it is increasingly difficult to detect LAEs due to attenuation of Ly $\alpha$  by the relatively neutral IGM. However, concentrated efforts with very deep photometric and spectroscopic surveys in recent years have led to detections of some Ly $\alpha$  emitting galaxies in the extreme redshift range of  $z = 6 - 9$  (Vanzella et al. 2011; Ono et al. 2012; Schenker et al. 2012; Shibuya et al. 2012; Finkelstein et al. 2013; Oesch et al. 2015; Konno et al. 2014; Zitrin et al. 2015; Song et al. 2016; Roberts-Borsani et al. 2016; Stark et al. 2017; Matthee et al. 2017; Songaila et al. 2018; Matthee et al. 2018; Itoh et al. 2018; Jung et al. 2019; Matthee et al. 2020; Meyer et al. 2021). The upcoming JWST surveys are expected to discover many more such galaxies in the EoR soon.

The possibility of the relatively intense Ly $\alpha$  radiation from galaxies being a tracer of LyC emission has become much studied in the past few years. Verhamme et al. (2015) explored the escape of Ly $\alpha$  and LyC in idealized galaxy models and found that Ly $\alpha$  line profiles show distinct signatures (strong, narrow peak and narrow peak separation if it is double-peaked) if the ISM of galaxies is transparent to the LyC. Dijkstra et al. (2016) found similar results in a theoretical study of a suite of 2500 idealized models of a dusty and clumpy ISM with Ly $\alpha$  radiative transfer simulations.

Verhamme et al. (2017) performed an observational study of LyC leakers in the sample of Green Pea galaxies (the local analogs of high- $z$  LAEs) and found that in the 8 galaxies where it is possible to detect LyC emission<sup>1</sup> in addition to Ly $\alpha$ , the escape fractions of Ly $\alpha$  and LyC are indeed positively correlated. Recently Izotov et al. (2021) observed 9 more galaxies in both LyC and Ly $\alpha$  in the redshift range of  $\sim 0.30 - 0.45$  and found that similar correlations exist in this sample as well. Steidel et al. (2018) studied the KLCS (Keck Lyman Continuum Spectroscopic Survey) sample which included 15 (out of 124) galaxies detected in LyC at  $z \sim 3$  and found that the LyC escape fraction is well correlated with the equivalent width of the Ly $\alpha$  emission.

The correlation between Ly $\alpha$  and LyC radiation shows great promise, but to use this in the reionization era to estimate LyC from galaxies, we need to statistically analyze a large sample of EoR galaxies. Since LyC cannot be observed in this epoch, we need to explore it with simulations. Modelling Ly $\alpha$  and LyC radiation from a large sample of galaxies in simulations has been particularly challenging, because it requires simulations to overcome several technical challenges. Such simulations need to incorporate LyC radiation transfer on the fly, i.e. coupled at each hydrodynamical time step, to describe the ionization state of each cell in the simulation volume accurately. These simulations also need to account for the radiative transfer of Ly $\alpha$  which requires a massively parallel resonant scattering code. Finally, the production and scattering or absorption of Ly $\alpha$  and LyC photons happen at small scales in the ISM of galaxies and their eventual escape or absorption is at galactic and inter-galactic scales, so the simulation needs to sample both small and large scales correctly in order to predict reliable escape fractions and reionization topology. All of these requirements make such undertakings challenging and computationally expensive. Hence simulation studies of this kind so far have generally focused on either analyzing a small volume with high resolution, such as isolated galaxies (Verhamme et al. 2012; Behrens & Braun 2014), Ly $\alpha$  nebulae/blobs (Yajima et al. 2013; Trebitsch et al. 2016); molecular clouds (Kimm et al. 2019), zoom-in simulation of individual galaxies (Faucher-Giguère et al. 2010; Smith et al. 2019; Laursen et al. 2019), or large volumes but with comparatively poor resolution (Yajima et al. 2014; Inoue et al. 2018; Gronke et al. 2021).

With this in mind, the SPHINX simulations are an ideal choice for this study, being a state-of-the-art radiation hydrody-

namics (RHD) simulation and a good balance of a sufficiently large volume and high resolution hosting a large sample of well resolved galaxies. SPHINX is a suite of cosmological radiation-hydrodynamics simulations which reaches a resolution up to 10 pc in 10 co-moving Mpc (cMpc) wide volumes (Rosdahl et al. 2018). This allows us to investigate the Ly $\alpha$  and LyC properties of thousands of simulated galaxies at  $z > 6$  (see Rosdahl et al. 2018; Garel et al. 2021).

In this paper, we focus on the following questions:

- is there a correlation between Ly $\alpha$  and LyC emission at galaxy scale during the EoR?
- is it possible to predict the LyC emission of galaxies, knowing their physical and Ly $\alpha$  properties?

The paper is structured as follows. We discuss our methods in §2, where we describe the SPHINX simulation and the radiative transfer code that we use for Ly $\alpha$  post processing, and present our sample of simulated galaxies. In §3 we explore the relationship between LyC and Ly $\alpha$  luminosities and escape fractions and analyze the contribution of LAEs to reionization. In §4 we build a multivariate regression model where we use the physical and Ly $\alpha$  properties of galaxies to predict their intrinsic and escaping LyC luminosities and escape fractions, determine the most important variables required for each prediction and apply our models to observed data for comparison. In §5 we discuss the limitations of our study and in §6 we summarize our results.

## 2. Methods

In this section we present the simulation, the selection procedure to build our sample of galaxies, and our methods to calculate LyC and Ly $\alpha$  emissions from them.

### 2.1. The Reionization Simulation

SPHINX (Rosdahl et al. 2018) is a suite of cosmological hydrodynamical simulations of the epoch of reionization. In this study we analyze galaxies in the 10 cMpc wide SPHINX volume previously presented in Rosdahl et al. (2018) which uses the binary stellar population model from BPASS (Stanway et al. 2016).

SPHINX is run with the RAMSES-RT code (Teyssier 2002; Rosdahl et al. 2013). It simulates an average density patch of the Universe. The spatial resolution reaches 10.9 pc at  $z = 6$ , the dark matter mass resolution is  $2.5 \times 10^5 M_\odot$  per particle and the stellar mass resolution is  $10^3 M_\odot$  per stellar particle (we refer to Rosdahl et al. 2018, for details of the simulation). Within the simulation the radiation tracked is split into three photon groups, which encompass the ionization energies for HI, HeI, and HeII. These photons interact with hydrogen and helium in the simulation via photo-ionization, heating, and momentum transfer. The simulation is run until  $z = 6$  and it uses Planck results (Planck Collaboration et al. 2014) for the cosmological parameters, i.e.  $\Omega_\Lambda = 0.68$ ,  $\Omega_m = 0.32$ ,  $\Omega_b = 0.05$ ,  $h = 0.67$ , and  $\sigma_8 = 0.83$ .

### 2.2. Halo and Galaxy samples

We use the same halos and galaxies as described and analyzed in Rosdahl et al. (2018). In short, galaxies are detected in two stages. The group finder algorithm ADAPTAHOP (Aubert et al. 2004; Tweed et al. 2009) is run on the dark matter particles, and the overdense virialized regions are identified as halos (and sub halos, sub-sub-halos etc depending on their level of structure). Halos are considered to be resolved when they have virial masses ( $M_{vir}$ ) greater than 300 times the dark matter particle mass, i.e.  $M_{vir} > 7.4 \times 10^7 M_\odot$ . Then ADAPTAHOP is run on stellar particles, and it identifies the overdense groups with at least 10 stellar particles as galaxies. Finally, the most massive galaxy within  $0.3 R_{vir}$  is assigned to each halo to build the galaxy-halo catalog.

In our analysis, we select systems which have i.) stellar mass  $M_\star > 10^6 M_\odot$  (this is stellar mass within  $0.3R_{vir}$  of the halo)

<sup>1</sup> After reionization the Universe has stayed ionized, so that in the local Universe LyC photons can travel without being absorbed, unlike in high- $z$  where neutral hydrogen atoms can absorb them easily.

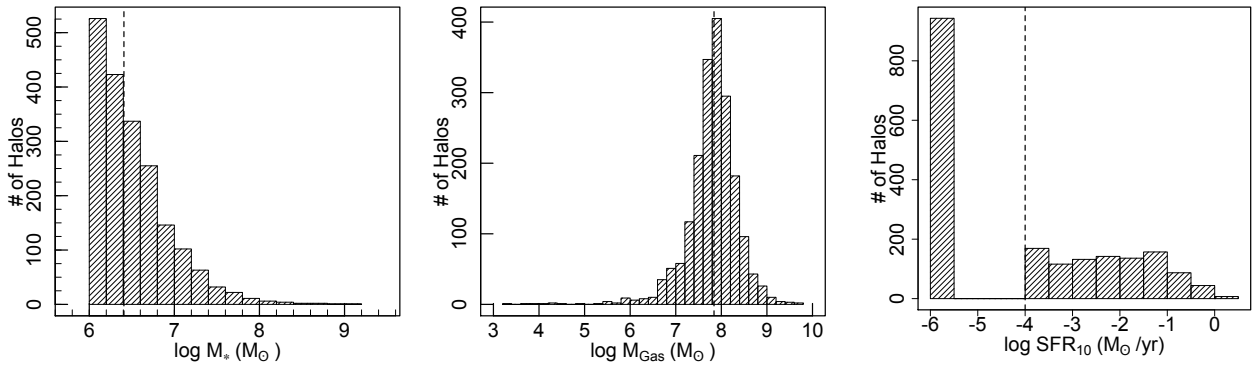


Fig. 1: Histograms of physical properties of the simulated galaxies in our sample. The histograms show the distribution of stellar mass (left), gas mass (middle) and SFR<sub>10</sub>(right) of the galaxies. The median values are shown by dashed black lines. The stellar mass histogram shows  $M_*$  within 30% of the halo virial radius. There are 10 galaxies with  $M_* > 10^8 M_\odot$ . The gas mass has a peaked distribution with few galaxies having very little gas (further discussion in § 5). There are 943 galaxies with zero SFR<sub>10</sub>, these galaxies are represented in the bar at  $10^{-6}$  (discussed in § 4.2.1).

and ii.) main halo is at level 1, i.e. they are not a substructure of a parent halo. We exclude less massive galaxies with  $M_* < 10^6 M_\odot$  from our sample and focus on bright galaxies that are potentially observable. This stellar mass limit also means that all of our galaxies contain at least  $10^3$  stellar particles, which ensures that the selected galaxies are reasonably well resolved.

We analyze snapshots of the SPHINX simulation at 5 different redshifts:  $z = 6, 7, 8, 9$ , and 10. We select all galaxies that satisfy our criterion described above. The numbers of selected galaxies at these redshifts are respectively, 674, 509, 362, 236 and 152. Among these galaxies at  $6 \leq z \leq 10$ , the maximum galaxy stellar mass is  $1.33 \times 10^9 M_\odot$ , and there are 10 galaxies with  $M_* > 10^8 M_\odot$ . We have compared the properties of the galaxies at these different redshifts and found that there is no significant evolution in terms of physical or radiative (Ly $\alpha$  or LyC) properties (more discussion at § A.1). Therefore, we combine our galaxy sample as a larger sample size can give better statistical significance for our understanding. Our final sample comprises of 1933 galaxies.

Figure 1 shows distributions of their stellar mass, gas mass and star formation rate. We recall that the stellar mass distribution in this figure shows the stellar mass within 30% halo virial radius. The median stellar mass is  $10^{6.41} M_\odot$ .

The gas mass shown in Figure 1 is the total gas mass of the halo, calculated by summing up the mass of all the gas cells inside  $R_{vir}$ . We find that the gas mass has a normal distribution (median mass  $10^{7.84} M_\odot$ ) with some halos containing very small amounts of gas, likely because a recent supernova or starburst has blown the gas away from these small systems.

The star formation rate, SFR<sub>10</sub> shows the star formation rate of the galaxy averaged over the last 10 Myrs. 10 Myrs is a typical lifetime of the massive stars, after which they undergo a supernova (the most massive stars live for about 3 Myr) and 10 Myr is also the typical timescale of the production of LyC and Ly $\alpha$ . In Figure 1 we show the distribution of the log SFR<sub>10</sub>. There are 943 galaxies in our sample that have SFR<sub>10</sub> = 0. We artificially set their SFR values equal to  $10^{-6} M_\odot/\text{yr}$  (which is lower than the lowest non-zero SFR) to show them in the histogram. The median value of SFR<sub>10</sub> is  $10^{-4} M_\odot/\text{yr}$ .

### 2.3. LyC emission from SPHINX galaxies

The production and escape of LyC photons in SPHINX has been described in Rosdahl et al. (2018). In short, the instantaneous escape fractions of LyC photons are calculated in post-processing, using RASCAS (Michel-Dansac et al. 2020). Rays are traced from every stellar particle inside a halo out to its virial radius. Along each ray, the optical depth ( $\tau$ ) is calculated for hydrogen

and helium. For each stellar particle, the escape fraction is the average of  $e^{-\tau}$  calculated with rays in 500 random directions. Then the global escape fraction of the halo ( $f_{esc}^{\text{LyC}}$ ) is the luminosity-weighted average escape fraction of all the stellar particles inside the halo. The LyC photons we consider range from 0–912Å and in the simulation they are described in 3 groups of photons: photons that ionize HI (UV<sub>HI</sub>, 912 - 504 Å, 13.6 - 24.59 eV), HeI (UV<sub>HeI</sub>, 504 - 228 Å, 24.59 - 54.42 eV) and HeII (UV<sub>HeII</sub>, 228 - 0 Å, 54.42 - ∞ eV). The distributions of intrinsic ( $L_{int}^{\text{LyC}}$ ) and escaping ( $L_{esc}^{\text{LyC}}$ ) LyC luminosities, and escape fractions for our galaxy sample are further described in § 3.2.

On the contrary, observations of LyC usually focus on a small part of the ionizing spectrum, close to the Lyman limit (912 Å). This observed LyC luminosity known as  $L_{esc}^{900}$ , i.e. the escaping LyC luminosity at 900Å, is defined as  $L_{esc}^{900} = L_{int}^{900} \times f_{esc}^{900}$  ( $L_{int}^{900}$  and  $f_{esc}^{900}$  are intrinsic luminosity and escape fraction at 900Å, respectively). So we perform additional LyC measurements more similar to what is done observationally. We can estimate the intrinsic LyC luminosity of the simulated galaxies at 900Å ( $L_{int}^{900}$ ), using the BPASS models (Stanway et al. 2016) that have been used in modeling the ionizing emission in the SPHINX simulation. Using RASCAS we distribute  $10^5$  photon packets with wavelengths between 10 - 912Å among the stellar particles and then transfer them until they are absorbed by HI, HeI, HeII, dust or escape the halo virial radius. Thereafter we have both the intrinsic and escaping spectral energy distribution from 10 - 912Å and this allows us to derive the LyC escaping luminosity and escape fraction over different wavelength ranges, e.g. 890-912 Å. The average luminosity in this range is the luminosity at 900Å (i.e.  $L_{int}^{900}$  and  $L_{esc}^{900}$  are in units of erg/s/Å).

Figure 2 (left and middle panel) shows the ratio of total LyC (i.e. 0 - 912Å) emission (intrinsic and escaping luminosities) to the LyC emission at 900Å as a function of their total escaping LyC ( $L_{esc}^{\text{LyC}}$ ) luminosities for all simulated galaxies. Since we integrate over a wavelength range 900 times larger for the total luminosity, we expect a rough ratio of around 900 between the two intrinsic luminosities. The ratio of the escaping luminosities is expected to be higher because the cross-section of Hydrogen photoionisation is approximately proportional to  $\lambda^3$  at  $\lambda < 912\text{Å}$ , so  $L_{esc}^{900}$  could be more attenuated than  $L_{esc}^{\text{LyC}}$ . Indeed, the median ratios for intrinsic and escaping luminosities are 1036 and 1536, respectively. We also find that this ratio for both intrinsic and escaping luminosities has a significant scatter, probably due to the particular star formation history and morphology of each galaxy.

In particular, we find that 242 galaxies, i.e. 12% of our galaxies, have  $L_{\text{esc}}^{900} = 0$ , i.e. for these galaxies  $L_{\text{esc}}^{\text{LyC}}/L_{\text{esc}}^{900} = \text{Inf}$ , these ratios are represented at a value of 4250 in the middle panel. Almost all of them are also faint in total LyC emission, with only 6 among them having  $L_{\text{esc}}^{\text{LyC}} > 10^{39}$  ergs/s. This result suggests that few LyC leakers can be missed by surveys probing only the flux close to the Lyman limit.

By comparing the number of intrinsic photons with the escaping photons, we also obtain the escape fraction at 900Å. We show the ratio of  $f_{\text{esc}}^{900}$  to  $f_{\text{esc}}^{\text{LyC}}$  as a function of  $f_{\text{esc}}^{\text{LyC}}$  in the right panel of figure 2 and find that, except very few galaxies (with low luminosities)  $f_{\text{esc}}^{900}$  is lower than  $f_{\text{esc}}^{\text{LyC}}$ . The overall median ratio is 0.69. Kimm et al. (2019) also finds similar ratio while investigating the escape of LyC radiation from turbulent clouds. We divide the (log)  $f_{\text{esc}}^{\text{LyC}}$  into groups of 0.5 dex each and find that the median ratio increases slightly with  $f_{\text{esc}}^{\text{LyC}}$ .

## 2.4. Ly $\alpha$ emission from SPHINX galaxies

In order to investigate the correlation between LyC leakage and the observable Ly $\alpha$  properties of galaxies, we now turn our interest to the Ly $\alpha$  post-processing of SPHINX galaxies.

A Ly $\alpha$  photon (wavelength - 1215.67 Å, energy - 10.2 eV) is emitted when a hydrogen electron jumps from the 2p to the 1s (ground) state. It is not only the hydrogen line with the largest flux, but also a resonant line. To obtain the Ly $\alpha$  properties of galaxies in the SPHINX simulation, we post process them using RASCAS (Michel-Dansac et al. 2020), which is a fully parallelized 3D radiative transfer code developed to perform the propagation of any resonant line in numerical simulations. It performs radiative transfer on an adaptive mesh using the Monte Carlo technique. We describe below the different steps of our implementation.

**Ly $\alpha$  intrinsic luminosities:** Ly $\alpha$  emission can be triggered by two processes, recombination and collisional de-excitation (Dijkstra 2014b). LyC photons from massive stars in galaxies ionize the neutral gas in their ISM and afterwards, the free proton and electron recombine. The electron can initially enter into any energy level, and then cascades to ground level with a probability of  $\approx 0.67$  to emit a Ly $\alpha$  photon (Partridge & Peebles 1967; Dijkstra 2014b). Alternatively, HI atoms can be excited collisionally, and when the electron returns to the ground state, a Ly $\alpha$  photon can be emitted. So, for any given halo in our sample, we track both recombinations and collisional excitations from all cells inside the halo virial radius to capture the intrinsic Ly $\alpha$  emission. For recombinations, the Ly $\alpha$  photon emission rate in each cell is (Cantalupo et al. 2008) :

$$N_{\gamma, \text{rec}} = n_e n_p \alpha_B(T) \epsilon_{\text{Ly}\alpha}^B(T) \times (\Delta x)^3 \quad (1)$$

where,  $n_e$  and  $n_p$  are the number density of electrons and protons respectively (these come from the simulation),  $\alpha_B(T)$  is the case-B recombination coefficient,  $\epsilon_{\text{Ly}\alpha}^B(T)$  is the fraction of recombination events that produces a Ly $\alpha$  photon eventually (at  $T = 10^4 \text{K}$ , it is 0.67) and  $(\Delta x)^3$  is the cell volume. For collisional excitation, the Ly $\alpha$  emission rate is given by (Goerdt et al. 2010) :

$$N_{\gamma, \text{col}} = n_e n_{\text{HI}} C_{\text{Ly}\alpha}(T) \times (\Delta x)^3 \quad (2)$$

where  $n_{\text{HI}}$  is the number density of neutral hydrogen, and  $C_{\text{Ly}\alpha}(T)$  is the rate of collisionally induced 1S-to-2P level transitions (we do not consider higher order transitions). We refer to Michel-Dansac et al. (2020) for a detailed description of how we fit each of the coefficients  $\alpha_B(T)$ ,  $\epsilon_{\text{Ly}\alpha}^B(T)$  and  $C_{\text{Ly}\alpha}(T)$ . Once these luminosities are known in each cell, we emit a total of  $10^5$  photon packets from the cells inside a galactic halo with the probability of a cell emitting a photon packet proportional to its luminosity. The number of photon packets has been chosen so as

to minimise the computational cost while preserving the accuracy of the Ly $\alpha$  angle-averaged escape fraction and luminosity. Performing convergence tests on the ten most massive galaxies in our sample, we find that these quantities are well converged using  $10^5$  photon packets.

**Ly $\alpha$  propagation and escape:** In each cell, we cast Ly $\alpha$  photons isotropically and propagate them through the halo with RASCAS code. Each Ly $\alpha$  photon can be scattered, i.e. absorbed and re-emitted, numerous times whenever they encounter HI atoms in the ISM, until they finally escape the halo or are absorbed by dust. The dust is modelled by specifying a cross section per hydrogen atom and a pseudo dust number density dependent on HI and HII density and metallicity (Michel-Dansac et al. 2020). The dust absorption coefficient in each cell is given by  $(n_{\text{HI}} + f_{\text{ion}} n_{\text{HII}}) \sigma_{\text{dust}}(\lambda) Z/Z_0$ , where  $f_{\text{ion}} = 0.01$  (abundance of dust in ionized gas),  $Z$  is the gas metallicity in that cell, the effective dust cross-section  $\sigma_{\text{dust}}$  and  $Z_0 (= 0.005)$  are normalized to the Small Magellanic Cloud (SMC) models following Laursen et al. (2009).

The boundary beyond which a Ly $\alpha$  photon can be considered as having escaped is not an obvious choice. At  $z \geq 6$  where reionization takes place, configuration of galaxies are complex, partly because in many cases galaxies are interacting or colliding with each other. So we perform convergence tests on the ten most massive galaxies in our sample. In each of them, we set the boundary at  $R_{\text{vir}}$ ,  $2R_{\text{vir}}$ , and  $3R_{\text{vir}}$  where  $R_{\text{vir}}$  is the corresponding halo virial radius, and run Ly $\alpha$  radiative transfer in each case. We find that beyond  $R_{\text{vir}}$ , the escape fraction converges, with only small increments in accuracy. So we fix  $R_{\text{vir}}$  to be the boundary of Ly $\alpha$  escape. Both the production and the propagation of photons are allowed within this radius, which encompass the main galaxy and in many cases, its satellites.

We use the core-skipping method to speed up the calculation (Michel-Dansac et al. 2020). We have tested the core-skipping method by simulating the Ly $\alpha$  radiation transfer in the 10 most massive galaxies in our simulation with and without core-skipping and found that the Ly $\alpha$  results, e.g. luminosities and escape fraction, are very similar (median 0.6% difference) and we gain significant (up to a factor of 100) speedup in the calculation. The distributions of intrinsic and escaping Ly $\alpha$  luminosities, and escape fractions, for our galaxy sample, are further described in § 3.2.

