

Mapping reionization bubbles in JWST era

II. Inferring the position and characteristic size of individual bubbles

Ivan Nikolić^{1,*}, Andrei Mesinger¹, Charlotte A. Mason^{2,3}, Ting-Yi Lu^{2,3}, Mengtao Tang⁵, David Prelogović^{1,4}, Samuel Gagnon-Hartman¹, and Daniel P. Stark⁵

¹ Scuola Normale Superiore, Piazza dei Cavalieri 7, 56125 Pisa, PI, Italy

² Cosmic Dawn Center (DAWN), Denmark

³ Niels Bohr Institute, University of Copenhagen, Jagtvej 128, 2200 Copenhagen N, Denmark

⁴ Scuola Internazionale Superiore di Studi Avanzati (SISSA), Via Bonomea 265, 34136 Trieste, Italy

⁵ Steward Observatory, University of Arizona, 933 N Cherry Ave, Tucson, AZ 85721, USA

Received 14 January 2025 / Accepted 27 May 2025

ABSTRACT

The James Webb Space Telescope (JWST) is discovering an increasing number of galaxies well into the early stages of the epoch of reionization (EoR). Many of these galaxies are clustered with strong Lyman-alpha ($\text{Ly}\alpha$) emission, which indicates surrounding cosmic HII regions that would facilitate $\text{Ly}\alpha$ transmission through the intergalactic medium (IGM). Detecting these HII bubbles would allow us to connect their growth to the properties of the galaxies inside them. We developed a new forward-modeling framework to estimate the size of the local HII region and its location based on $\text{Ly}\alpha$ spectra of galaxy groups in the early stages of the EoR. Our model uses the complementary information provided by neighboring sightlines through the IGM. Our forward models sample the main sources of uncertainty, including (i) the global neutral fraction, (ii) the EoR morphology, (iii) emergent $\text{Ly}\alpha$ emission, and (iv) NIRSpec instrument noise. Depending on the availability of complementary nebular lines, $\sim 0.006\text{--}0.01$ galaxies per cMpc^3 are required for a confidence of $\geq 95\%$ that the location and size of the HII bubble as recovered by our method is accurate to within ~ 1 comoving Mpc. This approximately corresponds to some dozens of galaxies at $z \sim 7\text{--}8$ in an $\sim 2 \times 2$ tiled pointing with JWST NIRSpec. A sample like this can be achieved with a targeted survey that is complete down to $M_{\text{UV}}^{\text{min}} \lesssim -19\text{--}-17$, depending on the overdensity of the field. We tested our method on 3D EoR simulations and on mispecified equivalent-width distributions. We accurately recovered the HII region that surrounded the targeted galaxy groups in both cases.

Key words. galaxies: high-redshift – intergalactic medium – dark ages, reionization, first stars

1. Introduction

The epoch of reionization (EoR) marks an important milestone in the evolution of the Universe. UV radiation from the first clustered galaxies reionized their surrounding intergalactic medium (IGM). These HII bubbles expanded, percolated, and eventually permeated all of space, completing the final phase change of our Universe. The timing and morphology of the EoR tell us which galaxies caused this, and it describes the role of the IGM clumps that regulated the end stages (e.g., McQuinn et al. 2007; Sobacchi & Mesinger 2014).

Lyman- α emission from galaxies is an especially useful tool for studying the early stages of the EoR, when HII regions were relatively small such that the neutral IGM left a strong imprint via damping wing absorption (e.g., see review in Dijkstra 2014). It is a common approach to estimate the mean neutral fraction of the IGM using a statistically large sample of galaxies (e.g., Mesinger & Furlanetto 2008a; Stark et al. 2010; Mesinger et al. 2015; Mason et al. 2018; Jung et al. 2020; Bolan et al. 2022; Jones et al. 2024; Nakane et al. 2024). Galaxies reside in biased regions of the IGM, however, and the connection of the corresponding damping wing signature to the mean neutral fraction depends very strongly on the model (e.g., Mesinger & Furlanetto 2008b; Lu et al. 2024). In contrast, unbiased probes such as the

Lyman- α forest are sourced from representatively-large volumes of the IGM, and can already tightly constrain the mean neutral fraction during the latter half of the EoR (Qin et al. 2021, 2025; Bosman et al. 2022).

In addition to estimating the global neutral fraction from the IGM $\text{Ly}\alpha$ damping wings, the presence (or absence) of an individual HII region surrounding an observed group of galaxies could be inferred (Tilvi et al. 2020; Endsley & Stark 2022; Jung et al. 2022; Hayes & Scarlata 2023). This might allow us to connect the growth of the local HII region to the properties of the galaxies inside it.

Several such estimates of HII bubble sizes would allow us to understand which type of galaxies drove reionization (e.g., faint/bright; McQuinn et al. 2007; Mesinger et al. 2016), and this would be possible well before the advent of tomographic 21 cm maps with the Square Kilometer Array (SKA). Fortunately, the James Webb Space Telescope (JWST) is providing spectra from an ever-increasing number of galaxy groups at high redshift that can be used for this purpose (e.g., Saxena et al. 2023; Witstok et al. 2025; Tang et al. 2023, 2024a; Chen et al. 2024; Umeda et al. 2024; Napolitano et al. 2024).

However, the interpretation of these observations has so far been fairly approximate. An IGM damping wing in each galaxy is estimated independently of its neighbors. This wastes invaluable information because neighboring galaxies provide

* Corresponding author.

complimentary sightlines into the local EoR morphology. The result is that the studies focusing on individual galaxies generally only predict lower limits for the radii of local HII regions. Furthermore, studies tend to ignore one or more than one important source of stochasticity in the EoR morphology, intrinsic galaxy emission, and/or telescope noise.

We develop a new framework to infer the local HII region size and location from Ly α observations of a galaxy group. Our formalism accounts for the relative position of each galaxy with respect to the host and nearby HII bubbles by creating self-consistent forward models of JWST/NIRSpec spectra for each galaxy. We account for the relevant sources of uncertainty (and/or stochasticity), including (i) the IGM mean neutral fraction, $\bar{x}_{\text{HI}}$; (ii) the EoR morphology, given $\bar{x}_{\text{HI}}$; (iii) the emergent Ly α emission, given the observed UV magnitudes; and (iv) the NIRSpec instrument noise. Unlike many previous studies, we made no assumptions about the unknown relative contribution of observed versus unobserved galaxies to the growth of the local HII region. We quantify how many galaxies are required to robustly detect individual ionized regions with an accuracy of $\leq 10\%$ in their inferred location and characteristic size during the early stages of EoR. This work is a companion to [Lu et al. \(2025\)](#), in which we presented a complementary formalism to detect edges of ionized regions using empirically calibrated relations.

This paper is structured as follows. In Section 2 we present our forward-modeling pipeline for Ly α galaxy spectra during the EoR. We introduce the procedure with which we infer the size and location of the surrounding HII region in Section 3. We apply our framework to mock data and show our main results in Section 4. We build further confidence by performing out-of-distribution tests in Section 5. In Section 6 we quantify observational requirements for detecting individual HII regions, and we conclude in Section 7. All quantities are presented in comoving units unless stated otherwise. We assume a standard Λ CDM cosmology ($\Omega_{\text{m}}, \Omega_{\text{b}}, \Omega_{\Lambda}, h, \sigma_8, n_s = 0.310, 0.049, 0.689, 0.677, 0.81, 0.963$), with parameters consistent with the latest estimates from [Planck Collaboration VI \(2020\)](#). All quantities are quoted in comoving units and are evaluated in the rest frame unless stated otherwise.

2. Observing Lyman- α spectra from galaxies during the EoR

Our fiducial setup is shown in Fig. 1. An HII region in an observed galaxy field is characterized as a sphere, with a center location (O_{b}) and characteristic radius (R_{b}). This is the local or target HII region whose properties we aim to infer. Nearby ionized regions are also present. Observed galaxies can be located both inside and outside HII regions; we denote three such galaxies with A, B, and C and highlight their sightlines toward the observer with red arrows.

Specifically, we wished to determine the conditional probability of the HII bubble center, O_{b} , and radius, R_{b} , based on the observed Ly α spectra of N_{gal} galaxies in a field with a central redshift z ,

$$\mathcal{P}(O_{\text{b}}, R_{\text{b}} | \mathbf{x}^i, f_{\alpha}^i(\lambda), M_{\text{UV}}^i, z), \quad i \in [1, N_{\text{gal}}]. \quad (1)$$

Here, $\mathbf{x}^i$, M_{UV}^i , and $f_{\alpha}^i(\lambda)$ are vectors of the Cartesian coordinates of the galaxies, the UV magnitudes, and the observed Ly α spectra. For each galaxy, the observed spectrum in the rest frame can be written as

$$f_{\alpha}(\lambda) = L_{\alpha} J(\lambda) e^{-\tau_{\text{EoR}}(\lambda)} + \mathcal{N}(\lambda), \quad (2)$$

where L_{α} is the emergent¹ Lyman- α luminosity of a galaxy, $J(\lambda)$ is the normalized, emergent Lyman- α profile, $\exp[-\tau_{\text{EoR}}(\lambda)]$ accounts for IGM attenuation, and $\mathcal{N}(\lambda)$ is the spectrograph noise.

In the schematic shown in Fig. 1, galaxy A lies outside of an ionized bubble, and its Lyman- α flux will therefore be strongly attenuated by the neutral IGM (i.e., it will have a large $\tau_{\text{EoR}}(\lambda)$). Lyman- α flux from galaxy A is only expected when it has a high emergent luminosity, L_{α} , and when its Ly α profile, $J(\lambda)$, is strongly redshifted from the systemic z (e.g., [Dijkstra 2014](#)). Galaxy B is close to the center of the local HII bubble and will have (on average, see the right panel of Fig. 5), which shows the lowest Ly α damping wing attenuation from the patchy EoR. The observed flux depends on all of the terms in Eq. (2), however, each of which can have a sizable stochasticity. Below, we detail how we compute each of these terms.

2.1. Emergent Lyman- α profile

We started with the Lyman-alpha profile emerging into the IGM, $J(\lambda)$, normalized to integrate to unity. In order to escape from the ISM of the galaxy, Lyman- α photons must diffuse spectrally, which leads to a double-peaked Lyman-alpha shape ([Neufeld 1990](#); [Hu et al. 2023](#); [Hutter et al. 2023](#); [Almada Monter & Gronke 2024](#)). Because the line is resonant, the blue peak is generally absorbed even by the ionized IGM at $z \gtrsim 5$ (but see [Meyer et al. 2021](#) for some putative rare counterexamples). Following [Mason et al. \(2018\)](#), we modeled the remaining red peak as a Gaussian,

$$J(\lambda) = \frac{2}{v_{\alpha}} \sqrt{\frac{\ln 2}{\pi}} \exp\left(-\frac{(\Delta v - v_{\alpha})^2}{v_{\alpha}^2 / (2 \ln 2)}\right), \quad (3)$$

where v_{α} represents the velocity offset from the systemic velocity of the center of the line, and $\Delta v = ((\lambda - \lambda_{\alpha})c / \lambda_{\alpha})$ is the velocity difference from the resonant wavelength of the line, $\lambda_{\alpha} = 1215.16 \text{ \AA}$. As in [Mason et al. \(2018\)](#), we assumed for simplicity that the FWHM of the line is equal to the velocity of the offset (e.g., [Verhamme et al. 2018](#)). Although these profiles are motivated by lower-redshift observations (e.g., [Yamada et al. 2012](#); [Orlitová et al. 2018](#); [Hu et al. 2023](#)), we note that our framework will easily accommodate any distribution for $J(\lambda)$ when we have better models for the emergent spectra. We also assumed that all Lyman-alpha photons with a velocity offset below the circular velocity, v_{circ} , of the host halo are absorbed by the CGM ([Dijkstra et al. 2011](#); [Laursen et al. 2011](#)). In Fig. 2 we show an example of the emergent profile in blue for a galaxy with an UV magnitude $M_{\text{UV}} = -20.0$ and a velocity offset $v_{\alpha} = 270 \text{ km/s}$, with a $v_{\text{circ}} = 160 \text{ km/s}$.

The PDF of the emergent velocity offset is well described by a log-normal distribution ([Steidel et al. 2014](#); [De Barros et al. 2017](#); [Stark et al. 2017](#); [Mason et al. 2018](#)):

$$\mathcal{P}(v_{\alpha} | M_{\text{UV}}) = \frac{1}{\sqrt{2\pi} \ln 10 v_{\alpha} \sigma_v} \exp\left(-\frac{(\log_{10} v_{\alpha} - \bar{v}_{\alpha}(M_{\text{UV}}))^2}{2\sigma_v^2}\right), \quad (4)$$

¹ Throughout we use the term “emergent” to refer to quantities describing values escaping from the galaxy into the IGM. Therefore the emergent amplitude, L_{α} , and profile, $J(\lambda)$, are determined by Lyman- α radiative transfer through the interstellar medium and the circumgalactic medium (e.g. [Neufeld 1990](#); [Laursen et al. 2011](#)). We did not model the details of this radiative transfer, but instead relied on empirical relations based on post-EoR observations to determine the conditional distributions of L_{α} and $J(\lambda)$.

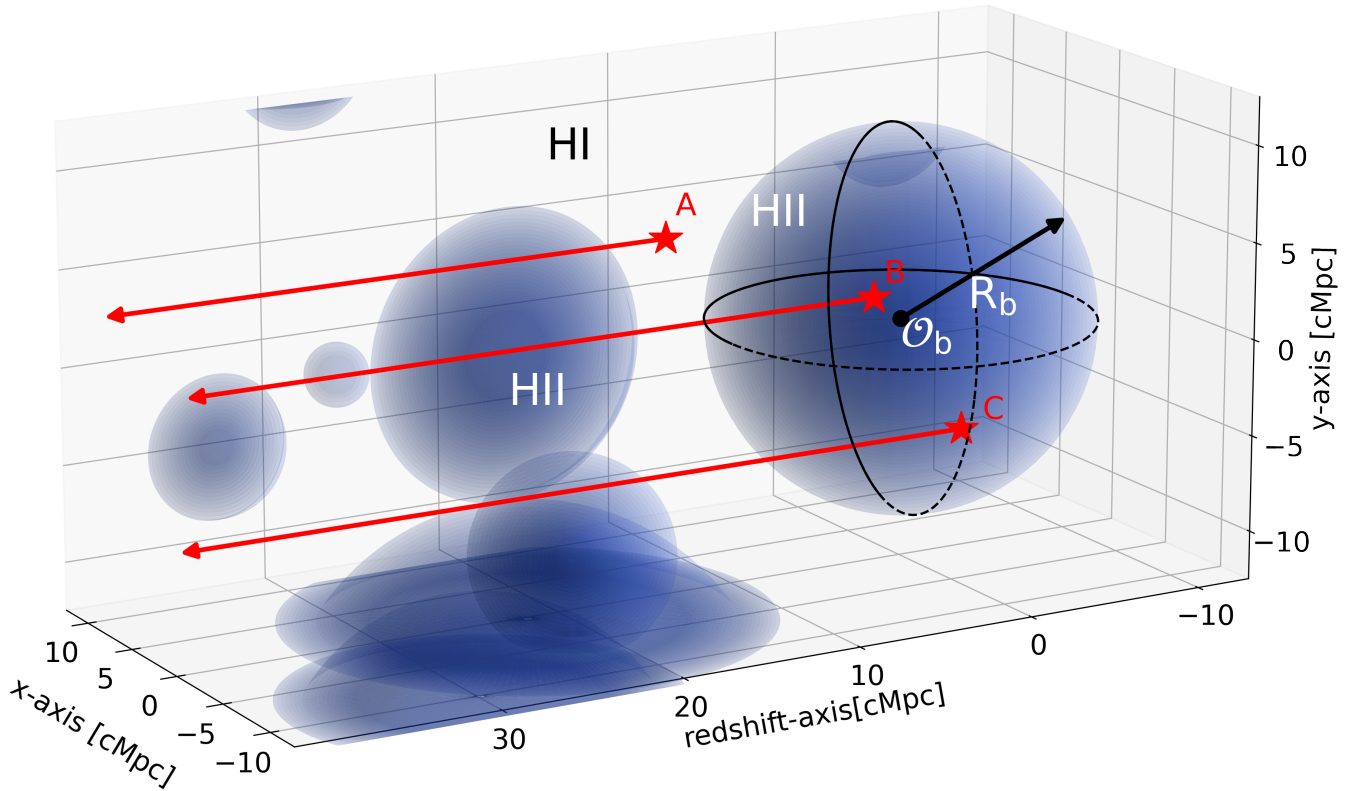


Fig. 1. Schematic of our framework. Cosmic HII regions are shown in blue, embedded in an otherwise neutral IGM. We observe Lyman- α spectra from a group of galaxies (three example galaxies are denoted A, B, and C), and we wish to infer the central HII bubble (characterized by a sphere with radius R_b and center O_b). Our framework combines the complimentary information provided by neighboring sightlines toward the galaxies (red arrows) to account for the main sources of stochasticity.

where the mean velocity offset is correlated with the UV magnitude (Stark et al. 2017, though see Bolan et al. 2024),

$$\log_{10} \bar{v}_\alpha(M_{UV}, z) = 0.32\gamma(M_{UV} + 20.0 + 0.26z) + 2.34 \quad (5)$$

and $\sigma_v = 0.24$, $\gamma = -0.3$ for $M_{UV} \geq -20.0 - 0.26z$, and $\gamma = -0.7$ otherwise. We show this distribution in the upper panel of Fig. 3. Although there is some indication of a mild redshift evolution in this distribution (e.g., Tang et al. 2024a; Witstok et al. 2024), we show in Section 5 that our results are insensitive to such changes.

2.2. Emergent Lyman- α luminosity

The absolute normalization of the profile discussed above (i.e., the emergent Lyman- α luminosity L_α) is generally defined via the so-called rest-frame equivalent width (e.g., Dijkstra & Wyithe 2012),

$$W = \frac{L_\alpha}{L_{1500,\nu}} \frac{\lambda_\alpha}{\nu_\lambda} \left(\frac{\lambda_{UV}}{\lambda_\alpha} \right)^{\beta+2}, \quad (6)$$

where $L_{1500,\nu}$ [erg/s/Hz] is the specific UV luminosity evaluated at $1500 \text{ \AA}$ obtained from the continuum flux $f_{1500,\nu}$: $L_{1500,\nu} = 4\pi d_L^2 f_{1500}$ where f_{1500} is given in units of [erg/s/cm²/Hz], where d_L is the luminosity distance to the source, $\nu_\lambda = 2.47 \times 10^{15} \text{ Hz}$, $\lambda_\alpha = 1215.67 \text{ \AA}$, $\lambda_{UV} = 1500 \text{ \AA}$ is the rest-frame wavelength at which the UV magnitude is measured and β is the UV slope (which we assumed to be $\beta = -2.0$ for simplicity).

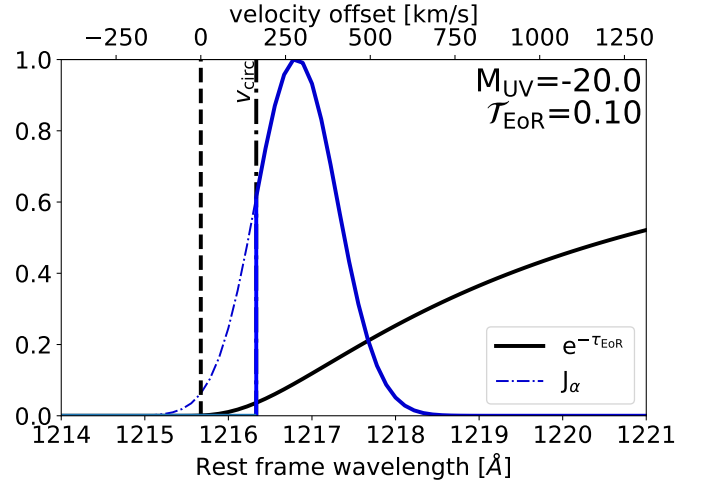


Fig. 2. Emergent Lyman-alpha line and IGM opacity as a function of wavelength. The solid blue line represents the normalized Lyman- α emergent profile, and the dot-dashed blue line demarcates the flux absorbed by the circumgalactic medium blueward of the circular velocity of the galaxy. The solid black line illustrates an IGM damping wing attenuation profile, taken from a random sightline at $\bar{x}_{HI} = 0.65$ and $R_b = 10 \text{ cMpc}$ (see text for details). The Lyman-alpha transmission integrated over all wavelengths for this example would be $\mathcal{T}_{EoR} \equiv \int d\lambda J(\lambda) e^{-\tau_{EoR}(\lambda)} = 0.10$.

For galaxies at $z < 6$, where we expect τ_{EoR} to be negligible, Mason et al. (2018) found the following fit based on data from

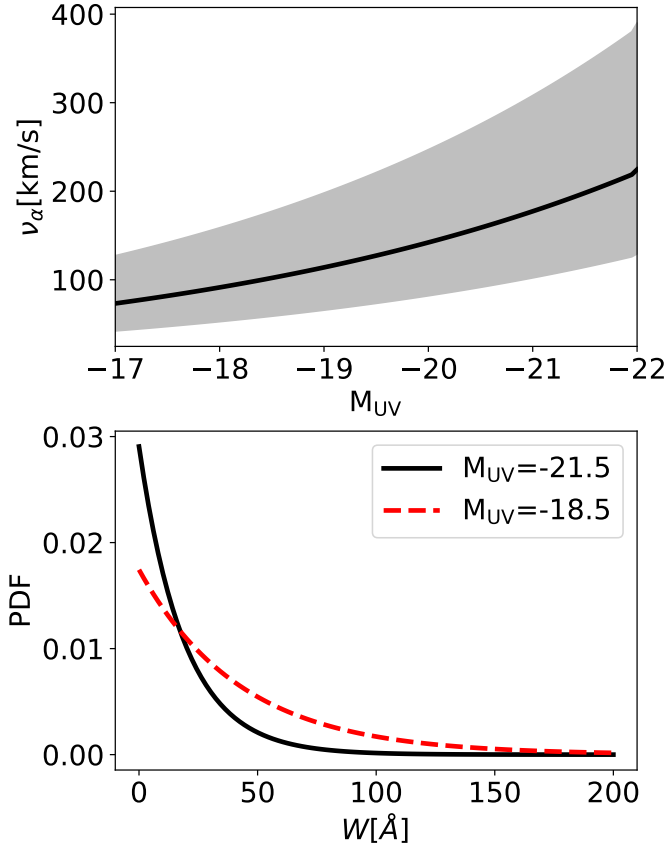


Fig. 3. Upper panel: Mean (black line) and 1σ (shaded region) of the velocity offset distribution as a function of UV magnitude (Eq. (4)). Lower panel: PDF of the Lyman-alpha rest-frame equivalent widths. The dashed black and red curves represent the distribution of the equivalent widths (W) for Lyman-alpha emitters (Eq. (7)). The equivalent widths of the nonemitters are described by a delta function at $W = 0 \text{ \AA}$ (not shown) normalized such that the integral of the PDF is 1. The distribution is shown for $M_{\text{UV}} = -18.5$ and -21.5 .

De Barros et al. (2017):

$$p_6(W|M_{\text{UV}}) = \frac{A(M_{\text{UV}})}{W_c(M_{\text{UV}})} e^{-\frac{W}{W_c(M_{\text{UV}})}} H(W) + [1 - A(M_{\text{UV}})] \delta(W), \quad (7)$$

where $A(M_{\text{UV}})$ is the fraction of intrinsic emitters for a given M_{UV} , and $W_c(M_{\text{UV}})$ is the characteristic scale of the distribution, which is anticorrelated with M_{UV} . $H(W)$ is the Heaviside step function, and $\delta(W)$ is a Dirac delta function. We used the following fit, as in Mason et al. (2018): $A = 0.65 + 0.1 \tanh[3(M_{\text{UV}} + 20.75)]$ and $W_c = 31 + 12 \tanh[4(M_{\text{UV}} + 20.25)] \text{ \AA}$. The distribution of emergent equivalent widths is shown in the lower panel of Figure 3 for two UV magnitudes. We note that our framework can easily accommodate different EW distributions (e.g., Treu et al. 2012; Lu et al. 2025; Tang et al. 2024a). We show in Section 5 that our results are not sensitive to the choice of distribution shape, however.

2.3. IGM damping wing absorption

During the EoR, the damping wing absorption from the residual HI patches along the line of sight can strongly attenuate the Ly α line (cf. the $\exp[-\tau_{\text{IGM}}]$ curve in Fig. 2). The optical depth of the damping wing is mostly sensitive to the distance to the nearest neutral HI patch (e.g., Miralda-Escudé 1998). For

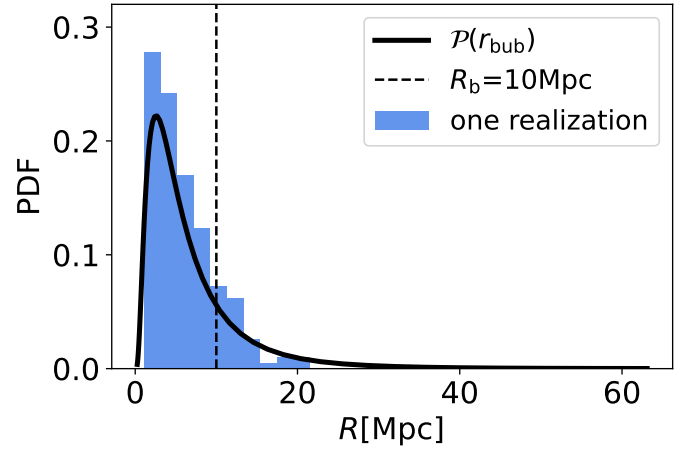


Fig. 4. Distribution of HII region radii used in our overlapping spheres algorithm for the large-scale EoR morphology (see text for details). The vertical dashed line marks $R_b = 10 \text{ cMpc}$ which is our fiducial size for the central ionized region whose properties we aim to infer (see Fig. 1).

this reason, we are able to infer the size of the local HII bubble in this work (see also the complementary empirical approach in Lu et al. 2025 based on empirical $\tau_{\text{IGM}}-R_b$ relations from an EoR simulation).