## 3. LyC - Ly $\alpha$ relationship

The goal of our study is to investigate the connection between the Ly $\alpha$  and LyC properties of galaxies in order to investigate if, or how, Ly $\alpha$  can trace the total ionizing radiation escaping from galaxies at EoR. To that end, in this section we first discuss the relationship between their intrinsic and escaping luminosities and then analyze their escape fractions.

### 3.1. Observed LyC emitters

Before we explore the relationship between the various Ly $\alpha$  and LyC properties, we review existing observed sample of LyC emitters (LCEs) in order to facilitate the comparison of our simulated galaxies with observed ones.

Although there are many observations of Ly $\alpha$  at different redshifts, it is difficult to observe LyC even at low redshift galaxies because earth's atmosphere blocks UV radiation, so no ground based observations are possible. However, in recent years it has become possible to obtain direct observations of LyC leakers using space-based facilities, e.g. HST (Verhamme et al. 2017; Izotov et al. 2016a,b, 2018a,b; Izotov et al. 2021). We compile these observations (23 galaxies) in Table 1, where we note their redshift, available physical properties, i.e. stellar mass, SFR, surface SFR density, escaping luminosity and escape fraction in LyC and Ly $\alpha$ . The SFR is derived from H $\beta$  observations and therefore cor-

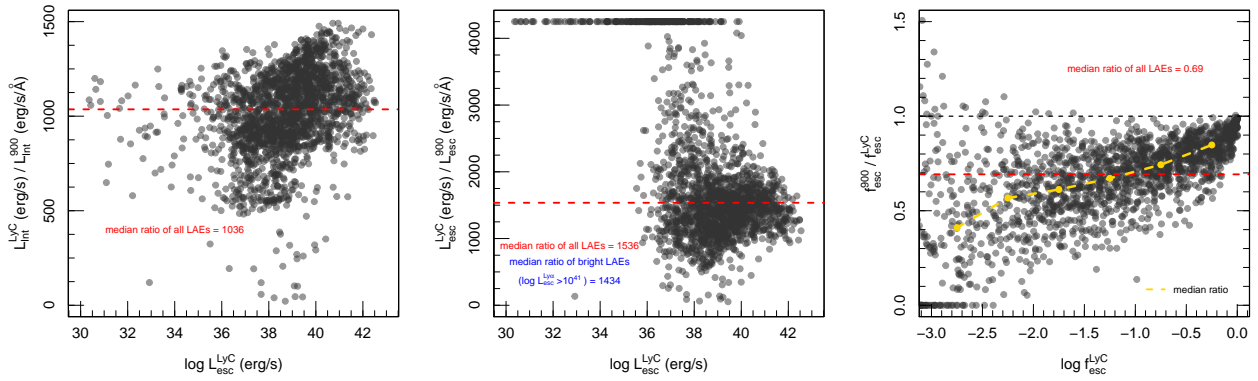


Fig. 2: Ratio of total intrinsic (left) / escaping (middle) LyC luminosity emitted over the range of 0 - 912 Å and the intrinsic / emitted LyC at 900 Å, as a function of their total escaping LyC luminosity. The median ratio calculated with all simulated galaxies is shown in red dashed line. Some of the galaxies have  $L_{\text{esc}}^{900} = 0$ , i.e. their ratios of  $L_{\text{esc}}^{\text{LyC}} / L_{\text{esc}}^{900} = \text{Inf}$ , these ratios are represented at a value of 4250 in the middle panel. The ratio of  $f_{\text{esc}}^{900} / f_{\text{esc}}^{\text{LyC}}$  as a function of the  $f_{\text{esc}}^{\text{LyC}}$  is shown at the right panel (note that the escape fraction ratio is 900 Å divided by total and the other two ratios are total divided by 900 Å quantities). The dashed yellow line shows the median ratio as a function of  $f_{\text{esc}}^{\text{LyC}}$  (we divide the  $\log f_{\text{esc}}^{\text{LyC}}$  between -3 to 0 into groups of 0.5 dex each and find the median ratios).

respond to SFR on a short time scale. They can thus be considered similar to the  $\text{SFR}_{10}$  in our simulated galaxies.

Figure 3 shows a comparison of the physical properties of these observed LCEs with our simulated sample. We find that Ly $\alpha$  luminosities of the SPHINX galaxies do not scale well with stellar masses, gas metallicities or galaxy sizes, whereas they correlate with recent star formation, as expected since higher SFR means more energetic photons are being produced which can be reprocessed in the ISM as Ly $\alpha$ . The  $\text{SFR}_{10}$  of the galaxies correlates weakly with the stellar mass and has a large scatter. Because of the finite volume of our simulation, our sample is restricted to relatively faint and low-mass galaxies such that most of the observed objects considered here are brighter, slightly more massive, slightly bigger and have higher star formation than our simulated sources. The observed galaxies also have higher metallicities compared to the simulated ones, which is perhaps not surprising as the observed sample is at a much lower redshift ( $z \sim 0.3$  compared to  $z \sim 6$ ), hence they can be more metal enriched. While this certainly represents a limitation of our study, investigating the LyC-Ly $\alpha$  connection in our sample can still be used to interpret available observational data and guide future surveys that will target galaxies more similar to our sample.

It is important to note that the LyC luminosity in the Table 1 is  $L_{\text{esc}}^{900}$ , i.e. the LyC luminosity at 900 Å. However, the escaping LyC luminosity that counts for reionization is the total luminosity of all photons that can ionize HI, i.e. all photons with  $\lambda = 0-912$  Å so we consider this total LyC throughout the paper. These two measures of LyC luminosities can be very different as discussed in § 2.3, and the contribution of the highly ionizing spectrum for observed LCEs ( $< 900$  Å) is still largely unknown. Since the observed LCEs are brighter in Ly $\alpha$  ( $> 10^{41}$  erg/s) than the bulk of our galaxies, we recalculate the median of the  $L_{\text{esc}}^{\text{LyC}} / L_{\text{esc}}^{900}$  ratio for bright LAEs and find the ratio to be 1434 (Figure 2, middle panel). This ratio can be used to convert observed 900 Å luminosities to total LyC luminosities, if needed.

As we are interested in investigating the global theoretical connection between Ly $\alpha$  and the ionizing radiation of galaxies in EoR, hereafter throughout this paper we consider the global Ly $\alpha$  and LyC photon budgets from galaxies, i.e. summed over all directions and relevant wavelengths (i.e. 0 - 912 Å for LyC luminosities and  $f_{\text{esc}}^{\text{LyC}}$ ) unless otherwise specified.

### 3.2. Distributions of Ly $\alpha$ and LyC properties

Figure 4 shows the distribution of the Ly $\alpha$  and LyC properties of our simulated galaxy sample, namely their intrinsic luminosities, escaping luminosities and their escape fractions. In all four cases (intrinsic and escaping for Ly $\alpha$  and LyC), the luminosities have a peaked distribution. The median values of  $L_{\text{int}}^{\text{Ly}\alpha}$  and  $L_{\text{esc}}^{\text{Ly}\alpha}$  are 39.88 and 39.51 erg/s (in log scale) respectively, with the maximum escaping luminosity at  $1.375 \times 10^{42}$  erg/s. The LyC luminosities show a similarly peaked distribution with median (log) values at 40.23 ( $L_{\text{int}}^{\text{LyC}}$ ) and 38.80 ( $L_{\text{esc}}^{\text{LyC}}$ ). We note that the maximum luminosities of simulated galaxies are a consequence of the finite volume of the simulation box and the low end of the luminosities are affected by galaxy mass selection and the mass resolution of the simulation (Garel et al. 2021).

In contrast,  $f_{\text{esc}}^{\text{Ly}\alpha}$  shows a bi-modal distribution with the major peak at 1 (the minor peak is at 0). We find that 32% of the sample has  $f_{\text{esc}}^{\text{Ly}\alpha} > 0.9$ . The distribution of  $f_{\text{esc}}^{\text{LyC}}$  shows that most galaxies have low  $f_{\text{esc}}^{\text{LyC}}$ , with 62% of galaxies with  $f_{\text{esc}}^{\text{LyC}} < 0.1$ . Since LyC can be absorbed by HI and HI is plentiful in the ISM, it is very hard for LyC to escape, resulting in very low  $f_{\text{esc}}^{\text{LyC}}$  in most galaxies. Ly $\alpha$  on the other hand is absorbed only by dust, so has a easier time to escape, which results in the peak around  $f_{\text{esc}}^{\text{Ly}\alpha} = 1$ .

Among these 6 quantities, only the escaping Ly $\alpha$  luminosity is observable at the EoR. Our sample is fainter than most available LAE data but it can still be compared with the faint LAEs from MUSE surveys. Therefore in the histogram of  $L_{\text{esc}}^{\text{Ly}\alpha}$  in Figure 4 we also show the distribution of  $L_{\text{esc}}^{\text{Ly}\alpha}$  from galaxies in MUSE GTO surveys. The MUSE data are taken from the MUSE-Deep survey (Drake et al. 2017) and MUSE Extremely Deep Field (MXDF) (Bacon et al. in prep). In total there are 892 MUSE galaxies in the redshift range of  $z = 2.92-6.64$  with luminosities  $10^{40.33-43}$  erg/s. Among these, 21 galaxies are at  $z > 6$ . We see that there is overlap between the most luminous end of our simulated galaxies and the faint end from MUSE, in the luminosity range of  $\sim 10^{40-42}$  erg/s. Our simulated luminosities are the total Ly $\alpha$  output of the galaxy in all directions, before IGM attenuation. The observed data is, of course, directional measurement after IGM attenuation. We discuss the potential observational biases towards bright galaxies, and the lack of very bright LAEs in our sample due to the simulation box size limit in § 5.

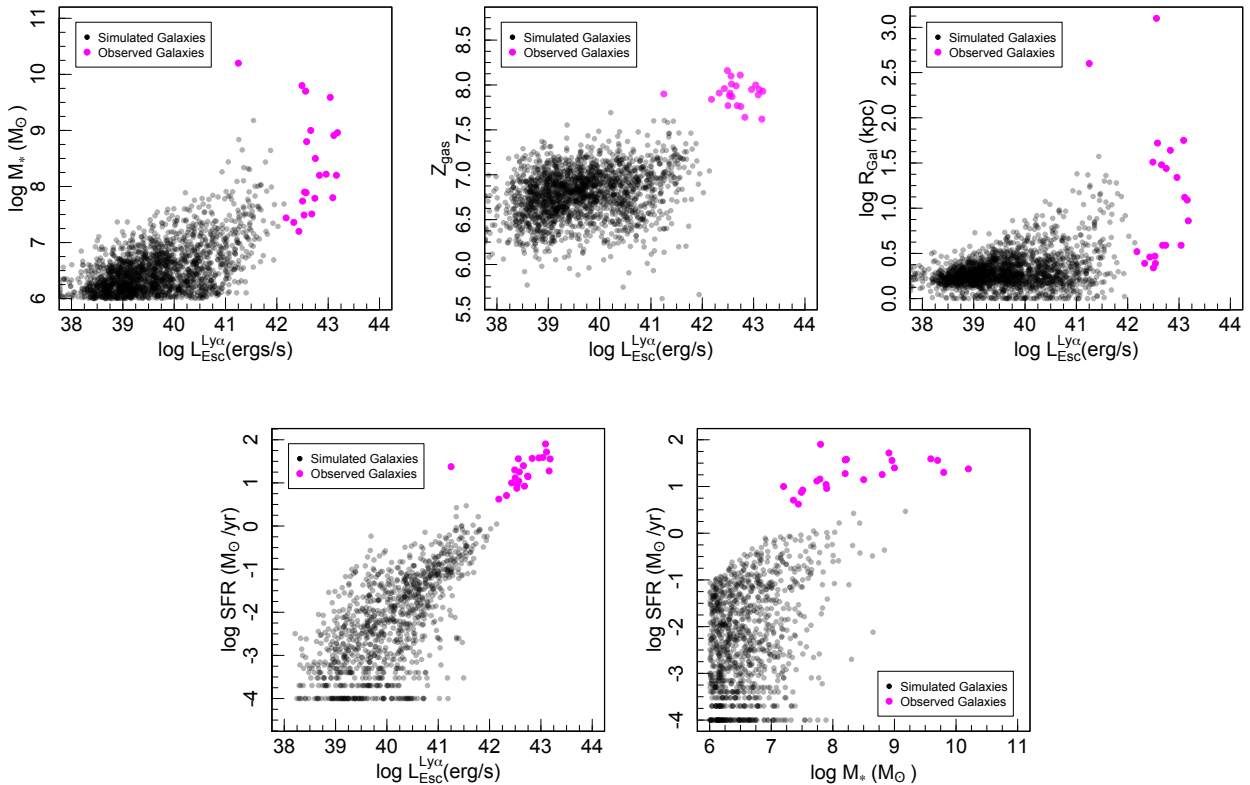


Fig. 3: Comparison of physical properties of observed LCEs (magenta points) and the simulated galaxies (black points). Here we show the stellar mass (top left), gas metallicity (oxygen abundances i.e.  $12+\log_{10}(\text{O}/\text{H})$  for observed galaxies, top middle), galaxy radius (galaxy virial radius for simulated ones and exponential disc scale length for observed ones, top right) and SFR ( $\text{SFR}_{10}$ ) as a function of their escaping  $\text{Ly}\alpha$  luminosities. The  $\text{Ly}\alpha$  luminosities of the SPHINX galaxies do not scale with stellar masses, metallicities or galaxy sizes, whereas they correlate with recent star formation, as expected. In the top right panel we show SFR ( $\text{SFR}_{10}$  for simulated galaxies) as a function of the stellar mass of galaxies. The properties of the observed LCEs are listed in Table 1 and more details can be found in their corresponding reference papers.

### 3.3. Investigating the $\text{Ly}\alpha$ - $\text{LyC}$ luminosity relationship

To assess possible correlations between the  $\text{LyC}$  and  $\text{Ly}\alpha$  radiation in galaxies, in the first step we analyze their intrinsic and escaping luminosities.

In Figure 5 we show the  $\text{LyC}$  luminosities of galaxies as a function of their  $\text{Ly}\alpha$  luminosities. We find that for intrinsic luminosities,  $\text{Ly}\alpha$  and  $\text{LyC}$  have a fairly tight positive correlation. The production of both  $\text{LyC}$  and  $\text{Ly}\alpha$  is strongly related to the star formation rate of the galaxy because massive stars directly emit  $\text{LyC}$  photons and these same photons generate  $\text{Ly}\alpha$  by photoionizing the  $\text{HI}$  in the ISM, which then can produce  $\text{Ly}\alpha$  through recombination.

Furthermore, we also show their intrinsic  $\text{LyC}$  luminosity at  $900\text{\AA}$  (as discussed in § 2.3), and find that the intrinsic luminosities of observed LCEs (derived as observed luminosity/escape fraction) also fall on the same tight correlation, though extending to higher luminosities. This suggests that the correlation between intrinsic  $\text{Ly}\alpha$  and  $\text{LyC}$  luminosities is valid over a large range of  $\text{Ly}\alpha$  luminosities.

In the same figure we also show predictions for intrinsic  $\text{Ly}\alpha$  luminosities from a simple model based on case B recombination (Spitzer 1978) given by,  $L_{\text{int}}^{\text{Ly}\alpha} = 0.67(1 - f_{\text{esc}}^{\text{LyC}})L_{\text{int}}^{\text{LyC}}$ . This model assumes that all  $\text{LyC}$  photons that do not escape the galaxy will ionize the neutral hydrogen gas in the ISM. It also assumes that 67% of them will be reprocessed as  $\text{Ly}\alpha$  photons through recombinations.

We find that the simulated data is generally matched well by this model. Some galaxies, especially among lower  $\text{Ly}\alpha$  luminosity galaxies, lie below the analytical relationship, implying that the contribution of collisions is increasingly important for faint and low mass  $\text{Ly}\alpha$  emitters. For example, we find that

in galaxies where  $L_{\text{int}}^{\text{Ly}\alpha} > 10^{42}$  erg/s, collisional emission contributes only  $\sim$  a few percent of the total  $\text{Ly}\alpha$  production, but it can rise to  $\sim 50\%$  in galaxies  $10^{38} \leq L_{\text{int}}^{\text{Ly}\alpha} \leq 10^{40}$  erg/s (see discussion and figure in A.2, also Rosdahl & Blaizot 2012). We also find that in all luminosity ranges, some galaxies fall above the analytical relationship, i.e. there are some galaxies which have less  $\text{Ly}\alpha$  production than estimated by the analytical equation. This is mainly due to the fact that a fraction of the most energetic photons go towards ionising He or HeI, rather than HI, and as a result they cannot be reprocessed as  $\text{Ly}\alpha$ . We also note that galaxies at the very faint end of  $\text{Ly}\alpha$  ( $L_{\text{int}}^{\text{Ly}\alpha} < 10^{38}$  erg/s) have  $\text{LyC}$  luminosity in the range of  $10^{38} - 10^{40}$  erg/s. These galaxies are extremely gas deficient, so they produce very little  $\text{Ly}\alpha$  and the stars in them continue to produce  $\text{LyC}$  for a long time (further discussed in §5).

Furthermore, from the right panel of Figure 5 we find that the escaping luminosity of  $\text{Ly}\alpha$  and  $\text{LyC}$  is also well correlated. The escaping  $\text{Ly}\alpha$  and  $\text{LyC}$  luminosities of the observed LCEs are also shown in this figure along with the  $L_{\text{esc}}^{900}$  of the simulated galaxies and these LCEs seem to follow the similar trend. We note that the correlation is tight at higher luminosities, although the scatter is overall larger compared to the correlation between the intrinsic luminosities. The scatter increases as the galaxies become fainter in  $\text{Ly}\alpha$  (or  $\text{LyC}$ ). This is mostly due to the fact that the faint LAEs have a very wide range of  $\text{Ly}\alpha$  and  $\text{LyC}$  escape fractions (discussed further in § 3.5, see also Fig 7, 8). Hence galaxies with similar intrinsic luminosities can end up with very different escaping luminosities, which scatters the points horizontally and vertically. The escape fractions of the galaxies depend on the structure of the ISM, in particular on the possibility of having holes or low HI column density channels in

Table 1: Observed data. The columns here denote the name of the galaxy, its redshift, Ly $\alpha$  and LyC escape fraction, stellar mass, star formation rate, star formation rate, oxygen abundance, Ly $\alpha$  and LyC luminosity (at 900Å) and the reference respectively. The uncertainties of the escape fractions are noted in the table. The typical uncertainty of luminosities is  $\sim 10\%$ .

Galaxy	z	$f_{\text{esc}}^{\text{Ly}\alpha}$	$f_{\text{esc}}^{\text{LyC}}$	$\log M_{\star}$ ( $\log M_{\odot}$ )	SFR ( $M_{\odot}/\text{yr}$ )	12+log(O/H)	$\log L_{\text{esc}}^{\text{Ly}\alpha}$ (erg/s)	$\log L_{\text{esc}}^{900}$ (erg/s/Å)	Ref
J0901 + 2119	0.2993	0.14±0.01	0.027±0.007	9.8	20	8.16	42.49	39.20	a, n
J0925 + 1409	0.3013	0.29±0.03	0.078±0.011	8.91	52.2	7.95	43.11	39.84	c, n
J1011 + 1947	0.3322	0.18±0.01	0.114±0.018	9.0	25	7.99	42.66	39.73	a, n
J1152 + 3400	0.3419	0.34±0.07	0.132±0.011	9.59	39	8.00	43.04	40.23	d, n
J1154 + 2443	0.3690	0.61±0.03	0.46±0.02	8.2	18.9	7.62	43.16	40.26	e, n
J1243 + 4646	0.4317	0.52±0.04	0.726±0.097	7.8	80	7.89	43.09	40.78	a, n
J1248 + 4259	0.3629	0.17±0.01	0.022±0.007	8.2	37	7.64	42.83	39.26	a, n
J1256 + 4509	0.3530	0.32±0.03	0.380±0.057	8.8	18	7.87	42.58	40.27	a, n
J1333 + 6246	0.3181	0.51±0.09	0.056±0.015	8.50	14	7.76	42.75	39.44	d, n
J1442 - 0209	0.2937	0.54±0.05	0.074±0.01	8.96	36	7.93	43.18	39.74	d, n
J1503 + 3644	0.3557	0.30±0.04	0.058±0.006	8.22	38	7.95	42.96	39.84	d
Tol1247 - 232	0.0488	0.10±0.02	0.045±0.012	9.7	36.2	8.1	42.56	40.4	f,g,i,n
Haro 11	0.021	0.04	0.032±0.012	10.2	23.8	7.9	41.25	39.60	f,h,j,l,o
J0232 - 0426	0.45236	0.425±0.053	< 0.04	7.49	7.5	7.88	42.53	38.73	m
J0919 + 4906	0.40512	0.687±0.089	0.162±0.059	7.51	8.4	7.77	42.68	39.63	m
J1046 + 5827	0.39677	0.318±0.043	< 0.02	7.89	11.0	8.01	42.57	38.69	m
J1121 + 3806	0.31788	0.432±0.052	0.35±0.056	7.20	10.0	7.96	42.43	39.85	m
J1127 + 4610	0.32230	0.397±0.085	0.111±0.040	7.44	4.2	7.84	42.18	39.05	m
J1233 + 4959	0.42194	0.412±0.039	0.121±0.034	7.79	14.4	8.11	42.74	39.72	m
J1349 + 5631	0.36366	0.403±0.044	< 0.07	7.36	5.1	7.91	42.33	38.72	m
J1355 + 1457	0.36513	0.231±0.028	< 0.01	7.74	13.1	7.77	42.50	38.72	m
J1455 + 6107	0.36793	0.365±0.045	< 0.01	7.90	9.1	7.91	42.54	38.68	m

a)Izotov et al. (2018b), b)Borthakur et al. (2014), c)Izotov et al. (2016a), d)Izotov et al. (2016b), e)Izotov et al. (2018a),  
f)Leitet et al. (2013), g)Verhamme et al. (2017), h)Leitet et al. (2011), i)Puschnig et al. (2017), j)Pardy et al. (2016),  
k)Heckman et al. (2015), l)<http://lasd.lyman-alpha.com/>, m)Izotov et al. (2021), n)Gazagnes et al. (2020)  
o)Micheva et al. (2010)

the ISM which can facilitate the escape of LyC. We discuss the escape fractions in more detail in the next section. We also show this figure color-coded with  $f_{\text{esc}}^{\text{Ly}\alpha}$  and  $f_{\text{esc}}^{\text{LyC}}$  in Figure A.3 and further discuss the relationship of escaping luminosities with escape fractions in § A.3. Moreover, we note that there are no galaxies with simultaneously very low Ly $\alpha$  and LyC luminosities. This is an effect of the stellar mass limit we imposed on our galaxies. We recall from § 2.2 that we analyze here all galaxies with  $M_{\star} > 10^6 M_{\odot}$ . We checked that if we do include less massive galaxies in our sample, they start to fill up this faint section of the plot, as they are very faint in both Ly $\alpha$  and LyC. The few extremely faint LAEs we do have in our sample are extremely gas deficient as we discussed in the previous paragraph, so it is easy for the LyC emission to escape from these systems, hence their intrinsic and the escaping LyC luminosities remain almost same.