Nevertheless, the surrounding EoR morphology beyond the local HII region does contribute to the total τ_{IGM} as an additional source of scatter (e.g. Mesinger & Furlanetto 2008b).

We generated an EoR morphology at a given $\bar{x}_{\text{HI}}$ by placing overlapping spherical HII regions with a radius distribution given by (cf. Fig. 1)

$$\mathcal{P}(r_{\text{bub}}|\bar{x}_{\text{HI}}) = \frac{1}{\sqrt{2\pi}\sigma_{\text{bub}}} \cdot \exp\left(-\frac{(\log r_{\text{bub}} - \mu_{\text{bub}})^2}{\sigma_{\text{bub}}^2}\right). \quad (8)$$

We sampled from the above distribution of radii and randomly chose the center location. We stopped when the volume-filling factor of ionized regions reached the desired value, $\bar{x}_{\text{HI}}$. Although this is simplistic, the overlapping ionized spheres resulted in a similar EoR morphology as is seen in cosmological radiative transfer simulations (e.g. Zahn et al. 2011; Mesinger et al. 2011; Ghara et al. 2018; Doussot & Semelin 2022). We did not assume that we know the true value $\bar{x}_{\text{HI}}^{\text{true}}$ a priori; instead, we sampled a prior distribution of $\bar{x}_{\text{HI}}$ from complimentary observations while performing inference (see Section 3 for more details).

For simplicity, we ignored in this proof-of-concept work the $\bar{x}_{\text{HI}}$ dependence of the bubble size distribution in Eq. (8) and took constant values of $\mu_{\text{bub}} = \log(5 \text{ cMpc})$ and $\sigma_{\text{bub}} = 0.5$. These choices in our algorithm roughly reproduce the ionized bubble scales seen in simulations during the early stages of the EoR (e.g., Mesinger & Furlanetto 2007; Giri et al. 2018; Lu et al. 2025; Doussot & Semelin 2022; Neyer et al. 2024). The r_{bub} from Eq. (8) does not directly translate into any of the metrics that are commonly used to characterize the EoR morphology (e.g. Lin et al. 2016; Giri et al. 2018), and comparisons must therefore be made a posteriori. In future work, we will calibrate Eq. (8) to EoR simulations and will also condition it on the matter field to account for the (modest) bias of HII regions at early times (e.g. Fig. 12 in Sobacchi & Mesinger 2014). We show our assumed bubble size distribution in Fig. 4 together with one realization in a $200 \times 80 \times 80 \text{ cMpc}^3$ volume.

For a given realization of the EoR morphology and galaxy field (cf. Fig. 1), we computed the optical depth of the damping

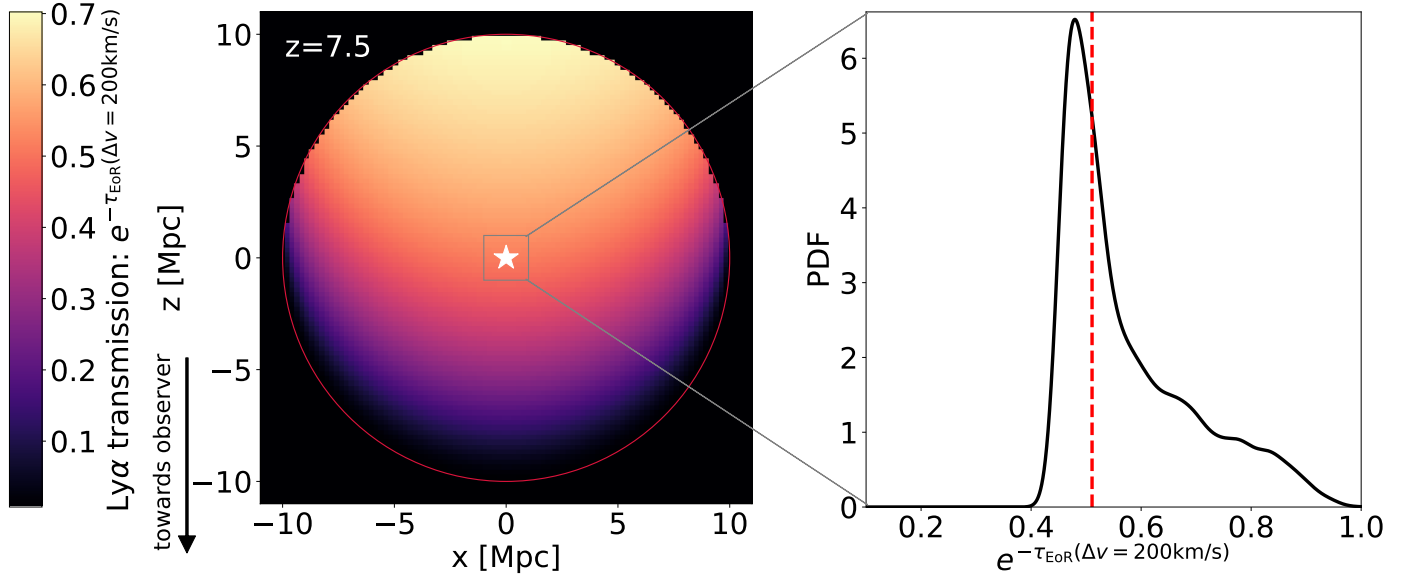


Fig. 5. Left panel: Mean Ly α transmission evaluated at $\Delta v = +200$ km/s ($e^{-\tau_{\text{EoR}}(\Delta v = +200)}$ km/s) as a function of the position inside an ionized bubble. The observer is located toward the bottom of the figure. We computed the mean transmission by averaging over 10 000 realizations of EoR morphologies, given an assumed neutral fraction of $\bar{x}_{\text{HI}} = 0.65$ at $z = 7.5$ (see text for details). Right panel: Distribution of Ly α transmission at $\Delta v = +200$ km/s from these realizations, evaluated at a fixed position inside the bubble denoted by the white star (see text for details). The mean is marked by the vertical dashed line.

wing by casting rays from the galaxy locations and summing the contributions from the optical depth from all HI patches along the LOS (e.g. [Miralda-Escudé 1998](#)),

$$\tau_{\text{EoR}}(\lambda_{\text{em}}) = \frac{\tau_{\text{GP}} R_\alpha}{\pi} \sum_i x_{\text{H},i} \left(\frac{1+z_{b,i}}{1+z_\lambda} \right)^{3/2} \left[I\left(\frac{1+z_{b,i}}{1+z_\lambda} \right) - I\left(\frac{1+z_{e,i}}{1+z_\lambda} \right) \right], \quad (9)$$

where $z_\lambda = \frac{\lambda_{\text{em}}}{\lambda_\alpha} (1+z_{\text{obs}}) - 1$, λ_{em} is the wavelength at which we evaluated the optical depth, $\lambda_\alpha = 1215.67 \text{ \AA}$ is the Lyman- α resonant wavelength, and z_{obs} is the redshift of the emitting galaxy. $\tau_{\text{GP}} \sim 7.16 \cdot 10^5 ((1+z_{\text{obs}})/10)^{3/2}$ is the Gunn-Peterson optical depth, $R_\alpha = \frac{\Lambda}{4\pi\nu_\alpha}$, $\Lambda = 6.25 \cdot 10^8 \text{ s}^{-1}$ is the decay constant, and $\nu_\alpha = 2.47 \cdot 10^{15} \text{ Hz}$ is the Lyman- α resonant frequency. In the above equation, $I(x)$ is given by

$$I(x) \equiv \frac{x^{9/2}}{1+x} + \frac{9}{7}x^{7/2} + \frac{9}{5}x^{5/2} + 3x^{3/2} + 9x^{1/2} - \ln \left| \frac{1+x^{1/2}}{1-x^{1/2}} \right|. \quad (10)$$

The summation accounts for every neutral patch encountered along the LOS, with a given patch, i , extending from $z_{b,i}$ to $z_{e,i}$. We assumed that ionized patches have no neutral hydrogen atoms and therefore do not contribute to the attenuation.

In the left panel of Fig. 5, we show the mean IGM transmission, evaluated at $\Delta v = +200$ km/s redward of the systemic redshift, as a function of the position inside the central HII bubble of $R_b = 10 \text{ cMpc}$. The observer is located toward the bottom of the figure. This mean transmission was computed by averaging over 10 000 realizations of EoR morphologies outside the central bubble, assuming $\bar{x}_{\text{HI}} = 0.65$ (e.g., one such realization is shown in Fig. 1). As expected, there is a clear trend of increased transmission for galaxies located at the far end of the central HII bubble. The mean transmission is a function of the distance of the galaxy to the bubble edge in the direction toward the observer. In [Lu et al. \(2025\)](#) we used these trends to define empirical edge-detection algorithms.

At every location in the bubble, however, there is sizable sightline-to-sightline scatter in the IGM transmission. We quantify this in the right panel of Fig. 5, where we show the transmission PDF constructed from the 10 000 realizations of the EoR morphology external to the central bubble. The sightlines we used to compute this PDF originated in the location marked by the white star in the left panel. The PDF is quite broad and asymmetric (see also Fig. 2 in [Mesinger & Furlanetto 2008a](#) and [Lu et al. 2024](#)). While it is difficult for the IGM to completely attenuate Ly α for a galaxy that is located at the center of this bubble, some morphologies can result in large stretches of ionized IGM in the direction of the observer, driving a high-transmission tail in the PDF. The width of this PDF highlights that it is important to account for stochasticity in the EoR morphology when galaxy Lyman- α observations are interpreted (e.g., [Mesinger & Furlanetto 2008b](#); [Mesinger et al. 2015](#); [Mason et al. 2018](#); [Bruton et al. 2023](#); [Keating et al. 2024](#); [Lu et al. 2025](#)).

2.4. Including NIRSpect noise

We took NIRSpect on the JWST for our fiducial spectrograph. It already measures Ly α spectra from galaxy groups during the EoR (e.g., [Tang et al. 2024b](#); [Witstok et al. 2025](#)).

To forward-model NIRSpect observations, we binned the observed flux $f_\alpha(\lambda) = L_\alpha J(\lambda) e^{-\tau_{\text{EoR}}(\lambda)}$ to $R \sim 2700$ (the high resolution of NIRSpect) and then added Gaussian noise $\mathcal{N}$ to each spectral bin with a standard deviation of $\sigma(\mathcal{N}) = 2 \times 10^{-20} \text{ ergs}^{-1} \text{ cm}^{-2} \text{ \AA}^{-1}$. This level of noise can be obtained with a few hours of integration on NIRSpect ([Bunker et al. 2023](#); [Saxena et al. 2023](#); [Tang et al. 2024b](#)), and it corresponds to an uncertainty on the integrated flux ($\mathcal{F}_{\text{int}} \equiv \int f_\alpha(\lambda) d\lambda$) of $\sigma(\mathcal{N}_{\text{int}}) = 1 \times 10^{-19} \text{ ergs}^{-1} \text{ cm}^{-2}$ (estimated assuming that the emission line is spectrally unresolved). This can be further translated into 5σ limiting equivalent widths of $W = 25 \text{ \AA}$ ($W = 60 \text{ \AA}$) for $M_{\text{UV}} = -18$ ($M_{\text{UV}} = -17$). The recovery is worse for shallower

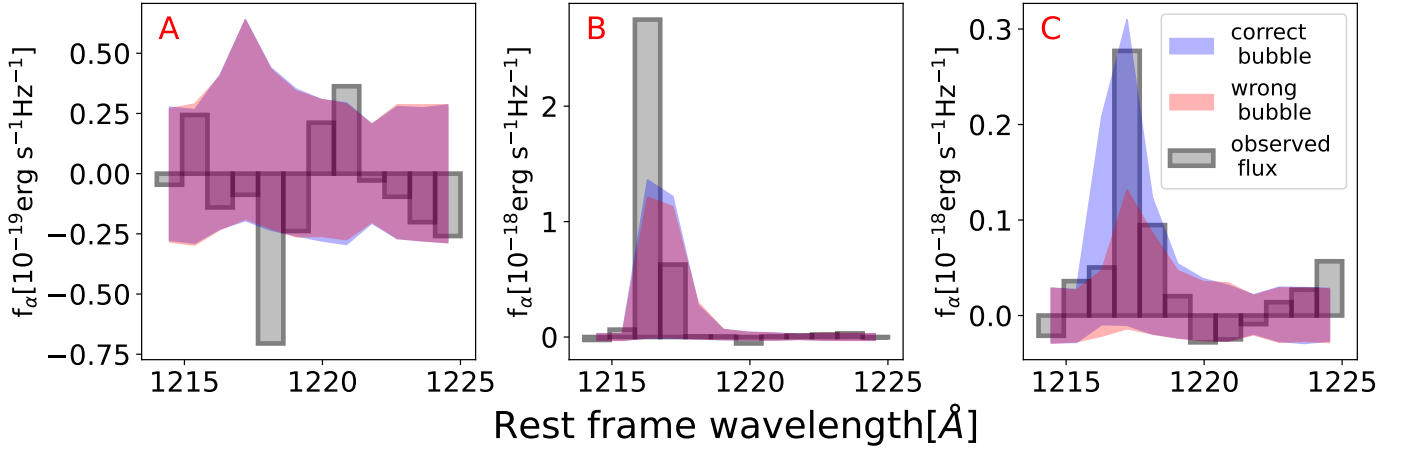


Fig. 6. Observed flux (gray) corresponding to the galaxies A, B, and C from the mock observation shown in Fig. 1. We show the 68% C.L. of the likelihood assuming the correct HII bubble location and radius in blue, $(R_b, O_b) = (R_b^{\text{true}}, O_b^{\text{true}})$. We show the 68% C.L. of the flux likelihood assuming the correct HII bubble location, but a slightly smaller radius in red, $R_b = 0.8R_b^{\text{true}}$.

observations, but our results did not improve significantly for integrations deeper than this fiducial value.

We rebinned the spectra to lower values of R and tested the inference with a coarser resolution. By performing additional binning, we lose some information on the observed line profile (e.g., Byrohl & Gronke 2020) but lower the dimensionality of our likelihood (see Section 3). In future work, we will explore more sophisticated inference approaches that can scale to high-dimensional likelihoods (Cranmer et al. 2020; Anau Montel et al. 2024). We empirically settled on rest frame $\Delta\lambda \sim 1$ Å as our bin width (i.e., $R \sim 1000$), which corresponds to the medium-resolution NIRSpec grating (Jakobsen et al. 2022, see Section 3).

3. Inferring the local HII bubble

With the above framework, we created mock observations and the corresponding forward models by sampling each of the terms in Eq. (2). We detail this procedure below (see Fig. 6).

3.1. Mock observations

We first constructed a mock observation of N_{gal} galaxies in a survey volume of $V_{\text{survey}} = 20 \times 20 \times 20 \text{ cMpc}^3$ ($\sim 7' \times 7' \times \Delta z = 0.07$) at $z = 7.5$. This FoV is roughly motivated by JWST (corresponding to roughly four NIRSpec pointings), although in practice, the forward-modeled volume should be tailored to the specific observation that is interpreted. We placed a bubble with radius $R_b = 10 \text{ cMpc}$ at the (arbitrarily chosen) center of the volume and constructed the surrounding EoR morphology out to distances of 200 cMpc^2 using the prescription from Sect. 2.3 and assuming $\bar{x}_{\text{HI}} = 0.65$. We demonstrate below that our results are insensitive to these fiducial choices.

We assigned random locations to the galaxies, $\mathbf{x}^i$ with $i \in [1, N_{\text{gal}}]$, using rejection sampling to ensure that 75% of the galaxies are inside HII regions on average. This number was motivated by the analysis reported by Lu et al. (2024), who showed that $\sim 25\%$ galaxies are found outside of ionized bubbles for $M_{\text{UV}} > -18$ at $\bar{x}_{\text{HI}} = 0.7$ (their Fig. 3). This is a very

approximate way of accounting for galaxy bias because galaxies and HII regions are both correlated to the large-scale matter field³.

We generated UV magnitudes for each galaxy by sampling the UV luminosity function (LF) from Park et al. (2019) down to a magnitude limit of $M_{\text{UV}} = -18.0$. Each galaxy was then assigned an emergent emission profile according to the procedure in the previous section, which was attenuated by its sight-line through the realization of the EoR morphology. Finally, a noise realization was added to the binned flux to create a mock spectrum for each galaxy (cf. Eq. (2)).

3.2. Maximum likelihood estimate of the bubble size and location

We then interpreted this mock observation by forward-modeling the observed flux for each galaxy and by varying (i) the position and radius of the central HII bubble; (ii) the surrounding EoR morphology; (iii) the neutral fraction of the Universe (within ± 0.1 of the truth, which we fixed conservatively to wider than current limits; Qin et al. 2025); (iv) the emergent Lyman-alpha flux given the observed M_{UV} of the galaxy (i.e., W and v_α); and (v) NIRSpec noise realizations.

We then interpreted this mock observation by forward-modeling the observed flux for each galaxy and by varying (i) the position and radius of the central HII bubble; (ii) the surrounding EoR morphology; (iii) the neutral fraction of the Universe (within ± 0.1 of the truth, which we fixed conservatively to wider than current limits; Qin et al. 2025); (iv) the emergent Lyman-alpha flux given the observed M_{UV} of the galaxy (i.e., W and v_α); and (v) NIRSpec noise realizations.

For each forward model, we computed the likelihood of the mock observation based on the location and radius of the central HII region. Our sampling procedure effectively marginalized over the unknowns (ii)–(v) from above. Because it is

³ This is a reasonable approximation, as evidenced by our results in Section 5.2, where we apply our framework to simulations that account for galaxy and HII region bias self-consistently. Observations of galaxies have demonstrated that the bias dominates clustering at larger scales (some dozen cMpc), while the smaller scales that are relevant for this work are dominated by Poisson noise (Bhowmick et al. 2018; Jespersen et al. 2025; Davies et al., in prep.).

² Neutral IGM at larger distances contributes a negligible amount to the total attenuation because the damping wing profile is steep (Mesinger & Furlanetto 2008b).

numerically challenging to map out the joint likelihood over all of the observed galaxies, we made the simplifying assumption that the likelihood of the observed flux from each galaxy, $f_\alpha^i(\lambda)$, is independent from the other galaxies. This allowed us to write the total likelihood of the observation as a product of the likelihoods of the individual galaxies,

$$\mathcal{L} = \prod_i^{N_{\text{gal}}} \mathcal{L}^i(f_\alpha^i(\lambda) | O_b, R_b, \mathbf{x}^i, M_{\text{UV}}^i, z). \quad (11)$$

While this assumption is clearly incorrect, we only present results here in terms of the maximum likelihood, $\hat{\mathcal{L}}$. We demonstrate below that Eq. (11) provides an unbiased estimate of $\hat{\mathcal{L}}$ ⁴.

It is important to note that *we did not assume a Gaussian likelihood for the observed flux at each wavelength*, as is commonly done. With high S/N spectra, the correlations between flux bins can be significant. We instead directly mapped out the *joint PDF of flux over all wavelength bins*, $\mathcal{L}^i(f_\alpha^i(\lambda_1), f_\alpha^i(\lambda_2), f_\alpha^i(\lambda_3), \dots | O_b, R_b, \mathbf{x}^i, M_{\text{UV}}^i, z)$ using a kernel density estimation over the forward-modeled spectra (see Appendix A for details). This preserved the covariances between the wavelength bins and is commonly known as an implicit likelihood or a simulation-based inference.

We demonstrate this procedure for the three galaxies shown in Fig. 1. For the observed fluxes from galaxies A, B, and C we show the 68% C.L. of the likelihood assuming the correct HII bubble location and radius, $(R_b, O_b) = (R_b^{\text{true}}, O_b^{\text{true}})$. We also show the 68% C.L. of the flux likelihood assuming the correct HII bubble location but a slightly smaller radius, $R_b = 0.8R_b^{\text{true}}$. Galaxy A in this mock observation is outside of the central HII bubble, and therefore the true and slightly smaller values of R_b result in the same likelihood. Galaxy B is located close to the center of the bubble, so both $R_b = R_b^{\text{true}}$ and $R_b = 0.8R_b^{\text{true}}$ give similar values of transmission $\exp[-\tau_{\text{EoR}}]$ (see Fig. 5). This results in only a slight preference for $R_b = R_b^{\text{true}}$. On the other hand, galaxy C is located close to the edge of the bubble in Fig. 1. For this galaxy, the observed spectrum in gray is more consistent with the correct likelihood in blue than with the incorrect likelihood in red. A smaller bubble would imply a stronger IGM attenuation on average at the location of this galaxy, which makes the observed strong Ly α emission less likely. In this specific realization, the joint likelihood (Eq. (11)) of the observed fluxes of A, B, and C is twice higher for the correct value of the bubble radius than for the incorrect value. When progressively more galaxies are included, the maximum likelihood increasingly peaks around the true values for the bubble size and location.

We illustrate this explicitly in Figure 7, in which we plot 2D slices through the log likelihood. The log likelihood normalized to the maximum value is indicated with the color bar. The vertical and horizontal axis in each panel correspond to the sampled bubble radius, R_b , and to the redshift axis of the center, $O_b = (x_b = x_b^{\text{true}}, y_b = y_b^{\text{true}}, z_b)$. The true values, $(R_b^{\text{true}}, z_b^{\text{true}}) = (10 \text{ cMpc}, 0 \text{ cMpc})$, are demarcated with the dashed lines. The columns correspond to increasing number density of observed

galaxies (left to right), and the rows correspond to different realizations of forward models.

Several interesting trends are evident in Figure 7. First, an increasing number of observed galaxies (from left to right in each row) results in an increasingly peaked likelihood, and the maximum settles on the true values for the bubble radius and location. Different realizations of the EoR morphology give different likelihood distributions when a small number of galaxies is observed. As the number density of the observed galaxies is increased, however, the likelihood distributions between different realizations converge (i.e., all sources of stochasticity are averaged out).

The redshift of the bubble center and its radius are also degenerate. Because the transmission is mostly determined by the distance to the bubble edge along the line of sight, a movement of the bubble center farther away from the observer at a fixed radius is roughly degenerate with a decrease in the radius at a fixed center location. This degeneracy is mitigated by a larger number of sightlines to observed galaxies, which allows us to constrain the curvature radius of the bubble.

4. How many galaxies do we need to confidently infer the local HII bubble?

In the previous section, we demonstrated that our framework gives an unbiased estimate of the HII region size and location when it is applied to a large galaxy sample. We quantify below just how large this galaxy sample needs to be in order for us to be confident in our results.

For this purpose, we defined two figures of merit,

$$\text{Err}_{\text{loc}} = \left| \hat{O}_b(n_{\text{gal}}) - O_b^{\text{true}} \right| / R_b^{\text{true}} \quad (12)$$

$$= \sqrt{(\hat{x}_b - x_b^{\text{true}})^2 + (\hat{y}_b - y_b^{\text{true}})^2 + (\hat{z}_b - z_b^{\text{true}})^2} / R_b^{\text{true}}$$

$$\text{Err}_{\text{rad}} = \left| \hat{R}_b(n_{\text{gal}}) - R_b^{\text{true}} \right| / R_b^{\text{true}}. \quad (13)$$

Here, $\hat{O}_b(n_{\text{gal}})$ and $\hat{R}_b(n_{\text{gal}})$ are the maximum likelihood estimates of the HII bubble center and radius computed from a galaxy field with a number density $n_{\text{gal}} = N_{\text{gal}}/V_{\text{survey}}$, and Err_{loc} and Err_{rad} are the corresponding fractional errors (normalized to the true bubble radius).

We calculated the likelihood on the grid, and our framework has therefore found the optimal bubble when the fiducial and inferred locations and radii coincide, that is, $\text{Err}_{\text{loc}} = \text{Err}_{\text{rad}} = 0$. In this case, the error of the location and radius is below the grid on which we calculated the likelihood (1.5 cMpc in our fiducial case, corresponding to a $\lesssim 15\%$ fractional error).

The solid black curve in Fig. 8 shows the change in these fractional errors with galaxy number density for a single realization. The realizations of the EoR morphology and observed galaxies are held constant here, with the maximum likelihood computed each time a new observed galaxy is added. The more galaxies we observe, the smaller the error on our inferred HII bubble location and radius. The sizable stochasticity in galaxy properties and sightline-to-sightline scatter in the IGM opacity manifest as noise in this evolution, making it nonmonotonic. Nevertheless, even a single galaxy is able to shrink the fractional error by a factor of two from our prior range, ruling out extreme values.