### 3.4. Fraction of LyC leakers in LAE samples

As shown in the previous sections, Ly $\alpha$  and LyC luminosities are correlated with one another. Hence, we can wonder what fraction of LAEs would be detectable as LyC leakers, assuming typical LyC and Ly $\alpha$  detection limits. To answer this question, we divide galaxies in our sample with  $L_{\text{esc}}^{\text{Ly}\alpha}$  between  $10^{38}$  to  $10^{42.5}$  erg/s into 9 equally logarithmically-spaced bins (bin width 0.5 dex). In each group we calculate the median Ly $\alpha$  luminosity and the fraction of galaxies that have their  $L_{\text{esc}}^{900}$  luminosity higher than a given threshold value and report these fractions against their median  $L_{\text{esc}}^{\text{Ly}\alpha}$  in Figure 6. We do this exercise for three different threshold values of escaping LyC lu-

minosity,  $L_{\text{Threshold}} = 10^{37}, 10^{38}$  and  $10^{39}$  erg/s and we find that as galaxies become brighter in Ly $\alpha$ , the fraction of galaxies with  $L_{\text{esc}}^{900} > L_{\text{Threshold}}$  increases. For example, given a threshold LyC luminosity of  $10^{38}$  erg/s, 65% of LAEs with luminosity  $L_{\text{esc}}^{\text{Ly}\alpha} = 10^{41-41.5}$  erg/s and 97% of LAEs with luminosity  $L_{\text{esc}}^{\text{Ly}\alpha} = 10^{41.5-42}$  erg/s are bright in LyC emission. Granted, our simulated galaxies are at high redshift ( $z = 6 - 10$ ) but these results could be useful at lower redshifts, where LyC emission can be detected. Katz et al. (2019, 2020) have shown that low metallicity LyC leakers at  $z \sim 3$  are good analogues of EoR galaxies. The observed  $L_{\text{esc}}^{900}$  limit around  $z = 3$  is  $\sim 1.61 \times 10^{39}$  erg/s (flux limit  $2 \times 10^{-20}$  erg/s/cm $^2$ /Å or  $5.5 \times 10^{-4}$   $\mu\text{Jy}$ , Kerutt et al, in prep). At this threshold LyC luminosity, our analysis highlights that among LAEs with luminosity  $10^{41.5-42}$  erg/s,  $\sim 15\%$  of galaxies will be detected as LyC emitters.

### 3.5. Escape fraction

Figure 7 shows the  $f_{\text{esc}}^{\text{Ly}\alpha} - f_{\text{esc}}^{\text{LyC}}$  relationship of our simulated galaxies. Here we have plotted galaxies with progressively brighter sample selections: all galaxies ( $N = 1933$ ), galaxies with  $L_{\text{esc}}^{\text{Ly}\alpha} > 10^{39}$  erg/s ( $N = 1396$ ),  $> 10^{40}$  erg/s ( $N = 598$ ), and finally  $> 10^{41}$  erg/s ( $N = 150$ ). We find that if we consider all 1933 galaxies, including the very faint ones, the escape fractions of Ly $\alpha$ - LyC are very scattered and not correlated. The escape fractions occupy the whole space above the equality line, with only a few galaxies with  $f_{\text{esc}}^{\text{Ly}\alpha} < f_{\text{esc}}^{\text{LyC}}$ . However, if we limit our sample to only Ly $\alpha$  bright galaxies, the dispersion decreases. If we



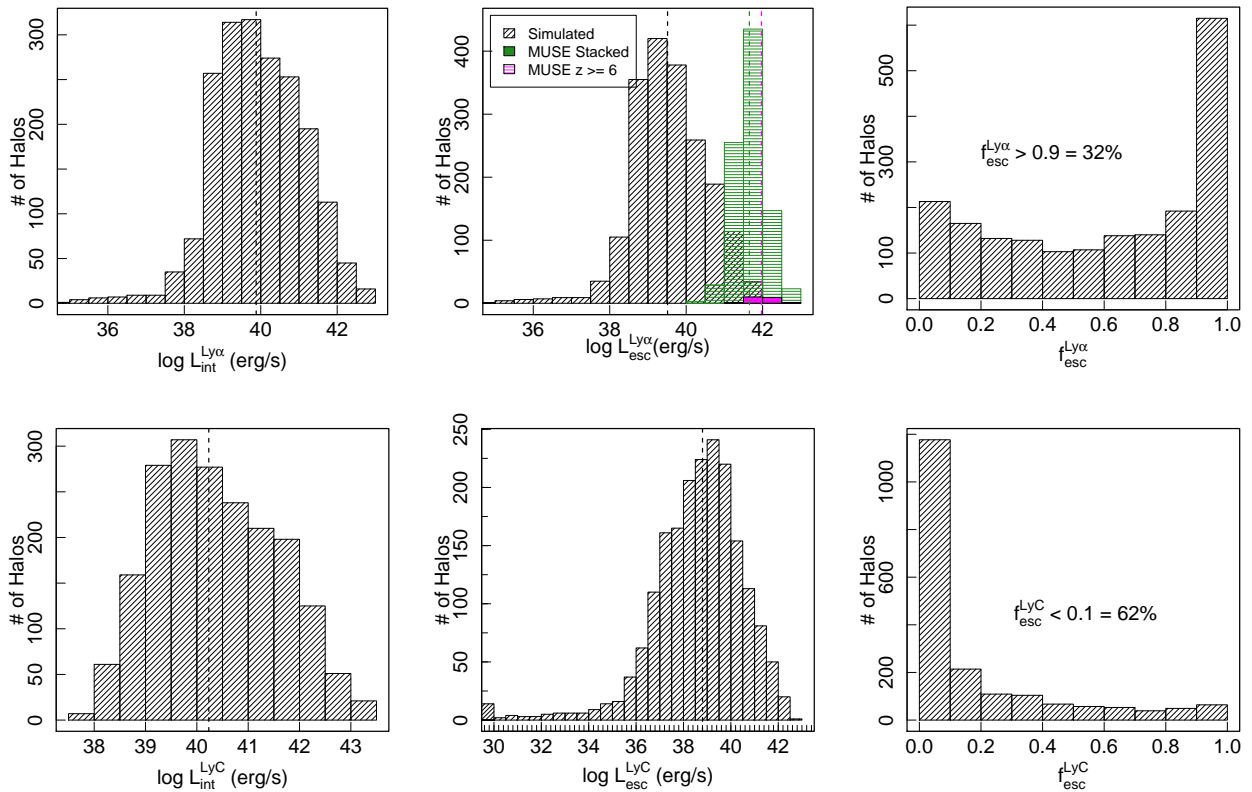


Fig. 4: Histograms of Ly $\alpha$  and LyC emission of our sample of 1933 galaxies. The top row shows the Ly $\alpha$  properties of our sample with intrinsic luminosity (left), escaping luminosity (middle) and escape fraction (right). The bottom row shows the same properties but for LyC radiation. In the middle panel of the top row, we also show the distribution of  $L_{\text{esc}}^{\text{Ly}\alpha}$  of galaxies observed in MUSE GTO surveys (MUSE galaxies are shown in green shade, among these, galaxies that are at  $z > 6$  are shown in pink). The dashed lines show their respective median values. The luminosities have a peaked distribution,  $f_{\text{esc}}^{\text{Ly}\alpha}$  has a bi-modal distribution with a strong peak at 1, and in most galaxies  $f_{\text{esc}}^{\text{LyC}}$  is very low.

include only the brightest galaxies with  $L_{\text{esc}}^{\text{Ly}\alpha} > 10^{41}$  erg/s, a positive correlation emerges between the two escape fractions. A linear regression of these bright galaxies yields the following model (with standard errors),  $f_{\text{esc}}^{\text{Ly}\alpha} = (1.02 \pm 0.07)f_{\text{esc}}^{\text{LyC}} + (0.24 \pm 0.02)$ . We find that the observed LCEs (Table 1) which are all bright LAEs ( $> 10^{41}$  erg/s), fall in the same escape fraction range as the simulated galaxies, which is an encouraging indication that escape fractions of our simulated galaxies are not significantly different from the escape fractions calculated from observed local LCEs. The correlation between  $f_{\text{esc}}^{\text{Ly}\alpha}$  and  $f_{\text{esc}}^{\text{LyC}}$  in the simulated bright galaxies and the observed ones is also very similar. This analysis indicates that the linear positive correlation of  $f_{\text{esc}}^{\text{Ly}\alpha}$  and  $f_{\text{esc}}^{\text{LyC}}$  that we find in observed LyC emitters (Verhamme et al. 2017) may be a selection bias which holds true only when we consider the brightest LAEs.

Additionally, in Figure 7 we find that in galaxies with very low  $f_{\text{esc}}^{\text{LyC}}$ , the  $f_{\text{esc}}^{\text{Ly}\alpha}$  can take any value between 0 to 1, but in galaxies with high  $f_{\text{esc}}^{\text{Ly}\alpha}$ , the  $f_{\text{esc}}^{\text{Ly}\alpha}$  is always very high. Conversely, galaxies with low  $f_{\text{esc}}^{\text{Ly}\alpha}$  always have low  $f_{\text{esc}}^{\text{LyC}}$ , but in galaxies with high  $f_{\text{esc}}^{\text{Ly}\alpha}$ ,  $f_{\text{esc}}^{\text{LyC}}$  can range from 0 to 1. Dijkstra et al. (2016) also found similar distributions using idealized models. We also note that  $f_{\text{esc}}^{\text{Ly}\alpha}$  is always greater than  $f_{\text{esc}}^{\text{LyC}}$ , except for a few outlier galaxies in our simulated sample where  $f_{\text{esc}}^{\text{Ly}\alpha} < f_{\text{esc}}^{\text{LyC}}$ . Theoretically it is expected that the Ly $\alpha$  escape fraction is greater than LyC because Ly $\alpha$  is only destroyed by dust while LyC can also be killed by HI atoms in the ISM. Ly $\alpha$  photons can scatter numerous times and have a greater possibility to find channels in the ISM with low HI column density through which they can escape the galaxy (Dijkstra et al. 2016). However in 6 out of 1933 (or 0.3%) of our galaxies we find that this is not the case.

Similarly for observed LCEs, although most of them have higher  $f_{\text{esc}}^{\text{Ly}\alpha}$ , in 2 out of 23 galaxies (8.7%),  $f_{\text{esc}}^{\text{Ly}\alpha}$  is less than  $f_{\text{esc}}^{\text{LyC}}$ . It is possible that in these systems there are dusty escape channels with low HI column density (the dust model allows for dust in ionized gas) such that it is optically thin to LyC photons but not to Ly $\alpha$ . We looked into these 6 simulated galaxies and found that these systems are comprised of interacting galaxies with complex configurations. The distribution of Ly $\alpha$  and LyC sources differ and they have escape channels of low density gas columns very close to the center where LyC production happens which can greatly aid its escape.

### 3.5.1. Median escape fraction at different Ly $\alpha$ luminosities

Since we have a large sample of galaxies with both Ly $\alpha$  and LyC radiative transfer, it is instructive to study how  $f_{\text{esc}}^{\text{Ly}\alpha}$  and  $f_{\text{esc}}^{\text{LyC}}$  correlate with the Ly $\alpha$  luminosity of galaxies. To analyze this, we have taken all galaxies in our sample with  $L_{\text{esc}}$  from  $10^{38}$  to  $10^{42.5}$  erg/s and divided the luminosities into 9 equally logarithmically-spaced bins (bin width 0.5 dex). We show the median escape fractions against median luminosities in Figure 8 and find that as the luminosity increases  $f_{\text{esc}}^{\text{Ly}\alpha}$  decreases. The drop in  $f_{\text{esc}}^{\text{Ly}\alpha}$  is fairly gradual and in our highest luminosity bins,  $10^{41.5-42.5}$  erg/s, the median value of  $f_{\text{esc}}^{\text{Ly}\alpha}$  is  $\sim 0.3$ . Brighter galaxies have higher mass in all components, including dust mass, and as dust content increases, more Ly $\alpha$  is absorbed by dust which reduces  $f_{\text{esc}}^{\text{Ly}\alpha}$ . At the bright end,  $L_{\text{esc}}^{\text{Ly}\alpha} \approx 10^{42}$  erg/s, our sample size decreases to only a couple of galaxies, owing to the limited simulation volume. Therefore, although the flattening of the median curves in bright LAEs suggest a similar value for

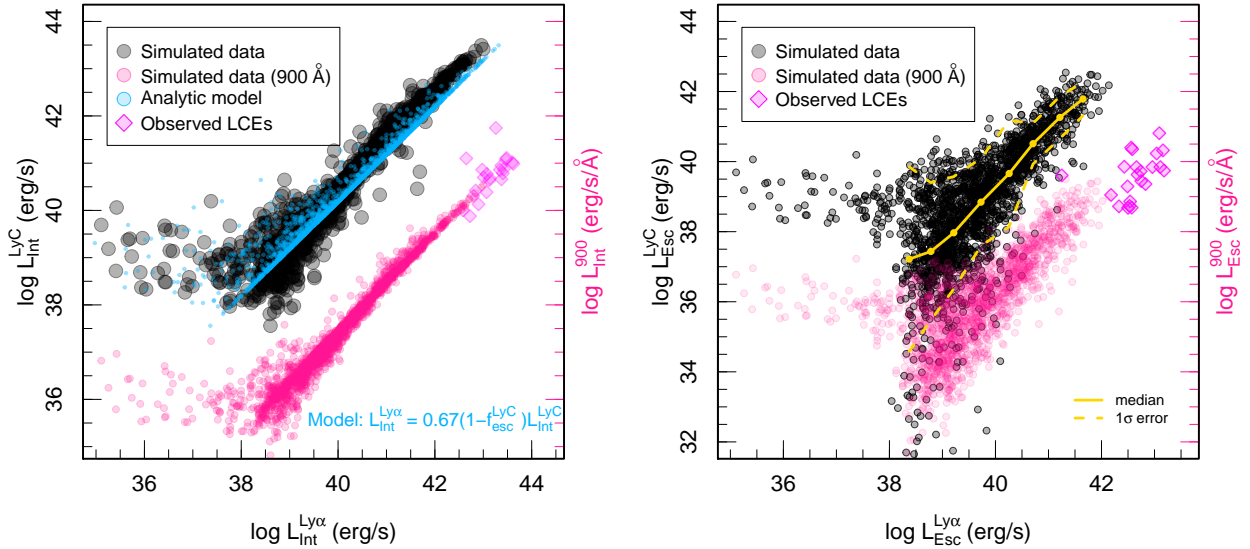


Fig. 5: Left- Intrinsic LyC luminosity of galaxies as a function their intrinsic Ly $\alpha$  luminosity. The black and the pink points show the total LyC luminosity (0 - 912Å) and the 900Å luminosity of the simulated galaxies respectively (§ 3.1). The diamond shaped magenta points show the observed LCEs described in Table 1. The sky blue points show the intrinsic luminosities derived from the analytic model described in the left figure. Right- Escaping LyC luminosity of galaxies as a function of their escaping Ly $\alpha$  luminosity. The solid yellow line shows the median  $L_{\text{esc}}^{\text{LyC}}$  as a function of  $L_{\text{esc}}^{\text{Ly}\alpha}$  (we divide the  $\log L_{\text{esc}}^{\text{Ly}\alpha}$  between 38 – 42 into groups of 0.5 dex each and find the median luminosities). The dashed line show  $1\sigma$  deviation from this, which illustrates the typical dispersion of the escaping luminosities.

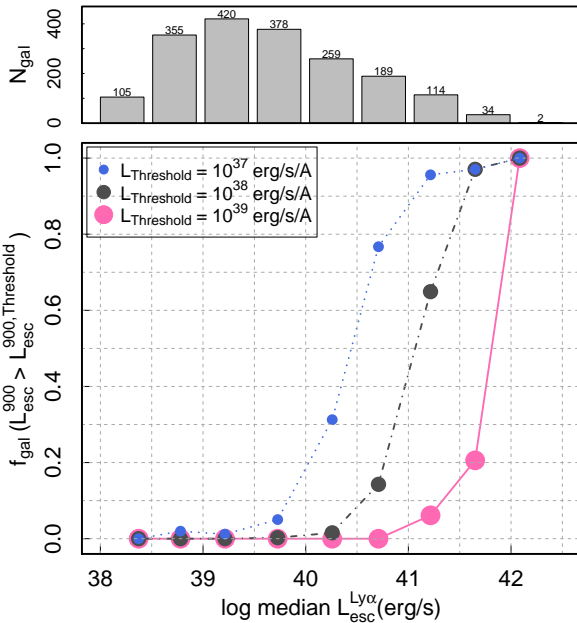


Fig. 6: Fraction of galaxies with  $L_{\text{esc}}^{900}$  luminosity above a threshold value against their median escaping Ly $\alpha$  Luminosity.

even brighter galaxies, we cannot make any concrete prediction for much brighter LAEs.

The median  $f_{\text{esc}}^{\text{LyC}}$  is low for all Ly $\alpha$  luminosities. In galaxies with  $L_{\text{esc}}^{\text{Ly}\alpha} < 10^{40.5}$  erg/s median  $f_{\text{esc}}^{\text{LyC}}$  is very low ( $\sim 0.02$ ), and in brighter galaxies it rises to  $\sim 0.1$ . A large fraction of faint LAEs have zero or very low  $f_{\text{esc}}^{\text{LyC}}$  as ionizing photons are absorbed by HI gas in the surrounding ISM which drives the median low (Chuniaux et al 2021, in prep).

The median Ly $\alpha$  luminosity of MUSE LAEs is around  $10^{41.6}$  erg/s, as shown in Figure 4. Our simulation predicts that the typical  $f_{\text{esc}}^{\text{Ly}\alpha}$  and  $f_{\text{esc}}^{\text{LyC}}$  of galaxies at this luminosity are around 0.3 and 0.1, respectively. Here we note that the Ly $\alpha$  luminosities of

MUSE galaxies are what we observe after Ly $\alpha$  has gone through IGM attenuation. The escaping Ly $\alpha$  luminosity of galaxies can be affected adversely by IGM attenuation, especially at  $z > 6$ . In our simulation, we have not considered the effects of IGM. Along the same lines, the observed luminosities of MUSE galaxies are what we measure along our line of sight whereas the simulated luminosities and escape fractions quoted here are global ones. We provide further discussion on the effects of IGM attenuation and line of sight variability in § 5.

### 3.6. Contribution of LAEs to reionization

In our analysis, we have both ionizing or LyC luminosities and the Ly $\alpha$  luminosities for a large sample of simulated galaxies in EoR, so we can investigate the role of LAEs as sources of cosmic reionization. Similar to the previous section, we take our sample of galaxies that have Ly $\alpha$  luminosities in the range  $10^{38} - 10^{42.5}$  erg/s range and divide them into 9 equally logarithmically-spaced bins (bin width 0.5 dex). For each group of galaxies we calculate their total escaping ionizing luminosities and plot it as a function of their median escaping Ly $\alpha$  luminosity in Figure 9 (left panel). We find that as galaxies become brighter, their total escaping LyC luminosity in each group increases. Since our 10 Mpc<sup>3</sup> simulation volume does not contain galaxies brighter than  $1.35 \times 10^{42}$  erg/s, (see also the Ly $\alpha$  luminosity functions, e.g. Fig 5, in Garel et al. 2021), there is a downward trend at the extreme bright end of our sample ( $10^{41.5-42}$ ). Therefore our sample size is too small to be conclusive about a peak at  $10^{41}$  erg/s. Nevertheless, the luminosity range of  $10^{38} - 10^{41}$  erg/s is well sampled, and we find that in this luminosity range, the brighter LAEs have higher total  $L_{\text{esc}}^{\text{LyC}}$ .

Now we calculate the total ionizing luminosity in the whole simulation box. We recall that our galaxy sample consists of galaxies with selection criterion provided in § 2.2, i.e. galaxies at level 1 and with  $M_{\star} > 10^6 M_{\odot}$ . Then, to be consistent in our comparisons, we estimate the total ionizing luminosity in the box by summing up the LyC luminosity ( $\Sigma L_{\text{esc}}^{\text{LyC}}$ ) of all galaxies at level 1, i.e. from a total of 8783 such galaxies in our simulation.

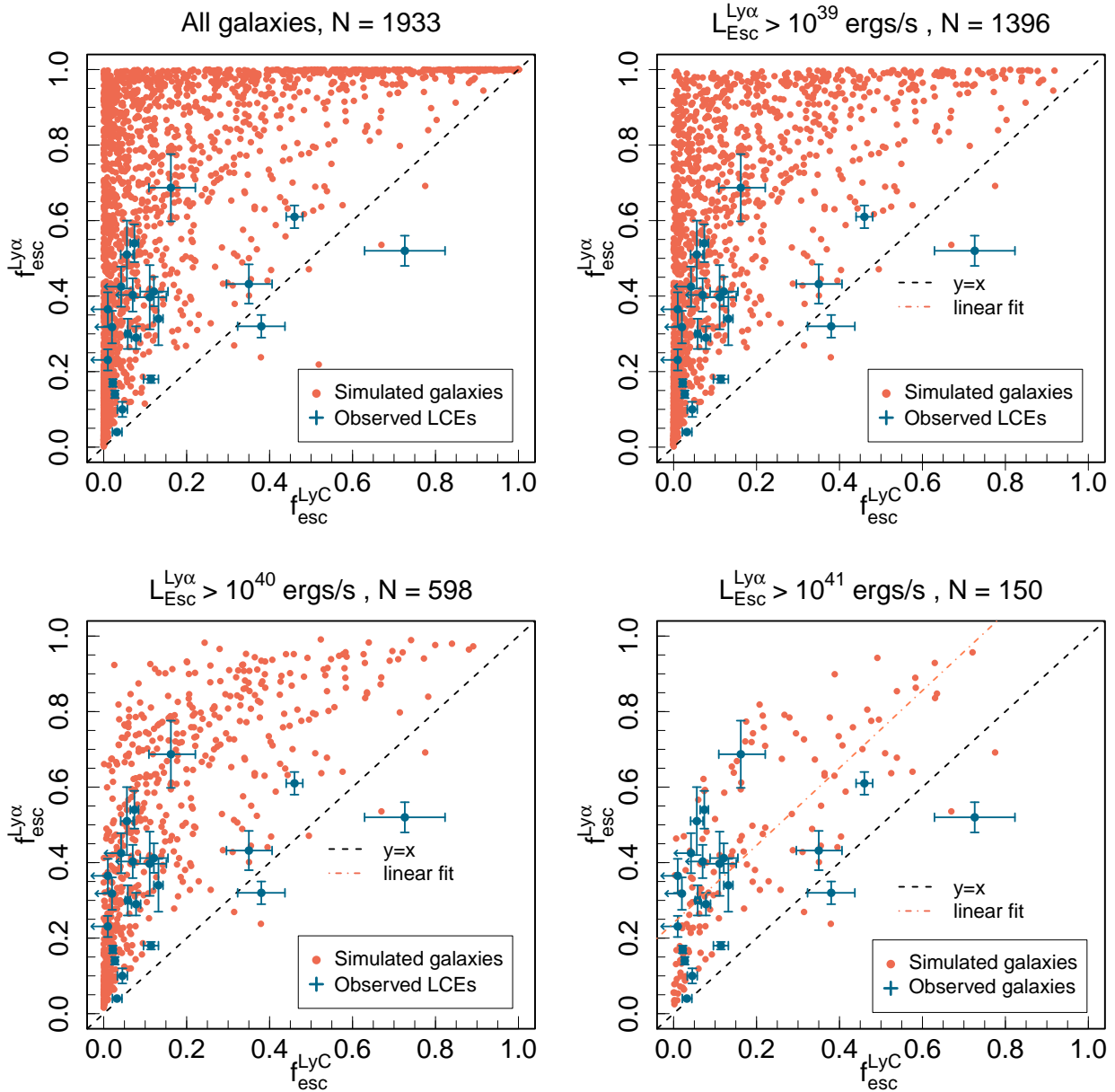


Fig. 7: Escape fractions of Ly $\alpha$  vs LyC. The plots here show progressively brighter sample selection, all galaxies (top left), and galaxies with  $L_{\text{esc}}^{\text{Ly}\alpha} > 10^{39}$  erg (top right),  $L_{\text{esc}}^{\text{Ly}\alpha} > 10^{40}$  erg (bottom left), and  $L_{\text{esc}}^{\text{Ly}\alpha} > 10^{41}$  erg (bottom right) respectively. In each plot we include the  $f_{\text{esc}}^{\text{Ly}\alpha}$  and  $f_{\text{esc}}^{\text{LyC}}$  of observed LCEs from Table 1 (blue points) with their error bars. The observed LCEs are all bright in Ly $\alpha$ , with  $L_{\text{esc}}^{\text{Ly}\alpha} > 10^{41}$  erg/s. For few galaxies the observed  $f_{\text{esc}}^{\text{LyC}}$  is an upper limit, these are marked by blue arrows. The dashed black line shows the  $y=x$  or equality line. The orange dashed line in the bottom right plot shows a linear fit of the simulated galaxies, which yields a slope of 1.02. This plot shows that if we include all galaxies, including very faint ones,  $f_{\text{esc}}^{\text{Ly}\alpha}$  and  $f_{\text{esc}}^{\text{LyC}}$  are very scattered and not correlated, but as we restrict our sample to progressively brighter LAEs, a correlation emerges. In the last panel, the simulated galaxies are in the same luminosity range as the observed ones ( $> 10^{41}$  erg/s), and they both show similar correlation.