We repeated this calculation with 100 different realizations of the mock observation (EoR morphology and observed galaxy samples). In Fig. 8 we show the resulting 68% and 95% C.L. on the fractional errors as increasingly more galaxies are added

⁴ Ideally, we would want to map out the full posterior:

$$\mathcal{P}(O_b, R_b | \mathbf{x}, f_\alpha(\lambda), M_{\text{UV}}, z) \propto \mathcal{L}(f_\alpha(\lambda) | O_b, R_b, \mathbf{x}, M_{\text{UV}}, z) \pi(O_b, R_b),$$

where $\pi(O_b, R_b)$ is a prior for the center location and radius of the HII bubble. However, including the correlations of the (non-Gaussian) likelihoods at the location of every galaxy is analytically not tractable, and would require high dimensional simulation based inference (e.g., de Santi et al. 2025; Lemos et al. 2023). We save this for future work.

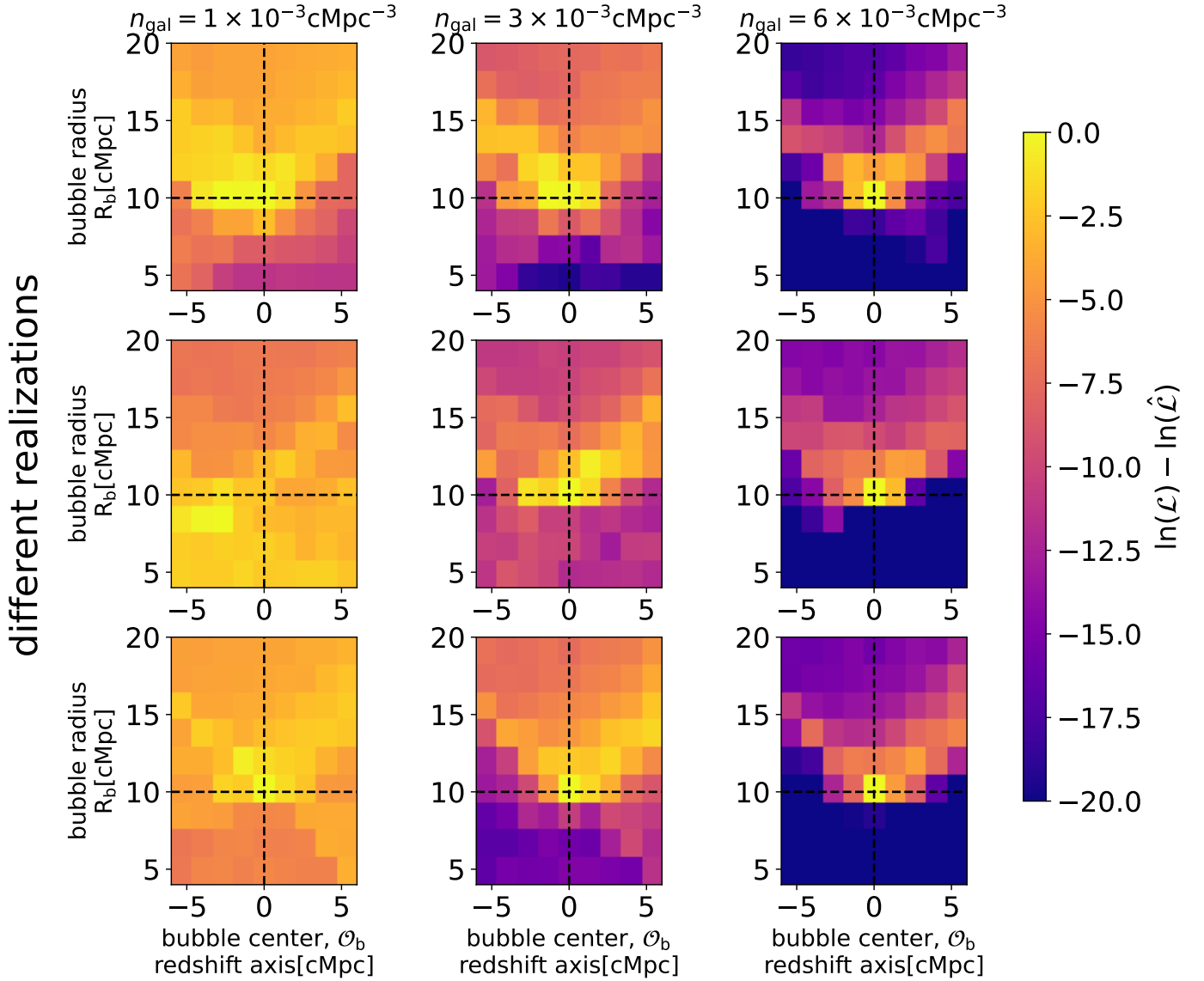


Fig. 7. Two-dimensional slices through the log likelihood, normalized to the maximum value in each panel. The vertical and horizontal axis in each panel correspond to the bubble radius, R_b , and redshift axis of the center, $O_b = (x_b = x_b^{\text{true}}, y_b = y_b^{\text{true}}, z_b)$. The true values, $(R_b^{\text{true}}, z_b^{\text{true}}) = (10 \text{ cMpc}, 0 \text{ cMpc})$, are demarcated with the dashed lines. The mock observation in this example was made at $z = 7.5$, assuming $\bar{x}_H = 0.65$. The different columns represent different numbers of galaxies that were used for inference (8 galaxies on the left, 24 galaxies in the middle, and 48 galaxies on the right). The different rows correspond to different realizations of the EoR morphology and galaxy samples.

to the field. In 68% (95%) of the cases, number densities of $7.7 \times (10.5 \times) 10^{-3} \text{ cMpc}^{-3}$ are sufficient to obtain an error of $\lesssim 15\%$ on the center position. The corresponding requirements are $4.2 \times (7.2 \times) 10^{-3} \text{ cMpc}^{-3}$ for an error of $\lesssim 15\%$ on the bubble radius (comparable to Lu et al. 2025). In other words, Ly α spectra from ~ 0.01 galaxies per cMpc^3 should be observed to be $\gtrsim 95\%$ confident that the HII bubble location and size recovered by our method is accurate at $\lesssim 1 \text{ cMpc}$. This approximately corresponds to 80 galaxies in 2×2 tiled pointings with JWST/NIRSpec.

4.1. Including a prior on the emergent Lyman- α

We assumed that the emergent Lyman- α emission follows post-EoR empirical relations, as described in Section 2.1. Beyond this, we assumed no prior knowledge on the intrinsic emission of each galaxy. Nebular lines with a lower opacity such as the Balmer lines ($H\alpha$, $H\beta$), however, can provide complimentary

estimates of the intrinsic production of Lyman- α photons (e.g., Hayes et al. 2023; Saxena et al. 2023; Chen et al. 2024). In this section, we repeat the analysis from above, but with a simple prior on the emergent Ly α emission.

Specifically, we applied a rejection sampling to keep only those forward models for which the transmission was within 20% of the true transmission (or placing an upper bound if $\mathcal{T}_{\text{EoR}} \leq 0.2$)⁵. Then, we calculated the likelihood in the same

⁵ A detection of Balmer lines places constraints on the production rate of Lyman-alpha photons, which then have to pass through the ISM, CGM, and IGM. Our distribution of Lyman- α equivalent widths and profile shapes described in Section 2.2 effectively already includes radiative transfer and escape through the ISM+CGM, which we assumed does not evolve significantly with redshift at $z \gtrsim 6$. Thus, our prior is effectively a prior on the IGM transmission. We leave a self-consistent treatment of the redshift evolution of the ISM+CGM escape fraction and IGM transmission for future work.

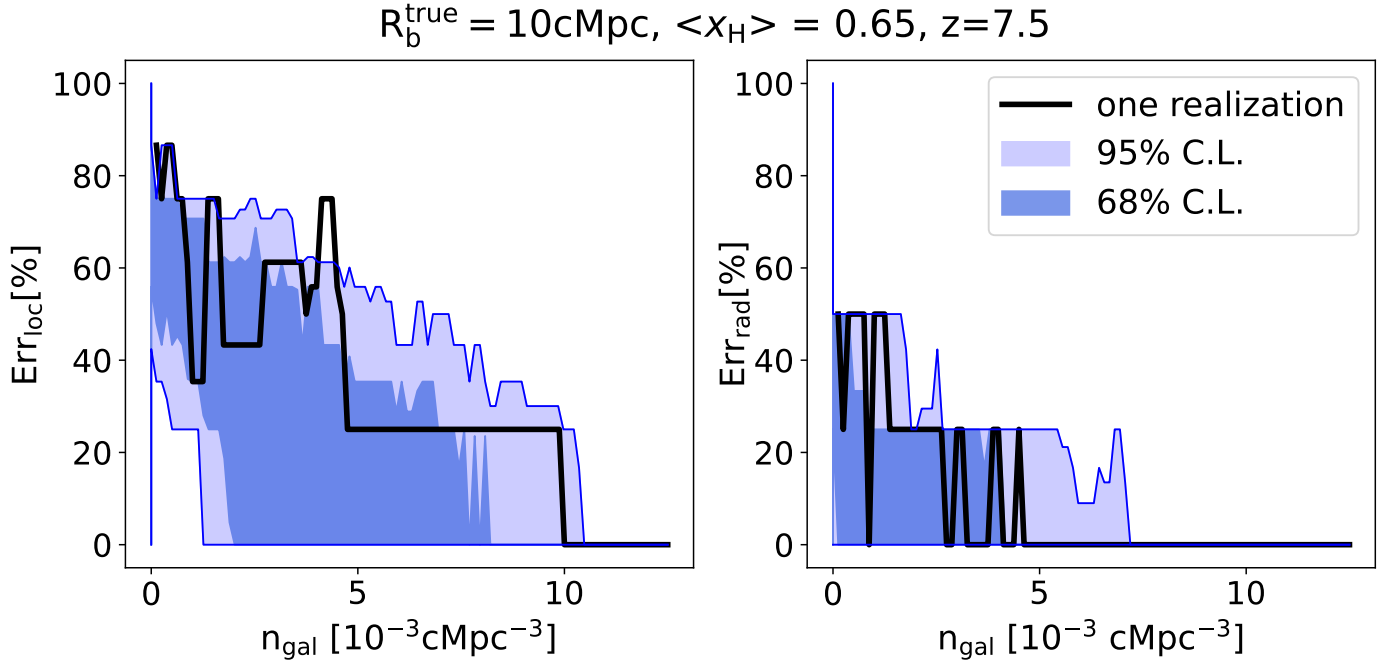


Fig. 8. Fractional errors of the maximum likelihood HII bubble center (left) and radius (right) as a function of the number density of the observed galaxies. The black curve corresponds to a single realization of the EoR morphology and galaxy samples. The light (dark) blue region corresponds to the 95th (68th) percentile of fractional errors obtained from 100 realizations. With $n_{\text{gal}} \gtrsim 0.01$ galaxies cMpc^{-3} , the fractional errors drop below the grid resolution we used to calculate the likelihoods (1.5 cMpc).

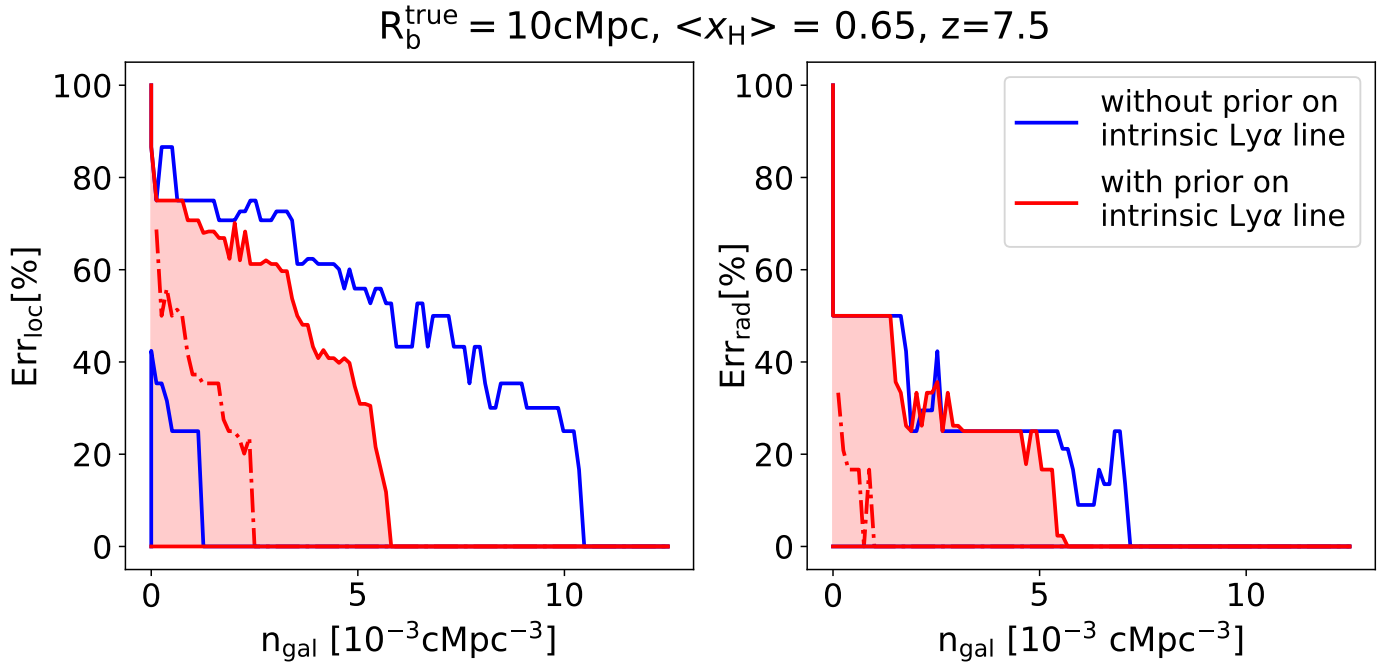


Fig. 9. Same as Fig. 8, but with the 95% C.L. on the fractional errors (red) assuming a prior on the emergent Lyman-alpha emission, motivated by observed Balmer lines (see text for details).

way as we did in Section 3. The width of the prior was motivated by current observations using JWST (cf. [Saxena et al. 2023](#) for $\text{H}\beta$ and [Lin et al. 2024](#) for $\text{H}\alpha$ from ground-based instruments). $\text{H}\alpha$ observations above $z > 7$ are not possible with NIRSpec. For higher redshifts, we therefore relied on a fainter $\text{H}\beta$. Despite this, $\text{H}\beta$ is regularly observed in high- z galaxies, and it is there-

fore reasonable to use the prior for all galaxies (e.g., [Meyer et al. 2024](#)).

In Fig. 9 we show the 95% C.L. on the fractional errors with (red) and without (blue; same as Fig. 8) the prior information on the emergent Lyman- α emission. Additional nebular lines to constrain the emergent Lyman- α emission can reduce the

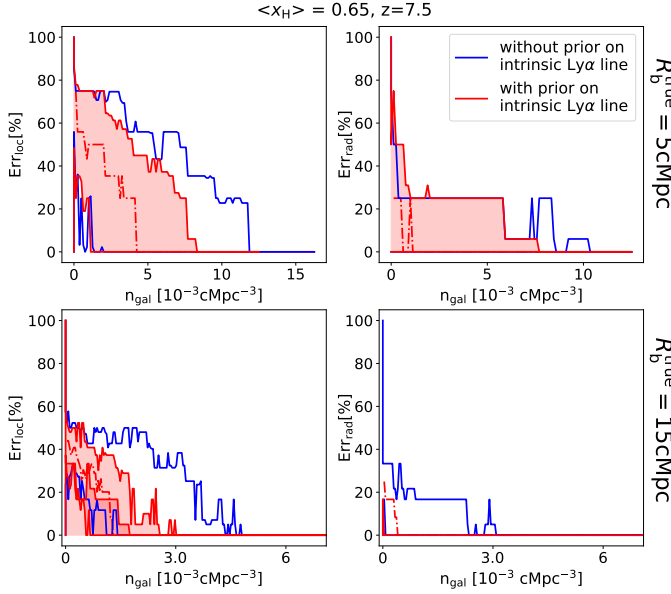


Fig. 10. Same as Fig. 9, but assuming $R_b^{\text{true}} = 5$ (15) cMpc in the top (bottom) row. Larger bubbles can be inferred with the same accuracy using a smaller number density of galaxies.

number of galaxies that are required to obtain the same constraints by a factor of about two.

4.2. Results for different bubble sizes

Our fiducial choice for the radius of the central HII bubble was $R_b = 10$ Mpc, which was motivated by HII bubble sizes in the early stages of the EoR. We test the performance of our pipeline for other radii of $R_b = 5$ and 15 Mpc below.

In the top (bottom) panels of Fig. 10, we show the analysis with and without the 20% prior on the emergent emission, assuming $R_b^{\text{true}} = 5$ Mpc ($R_b^{\text{true}} = 15$ Mpc). By comparing to the fiducial results in Fig. 9, we see that fewer galaxies are required to infer the size and location for larger HII bubbles. For $R_b^{\text{true}} = 15$ cMpc, about two to three times fewer galaxies are required to constrain the HII bubble with the same accuracy compared to $R_b^{\text{true}} = 5$ cMpc. The reason most likely is that the larger bubbles allow for a broader range of IGM opacities. Galaxies inside small bubbles have roughly the same attenuation regardless of their relative location inside the bubble. In contrast, there is a noticeable difference between the (average) attenuation on the near and far side of the bubble for larger bubbles, which allows us to constrain their geometry more easily (cf. the left panel of Fig. 5).

5. Building confidence in our framework

In this section, we explore the dependence of our results on our model assumptions. To do this, we applied our framework on a different reionization morphology and on different emergent EW distributions. Even though our framework is a proof of concept, it would help us to build confidence that it is not sensitive to uncertainties in the details of our model if it performed well in these out-of-distribution tests.

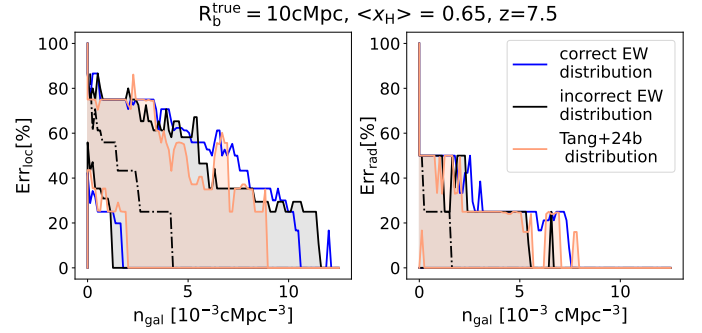


Fig. 11. Same as Fig. 8, but including in gray (orange) the 95% C.L. on the fractional errors assuming that the EW distribution we used to interpret the mock observation is a Gaussian constructed from the same data (log-normal distribution from Tang et al. 2024a). In other words, the mock observations and forward models are generated using different EW distributions (see text for details). The fact that the C.L. demarcated in gray, orange and blue roughly overlap indicates that our results are not sensitive to our choice of the emergent EW distribution.

5.1. Different EW distribution

First, we tested the performance of our pipeline assuming the emergent Lyman- α luminosities followed a different EW distribution than the we used in our forward models (Sect. 2.2). Specifically, we generated mock observations assuming the Gaussian distribution from Treu et al. (2012) (based on Stark et al. 2011, but with the same parameters as in Section 2.2),

$$p_{6,G}(W|M_{UV}) = \frac{A(M_{UV})}{W_c(M_{UV})} \sqrt{\frac{2}{\pi}} e^{-\frac{1}{2} \left(\frac{W}{W_c(M_{UV})} \right)^2} H(W) + [1 - A(M_{UV})] \delta(W). \quad (14)$$

We then interpreted these mock observations with our fiducial pipeline, which uses the exponential EW distribution from Eq. (7).

The resulting 95% C.L. on the fractional errors are highlighted in gray in Fig. 11. In blue, we show the same 95% C.L. from Fig. 8, which use the same EW distribution for the forward models and the mock observation. The fact that the gray and blue regions demarcate roughly the same fractional errors suggests that our analysis is not sensitive to the choice of EW distribution.

As a further test, we repeated the analysis, but with the log-normal distribution from Tang et al. (2024a) that was constructed using completely separate data. The distribution is given with

$$p(W|M_{UV}) = \frac{1}{\sqrt{2\pi} \cdot \sigma_W \cdot W} \exp \left(-\frac{(\log(W) - \mu_W)^2}{2 \cdot \sigma_W^2} \right), \quad (15)$$

with $\mu_W = \log(5.0)$ and $\sigma_W = 1.74$. The results are shown in orange in Fig. 11. The blue and orange contours again follow each other closely, indicating that our pipeline is robust to the changes in the EW distribution.

5.2. Demonstration on a 3D reionization simulation

Because this work is intended as a proof of concept, we made several simplifying assumptions throughout. For example, our reionization morphology was generated by overlapping ionized spheres, and we treated the galaxy – HII bubble bias in a simplified way. A more realistic bias for the galaxies and HII regions

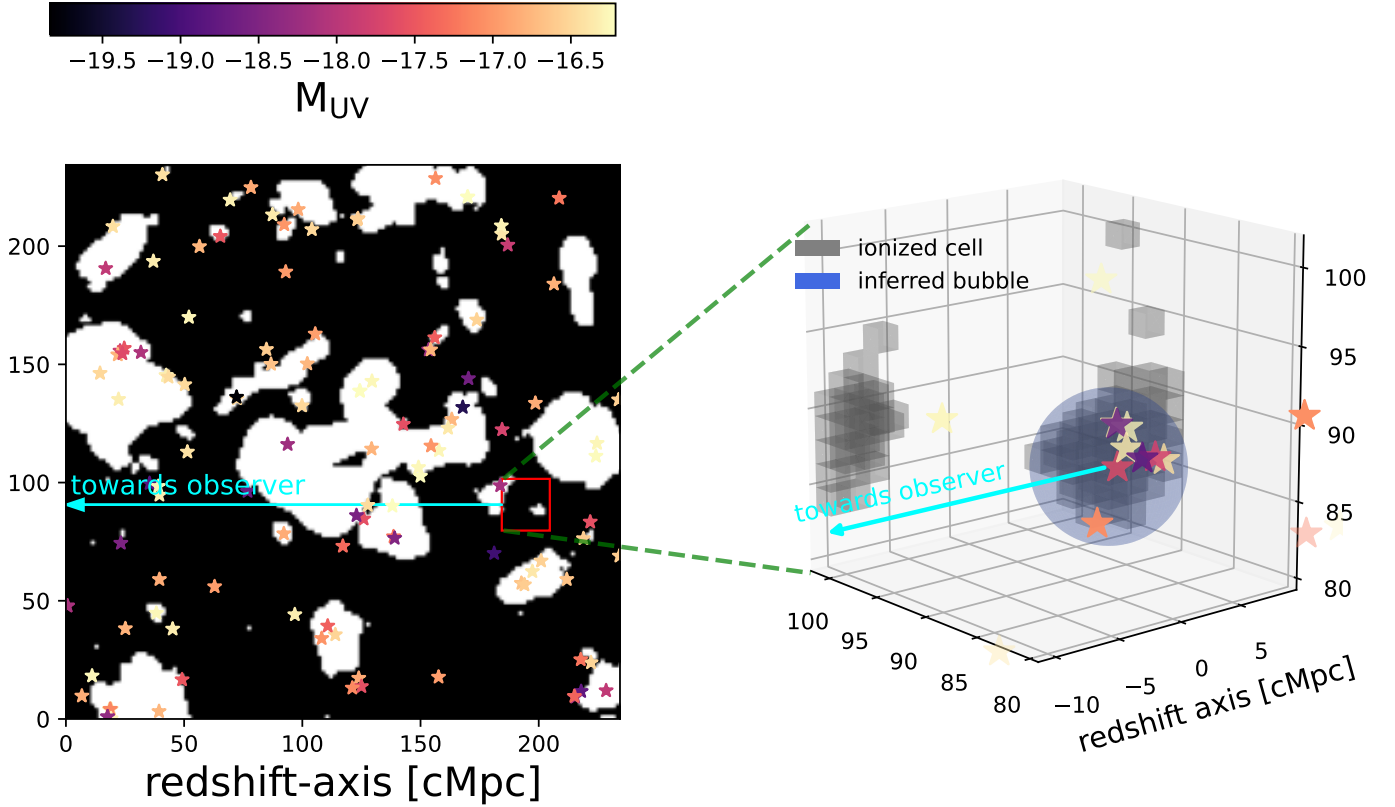


Fig. 12. Application of our bubble-finding procedure to a 3D simulation. Left: 2D slices through the ionization field. Ionized/neutral regions are shown in white/black. The brightest galaxies in this 1.5 cMpc slice are denoted with stars, colored according to their UV magnitudes. The line-of-sight direction used to compute the IGM opacity for each observed galaxy is indicated with the arrow. The $20 \times 20 \times 20$ cMpc subvolume used for the mock survey is denoted with the red square. Right: Zoom-in of the mock survey. A random subsample (to avoid overcrowding) of the observed galaxies is shown with stars. Ionized voxels are shown in gray. Our maximum likelihood solution for the central HII bubble is overlaid with the blue sphere. The reasonable overlap of the central gray blob with the blue sphere shows that our framework is reliable.

could be included straightforward analytically, but the question is whether this is necessary. We tested the performance of our simple model on self-consistent 3D simulations of the reionization.