We compute the contribution of galaxies with Ly $\alpha$  luminosity brighter than some limit to the total ionizing luminosity emitted by all simulated galaxies. The result of this is shown in the right panel of Figure 9. We find that simulated LAEs brighter than  $10^{40}$  erg/s ( $N = 598$ ) contribute more than 90% to the total ionizing luminosity of the box, even though the number of faint LAEs is much larger than bright ones. So 6.8% (598 out of 8783 galaxies) of the galaxies, which hosts 37% of total stellar mass, are responsible for more than 90% of the escaping ionizing radiation. Including all LAEs brighter than  $10^{38}$  erg/s ( $N = 1856$ ) can account for  $\approx 95\%$  of the total LyC luminosity.

In the MUSE Ultra Deep Field survey (Drake et al. 2017, Figure 5) at  $z=3$  the Ly $\alpha$  luminosity limit is  $10^{41.25}$  erg/s (50% completeness). Our analysis suggests that the LAEs with  $L_{\text{esc}}^{\text{Ly}\alpha} >$

$10^{41.25}$  erg/s at EoR could have contributed  $\sim 57\%$  of the ionizing radiation budget.

The faint LAEs produce a small amount of LyC intrinsically, compared to the bright LAEs (§ 3.3). From our analysis of escape fractions in the previous section we know that the median  $f_{\text{esc}}^{\text{LyC}}$  of all galaxies is rather low. Consequently the escaping LyC luminosities of faint LAEs is generally low. Therefore, we find that although faint LAEs are far more numerous, brighter LAEs as a group contribute more to the escaping ionizing luminosity. We have also explored the effect of the lower mass limit of the galaxies (discussed further in § A.4) on this reionization study and found that if we lower the mass limit of our galaxies from  $10^6$  to  $10^5 M_{\odot}$ , LAEs brighter than  $10^{40}$  contribute 97% of the total ionizing radiation (Fig A.4). This shows that although lowering the mass limit slightly increase these fractions, the differences are

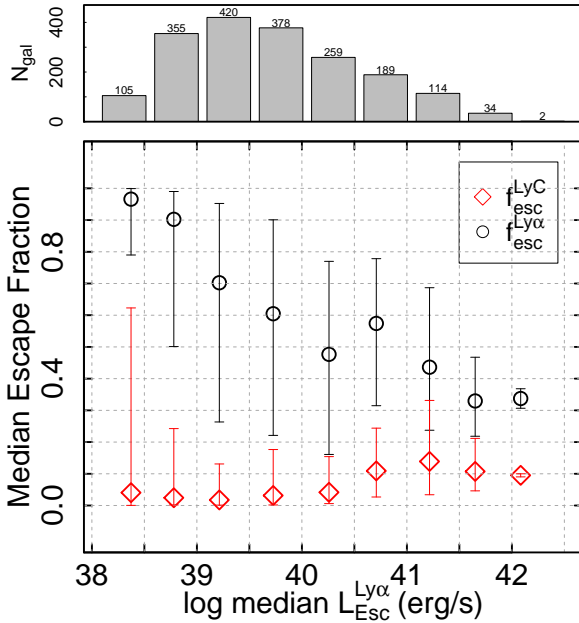


Fig. 8: Median  $f_{\text{esc}}^{\text{Ly}\alpha}$  and  $f_{\text{esc}}^{\text{LyC}}$  of galaxies in different escaping Ly $\alpha$  luminosity groups, plotted against the median Ly $\alpha$  escaping luminosity of the group. The vertical lines through each median point connect the 1st quantile (25%) and the 3rd quantile (75%) of the escape fraction distribution in that luminosity bin. The histogram shown on the top indicates the number of galaxies in the respective luminosity bins.

very small, our results thus converge. Therefore, we conclude that the primary sources of reionization are likely bright LAEs with  $L_{\text{esc}}^{\text{Ly}\alpha} > 10^{40}$  erg/s.

#### 4. Predicting LyC luminosities and escape fractions

The major goal towards studying the connection between Ly $\alpha$  and LyC emission from galaxies is to discover a correlation or develop a model that can estimate the LyC emission of EoR galaxies using the observable properties of galaxies, as the ionizing photons themselves cannot be observed.

In the previous section (§ 3) we have found that the escape fractions of Ly $\alpha$  and LyC are correlated in bright LAEs during the EoR, as observations have suggested, but when we include all LAEs in our sample, including the fainter ones, there is no correlation, which implies that the observed relation may be due to a selection bias. We also found that the intrinsic luminosities in Ly $\alpha$  and LyC are well correlated, whereas the escaping luminosities have a positive correlation but with much more dispersion, especially at the faint end. Thus in the quest for predicting the LyC emission, it is important to explore beyond the simple 1:1 correlation. Since we have a large dataset of galaxies with a number of their physical, Ly $\alpha$  and LyC properties we now investigate if it is possible to construct a statistical model that predicts the LyC emission using other properties, e.g. mass, SFR and Ly $\alpha$ .

Our galaxy sample is generally fainter (highest  $L_{\text{esc}}^{\text{Ly}\alpha} \sim 1.37 \times 10^{42}$  erg/s) and less massive (highest stellar mass  $M_{\star} \sim 1.33 \times 10^9 M_{\odot}$ ) than typical observed LAEs. The model that we can build with this data can be best applied to galaxies with properties similar to SPHINX galaxies. Whether this model can be applied to more massive or more luminous galaxies cannot be conclusively determined based on this study alone. Nevertheless, building such a predictive model for LyC using our data is an important first step towards a quantitative understanding of the contribution of galaxies to reionization. This analysis will also identify which galaxy properties are the main predictors of LyC emission and this can help identify strong LyC emitters among observed samples of EoR galaxies and guide future surveys.

#### 4.1. Multivariate model: A general framework

In our simulation we have a large data-set of hundreds of galaxies each with several physical and radiative properties that can be measured in their real-world counterparts. Given the large number of variables available, we aim to build a model that can be interpreted easily. Multivariate linear regression is a common statistical method for building such models, it is also straightforward to interpret and gain insights from the final model.

Recently Runnholm et al. (2020) did an analysis where they applied multivariate linear regression to observed galaxies at low-redshift to predict escaping Ly $\alpha$  luminosities using observed galaxy properties. In this study, they have analyzed galaxies in the Lyman Alpha Reference Sample (LARS) and extended LARS (e-LARS) containing 14 and 28 galaxies respectively, within a redshift range of  $0.028 \leq z \leq 0.18$  and found that using either observed or derived physical quantities it is possible to predict Ly $\alpha$  luminosities of galaxies accurately with their multivariate regression method. Keeping these considerations in mind, we choose to use multivariate linear regression for predicting LyC and Ly $\alpha$  properties of  $z \geq 6$  galaxies.

A multivariate linear regression model can be written as follows:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n \quad (3)$$

where  $x_1, x_2, \dots, x_n$  are independent variables or predictor variables (which would be a set of known properties of the galaxy) and  $y$  is the dependent variable or response variable that we want to predict, which in our case are LyC luminosities (intrinsic and escaping) and LyC escape fraction. The resulting model is characterized by the values of the coefficients in the equation, i.e.,  $\beta_0, \beta_1, \beta_2, \dots, \beta_n$ .

##### 4.1.1. Variables in the model

For our model building purpose, we explore various galaxy properties and we feed different combinations of them into the linear regression method.

Here we list the properties of galaxies that can be considered as  $x$ -variables or known variables and ones that are response or  $y$  variables.

1.  $M_{\text{Gas}}$  - Total gas mass of the halo. The gas mass is calculated by summing up the mass of all the gas cells inside halo radius. In our sample,  $M_{\text{Gas}}$  values ranges from  $10^{3.2} - 10^{9.7} M_{\odot}$ .
2.  $M_{\star}$  - Total stellar mass within  $0.3R_{\text{vir}}$  of the halo.
3. Galaxy  $R_{\text{vir}}$  - Virial radius of the main galaxy associated with halo. The median radius is  $\sim 0.3$  kpc (median halo  $R_{\text{vir}}$  is 3.9 kpc).
4. SFR<sub>10</sub> - Star formation rate of the halo averaged over last 10 Myr.
5. SFR<sub>100</sub> - SFR of the halo averaged over last 100 Myr.
6.  $\tau_{\star}$  - Mass-weighted mean stellar age of all stellar populations within 30% of the halo virial radius (median age  $\sim 102$  Myr).
7.  $Z_{\star}$  - Mass-weighted metallicity of stars within 30% of the halo virial radius.
8.  $Z_{\text{gas}}$  - Mass-weighted metallicity of gas within the halo virial radius.
9.  $L_{\text{int}}^{\text{Ly}\alpha}$  - Intrinsic Ly $\alpha$  Luminosity.
10.  $L_{\text{esc}}^{\text{Ly}\alpha}$  - Escaping Ly $\alpha$  Luminosity.
11.  $f_{\text{esc}}^{\text{Ly}\alpha}$  - Ly $\alpha$  Escape fraction, defined as the ratio of the escaping and intrinsic Ly $\alpha$  luminosity.
12.  $L_{\text{int}}^{\text{LyC}}$  - Intrinsic ionizing Luminosity.
13.  $L_{\text{esc}}^{\text{LyC}}$  - Escaping ionizing Luminosity.
14.  $f_{\text{esc}}^{\text{LyC}}$  - LyC Escape fraction.

We show the histogram of these variables for our sample of galaxies used in building multivariate models in Figure A.5.

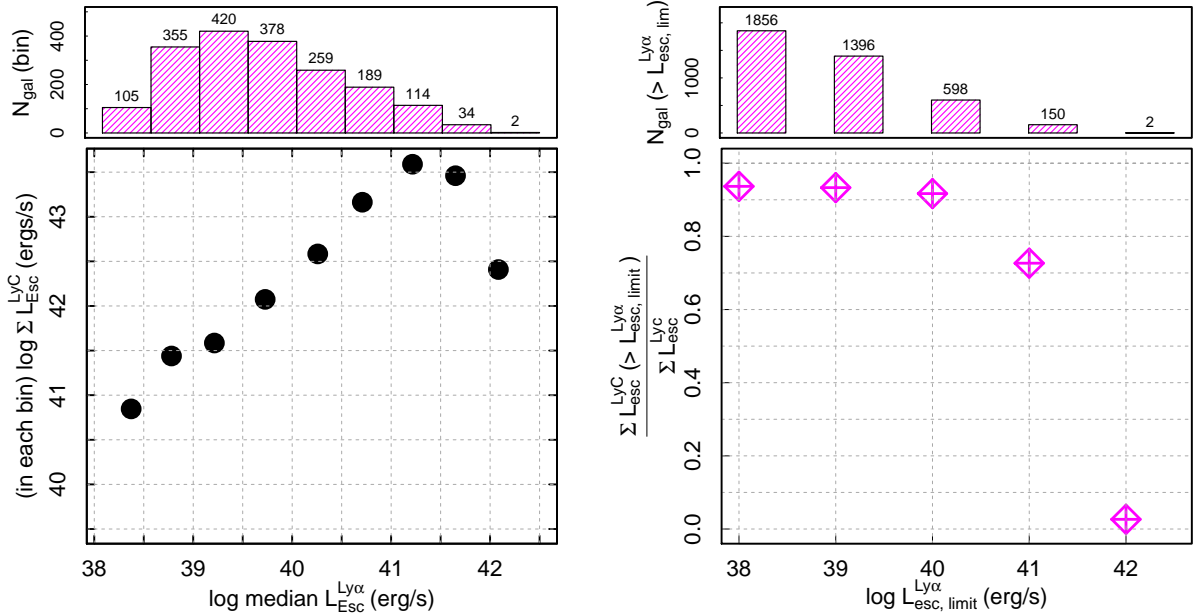


Fig. 9: Left : Total escaping LyC luminosity of galaxies grouped by their Ly $\alpha$  luminosities as a function of their median Ly $\alpha$  luminosity. The histogram above shows the number of galaxies in the corresponding bins below. Right : Conditional total escaping LyC luminosity of galaxies brighter than a given Ly $\alpha$  luminosity limit as a function of the Ly $\alpha$  luminosity limit. The histogram above indicates the number of galaxies with  $L_{\text{Esc}}^{\text{Ly}\alpha} > L_{\text{Esc, limit}}^{\text{Ly}\alpha}$  the respective Ly $\alpha$  luminosity limits below, so these are numbers of galaxies that have been used to calculate the respective fractions.

#### 4.1.2. Preparing the data

When we use multivariate methods for constructing a predictive model, it is important that all variables involved in the model have the same order of magnitude. However, standardizing the measurement scales has no impact on the validation and interpretation of the models. Data standardizing comprises of various techniques, for example, z-score standardization where if the data is Gaussian it is shifted so that the new dataset is centered around 0 and has a standard deviation of 1 ( $z = \frac{x-\mu}{\sigma}$ , where  $z$  = new data,  $x$  = old data,  $\mu$  =  $\langle x \rangle$  and  $\sigma$  = standard deviation of  $x$ ) or min-max standardization where the data is scaled between 0 to 1 ( $z = \frac{x-x_{\min}}{x_{\max}-x_{\min}}$ ). In our analysis, not all of the galaxy properties have a Gaussian distribution (as can also be seen from Figure A.1). More importantly, our variables typically cover many orders of magnitudes in range. So for standardizing our data, we first take logarithmic values of all variables and then subtract the median value from them to center them. So for any variable  $x$  we scale it to  $x_{\text{scaled}}$  or  $x_s$  by,

$$x_s = \log(x) - \text{median}(\log(x)) \quad (4)$$

The next steps for constructing the model are carried out with these scaled variables (equation 3 will be applied on scaled variables for building the models). The variables we have plotted in Figure 10 (and A.6) and discussed in §4.2 are these scaled variables.

#### 4.1.3. Estimating the quality of the fit

There are several metrics that can be used to quantify how suitable the model is or how well it fits the data. A popular statistical metric for the multivariate regression model is the  $R^2$  which is a measure of how much of the response variance is explained by the model, i.e. the linear combination of the predictors. It is mathematically defined as:

$$R^2 = 1 - \frac{\sum (y_i - f_i)^2}{\sum (y_i - \bar{y})^2} \quad (5)$$

where  $y_i$  is the actual  $y$  value, i.e.  $y$  value from our simulation of  $i$ -th halo,  $\bar{y}$  is the mean value of these  $y$  values, and  $f_i$  is the

predicted value for the  $i$ -th halo computed using the model.  $R^2 = 0$  means that the model explains no response variance and  $R^2 = 1$  means that the model explains all the response variance, i.e. it can predict  $y$  exactly. So the closer the  $R^2$  value is to 1, the better the model.

Although  $R^2$  is a widely used metric of model performance, it should be noted that the value of  $R^2$  always increases, however slightly, when more and more variables are added to the model. Therefore, in models where the number of  $x$ -variables is large,  $R^2$  may slightly overestimate the model performance. To ensure that our metric does not depend on the number of  $x$ -variables, we define the adjusted  $R^2_{\text{adj}}$  as,

$$R^2_{\text{adj}} = 1 - (1 - R^2) \frac{n - 1}{n - p - 1} \quad (6)$$

where  $n$  is the number of data points (galaxies) and  $p$  is the number of  $x$ -variables in the model (see e.g. Feigelson & Babu 2012). The adjusted  $R^2$  increases only when the addition of a  $x$ -variable increases the  $R^2$  more than it would just by chance. The value of  $R^2_{\text{adj}}$  will always be equal to or less than  $R^2$ . From here onward, whenever we mention  $R^2$  and its values, either in text or in figures, we mean the  $R^2_{\text{adj}}$ , unless otherwise specified.

#### 4.1.4. Finding the most important predictors

We perform a stepwise forward and backward selection method to determine which  $x$  variables are the most important for predicting  $y$ . In forward selection, the model takes the  $x$ -variables one by one, and inspects which of them lead to the largest value of  $R^2$  by itself and classifies that as the most important  $x$ -variable (rank 1). Then the model adds each of the remaining  $x$ -variables one by one to rank 1, and the variable that produces the largest increase in  $R^2$  value is the second most important  $x$ -variable (rank 2). This continues until all the variables have been added and a ranked choice of  $x$ -variables has been made. In the backward selection method, the model starts with all  $x$ -variables and then determines which one variable removal decreases the value of  $R^2$  the least, this is least important variable. The process continues until all but one variable have been removed and a ranking has been generated. We use both methods on our data-set.

#### 4.1.5. Validating the models

After building the regression models it is important to estimate the performance of the model on various datasets. To do so, we use repeated k-fold cross validation method to test the model performance. First the entire dataset is randomly divided in k subsets, where the number k (typically 5 or 10) can be specified. Then we reserve one subset as test data and estimate the model using the rest of the subsets which act as training data. We then use this estimated model on the test data in order to calculate the fit/error indicators (which can be  $R_{\text{adj}}^2$  or root-mean-square-estimate or mean absolute error). We repeat this process k-times and ensure that each of the subset has acted as the test dataset once. Then we calculate the average of these indicators from these k measurements of errors. This whole process of dividing into test-train datasets and computing the average indicators is then performed multiple times and finally we average all the indicators corresponding to each model and compare it with the  $R_{\text{adj}}^2$  of the full model.

#### 4.2. Application to SPHINX galaxies

From equation 3, we can deduce that the multivariate linear model is suitable if some (or all)  $x$ -variables individually vary linearly with  $y$ , i.e. at least for some variables  $y \propto x_i$ . If none of the  $x$ -variables have any linear correlation with  $y$ , it is unlikely that a linear combination of them can determine  $y$ . Therefore, we first explore if individual correlation between  $y$  and any  $x$ -variable exists.

Such an exploratory plot is shown in Figure 10 where we plot the response variable  $L_{\text{esc}}^{\text{LyC}}$  vs each of the galaxy properties. From this figure we find that  $L_{\text{esc}}^{\text{LyC}}$  correlates well with  $L_{\text{esc}}^{\text{Ly}\alpha}$  and SFR, along with some other weaker correlations. Similar plot for  $L_{\text{int}}^{\text{LyC}}$  and  $f_{\text{esc}}^{\text{LyC}}$  is provided in appendix (Figure A.6 and A.7) where we see that  $L_{\text{int}}^{\text{LyC}}$  is strongly correlated to SFR and  $L_{\text{esc}}^{\text{Ly}\alpha}$  and weakly correlated to mass and stellar age and  $f_{\text{esc}}^{\text{LyC}}$  is correlated to  $L_{\text{esc}}^{\text{Ly}\alpha}$ . This preliminary inspection shows that a multivariate linear regression can be a good model for predicting LyC.

##### 4.2.1. Sample Selection

Before we delve into regression modelling, we examine the galaxy dataset to select a galaxy sample that can be used for building the model. The initial dataset contains 1933 galaxies, which is the sample of all galaxies with stellar mass  $\geq 10^6 M_{\odot}$  at  $z = 6, 7, 8, 9$ , and 10 (§2.2).

As discussed above and from Figure A.6 it is clear that the star formation rate, especially recent (over last 10 Myr) SFR or SFR<sub>10</sub>, has a strong linear correlation with intrinsic LyC luminosity, and it is also correlated well with the escaping luminosity of LyC (Figure 10). This is also expected from theoretical studies (Stanway et al. 2016; Raiter et al. 2010; Schaerer 2003; Partridge & Peebles 1967) that show star formation is the main driver for the production of both Ly $\alpha$  and LyC photons. In our dataset, there are some galaxies (943 out of 1933, most of which are faint LAEs) which have no recent star formation, i.e., the average star formation rate over the last 10 Myr, SFR<sub>10</sub> = 0 (Figure 4, right panel). So for building our models, we exclude these non star-forming galaxies and with this criterion, there are 990 galaxies left in our dataset.

Next we investigate this modified dataset (galaxies with non-zero SFR) for any significant outliers. We find that there are some clear outliers in the distribution of  $f_{\text{esc}}^{\text{LyC}}$  with values as low as  $10^{-20}$ . We remove galaxies with  $f_{\text{esc}}^{\text{LyC}} < 10^{-6}$  from the dataset, after which the  $f_{\text{esc}}^{\text{LyC}}$  distribution is free of outliers. This leads to a dataset of 940 galaxies. We find no significant outliers in other galaxy properties. Incidentally, we note that all of the galaxies in

this final dataset of 940 galaxies have both intrinsic and escaping Ly $\alpha$  luminosities  $> 10^{38}$  erg/s.

##### 4.2.2. Building the models and the most important variables

Our main goal is to predict the LyC luminosities and  $f_{\text{esc}}^{\text{LyC}}$  using other properties. However, for many galaxies at high redshift the observation of Ly $\alpha$  luminosity can also be difficult, owing to increasing IGM opacity. Moreover, estimation of the intrinsic Ly $\alpha$  luminosity and hence  $f_{\text{esc}}^{\text{Ly}\alpha}$  is also challenging at all redshifts, as these are not observables and must be derived using stellar models which can have many underlying assumptions. So it can be useful to also build models for predicting these Ly $\alpha$  emissivities which may complement existing methods.