We applied our framework on the galaxy catalogs and ionization maps from the simulations of [Lu et al. \(2025\)](#), which were generated with the public 21cmFAST code ([Mesinger & Furlanetto 2007](#); [Mesinger et al. 2011](#); [Murray et al. 2020](#)). These simulations capture the complex morphology of the reionization, which is self-consistently generated from the underlying galaxy fields. We processed the galaxies with the procedure outlined in Section 2 to create mocks on which we performed the inference.

We used two ionization boxes, one at $\bar{x}_{HI} = 0.69$ and $\bar{x}_{HI} = 0.90$. We selected ionized bubbles and associated galaxies that matched the volumes we used in our fiducial setup, specifically, $20 \times 20 \times 20$ cMpc. When the inference framework is applied to real data, the forward models must be fine-tuned to match the specific details of the survey.

We illustrate the results in Figures 12 and 13 for some example surveys at $\bar{x}_{HI} = 0.69$ and $\bar{x}_{HI} = 0.9$, respectively. In the left panels, we show 1.5 cMpc thick slices through the simulation boxes. Ionized/neutral regions are shown in white/black. The brightest galaxies are shown with stars, whose colors correspond to their UV magnitudes. The line-of-sight direction used to compute the IGM opacity for each observed galaxy is indicated with the arrow.

The volumes of the mock surveys are illustrated with the zoom-ins on the right. Here, gray cells show the ionized voxels of the simulation, and the blue sphere is the maximum likelihood solution for the central HII region. We assumed all galaxies inside this $20 \times 20 \times 20$ cMpc volume brighter than $M_{UV} < -17.0$ are observed with NIRSpect. This corresponds to $\sim 50(30)$ galaxies at $\bar{x}_{HI} = 0.69$ ($\bar{x}_{HI} = 0.9$).

The zoom-ins in these figures show that the inferred HII regions in blue provide a reasonable characterization of the true HII region in gray. This gives us confidence in our simplified treatment of the EoR morphology and the associated spatial correlation with galaxies. As mentioned above, we will tailor future applications to actual NIRSpect surveys of galaxy groups.

6. JWST observational requirements

We currently have several spectroscopically-targeted fields containing groups of galaxies with detection in Lyman- α , such as COSMOS ([Endsley & Stark 2022](#); [Witten et al. 2024](#)), EGS ([Oesch et al. 2015](#); [Zitrin et al. 2015](#); [Roberts-Borsani et al. 2016](#); [Tilvi et al. 2020](#); [Leonova et al. 2022](#); [Larson et al. 2022](#); [Jung et al. 2022](#); [Tang et al. 2023, 2024a](#); [Chen et al. 2024](#); [Napolitano et al. 2024](#)), BDF ([Castellano et al. 2016, 2018](#)), GOODS-N ([Oesch et al. 2016](#); [Eisenstein et al. 2023](#); [Bunker et al. 2023](#); [Tacchella et al. 2023](#)), and GOODS-S ([Witstok et al. 2024](#); [Tang et al. 2024a](#)). Although the number densities are lower by a factor of a few ($\sim 1.0 \times 10^{-3}$ cMpc $^{-3}$) than

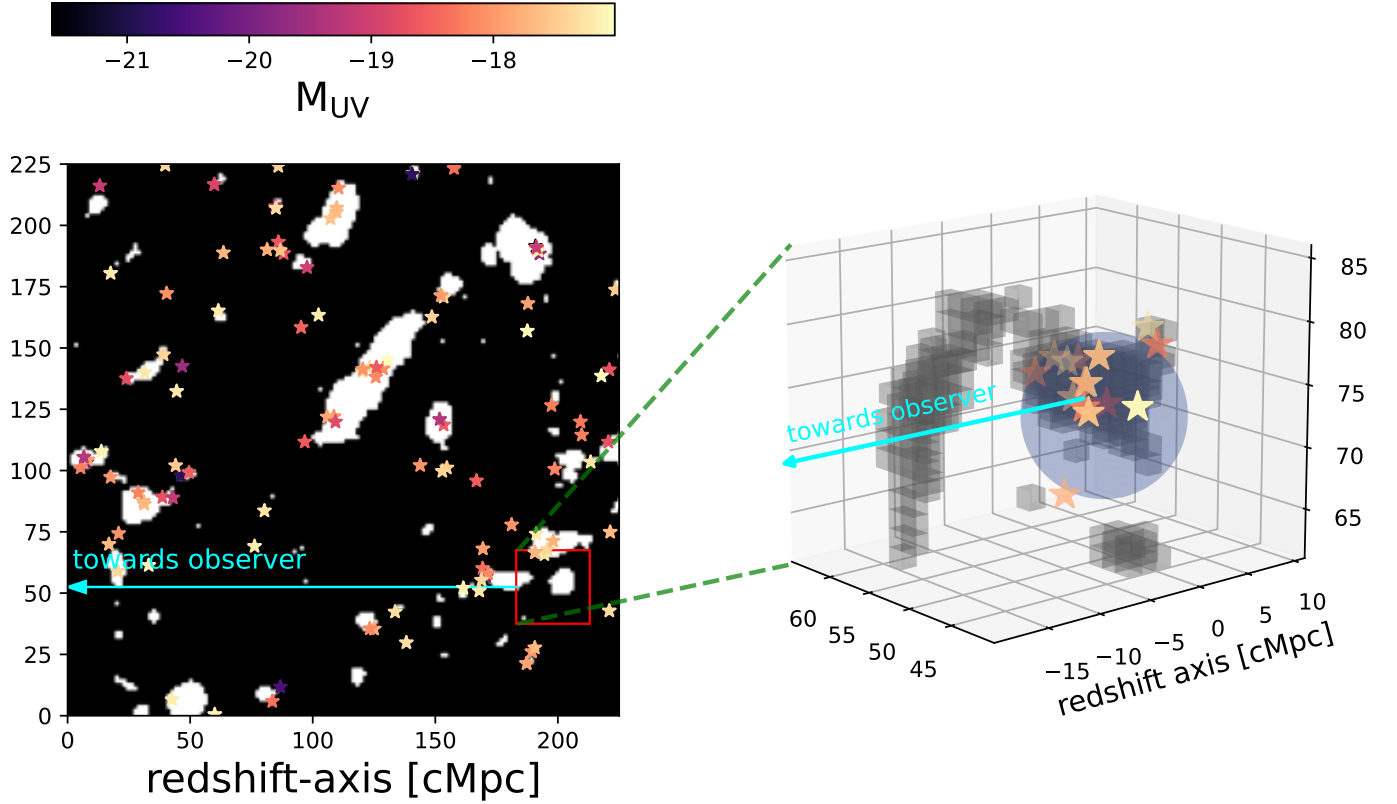


Fig. 13. Same as Fig. 12, but for $\bar{x}_{\text{HI}} = 0.9$.

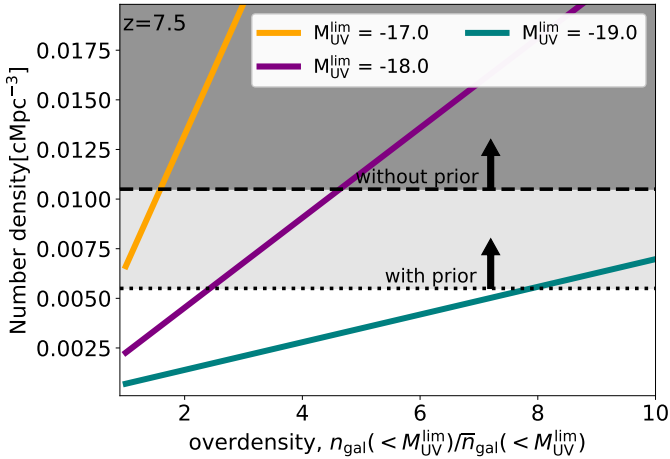


Fig. 14. Cumulative number density of galaxies down to various magnitude limits as a function of galaxy overdensity. The dotted (dashed) lines demarcate the minimum number densities we found for an accurate HII bubble recovery, with (without) a prior on the emergent Lyman- α emission.

the required values in Sect. 4, the data sets are expanding rapidly. Ongoing and proposed programs are extending fields and reach greater depths, which results in larger areas and higher number densities. We quantify the JWST survey requirements in more detail in order to be able to robustly constrain HII bubbles with our procedure.

In Fig. 14 we show the number density of galaxies as a function of limiting UV magnitude, $M_{\text{UV}}^{\text{lim}}$, and galaxy overdensity, $n(< M_{\text{UV}}^{\text{lim}}/\bar{n}(< M_{\text{UV}}^{\text{lim}}))$. We plot three curves corresponding to

$M_{\text{UV}}^{\text{lim}} = -17, -18$, and -19 . We demarcate in gray the required number density for accurate HII bubble recovery we found in the previous section; the lower/upper range correspond to including/not including a prior on the emergent Ly α emission from Balmer lines (cf. Section 4).

At the mean number density, we would require the photometric survey to detect galaxies down to $M_{\text{UV}}^{\text{lim}} \lesssim -17.0$ to -17.5 . The quoted range spans what can be achieved with or without a prior on the Lyman- α production rate from Balmer lines. In an unlensed field, approximately 600 hours of integration with NIRCcam on the JWST (estimated based on exposure times in Morishita et al. 2025, for 4 pointings) would be required to obtain these number densities, which is achievable, but ambitious.

Several current fields are known to contain overdensities, however. For example, COSMOS contains a 140 pMpc^3 volume that is estimated to be three times overdense at $z \sim 6.8$ (Endsley & Stark 2022), while EGS is estimated to contain a 12 pMpc^3 volume that is also three times overdense at $z \sim 8.7$ (Zitrin et al. 2015; Leonova et al. 2022; Larson et al. 2022; Tang et al. 2023; Chen et al. 2024; Lu et al. 2024). Moreover, GOODS-S and GOODS-N contain four and eight times overdensities over volumes that exceed the fiducial volume we used here ($62 \text{ arcmin}^2 \times \Delta z \sim .2$ Tang et al. 2024a). Furthermore, one of the most distant observed LAE is also thought to be located in an overdensity (up to $24\times$ overdense at $z = 10.6$ in 2.6 pMpc^3 volume; Oesch et al. 2016; Bunker et al. 2023; Tacchella et al. 2023, see also Lu et al. 2024 for other examples). Observing a field with an overdensity of three (eight) would require photometric completeness down to $M_{\text{UV}}^{\text{lim}} \lesssim -18.5$ to -18 (-19 to -18.5). This would require only 50 (4.2) hours of integration with NIRCcam. We assumed that the emergent Ly- α properties remain the same depending on the

overdensity (i.e., there are no environment dependences in Sections 2.1 and 2.2).

These candidates then require spectroscopic follow-up. The best candidate for a spectroscopic confirmation is the [OIII]5007 Å emission line. The [OIII] emission line is routinely observed with the JWST (e.g., [Endsley et al. 2023, 2024; Meyer et al. 2024](#)). We estimated the required exposure time for detecting [OIII] line by assuming the M_{UV} dependent distribution of the equivalent widths from [Endsley et al. \(2024\)](#). For this distribution, the equivalent-width limit for detecting 90% of the galaxies at $M_{UV} = -18.0$ is 120 Å. Eighteen hours per pointing are required to obtain this equivalent width (at 5σ) with the G395M filter. The number density requirement (and thus exposure time) would increase if we were not to detect H β , as required to place a prior on the intrinsic Lyman- α emission. It would also increase if the [OIII] distribution were lower at lower magnitudes.

Following this, 23 hours with G140M per pointing are required to detect Lyman α at a noise level of $5 \times 10^{-19} \text{ erg s}^{-1} \text{ cm}^{-2}$ (corresponding to a detection of a line of $W \approx 25$ Å). These numbers are comparable to existing large JWST surveys ([Eisenstein et al. 2023](#)).

We also note that the volume of the potential JWST surveys used for this analysis has a direct impact on the sizes of the HII regions that can be inferred (see also [Lu et al. 2025](#)). Tiled surveys of the same depth but with larger volumes could hope to detect correspondingly larger HII regions. The fact that larger HII regions occur later in the EoR, with correspondingly higher galaxy number densities, might mitigate the required integration time. We postpone a more systematic survey design to subsequent work in which we apply our framework to simulated 3D EoR light cones.

Another interesting approach might be to couple the framework from [Lu et al. \(2025\)](#) with the framework outlined here. The large-scale empirical method from [Lu et al. \(2025\)](#) could isolate interesting subvolumes that could then be analyzed with the quantitative inference approach presented here. Finally, the pipeline we developed here could be used to target multiple fields and can be used to infer the sizes of multiple ionized bubbles, which would help us to determine the bubble size distribution. Constraining it would in turn allow us determine the sources that ionized the Universe.