Therefore, we explore the full predictive power of multiple linear regression models with our dataset and we aim to build models to predict the following 6 quantities:

- $L_{\text{int}}^{\text{LyC}}, L_{\text{esc}}^{\text{LyC}}, \text{ and } f_{\text{esc}}^{\text{LyC}},$
- $L_{\text{int}}^{\text{Ly}\alpha}, L_{\text{esc}}^{\text{Ly}\alpha}, \text{ and } f_{\text{esc}}^{\text{Ly}\alpha}.$

We investigate several combinations of physical parameters that we can access in the simulation to build a good predictive model. We calculate the performances of these models using the metric  $R_{\text{adj}}^2$  and our most relevant model results are summarized in Table 2.

**Model 1** - In Model 1, as predictors we supply all physical galaxy properties (GP), i.e. items 1 - 8 from our list in § 4.1.1, namely gas mass, stellar mass, galaxy radius, SFR<sub>10</sub>, SFR<sub>100</sub>, stellar age, stellar and gas metallicity. We find that given only the physical properties of galaxies, we can predict the intrinsic LyC luminosity quite accurately ( $R^2 = 0.87$ ) but the emerging luminosity and the escape fraction cannot be modeled very well ( $R^2 = 0.53$  and  $0.26$  respectively). Conversely, both Ly $\alpha$  intrinsic ( $R^2 = 0.88$ ) and escaping ( $R^2 = 0.71$ ) luminosities can be predicted quite well with galaxy properties.

**Model 2** - When we add Ly $\alpha$  escaping luminosity to our input list of predictors, (Model 2 in Table 2) we find that in addition to the intrinsic luminosities, now the LyC escaping luminosity is also predicted with high accuracy, with  $R^2 = 0.85$ . The average error (root mean square error, RMSE, is the average difference between the predicted and actual value) in predicting the  $L_{\text{esc}}^{\text{LyC}}$  is a factor of  $\sim 4$  (RMSE = 0.62 in log scale, Figure 11). Both  $f_{\text{esc}}^{\text{Ly}\alpha}$  and  $f_{\text{esc}}^{\text{LyC}}$  are also fairly well predicted with this model, with an  $R^2$  value of 0.69 and 0.64 respectively.

We consider Model 2 as our fiducial model and we show the predicted intrinsic and escaping LyC luminosities and  $f_{\text{esc}}^{\text{LyC}}$  from Model 2 in Figure 11 against the observed values from the simulation. In each of these plots we also show the 95% confidence interval and the 95% prediction interval. The confidence interval signifies that given a set of predictor values, i.e.  $x$ -values, the mean of the response variable or  $y$ , will fall within this interval with 95% confidence. On the other hand, the prediction interval tells us where the next individual  $y$  value will fall. Given a set of  $x$ -values, an individual  $y$  value will fall within the predictor interval with 95% confidence. The prediction interval accounts for both the uncertainty of the estimation of population mean as well as the variation of the individual  $y$ -values. Hence, the predictor interval is always wider than the confidence interval. We see in Figure 11 that most of the observed (in our simulation) values of  $y$  do indeed lie within the 95% predictor interval of our model.

Figure 11 shows both intrinsic and escaping luminosities are well predicted. We give here the equation for predicting  $L_{\text{esc}}^{\text{LyC}}$  obtained using this model:

$$\begin{aligned} \log_{10} L_{\text{esc}}^{\text{LyC}} = & 39.30 + 2.08 \log_{10}(L_{\text{esc}}^{\text{Ly}\alpha}/10^{40}) \\ & - 1.11 \log_{10}(M_{\text{Gas}}/10^8) + 0.85 \log_{10}(Z_{\text{Gas}} \times 10^3) \\ & - 0.20 \log_{10}(SFR_{10} \times 10^2) - 0.21 \log_{10}(Z_{\star} \times 10^3) \\ & + 0.16 \log_{10}(M_{\star}/10^6) - 0.16 \log_{10}(\text{Age}_{\star}/10^2) \end{aligned} \quad (7)$$

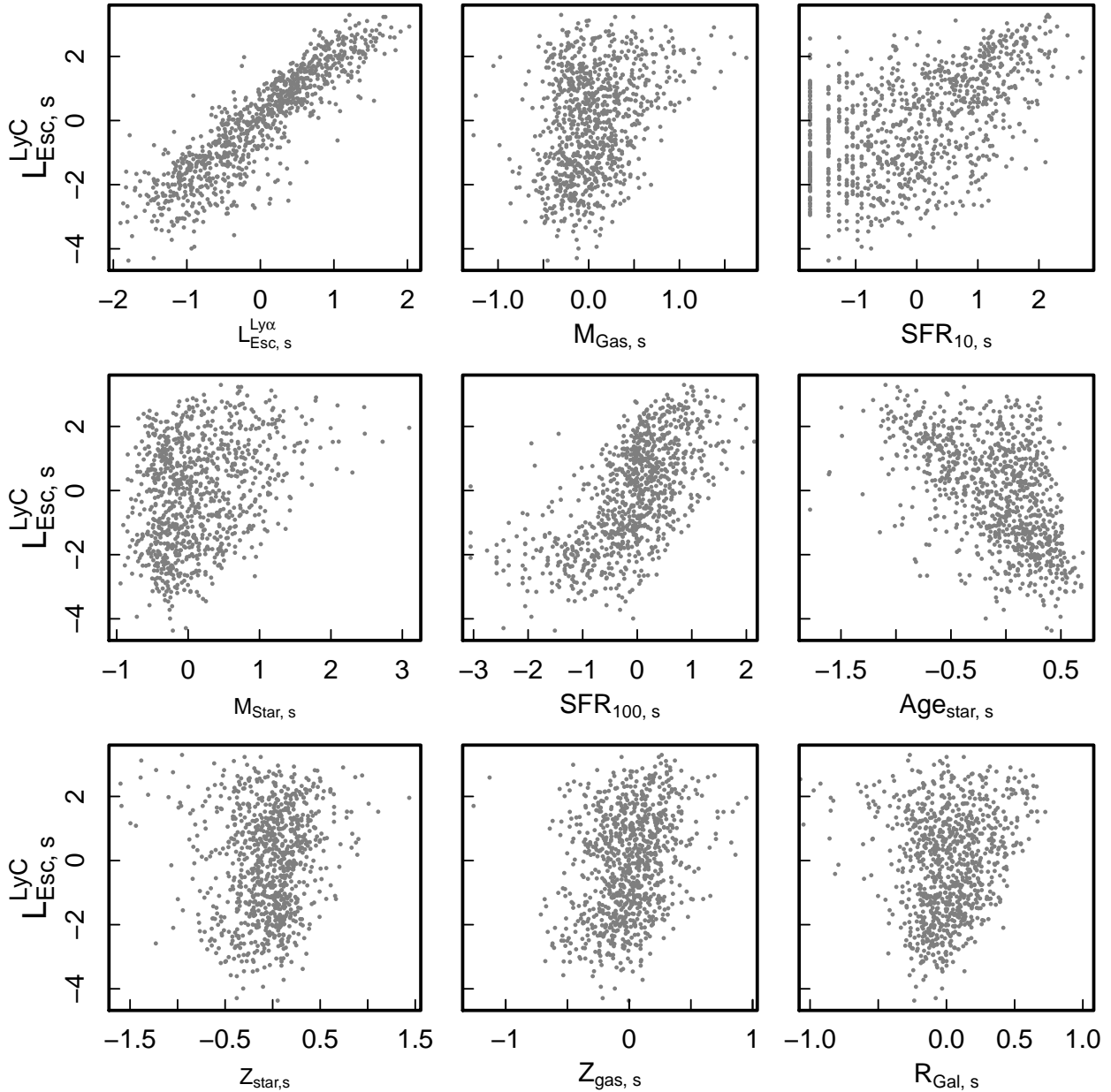


Fig. 10: LyC escaping luminosity vs. each of the 9 galaxy variables. All variables plotted here are scaled (as denoted by the subscript  $s$ ) using equation 4, as described in § 4.1.2. The particular definitions of the parameters are as follows:  $L_{\text{Esc},s}^{\text{LyC}} = \log(L_{\text{esc}}^{\text{LyC}}/\text{erg s}^{-1}) - 39.25$ ,  $L_{\text{Esc},s}^{\text{Ly}\alpha} = \log(L_{\text{esc}}^{\text{Ly}\alpha}/\text{erg s}^{-1}) - 40.11$ ,  $M_{\text{Gas},s} = \log(M_{\text{Gas}}/M_{\odot}) - 8.01$ ,  $SFR_{10,s} = \log(\text{SFR}_{10}/M_{\odot} \text{ yr}^{-1}) + 2.25$ ,  $M_{\text{Star},s} = \log(M_{\text{Star}}/M_{\odot}) - 6.03$ ,  $SFR_{100,s} = \log(\text{SFR}_{100}/M_{\odot} \text{ yr}^{-1}) - 0.01$ ,  $\text{Age}_{\text{Star},s} = \log(\text{Age}_{\text{Star}}/\text{Myr}) - 2.01$ ,  $Z_{\text{star},s} = \log(Z_{\text{star}}/Z_{\odot}) + 3.59$ ,  $Z_{\text{gas},s} = \log(Z_{\text{gas}}/Z_{\odot}) + 3.56$ , and  $R_{\text{Gal},s} = \log(R_{\text{vir}}/\text{kpc}) - 0.29$ .

Table 2:  $R_{\text{adj}}^2$  for predicting different variables with different models. Here Galaxy Properties (GP) refers to physical properties of the galaxies described in § 4.1.1 (i.e. items 1 - 7).

Model	$L_{\text{int}}^{\text{Ly}\alpha}$	$L_{\text{esc}}^{\text{Ly}\alpha}$	$f_{\text{esc}}^{\text{Ly}\alpha}$	$L_{\text{int}}^{\text{LyC}}$	$L_{\text{esc}}^{\text{LyC}}$	$f_{\text{esc}}^{\text{LyC}}$
1. GP (Galaxy Properties)	0.8758	0.7061	0.2812	0.8665	0.5336	0.2631
2. GP + $L_{\text{esc}}^{\text{Ly}\alpha}$	0.9031	NA	0.6886	0.8969	0.8516	0.6561

Here the luminosity is in  $\text{erg/s}$ , mass is in  $M_{\odot}$ , SFR unit is  $M_{\odot}/\text{yr}$ , stellar age is in Myr and metallicity unit is solar metallicity.

Here we note that in our models we have included both the gas mass and the gas metallicity. Since the dust content is modelled by using these factors (as described in § 2.4), including the dust in addition to the other parameters does not give us additional information (we tested this and found that inclusion of dust changes the  $R_{\text{adj}}^2$  by less than 0.01%).

**Most important variables:** In the models described above, we have supplied 7 or 8 galaxy properties for predicting various Ly $\alpha$  and LyC quantities. However, observing and determining many galaxy properties at high redshift can be extremely challenging. Thus it is necessary to identify which of the  $x$ -variables is the most important in predicting  $y$ . Here we discuss the ranking of most important predictors in the context of Model 2 and the response variables  $L_{\text{int}}^{\text{LyC}}$ ,  $L_{\text{esc}}^{\text{LyC}}$  and  $f_{\text{esc}}^{\text{LyC}}$ .

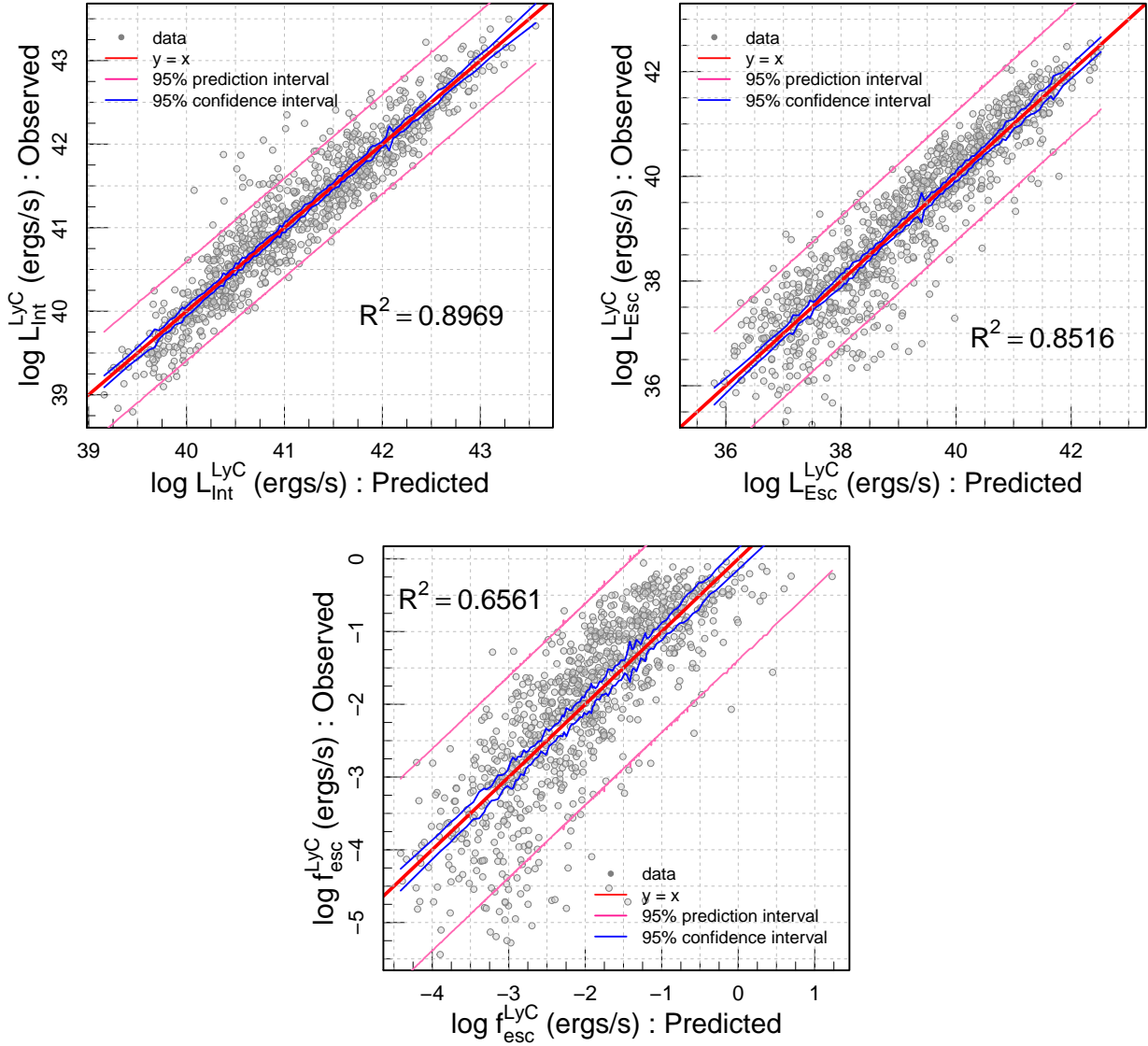


Fig. 11: Prediction of intrinsic luminosity, escaping luminosity and escape fraction of LyC from Model 2, where the input variables are the physical galaxy properties and the escaping Ly $\alpha$  luminosity. The  $R^2$  value for each fitting is noted in the plots. The red lines show the 1:1 correlation or  $y=x$  line. The pink lines show the 95% prediction interval and the blue lines show the 95% confidence interval.

We present the results of the ranking process described in § 4.1.4 in Table 3, listing the most important variables with their ranks and their  $R^2_{\text{adj}}$  value. The  $R^2_{\text{adj}}$  value associated with the  $n$ -th rank variable is the  $R^2_{\text{adj}}$  the model produces including the first to  $n$ -th rank variable. The adjusted  $R^2$  increases only when the addition of a  $x$ -variable increases the  $R^2$  more than it would be just by chance, otherwise it actually decreases with variable addition (§4.1.3). In a ranking table, such as Table 3, when the adjusted  $R^2$  reaches a peak value, the model has reached its best predictive power. We perform both stepwise forward selection and backward selection for the ranking (§4.1.4), and find that both processes give the same ranking in all cases, which suggests that our ranking is stable.

We find that 88% of the variance in  $L_{\text{int}}^{\text{LyC}}$  can be explained if we only use  $\text{SFR}_{10}$  and  $L_{\text{esc}}^{\text{Ly}\alpha}$ , such that these two values alone can provide a reliable prediction of the intrinsic ionizing power of galaxies. For escaping LyC, knowing the escaping Ly $\alpha$  luminosity is the most important factor and combining this with gas mass, gas metallicity and  $\text{SFR}_{10}$  can account for 85% of the variance. Lastly, the most important three predictors of  $f_{\text{esc}}^{\text{LyC}}$  are  $L_{\text{esc}}^{\text{Ly}\alpha}$ ,  $\text{SFR}_{10}$  and gas mass, as these three can explain 63% of the response variance. In the case of  $f_{\text{esc}}^{\text{LyC}}$ , variables with rank 1 - 6

increase the  $R^2_{\text{adj}}$  (up to 0.6569), but the addition of more properties decreases the model performance. Similarly, we find that in models for predicting  $L_{\text{int}}^{\text{LyC}}$ , galaxy radius (rank 9) and for predicting  $L_{\text{esc}}^{\text{LyC}}$ ,  $\text{SFR}_{100}$  and radius (rank 8 and 9) are not important.

#### 4.2.3. Minimal model

Going one step further, we note that it is extremely difficult to observe gas properties in reionization era galaxies. Among the rest of the predictors used in our models so far, the observed LCEs we have discussed in § 3.1 and listed in Table 1 have only 3 predictors available, namely  $L_{\text{esc}}^{\text{Ly}\alpha}$ ,  $\text{SFR}_{10}$  and  $M_{\star}$ . It is now interesting to explore if a model built with only these 3 predictors can predict LyC quantities. We build a minimal model with these three predictors only (Model 3) and list the resulting model performances in Table 4. We find that here also  $L_{\text{esc}}^{\text{LyC}}$  is predicted with a high accuracy,  $R^2 = 0.80$  and the average error is a factor of  $\text{RMSE} \sim 5.24$ . The intrinsic luminosities are also predicted



very well, with fair performances for escape fractions. The equation for  $L_{\text{esc}}^{\text{LyC}}$  we get with this model is:

$$\begin{aligned} \log_{10} L_{\text{esc}}^{\text{LyC}} &= 38.94 \\ &+ 2.03 \log_{10}(L_{\text{esc}}^{\text{Ly}\alpha} / 10^{40}) \\ &- 0.15 \log_{10}(M_{\star} / 10^6) \\ &- 0.23 \log_{10}(\text{SFR}_{10} \times 10^2) \end{aligned} \quad (8)$$

The equation for  $L_{\text{int}}^{\text{LyC}}$  from this model can be written as:

$$\begin{aligned} \log_{10} L_{\text{int}}^{\text{LyC}} &= 40.96 \\ &+ 0.49 \log_{10}(L_{\text{esc}}^{\text{Ly}\alpha} / 10^{40}) \\ &- 0.08 \log_{10}(M_{\star} / 10^6) \\ &- 0.49 \log_{10}(\text{SFR}_{10} \times 10^2) \end{aligned} \quad (9)$$

The units of the quantities are the same as described in § 4.2.2. The ranking of the most important predictors for  $L_{\text{esc}}^{\text{LyC}}$  with this minimal model is shown in Table 5, where we find that  $L_{\text{esc}}^{\text{Ly}\alpha}$  is has rank 1, followed by  $\text{SFR}_{10}$  and  $M_{\star}$ , same results as we found with Model 2 (§ 4.2.2) also.

#### 4.3. Fitting the model to observed data

Finally, we explore if such a model can be fitted to real observed data. In § 3 we have listed the properties of known Ly $\alpha$  and LyC emitters in Table 1. These galaxies have observations of their stellar mass, star formation rate, Ly $\alpha$  luminosity and their LyC luminosity at 900Å. As discussed in § 3.3 and shown in Figure 3 these observed LCEs are more massive, have higher SFR and they are brighter in Ly $\alpha$  and LyC than the SPHINX galaxies. They are also observed at low redshifts,  $z \leq 0.45$  whereas the SPHINX galaxies are at  $z = 6 - 10$ . The simulated luminosities and escape fractions are angle-averaged quantities whereas observations are, of course, directional (further discussion in § 5). Nevertheless, this is the only observed sample we currently have with both LyC and Ly $\alpha$  observations, so we evaluate our predictive model on these galaxies.

Since the observed galaxies have only 3 predictors available, we use our minimal model (model 3) described in § 4.2.3 and use Equation 8 for predicting the  $L_{\text{esc}}^{\text{LyC}}$  of these observed LCEs. In Figure 12 we show the predicted  $L_{\text{esc}}^{\text{LyC}}$  from this model vs. the LyC luminosity that is derived from observations of  $L_{\text{esc}}^{900}$  (by multiplying the observed  $L_{\text{esc}}^{900}$  with a factor of 1434, the median value of the ratios  $L_{\text{esc}}^{\text{LyC}} / L_{\text{esc}}^{900}$  derived from our simulation, § 3.3). We find that the predicted luminosities are generally close to the observed values. In some cases the model over predicts the escaping luminosity, probably due to the differences in the physical properties between these observations and the SPHINX galaxy sample. Models performs best when the given input properties are inside the range of the training data (the ranges of properties for our SPHINX sample are shown in Figure A.5), otherwise it needs to be extrapolated. The outlier with low predicted LyC is the galaxy Haro 11 which is located at  $z = 0.021$ , much closer than other observations which may affect the galaxy properties.

##### 4.3.1. Cross Validation of the models

We have built these models using all 940 eligible galaxies available in our simulation dataset. To check the model validity, we need to estimate the accuracy of these models when applied on other, new data, that is not part of the dataset used in building the models. The most straightforward way to do this is to apply this model to other new datasets where all of our desired input and output variables are available in order to readily test the difference between the prediction from models and the actual values.

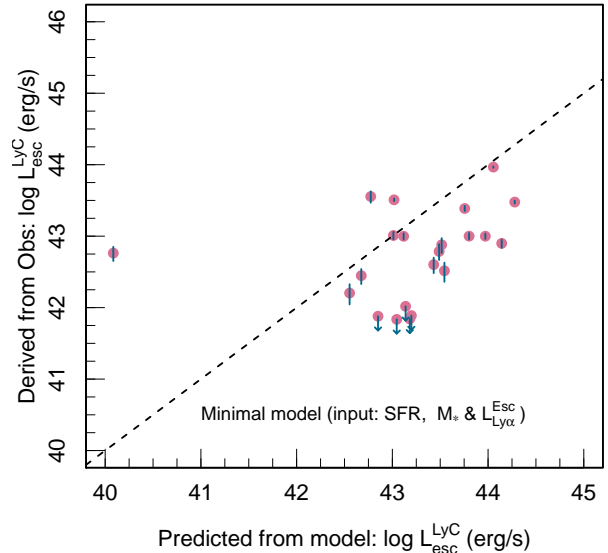


Fig. 12: Predicted  $L_{\text{esc}}^{\text{LyC}}$  from the minimal model (model 3, described in § 4.2.3) vs the  $L_{\text{esc}}^{\text{LyC}}$  derived from observations for the observed LCEs listed in Table 1. The values in the y axis are derived by multiplying the observed  $L_{\text{esc}}^{\text{LyC},900}$  with the ratio  $L_{\text{esc}}^{\text{LyC}} / L_{\text{esc}}^{900}$  derived from our simulation (§ 2.3). The observational luminosity error bars are shown in skyblue. For few galaxies the observed  $L_{\text{esc}}^{\text{LyC}}$  is an upper limit, these are marked by skyblue arrows.

However, such full datasets can only be obtained from high resolution reionization simulations and currently we do not have other datasets. Instead, we can use repeated k-fold cross validation method (described in 4.1.5) to gauge the performance of our models.

In this work we have used  $k = 10$ , so we divide the dataset into 10 random subsets and calculate the average  $R_{\text{adj}}^2$  for our response variables. We repeat this process 3 times and get an average of  $R_{\text{adj}}^2$  from these runs. We have calculated the k-fold  $R_{\text{adj}}^2$  for each model and found that the  $R_{\text{adj}}^2$  from the k-fold test is always very similar to the  $R_{\text{adj}}^2$  we calculated when building the model with our whole dataset. For example, when we perform the cross validation for Model 2, for predicting  $L_{\text{int}}^{\text{Ly}\alpha}$ ,  $L_{\text{int}}^{\text{LyC}}$  and  $L_{\text{esc}}^{\text{LyC}}$  we get an average  $R_{\text{adj}}^2$  of 0.8996, 0.8945 and 0.8471 respectively, compared to 0.9006, 0.8956 and 0.8466 from our full model, as shown in Table 2. These respective  $R_{\text{adj}}^2$  values are very close to each other which shows that the our proposed models are indeed stable.

## 5. Discussion

In this study we have explored the relationship between Ly $\alpha$  and LyC emission from simulated EoR galaxies and we have shown that it is possible to predict LyC emission of galaxies using their physical and Ly $\alpha$  properties. However, there are some important limitations of this study that we discuss below.

**Limitations of the simulation:** Our simulation has a box size of 10 Mpc and the most luminous LAE in our sample of 1933 galaxies has a luminosity of  $L_{\text{esc}}^{\text{Ly}\alpha} = 1.37 \times 10^{42}$  erg/s. As we have discussed in § 3.3 and shown in figures 4 and 5, recent observations of MUSE LAEs and low redshift LyC leakers (table 1) are starting to overlap with the brightest end of our sample of simulated galaxies. However, our sample is at  $z \geq 6$  and at these very high redshifts, the lower limit of observed Ly $\alpha$  luminosity is around  $\sim 10^{43}$  erg/s, still more luminous than our brightest

Table 3: Most important variables for predicting LyC luminosities and escape fractions using Model 2 ( $GP + L_{\text{esc}}^{\text{Ly}\alpha}$ )

$L_{\text{int}}^{\text{LyC}}$			$L_{\text{esc}}^{\text{LyC}}$			$f_{\text{esc}}^{\text{LyC}}$		
Rank	Variable	Adjusted $R^2$	Rank	Variable	Adjusted $R^2$	Rank	Variable	Adjusted $R^2$
1	SFR <sub>10</sub>	0.7762	1	$L_{\text{esc}}^{\text{Ly}\alpha}$	0.7877	1	$L_{\text{esc}}^{\text{Ly}\alpha}$	0.2983
2	$L_{\text{esc}}^{\text{Ly}\alpha}$	0.8856	2	$M_{\text{Gas}}$	0.8243	2	SFR <sub>10</sub>	0.5397
3	SFR <sub>100</sub>	0.8911	3	$Z_{\text{Gas}}$	0.8401	3	$M_{\text{Gas}}$	0.6269
4	$Z_{\text{Gas}}$	0.8923	4	SFR <sub>10</sub>	0.8493	4	$Z_{\text{Gas}}$	0.6545
5	$M_{\text{Gas}}$	0.8937	5	$Z_{\star}$	0.8507	5	$Z_{\star}$	0.6563
6	Stellar Age	0.8954	6	$M_{\star}$	0.8511	6	SFR <sub>100</sub>	0.6569
7	$M_{\star}$	0.8968	7	Stellar Age	0.8519	7	Stellar Age	0.6568
8	$Z_{\star}$	0.8969	8	SFR <sub>100</sub>	0.8517	8	$M_{\star}$	0.6565
9	$R_{\text{Gal}}$	0.8969	9	$R_{\text{Gal}}$	0.8516	9	$R_{\text{Gal}}$	0.6561

Table 4:  $R^2$  for predicting different variables with the minimal model (Model 3)

Model	$L_{\text{int}}^{\text{Ly}\alpha}$	$L_{\text{esc}}^{\text{Ly}\alpha}$	$f_{\text{esc}}^{\text{Ly}\alpha}$	$L_{\text{int}}^{\text{LyC}}$	$L_{\text{esc}}^{\text{LyC}}$	$f_{\text{esc}}^{\text{LyC}}$
3. $M_{\star} + SFR_{10} + L_{\text{esc}}^{\text{Ly}\alpha}$	0.8827	NA	0.6242	0.8877	0.8030	0.5498

Table 5: Most important variables for predicting  $L_{\text{esc}}^{\text{LyC}}$  with Minimal Model (Model 3)

$L_{\text{esc}}^{\text{LyC}}$		
Rank	Variable	Adjusted $R^2$
1	$L_{\text{esc}}^{\text{Ly}\alpha}$	0.7877
2	SFR <sub>10</sub>	0.8007
3	$M_{\star}$	0.8030

galaxies. These detections are probably not representative of the underlying LAEs populations. Although they may play a central role in reionizing the Universe, as demonstrated by the recent discovery of an extremely bright LCE at  $z \sim 3$  (Marques-Chaves et al. 2021), the lack of very bright LAEs in our sample prevents us from making quantitative predictions for the contribution of very bright LAEs to reionization. As a consequence, our estimate of the fraction of the ionizing photons budget provided by galaxies with  $L_{\text{esc}}^{\text{Ly}\alpha} > 10^{41} \text{ erg.s}^{-1}$  in § 3.6 could well be a lower limit. In order to directly compare our predictions with observational data and to make better statistical predictions for bright galaxies, we need to analyze more luminous galaxies, for which we need to simulate a larger volume. The next generation of SPHINX will simulate a volume eight times larger than in the current study (i.e. 20 cMpc in width), which will include halos with stellar masses (virial masses) up to about  $10^{10} M_{\odot}$  ( $10^{11} M_{\odot}$ ) at  $z=6$ .

**IGM attenuation:** In this work we have not considered the effects of the IGM absorption. The IGM is an important factor in determining the observability of Ly $\alpha$  emission at these high redshifts, because in order to be observable LAEs, Ly $\alpha$  must be transmitted through a partially neutral IGM which can easily scatter Ly $\alpha$  photons off the line-of-sight. This can considerably reduce the visibility of LAEs during the EoR, as hinted by the drop of the LAE fraction at  $z > 6$  (Schenker et al. 2014; Kusakabe et al. 2020; Garel et al. 2021). Our results in this paper depict both Ly $\alpha$  and LyC luminosities as they would be observed just outside of the halo virial radius. In practice some correction for IGM can be applied to the data before applying our model to estimate  $L_{\text{esc}}^{\text{LyC}}$  of galaxies. Furthermore, the absence of IGM absorption has allowed us to compare our simulation results to low redshift observations of LCEs. For more realistic modeling and direct comparison with high-redshift observations, we need to consider IGM absorption. Garel et al. (2021) predicts that the

IGM transmission in Sphinx decreases from a factor of  $\sim 2$  at  $z=6$  to  $\sim 10$  at  $z=9$ . Nevertheless, this study is the first necessary step to assess the link between Ly $\alpha$  and LyC escape from galaxies. Since there are known LAEs at  $z > 6$  (e.g. Meyer et al. (2021) and references therein), depending on the topology of the reionization, Ly $\alpha$  emission may still go through large ionized bubbles at high redshift (Dijkstra 2014a; Mason & Gronke 2020; Gronke et al. 2021), and could serve as a tracer for LyC escape from galaxies at the cosmic dawn.

**Directional variation:** In this study, we chose to consider global, theoretical, estimates of the Ly $\alpha$  and LyC quantities, since they are the quantities which matter to determine ionizing photons budget, and study the process of reionization.

However, when we observe galaxies we will, of course, only be able to observe them from one direction (along our line-of-sight). Furthermore, the Ly $\alpha$  and LyC luminosities and escape fraction of the same galaxy can differ significantly from direction to direction (Cen & Kimm 2015; Mauerhofer et al. 2021, Chuniaud et al. in prep). To capture this added complexity, we will need to do directional analysis of our galaxies. As a first attempt to quantify the angular variations of Ly $\alpha$  and LyC luminosities escaping from our simulated galaxies, we imagine a sphere around a halo at the halo virial radius and divide the surface area of the sphere into 1728 equal area pixels. We then count the Ly $\alpha$  and LyC photons that escape each of these pixels and calculate the  $L_{\text{esc}}^{\text{Ly}\alpha}$  and  $L_{\text{esc}}^{\text{LyC}}$  through each of them. For each pixel direction then we have the directional luminosity ( $L_{\text{esc}}^{\text{Ly}\alpha \text{ (or LyC), \text{directional}}} = 1728 \times L_{\text{esc}}^{\text{Ly}\alpha \text{ (or LyC), \text{pixel}}}$ ). In Figure 13 we show the distribution of the directional Ly $\alpha$  and LyC luminosities (1728 directions for each galaxy) of the 1933 sphinx galaxies as a function of their actual global luminosities. Interestingly, we find that Ly $\alpha$ -bright galaxies can vary up to a factor of  $\sim 100$  compared to their angle-averaged  $L_{\text{esc}}^{\text{Ly}\alpha}$ , whereas faint galaxies are more isotropic. On the other hand, the directional LyC luminosities vary quite a lot at all angle-averaged LyC luminosities. As we discussed in § 3.5, Ly $\alpha$  photons can scatter numerous times before escaping, hence they have a higher chance of finding channels in the ISM with low column density, hence their directional distribution is generally more isotropic. Conversely LyC photons generally escape close to the galaxy center where they are mainly produced, so they have lower probabilities of finding many channels, which can result in a more anisotropic distribution of directional luminosities.

The broad variety of Lyman alpha spectral shapes and strengths observed from galaxies is also one of the main probes

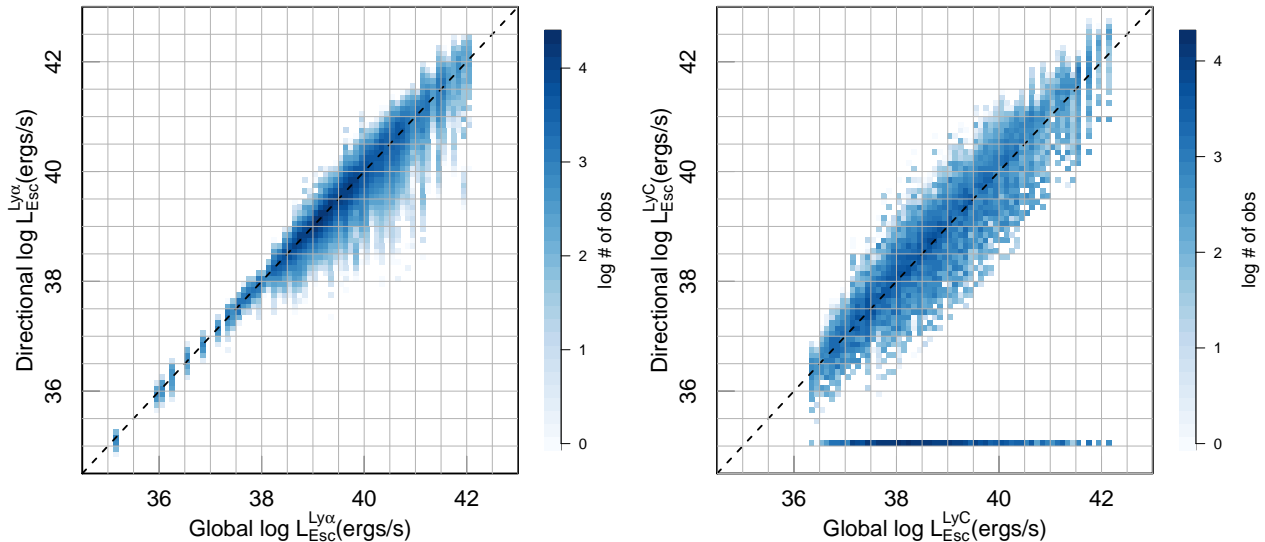


Fig. 13: Left: Directional  $L_{\text{esc}}^{\text{Ly}\alpha}$  of 1933 sphinx galaxies as a function of their real global  $L_{\text{esc}}^{\text{Ly}\alpha}$ . We imagine a sphere around a halo at the halo virial radius and divide the surface area of the sphere into 1728 equal area pixels. We calculate the  $L_{\text{esc}}^{\text{Ly}\alpha}$  through each of them and for each pixel direction then we have the directional luminosity,  $L_{\text{esc}}^{\text{Ly}\alpha, \text{directional}} = 1728 \times L_{\text{esc}}^{\text{Ly}\alpha, \text{pixel}}$ . Right: Same plot but for LyC where we show directional  $L_{\text{esc}}^{\text{LyC}}$  vs the global  $L_{\text{esc}}^{\text{LyC}}$ . The directions with no LyC escape are indicated with an artificial LyC luminosity of  $10^{35}$  erg/s.

of strong directional variations. Indeed, several recent observational studies (Verhamme et al. 2017; Steidel et al. 2018; Izotov et al. 2021) have found that spectral features of Ly $\alpha$  line profiles, such as high rest-frame equivalent width and a narrow separation between the blue and red peak of Ly $\alpha$  spectra correlates positively with escape of LyC. Testing these directional spectral features are beyond the scope of this article. But these two approaches are complementary of each other and we would ideally need both to get a complete picture of the contribution of the galaxies along our line of sight, and globally, to the reionization process. To that end, in the next step, we will employ peeling off algorithms on our galaxies and observe them from several directions. Then we can build mock observations to compare directly with existing and future observations and comment on how to employ our predictive models based on observed directional properties.

#### Uncertainties in the intrinsic LyC spectral distributions:

The shape of the ionizing spectrum of galaxies is still poorly constrained. The LCEs detected so far have all been observed close to the Lyman limit (Steidel et al. 2018; Izotov et al. 2021; Flury et al. 2022, e.g.). The only exception is the recent discovery of a  $z \sim 1.4$  galaxy leaking ionizing radiation at  $600\text{\AA}$  rest-frame with the Astrosat (Saha et al. 2020). The theoretical predictions from population synthesis models is also a debated topic so far. The SPHINX simulation uses BPASS models (Stanway et al. 2016) to build the SEDs of galaxies, and in this version of SPHINX, all stars are binary systems. The binary star systems can emit more LyC photons for a longer time compared to single stellar populations, which impacts the full reionization history (Rosdahl et al. 2018). While binaries appear as a central ingredient in stellar radiation modelling at the EoR, the fraction of binary stars in the early Universe remains uncertain, as well as their exact spectral contribution.

While discussing the relationship of Ly $\alpha$  and LyC intrinsic luminosities in § 3.3 we have noted in Figure 5 that galaxies (77/1933 or 3.98% of the population) at the very faint end of Ly $\alpha$  ( $L_{\text{int}}^{\text{Ly}\alpha} < 10^{38}$  erg/s) have LyC luminosity in the range of  $10^{38} - 10^{40}$  erg/s. These faint LAEs are extremely gas deficient compared to the rest of the population, as shown in Figure 14. So we find that in these systems there is not enough gas in the ISM

to produce Ly $\alpha$  photons, resulting in very low  $L_{\text{int}}^{\text{Ly}\alpha}$ . In contrast, these galaxies do have some residual LyC production although there have been no star formation in them in the last 10 Myrs (i.e.  $\text{SFR}_{10} = 0$ ). We show the stellar ages of these systems in Figure 14 and find that their median ages range from 100 - 300 Myrs and even their minimum stellar ages are very high. Furthermore, in all of them the 25th, 50th, and 75th percentile of ages are very close in values. This indicates that these systems are very old and their star formation finished within a short amount of time. Stanway et al. (2016) (Figure 1) demonstrates that, for binary populations in BPASS models with an instantaneous star formation model, it is possible for stellar populations to emit  $\sim 10^{49}$  LyC photons/s at an age of 100 Myr. So in these faint LAEs, it is feasible that even though the galaxies have very old stellar systems, the LyC production is non-negligible. If these simulated galaxies exist in the real Universe, their LyC contribution to the re-ionization photons budget cannot be captured by their Ly $\alpha$  emission, and they will be missed by our prediction models.

## 6. Summary

We explore the connection between LyC and Ly $\alpha$  emission from EoR galaxies using a sample of 1933 simulated galaxies in the SPHINX radiation hydrodynamical simulation. We post process these galaxies using the radiative transfer code RASCAS to obtain their Ly $\alpha$  emission properties.

We first investigate the link between Ly $\alpha$  and LyC radiation from galaxies and our main results are as follows:

- The intrinsic Ly $\alpha$  and LyC luminosities are strongly correlated. The total LyC (0 -  $912\text{\AA}$ ) escaping luminosities, are also correlated with escaping Ly $\alpha$  luminosity, although the dispersion is higher, especially in faint LAEs.
- Given a threshold in observed LyC luminosity, as galaxies become brighter in Ly $\alpha$ , the fraction of observable LCEs among LAE samples increases.
- In bright LAEs ( $L_{\text{esc}}^{\text{Ly}\alpha} > 10^{41}$  erg/s) escape fractions of Ly $\alpha$  and LyC are correlated, and in good agreement with the observed LCEs. However, when we consider all galaxies, including the fainter ones, there is no correlation, which suggests that the observed correlation is likely a selection effect.

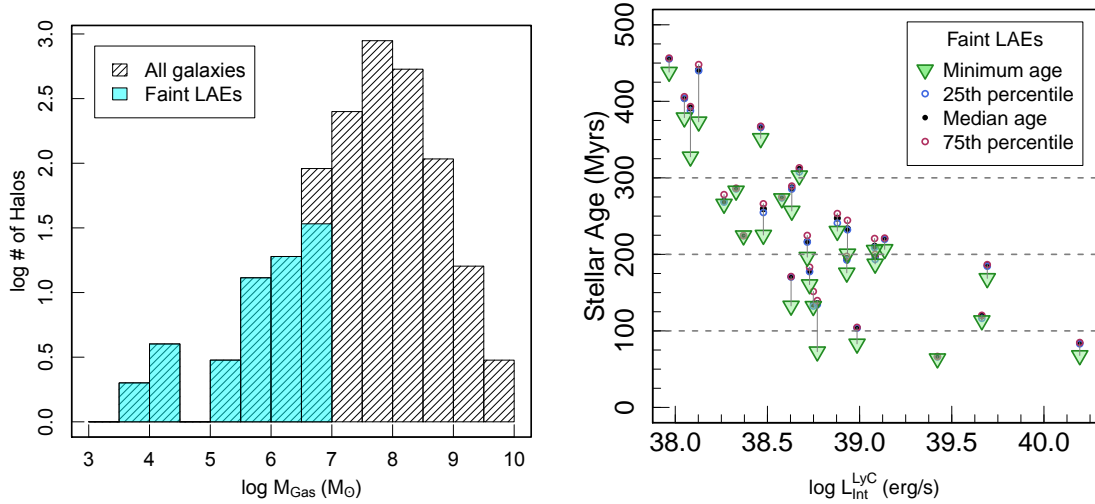


Fig. 14: Left : Distribution of gas mass in the faint LAEs ( $L_{\text{int}}^{\text{Ly}\alpha} < 10^{38}$  erg/s, in sky blue) and the whole sample (shaded). Right : Stellar ages of these faint LAEs as a function of their intrinsic LyC luminosity. Figure shows the minimum stellar age, 25th percentile, 75th percentile and the median age of the stars for each of these faint galaxies.

- The median  $f_{\text{esc}}^{\text{Ly}\alpha}$  of galaxies gradually decreases with their Ly $\alpha$  luminosity and at the bright end with  $L_{\text{esc}}^{\text{Ly}\alpha} \approx 10^{41.5} - 10^{42}$  erg/s, the median  $f_{\text{esc}}^{\text{Ly}\alpha} \approx 0.3$ . Median value of  $f_{\text{esc}}^{\text{LyC}}$  is low for all Ly $\alpha$  luminosities with the bright LAEs ( $L_{\text{esc}}^{\text{Ly}\alpha} > 10^{40.5}$  erg/s) having median  $f_{\text{esc}}^{\text{LyC}} \sim 0.1$ .
- Although very faint galaxies are more numerous, the relatively bright LAEs contribute more to reionization. In our SPHINX volume, LAEs with  $L_{\text{esc}}^{\text{Ly}\alpha} \geq 10^{40}$  erg/s account for about 90% of the total ionizing luminosity in the simulation box, even though they are only 6.8% of the population.

We explored models for predicting LyC emission from galaxies using their physical and Ly $\alpha$  properties. We apply multivariate linear models on our sample of simulated galaxies and the main results are summarized below:

- We build a set of models using different sets of galaxy properties as input parameters and predict LyC luminosities and escape fraction. In our fiducial model (Model 2) we give 8 galaxy physical properties (gas mass, stellar mass, galaxy  $R_{\text{vir}}$ , SFR $_{10}$ , SFR $_{100}$ , stellar age and stellar and gas metallicity) and  $L_{\text{esc}}^{\text{Ly}\alpha}$  as input parameters. The resulting model can predict  $L_{\text{int}}^{\text{LyC}}$  and  $L_{\text{esc}}^{\text{LyC}}$  very well, with high (adjusted)  $R^2$  values of 0.8969 and 0.8516 respectively. The  $f_{\text{esc}}^{\text{LyC}}$  is also predicted fairly well.
- We also determine the most important input variables for predicting LyC and find that the top four predictors of  $L_{\text{esc}}^{\text{LyC}}$  are  $L_{\text{esc}}^{\text{Ly}\alpha}$ , gas mass, gas metallicity, and SFR $_{10}$ .

These results and the predictive models can be very useful in predicting the LyC emission from EoR galaxies and thus help us to determine the primary sources of reionization. We can apply these models on the upcoming EoR galaxy observations of JWST and other future surveys. They can also facilitate the selection and detection of LyC leakers. These models can be helpful to plan future direct LCE observation missions at lower redshifts. In a future work, we will investigate the effect of directional variation of Ly $\alpha$  and LyC escape from galaxies, and IGM attenuation, on our predictions.

## 7. Acknowledgement

We thank the anonymous referee for valuable comments and suggestions that have substantially improved the paper.

MM, AV and TG are supported by the ERC Starting grant 757258 ‘TRIPLE’. AV acknowledges support from SNF Professorship PP00P2\_176808. TK was supported by the National Research Foundation of Korea (NRF-2019K2A9A1A0609137711 and NRF-2020R1C1C1007079).

We have performed the radiative transfer calculations in the LESTA and BAOBAB high-performance computing clusters of University of Geneva, and the RT post-processing for 1933 halos took approximately  $\sim 37000$  CPU hours. The SPHINX simulation results of this research have been achieved using the PRACE Research Infrastructure resource SuperMUC based in Garching, Germany, under PRACE grant 2016153539. We additionally acknowledge support and computational resources from the Common Computing Facility (CCF) of the LABEX Lyon Institute of Origins (ANR-10-LABX-66).

## References

- Aubert, D., Pichon, C., & Colombi, S. 2004, MNRAS, 352, 376  
 Bacon, R., Brinchmann, J., Richard, J., et al. 2015, A&A, 575, A75  
 Bassett, R., Ryan-Weber, E. V., Cooke, J., et al. 2019, MNRAS, 483, 5223  
 Behrens, C. & Braun, H. 2014, A&A, 572, A74  
 Borthakur, S., Heckman, T. M., Leitherer, C., & Overzier, R. A. 2014, Science, 346, 216  
 Cantalupo, S., Porciani, C., & Lilly, S. J. 2008, ApJ, 672, 48  
 Cen, R. & Kimm, T. 2015, ApJ, 801, L25  
 Chisholm, J., Orlicová, I., Schaerer, D., et al. 2017, A&A, 605, A67  
 Cowie, L. L., Barger, A. J., & Trouille, L. 2009, ApJ, 692, 1476  
 Dijkstra, M. 2014a, PASA, 31, e040  
 Dijkstra, M. 2014b, PASA, 31, e040  
 Dijkstra, M., Gronke, M., & Venkatesan, A. 2016, ApJ, 828, 71  
 Drake, A. B., Garel, T., Wisotzki, L., et al. 2017, A&A, 608, A6  
 Erb, D. K. 2015, Nature, 523, 169  
 Erb, D. K., Bogosavljević, M., & Steidel, C. C. 2011, ApJ, 740, L31  
 Faucher-Giguère, C.-A. 2020, MNRAS, 493, 1614  
 Faucher-Giguère, C.-A., Kereš, D., Dijkstra, M., Hernquist, L., & Zaldarriaga, M. 2010, ApJ, 725, 633  
 Feigelson, E. D. & Babu, G. J. 2012, Modern Statistical Methods for Astronomy: With R Applications (Cambridge University Press)  
 Finkelstein, S. L., Papovich, C., Dickinson, M., et al. 2013, Nature, 502, 524  
 Flury, S. R., Jaskot, A. E., Ferguson, H. C., et al. 2022, accepted to ApJS, arXiv:2201.11716  
 Fontanot, F., Cristiani, S., Pfrommer, C., Cupani, G., & Vanzella, E. 2014, MNRAS, 438, 2097  
 Fontanot, F., Cristiani, S., & Vanzella, E. 2012, MNRAS, 425, 1413  
 Garel, T., Blaizot, J., Rosdahl, J., et al. 2021, MNRAS, 504, 1902  
 Gazagnes, S., Chisholm, J., Schaerer, D., Verhamme, A., & Izotov, Y. 2020, A&A, 639, A85  
 Goerdt, T., Dekel, A., Sternberg, A., et al. 2010, MNRAS, 407, 613  
 Gronke, M., Ocvirk, P., Mason, C., et al. 2021, MNRAS, 508, 3697  
 Hayes, M., Östlin, G., Schaerer, D., et al. 2013, ApJ, 765, L27

- Heckman, T. M., Alexandroff, R. M., Borthakur, S., Overzier, R., & Leitherer, C. 2015, *ApJ*, 809, 147
- Heckman, T. M., Borthakur, S., Overzier, R., et al. 2011, *ApJ*, 730, 5
- Henry, A., Scarlata, C., Martin, C. L., & Erb, D. 2015, *ApJ*, 809, 19
- Inoue, A. K., Hasegawa, K., Ishiyama, T., et al. 2018, *PASJ*, 70, 55
- Inoue, A. K., Shimizu, I., Iwata, I., & Tanaka, M. 2014, *MNRAS*, 442, 1805
- Itoh, R., Ouchi, M., Zhang, H., et al. 2018, *ApJ*, 867, 46
- Izotov, Y. I., Orlitová, I., Schaerer, D., et al. 2016a, *Nature*, 529, 178
- Izotov, Y. I., Schaerer, D., Thuan, T. X., et al. 2016b, *MNRAS*, 461, 3683
- Izotov, Y. I., Schaerer, D., Worseck, G., et al. 2018a, *MNRAS*, 474, 4514
- Izotov, Y. I., Worseck, G., Schaerer, D., et al. 2021, *MNRAS*, 503, 1734
- Izotov, Y. I., Worseck, G., Schaerer, D., et al. 2018b, *MNRAS*, 478, 4851
- Jaskot, A. E. & Oey, M. S. 2013, *ApJ*, 766, 91
- Jung, I., Finkelstein, S. L., Dickinson, M., et al. 2019, *ApJ*, 877, 146
- Katz, H., Galligan, T. P., Kimm, T., et al. 2019, *MNRAS*, 487, 5902
- Katz, H., Āurovčíková, D., Kimm, T., et al. 2020, *MNRAS*, 498, 164
- Kimm, T., Blaizot, J., Garel, T., et al. 2019, *MNRAS*, 486, 2215
- Konno, A., Ouchi, M., Ono, Y., et al. 2014, 797, 16
- Kulkarni, G., Worseck, G., & Hennawi, J. F. 2019, *MNRAS*, 488, 1035
- Kusakabe, H., Blaizot, J., Garel, T., et al. 2020, *A&A*, 638, A12
- Laursen, P., Sommer-Larsen, J., & Andersen, A. C. 2009, *ApJ*, 704, 1640
- Laursen, P., Sommer-Larsen, J., Mølvang-Jensen, B., Fynbo, J. P. U., & Razoumov, A. O. 2019, *A&A*, 627, A84
- Leitet, E., Bergvall, N., Hayes, M., Linné, S., & Zackrisson, E. 2013, *A&A*, 553, A106
- Leitet, E., Bergvall, N., Piskunov, N., & Andersson, B. G. 2011, *A&A*, 532, A107
- Loeb, A. & Barkana, R. 2001, *ARA&A*, 39, 19
- Madau, P. 1995, *ApJ*, 441, 18
- Marques-Chaves, R., Schaerer, D., Álvarez-Márquez, J., et al. 2021, *MNRAS*, 507, 524
- Mason, C. A. & Gronke, M. 2020, *MNRAS*, 499, 1395
- Matthee, J., Pezzulli, G., Mackenzie, R., et al. 2020, *MNRAS*, 498, 3043
- Matthee, J., Sobral, D., Darvish, B., et al. 2017, *MNRAS*, 472, 772
- Matthee, J., Sobral, D., Gronke, M., et al. 2018, *A&A*, 619, A136
- Mauerhofer, V., Verhamme, A., Blaizot, J., et al. 2021, *A&A*, 646, A80
- Meyer, R. A., Laporte, N., Ellis, R. S., Verhamme, A., & Garel, T. 2021, *MNRAS*, 500, 558
- Michel-Dansac, L., Blaizot, J., Garel, T., et al. 2020, *A&A*, 635, A154
- Micheva, G., Zackrisson, E., Östlin, G., Bergvall, N., & Pursimo, T. 2010, *MNRAS*, 405, 1203
- Nakajima, K. & Ouchi, M. 2014, *MNRAS*, 442, 900
- Ocvirk, P., Gillet, N., Shapiro, P. R., et al. 2016, *MNRAS*, 463, 1462
- Oesch, P. A., van Dokkum, P. G., Illingworth, G. D., et al. 2015, *ApJ*, 804, L30
- Ono, Y., Ouchi, M., Mobasher, B., et al. 2012, *ApJ*, 744, 83
- Östlin, G., Hayes, M., Duval, F., et al. 2014, *ApJ*, 797, 11
- Ouchi, M., Harikane, Y., Shibuya, T., et al. 2018, *PASJ*, 70, S13
- Pardy, S. A., Cannon, J. M., Östlin, G., Hayes, M., & Bergvall, N. 2016, *AJ*, 152, 178
- Partridge, R. B. & Peebles, P. J. E. 1967, *ApJ*, 147, 868
- Planck Collaboration, Ade, P. A. R., Aghanim, N., et al. 2014, *A&A*, 571, A16
- Puschnig, J., Hayes, M., Östlin, G., et al. 2017, *MNRAS*, 469, 3252
- Raiter, A., Fosbury, R. A. E., & Teimoorinia, H. 2010, *A&A*, 510, A109
- Roberts-Borsani, G. W., Bouwens, R. J., Oesch, P. A., et al. 2016, *ApJ*, 823, 143
- Rosdahl, J. & Blaizot, J. 2012, *MNRAS*, 423, 344
- Rosdahl, J., Blaizot, J., Aubert, D., Stranex, T., & Teysier, R. 2013, *MNRAS*, 436, 2188
- Rosdahl, J., Katz, H., Blaizot, J., et al. 2018, *MNRAS*, 479, 994
- Runholm, A., Hayes, M., Melinder, J., et al. 2020, *ApJ*, 892, 48
- Saha, K., Tandon, S. N., Simmonds, C., et al. 2020, *Nature Astronomy*, 4, 1185
- Schaerer, D. 2003, *A&A*, 397, 527
- Schaerer, D., Izotov, Y. I., Verhamme, A., et al. 2016, *A&A*, 591, L8
- Schenker, M. A., Ellis, R. S., Konidaris, N. P., & Stark, D. P. 2014, *ApJ*, 795, 20
- Schenker, M. A., Stark, D. P., Ellis, R. S., et al. 2012, *ApJ*, 744, 179
- Shibuya, T., Kashikawa, N., Ota, K., et al. 2012, *ApJ*, 752, 114
- Shibuya, T., Ouchi, M., Konno, A., et al. 2018, *PASJ*, 70, S14
- Smith, A., Ma, X., Bromm, V., et al. 2019, *MNRAS*, 484, 39
- Song, M., Finkelstein, S. L., Livermore, R. C., et al. 2016, *ApJ*, 826, 113
- Songailla, A., Hu, E. M., Barger, A. J., et al. 2018, *ApJ*, 859, 91
- Spitzer, L. 1978, *Physical processes in the interstellar medium*
- Stanway, E. R., Eldridge, J. J., & Becker, G. D. 2016, *MNRAS*, 456, 485
- Stark, D. P. 2016, *ARA&A*, 54, 761
- Stark, D. P., Ellis, R. S., Charlot, S., et al. 2017, *MNRAS*, 464, 469
- Steidel, C. C., Bogosavljević, M., Shapley, A. E., et al. 2018, *ApJ*, 869, 123
- Teyssier, R. 2002, *A&A*, 385, 337
- Trainor, R. F., Steidel, C. C., Strom, A. L., & Rudie, G. C. 2015, *ApJ*, 809, 89
- Trebtsch, M., Verhamme, A., Blaizot, J., & Rosdahl, J. 2016, *A&A*, 593, A122
- Trebtsch, Maxime, Dubois, Yohan, Volonteri, Marta, et al. 2021, *A&A*, 653, A154
- Tweed, D., Devriendt, J., Blaizot, J., Colombi, S., & Slyz, A. 2009, *A&A*, 506, 647
- Urrutia, T., Wisotzki, L., Kerutt, J., et al. 2019, *A&A*, 624, A141
- Vanzella, E., Pentericci, L., Fontana, A., et al. 2011, *ApJ*, 730, L35
- Verhamme, A., Dubois, Y., Blaizot, J., et al. 2012, *A&A*, 546, A111
- Verhamme, A., Orlicová, I., Schaerer, D., & Hayes, M. 2015, *A&A*, 578, A7
- Verhamme, A., Orlicová, I., Schaerer, D., et al. 2017, *A&A*, 597, A13
- Wise, J. H. 2019, *Contemporary Physics*, 60, 145–163
- Yajima, H., Li, Y., & Zhu, Q. 2013, *ApJ*, 773, 151
- Yajima, H., Li, Y., Zhu, Q., et al. 2014, *MNRAS*, 440, 776
- Yang, H., Malhotra, S., Gronke, M., et al. 2017, *ApJ*, 844, 171
- Zitrin, A., Labbé, I., Belli, S., et al. 2015, *ApJ*, 810, L12

## Appendix A: Supplementary figures

### Appendix A.1: Comparing $z=6$ sample to the stacked sample

In the main text, we combine our galaxy samples at different redshifts and explored the connection between LyC and Ly $\alpha$  emission from galaxies. Herein we inspect if the selected populations of galaxies at different redshifts have significantly different properties. We compare two samples specifically, 674 galaxies at  $z = 6$ , and the stacked sample of 1933 galaxies that combines all galaxies in all of the 5 redshifts ( $z = 6, 7, 8, 9, 10$ ).

We compare the physical properties, Ly $\alpha$  properties and LyC properties of these two samples and present the results in Figure A.1. In the top row, it shows comparisons of three physical galaxy properties, stellar mass, gas mass and SFR calculated over the last 10 Myrs (SFR<sub>10</sub>). We find that in each case, the distributions are very similar and the median value of the mass and SFR<sub>10</sub> is also almost the same. We have also compared the halo mass, size of the galaxy ( $R_{\text{vir,gal}}$ ) and halo ( $R_{\text{vir,halo}}$ ) and SFR calculated over the last 100 Myr (SFR<sub>100</sub>) and found that for each of these properties, the two samples have very similar values. Here we choose to show only the three properties mentioned as representative plots for brevity's sake.

In the second and third row of Figure A.1, we have compared the Ly $\alpha$  and LyC properties of the two samples, showing for each intrinsic luminosity, escaping luminosity and the escape fraction. The plots clearly show that for both intrinsic and escaping luminosity, the distributions are again very similar with almost the same median values.

For  $f_{\text{esc}}^{\text{Ly}\alpha}$  and  $f_{\text{esc}}^{\text{LyC}}$  comparisons, we find that the distribution for either samples is not single peaked or gaussian like the other properties. The  $f_{\text{esc}}^{\text{Ly}\alpha}$  distribution is close to a binomial with values biased towards close to either 0 or 1.  $f_{\text{esc}}^{\text{LyC}}$  distribution is also biased towards values close to 0.

Since a median of these two distributions would not be very meaningful, we calculate the percentage of the population that have very high  $f_{\text{esc}}^{\text{Ly}\alpha}$ , defined as  $f_{\text{esc}}^{\text{Ly}\alpha} > 0.9$  and find that in  $z = 6$  sample 31% fall in this category, whereas in the stacked sample the population is 32%. For  $f_{\text{esc}}^{\text{LyC}}$ , the distribution peaks towards extremely low values, so we calculate the percentage of population with  $f_{\text{esc}}^{\text{LyC}} < 0.1$  and find it to be 66.7% and 61% for  $z = 6$  and stacked sample respectively. We find that the stacked sample is very similar to the  $z = 6$  sample of galaxies and there are no large systematic differences between them in terms of their physical or radiative properties. We note that the age of the Universe at  $z = 6$  is 927 Myr and at  $z = 10$  it is 470 Myr, so between the redshift range of 6 - 10, only 457 Myr pass. So it is not surprising that we find the statistical properties of the galaxies within this time frame do not change significantly in our simulation. Our results suggest that we can use our stacked sample of 1933 galaxies for our Ly $\alpha$  and LyC analysis to study reionization era galaxies.

### Appendix A.2: Contribution of Recombination and Collision to Ly $\alpha$ production

Figure A.2 shows the fraction of intrinsic Ly $\alpha$  that comes from recombination and collision respectively. We find that in bright LAEs almost all of the  $L_{\text{int}}^{\text{Ly}\alpha}$  is generated from recombination. However, the contribution of collision becomes higher as galaxies becomes fainter. For example, in galaxies where  $L_{\text{int}}^{\text{Ly}\alpha} > 10^{42}$  erg/s collisions contribute  $\sim$  a few percent, but it can rise to  $\sim 50\%$  in galaxies  $10^{38} > L_{\text{int}}^{\text{Ly}\alpha} > 10^{40}$  erg/s.

### Appendix A.3: Variation of escape fractions with escaping luminosities

We have discussed the relationship between Ly $\alpha$  and LyC luminosities and escape fractions in § 3.3 and § 3.5 respectively. Here we revisit them and discuss how galaxy escape fractions

vary with their luminosities. In fig A.3 we show  $L_{\text{esc}}^{\text{LyC}}$  as a function of their  $L_{\text{esc}}^{\text{Ly}\alpha}$ , similar to Fig 5, but here colored by their  $f_{\text{esc}}^{\text{LyC}}$  and  $f_{\text{esc}}^{\text{Ly}\alpha}$ . We find that most of the galaxies have low  $f_{\text{esc}}^{\text{LyC}}$  and there is a clear trend that for a given  $L_{\text{esc}}^{\text{LyC}}$ , brighter LAEs have lower  $f_{\text{esc}}^{\text{LyC}}$ . When  $f_{\text{esc}}^{\text{LyC}}$  is high, most of the LyC is escaping, so there are few LyC photons available to produce Ly $\alpha$ , hence Ly $\alpha$  luminosity is low. As  $f_{\text{esc}}^{\text{LyC}}$  decreases, more and more LyC photons are reprocessed into Ly $\alpha$ , and Ly $\alpha$  luminosity increases. On the other hand, most of the galaxies have high  $f_{\text{esc}}^{\text{Ly}\alpha}$ . In general, faint LAEs have high Ly $\alpha$  escape fraction, but there is significant scatter at each luminosities.

### Appendix A.4: Reionization accounting with lower mass limit

In § 3.6 we have discussed the contribution of LAEs towards reionization and found that LAEs brighter than  $10^{40}$  erg/s can account for 95% of the total ionizing luminosity in the simulation, suggesting that bright LAEs may be the most important sources of reionization. However, in this analysis while counting the LyC contribution of LAEs, following our galaxy selection criterion in § 2.2 we have considered all galaxies with  $M_{\star} > 10^6 M_{\odot}$ . It will be instructive to explore how the results will change if we impose a lower mass limit, e.g.  $10^5 M_{\odot}$ . In order to investigate this, we need to first run the Ly $\alpha$  radiative transfer on all galaxies with  $M_{\star} > 10^5 M_{\odot}$ . Since the number of galaxies within  $10^5 - 10^6 M_{\odot}$  range is very high, post-processing all of them in the full stacked sample will be very expensive. Hence, we limit our investigation to galaxies in  $z = 6$  snapshot only. At  $z=6$ , there are 674 and 1495 galaxies with  $M_{\star} > 10^6 M_{\odot}$  and  $M_{\star} > 10^5 M_{\odot}$ , respectively.

Similar to our analysis in § 3.6, we first calculate the total LyC luminosity emitted by all (level 1) galaxies at  $z = 6$ . Then we calculate how much of this total LyC is emitted by galaxies with  $L_{\text{esc}}^{\text{Ly}\alpha} > 10^{38}, 10^{39}, 10^{40}, 10^{41}$  and  $10^{42}$  erg/s using samples with both stellar mass limits of  $10^6$  and  $10^5 M_{\odot}$ . Figure A.4 show this cumulative fraction against the limiting Ly $\alpha$  luminosity of the galaxies. We find that LAEs brighter than  $10^{40}$  erg/s can account for 95% of total LyC when counting only  $M_{\star} > 10^6 M_{\odot}$  galaxies, and if we lower the mass limit to  $10^5 M_{\odot}$ , this fraction increases to 97%. At the low luminosity limit, LAEs brighter than  $10^{38}$  erg/s contribute 97% (99%) of the re-ionizing radiation. This results show that although lowering the mass limit slightly increase these fractions, the differences are very small. This indicates that the reionization accounting we have done in § 3.6 with  $10^6 M_{\odot}$  mass limit is reasonably accurate.

### Appendix A.5: Multivariate model: more exploratory analysis

We show the histograms for the galaxy properties used in building our models (as listed in § 4.1.1) for our sample of 940 galaxies (§ 4.2.1) in Figure A.5. We have discussed in § 4.1 that the before building a multivariate linear model to predict LyC properties, it is important to check if any of the proposed  $x$ -variables or input variables have any correlation with the  $y$ -variable or response variable. Figure A.6 and Figure A.7 show such exploratory plot of the response variable  $L_{\text{int}}^{\text{LyC}}$  and  $f_{\text{esc}}^{\text{LyC}}$  vs various galaxy properties, respectively. We find that several properties, especially, SFR<sub>10</sub> and  $L_{\text{esc}}^{\text{Ly}\alpha}$  correlates very well with  $L_{\text{int}}^{\text{LyC}}$ . There are also weak correlations with gas mass, SFR<sub>100</sub> and stellar age.  $f_{\text{esc}}^{\text{LyC}}$  is also correlated with  $L_{\text{esc}}^{\text{Ly}\alpha}$ . These suggests that the multivariate linear regression model can be a good choice for predicting LyC emission from galaxies using these properties.

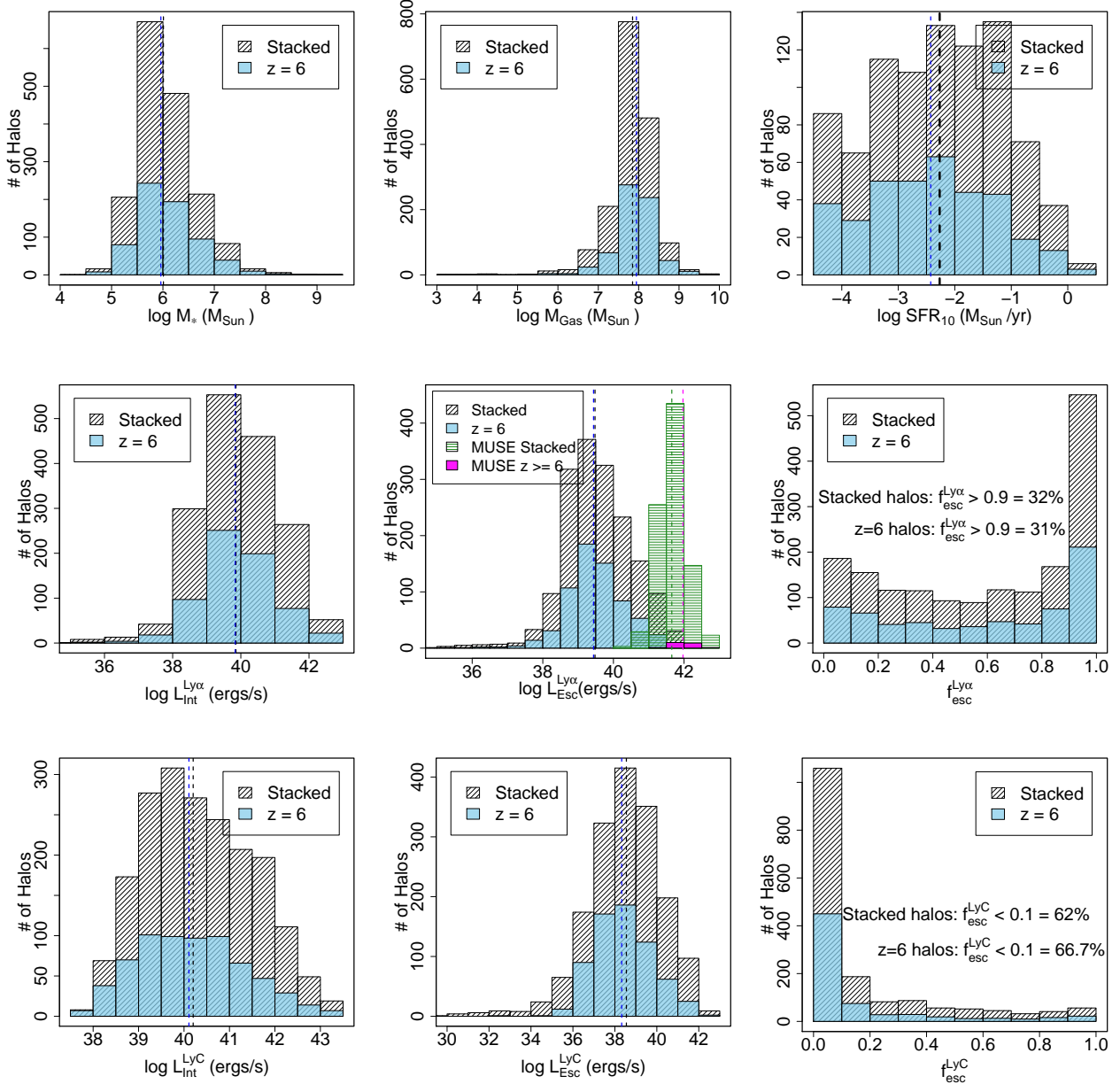


Fig. A.1: Comparing the physical, Ly $\alpha$  and LyC properties of galaxies of the stacked sample (gray) with  $z=6$  sample (blue). The top row shows compares stellar mass (left), gas mass (middle) and  $\text{SFR}_{10}$  (right). The middle row compares Ly $\alpha$  properties of the two samples with intrinsic luminosity (left), escaping luminosity (middle) and escape fraction (right). The bottom row shows the same properties but for LyC radiation. The dashed lines show the median value of the properties for both stacked (black) and  $z=6$  sample (blue).

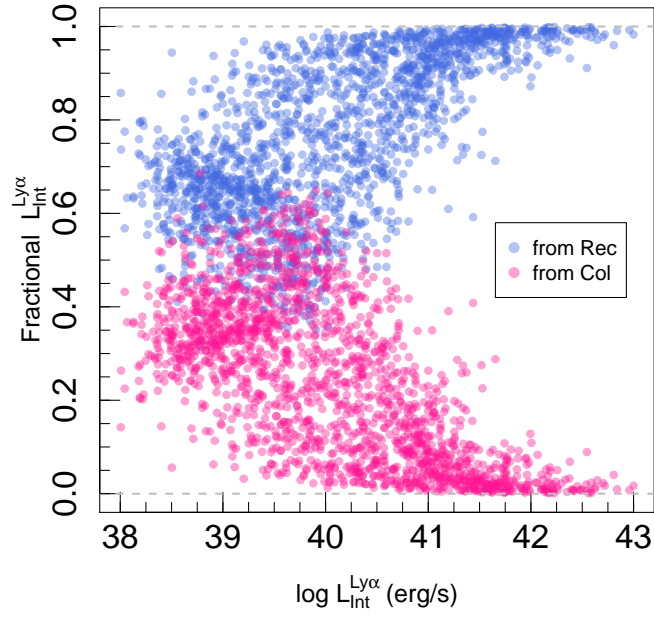


Fig. A.2: Fraction of intrinsic Ly $\alpha$  luminosity generated by recombination (blue) and collision (pink) as a function of  $L_{\text{int}}^{\text{Ly}\alpha}$ .

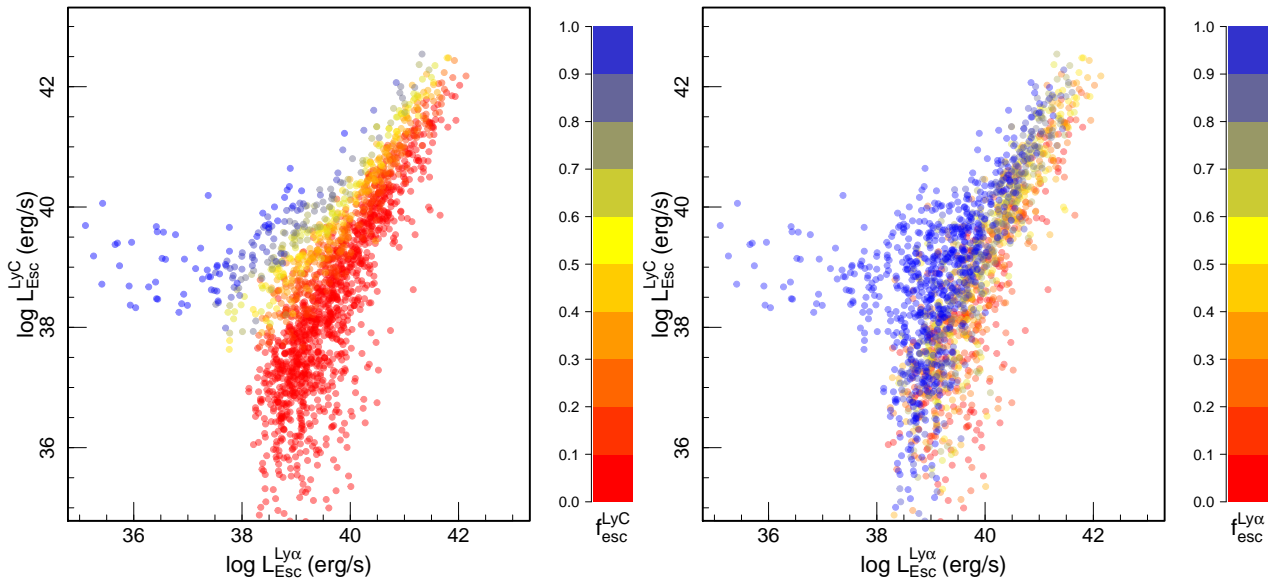


Fig. A.3: Escaping LyC luminosity of galaxies as a function of their escaping Ly $\alpha$  luminosity. This is same as Fig 5, but the points here are colored by their  $f_{\text{esc}}^{\text{LyC}}$  (left) and  $f_{\text{esc}}^{\text{Ly}\alpha}$  (right).



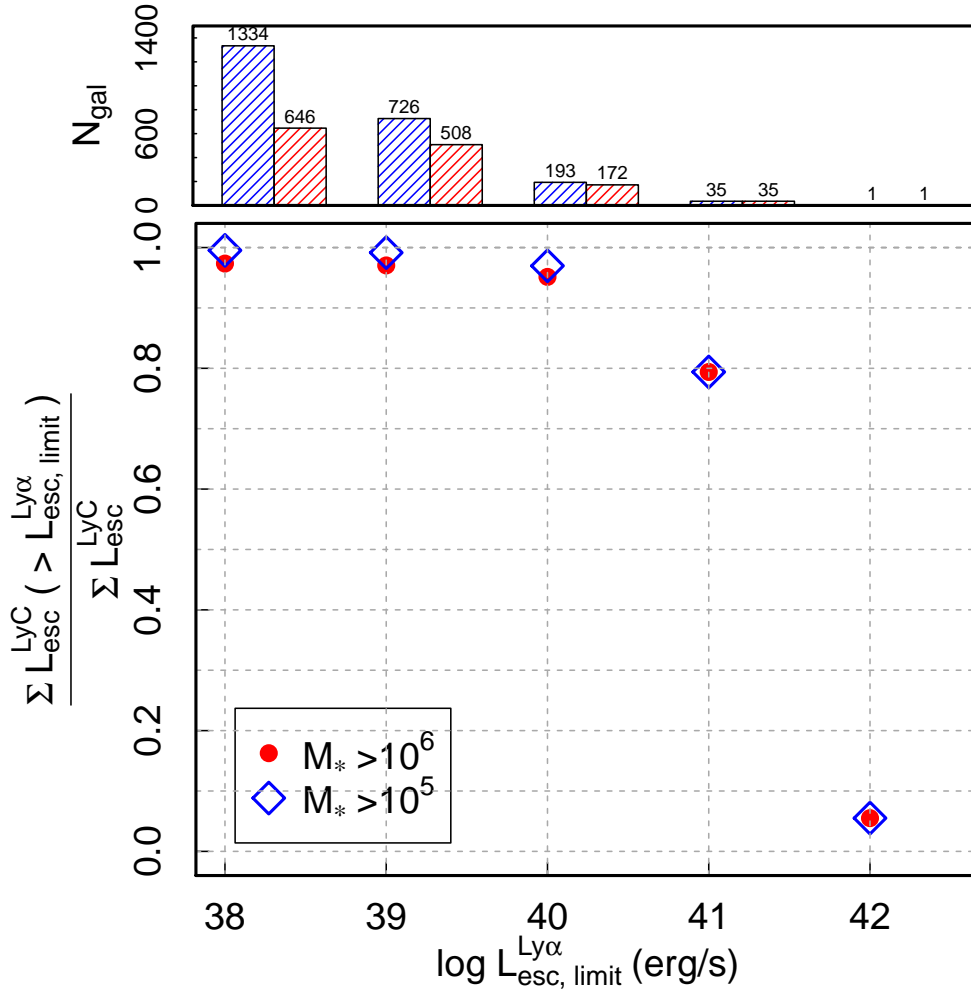


Fig. A.4: Fraction of the total escaping LyC luminosity emitted by galaxies brighter than a given Ly $\alpha$  luminosity limit as a function of the Ly $\alpha$  luminosity limit. Here we compare this fraction for two sets of galaxy sample: all galaxies at level 1 with  $M_\star > 10^6 M_\odot$  (red points), and all galaxies at level 1 with  $M_\star > 10^5 M_\odot$  (blue points). These galaxies are all taken from  $z=6$  snapshot. So the denominator of the fraction is same in both cases, the total LyC emission by all galaxies (at level 1) at  $z = 6$ . The numerator calculates the total LyC luminosity of the galaxies brighter than a given Ly $\alpha$  luminosity limit with the two samples, e.g. the total LyC emitted by all galaxies (at level 1) with  $M_\star > 10^6$  (or  $10^5$ )  $M_\odot$  and  $L_{\text{esc}}^{\text{Ly}\alpha} > 10^{40}$  erg/s. The histograms above show the number of galaxies brighter than the corresponding Ly $\alpha$  luminosity limit, e.g. the number of galaxies with  $L_{\text{esc}}^{\text{Ly}\alpha} > 10^{40}$  for the two mass limits. This is also the number of galaxies used to calculate the corresponding fractions shown in the main plot. We find that when we take all galaxies with  $M_\star > 10^6$  ( $10^5$ )  $M_\odot$ , LAEs brighter than  $10^{40}$  erg/s can account for 95% (97%) of the total ionizing luminosity.

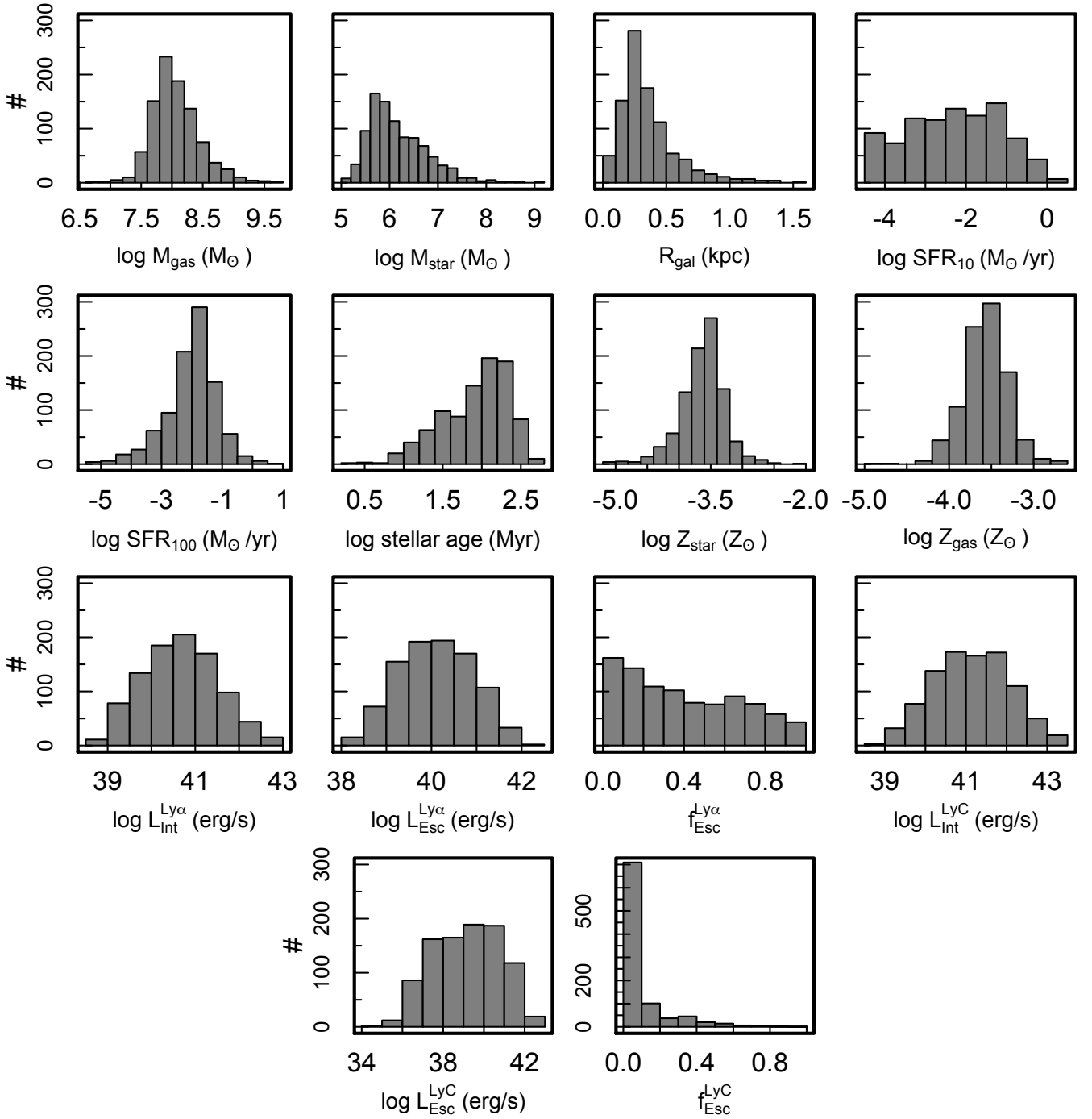


Fig. A.5: Histogram of the 14 galaxy properties (gas mass, stellar mass, galaxy radius,  $\text{SFR}_{10}$ ,  $\text{SFR}_{100}$ , stellar age, stellar and gas metallicity, intrinsic and escaping luminosities and escape fractions of  $\text{Ly}\alpha$  and  $\text{LyC}$ , as described in § 4.1.1) for our sample of 940 galaxies (§ 4.2.1) that were used to build the predictive models (§ 4.2.2).

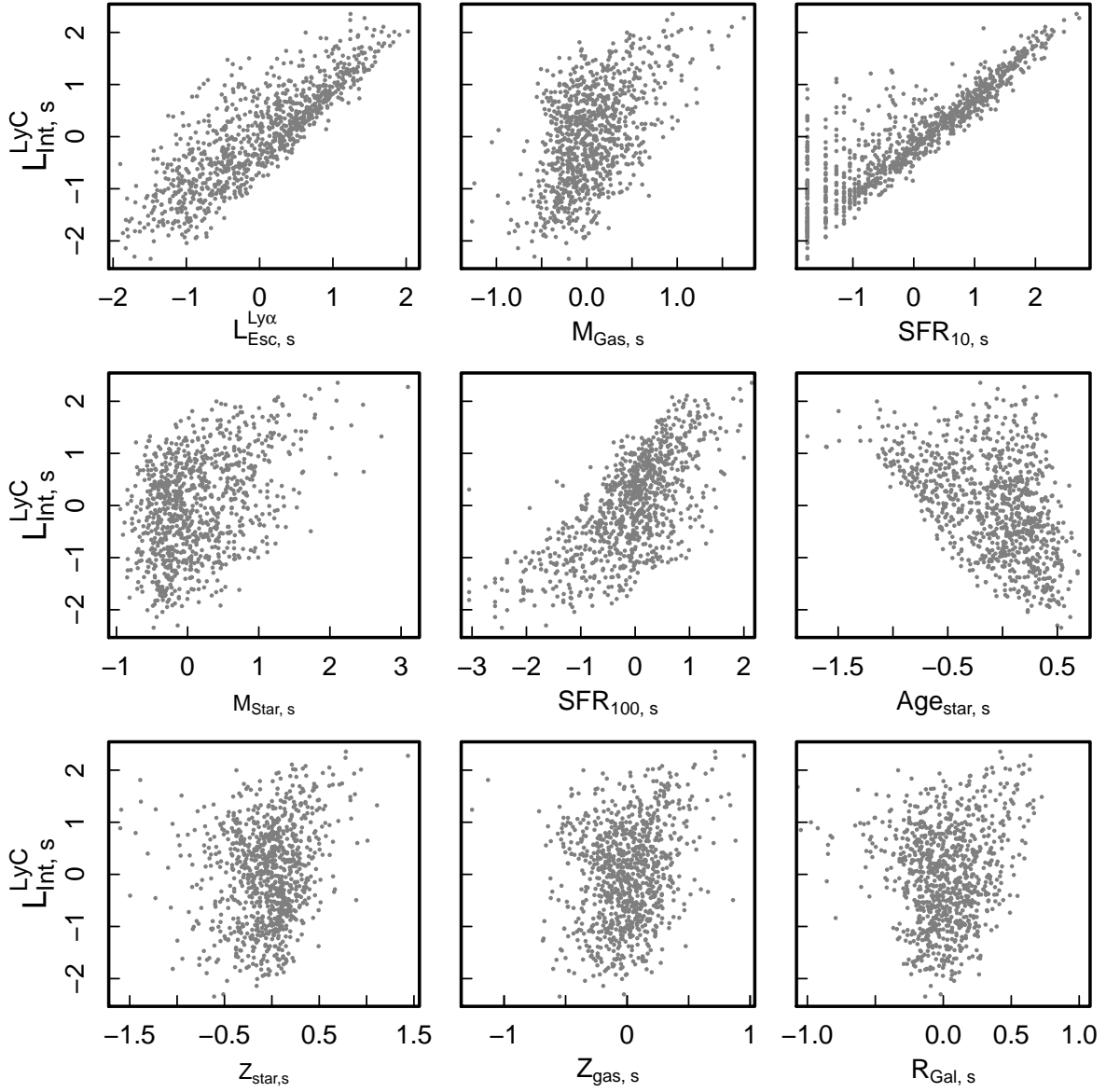


Fig. A.6:  $L_{\text{int}}^{\text{LyC}}$  vs. all  $x$ -variables in our model. All quantities here are scaled as prescribed in § 4.1.1. We find that several properties, especially,  $\text{SFR}_{10}$  and  $L_{\text{esc}}^{\text{Ly}\alpha}$  correlate very well with  $L_{\text{int}}^{\text{LyC}}$ , which suggests that the multivariate linear regression model will be a choice for predicting  $L_{\text{int}}^{\text{LyC}}$ .

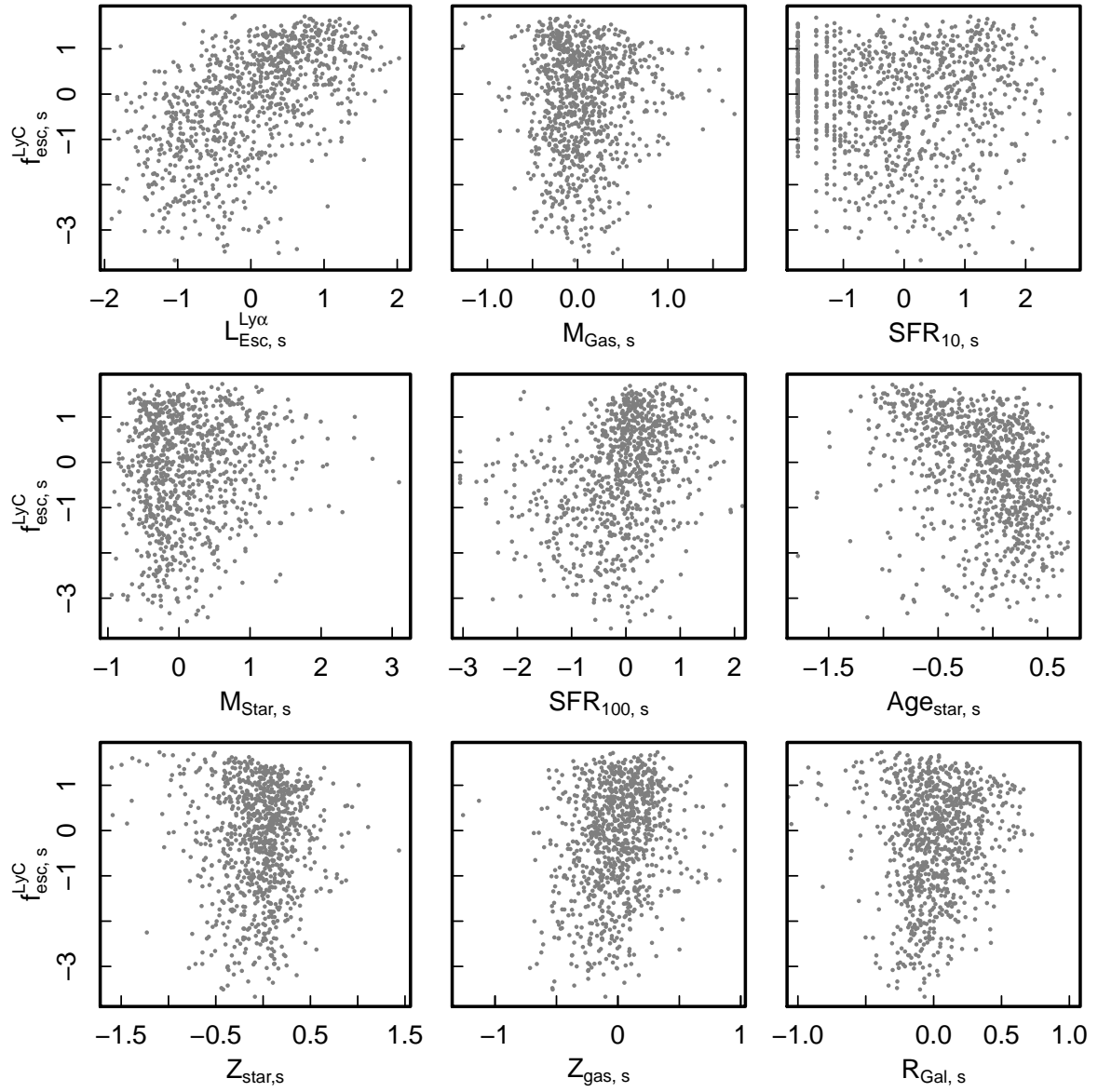


Fig. A.7: Same as A.6 but for the response variable  $f_{\text{esc}}^{\text{LyC}}$ .