7. Conclusions

The morphology of the reionization tells us which galaxies dominated the epoch of cosmic reionization. Individual bubbles surrounding groups of galaxies encode information on the contribution of unseen, faint sources and allow us to correlate the properties of bubbles and the galaxies they host.

We built a framework to use Lyman- α observations from JWST NIRSpec to constrain the ionization morphology around a group of galaxies. Our framework used for the first time complementary information from sightlines to neighboring galaxies, and it sampled all important sources of stochasticity to robustly place constraints on the size and position of local HII bubbles.

We found that Ly α spectra from ~ 0.01 galaxies per cMpc^3 are required for a confidence of $\geq 95\%$ that the HII bubble location and size recovered by our method is accurate to better than $\leq 1.5 \text{ cMpc}$. This approximately corresponds to 80 galaxies in 2×2 tiled pointings with JWST/NIRSpec. These requirements can be reduced by using additional nebular lines (e.g., H β) to constrain the intrinsic Lyman- α emission. A simple prior on

the emergent Lyman- α emission reduces the number of galaxies required to obtain the same constraints by a factor of about two. These number densities can be achieved with a targeted survey that is complete down to $M_{UV} = -17$ to -19 , depending on the overdensity of the field.

We demonstrated that our framework is not sensitive to the assumed distribution for the emergent Lyman- α emission. We also accurately recovered ionized bubbles when we applied our framework to 3D EoR simulations.

Our pipeline can be applied to existing observations of Lyman- α spectra from galaxy groups. Additionally, the observational requirements for a statistical detection of local HII regions presented here can be used to design complimentary new JWST surveys. The application of our framework to multiple independent fields would allow us to constrain the distribution of the bubble sizes, which helps us to understand which galaxies reionized the Universe.

Data availability

The code related to the work is publicly available at [IvanNikolic21/Lyman- \$\alpha\$ bubbles](https://github.com/IvanNikolic21/Lyman-alpha_bubbles).

Acknowledgements. We gratefully acknowledge computational resources of the Center for High Performance Computing (CHPC) at SNS. AM acknowledges support from the Italian Ministry of Universities and Research (MUR) through the PRIN project “Optimal inference from radio images of the epoch of reionization”, the PNRR project “Centro Nazionale di Ricerca in High Performance Computing, Big Data e Quantum Computing”. CAM and TYL acknowledge support by the VILLUM FONDEN under grant 37459. CAM acknowledges support from the Carlsberg Foundation under grant CF22-1322.

References

- Almada Monter, S., & Gronke, M. 2024, *MNRAS*, **534**, L7
- Anau Montel, N., Alvey, J., & Weniger, C. 2024, *MNRAS*, **530**, A107
- Bhowmick, A. K., Campbell, D., Di Matteo, T., & Feng, Y. 2018, *MNRAS*, **480**, 3177
- Bolan, P., Lemaux, B. C., Mason, C., et al. 2022, *MNRAS*, **517**, 3263
- Bolan, P., Bradăc, M., Lemaux, B. C., et al. 2024, *MNRAS*, **531**, 2998
- Bosman, S. E. I., Davies, F. B., Becker, G. D., et al. 2022, *MNRAS*, **514**, 55
- Bruton, S., Scarlata, C., Haardt, F., et al. 2023, *ApJ*, **953**, 29
- Bunker, A. J., Saxena, A., Cameron, A. J., et al. 2023, *A&A*, **677**, A88
- Byrohl, C., & Gronke, M. 2020, *A&A*, **642**, L16
- Castellano, M., Dayal, P., Pentericci, L., et al. 2016, *ApJ*, **818**, L3
- Castellano, M., Pentericci, L., Vanzella, E., et al. 2018, *ApJ*, **863**, L3
- Chen, Z., Stark, D. P., Mason, C., et al. 2024, *MNRAS*, **528**, 7052
- Cranmer, K., Brehmer, J., & Louppe, G. 2020, *Proc. Natl. Acad. Sci.*, **117**, 30055
- De Barros, S., Pentericci, L., Vanzella, E., et al. 2017, *A&A*, **608**, A123
- de Santi, N. S. M., Villaescusa-Navarro, F., Raul Abramo, L., et al. 2025, *JCAP*, **2025**, 082
- Dijkstra, M. 2014, *PASA*, **31**, e040
- Dijkstra, M., & Wyithe, J. S. B. 2012, *MNRAS*, **419**, 3181
- Dijkstra, M., Mesinger, A., & Wyithe, J. S. B. 2011, *MNRAS*, **414**, 2139
- Doussot, A., & Semelin, B. 2022, *A&A*, **667**, A118
- Eisenstein, D. J., Willott, C., Alberts, S., et al. 2023, *ApJS*, submitted [arXiv:2306.02465]
- Endsley, R., & Stark, D. P. 2022, *MNRAS*, **511**, 6042
- Endsley, R., Stark, D. P., Whitler, L., et al. 2023, *MNRAS*, **524**, 2312
- Endsley, R., Stark, D. P., Whitler, L., et al. 2024, *MNRAS*, **533**, 1111
- Ghara, R., Mellema, G., Giri, S. K., et al. 2018, *MNRAS*, **476**, 1741
- Giri, S. K., Mellema, G., Dixon, K. L., & Iliev, I. T. 2018, *MNRAS*, **473**, 2949
- Hayes, M. J., & Scarlata, C. 2023, *ApJ*, **954**, L14
- Hayes, M. J., Runholm, A., Scarlata, C., Gronke, M., & Rivera-Thorsen, T. E. 2023, *MNRAS*, **520**, 5903
- Hu, W., Martin, C. L., Gronke, M., et al. 2023, *ApJ*, **956**, 39
- Hutter, A., Trebitsch, M., Dayal, P., et al. 2023, *MNRAS*, **524**, 6124
- Jakobsen, P., Ferruit, P., Alves de Oliveira, C., et al. 2022, *A&A*, **661**, A80
- Jespersen, C. K., Steinhardt, C. L., Somerville, R. S., & Lovell, C. C. 2025, *ApJ*, **982**, 23
- Jones, G. C., Bunker, A. J., Saxena, A., et al. 2024, *A&A*, **683**, A238

- Jung, I., Finkelstein, S. L., Dickinson, M., et al. 2020, [ApJ](#), **904**, 144
- Jung, I., Finkelstein, S. L., Larson, R. L., et al. 2022, [ApJ](#), submitted [arXiv:2212.09850]
- Keating, L. C., Puchwein, E., Bolton, J. S., Haehnelt, M. G., & Kulkarni, G. 2024, [MNRAS](#), **531**, L34
- Larson, R. L., Finkelstein, S. L., Hutchison, T. A., et al. 2022, [ApJ](#), **930**, 104
- Laursen, P., Sommer-Larsen, J., & Razoumov, A. O. 2011, [ApJ](#), **728**, 52
- Lemos, P., Parker, L. H., Hahn, C., et al. 2023, in [Machine Learning for Astrophysics](#), 18
- Leonova, E., Oesch, P. A., Qin, Y., et al. 2022, [MNRAS](#), **515**, 5790
- Lin, Y., Oh, S. P., Furlanetto, S. R., & Sutter, P. M. 2016, [MNRAS](#), **461**, 3361
- Lin, X., Cai, Z., Wu, Y., et al. 2024, [ApJS](#), **272**, 33
- Lu, T.-Y., Mason, C. A., Hutter, A., et al. 2024, [MNRAS](#), **528**, 4872
- Lu, T.-Y., Mason, C. A., Mesinger, A., et al. 2025, [A&A](#), **697**, A69
- Mason, C. A., Treu, T., Dijkstra, M., et al. 2018, [ApJ](#), **856**, 2
- McQuinn, M., Lidz, A., Zahn, O., et al. 2007, [MNRAS](#), **377**, 1043
- Mesinger, A., & Furlanetto, S. 2007, [ApJ](#), **669**, 663
- Mesinger, A., & Furlanetto, S. R. 2008a, [MNRAS](#), **386**, 1990
- Mesinger, A., & Furlanetto, S. R. 2008b, [MNRAS](#), **385**, 1348
- Mesinger, A., Furlanetto, S., & Cen, R. 2011, [MNRAS](#), **411**, 955
- Mesinger, A., Aykutalp, A., Vanzella, E., et al. 2015, [MNRAS](#), **446**, 566
- Mesinger, A., Greig, B., & Sobacchi, E. 2016, [MNRAS](#), **459**, 2342
- Meyer, R. A., Laporte, N., Ellis, R. S., Verhamme, A., & Garel, T. 2021, [MNRAS](#), **500**, 558
- Meyer, R. A., Oesch, P. A., Giovanazzo, E., et al. 2024, [MNRAS](#), **535**, 1067
- Miralda-Escudé, J. 1998, [ApJ](#), **501**, 15
- Morishita, T., Mason, C. A., Kreilgaard, K. C., et al. 2025, [ApJ](#), **983**, 152
- Murray, S., Greig, B., Mesinger, A., et al. 2020, [J. Open Source Softw.](#), **5**, 2582
- Nakane, M., Ouchi, M., Nakajima, K., et al. 2024, [ApJ](#), **967**, 28
- Napolitano, L., Pentericci, L., Santini, P., et al. 2024, [A&A](#), **688**, A106
- Neufeld, D. A. 1990, [ApJ](#), **350**, 216
- Neyer, M., Smith, A., Kannan, R., et al. 2024, [MNRAS](#), **531**, 2943
- Oesch, P. A., van Dokkum, P. G., Illingworth, G. D., et al. 2015, [ApJ](#), **804**, L30
- Oesch, P. A., Brammer, G., van Dokkum, P. G., et al. 2016, [ApJ](#), **819**, 129
- Orlitová, I., Verhamme, A., Henry, A., et al. 2018, [A&A](#), **616**, A60
- Park, J., Mesinger, A., Greig, B., & Gillet, N. 2019, [MNRAS](#), **484**, 933
- Planck Collaboration VI. 2020, [A&A](#), **641**, A6
- Qin, Y., Mesinger, A., Bosman, S. E. I., & Viel, M. 2021, [MNRAS](#), **506**, 2390
- Qin, Y., Mesinger, A., Prelogović, D., et al. 2025, [PASA](#), **42**, e049
- Roberts-Borsani, G. W., Bouwens, R. J., Oesch, P. A., et al. 2016, [ApJ](#), **823**, 143
- Saxena, A., Robertson, B. E., Bunker, A. J., et al. 2023, [A&A](#), **678**, A68
- Sobacchi, E., & Mesinger, A. 2014, [MNRAS](#), **440**, 1662
- Stark, D. P., Ellis, R. S., Chiu, K., Ouchi, M., & Bunker, A. 2010, [MNRAS](#), **408**, 1628
- Stark, D. P., Ellis, R. S., & Ouchi, M. 2011, [ApJ](#), **728**, L2
- Stark, D. P., Ellis, R. S., Charlot, S., et al. 2017, [MNRAS](#), **464**, 469
- Steidel, C. C., Rudie, G. C., Strom, A. L., et al. 2014, [ApJ](#), **795**, 165
- Tacchella, S., Eisenstein, D. J., Hainline, K., et al. 2023, [ApJ](#), **952**, 74
- Tang, M., Stark, D. P., Chen, Z., et al. 2023, [MNRAS](#), **526**, 1657
- Tang, M., Stark, D. P., Topping, M. W., Mason, C., & Ellis, R. S. 2024a, [ApJ](#), **975**, 208
- Tang, M., Stark, D. P., Ellis, R. S., et al. 2024b, [MNRAS](#), **531**, 2701
- Tilvi, V., Malhotra, S., Rhoads, J. E., et al. 2020, [ApJ](#), **891**, L10
- Treu, T., Trenti, M., Stiavelli, M., Auger, M. W., & Bradley, L. D. 2012, [ApJ](#), **747**, 27
- Umeda, H., Ouchi, M., Nakajima, K., et al. 2024, [ApJ](#), **971**, 124
- Verhamme, A., Garel, T., Ventou, E., et al. 2018, [MNRAS](#), **478**, L60
- Witstok, J., Smit, R., Saxena, A., et al. 2024, [A&A](#), **682**, A40
- Witstok, J., Maiolino, R., Smit, R., et al. 2025, [MNRAS](#), **536**, 27
- Witten, C., Laporte, N., Martin-Alvarez, S., et al. 2024, [Nat. Astron.](#), **8**, 384
- Yamada, T., Matsuda, Y., Kousai, K., et al. 2012, [ApJ](#), **751**, 29
- Zahn, O., Mesinger, A., McQuinn, M., et al. 2011, [MNRAS](#), **414**, 727
- Zitrin, A., Labbé, I., Belli, S., et al. 2015, [ApJ](#), **810**, L12

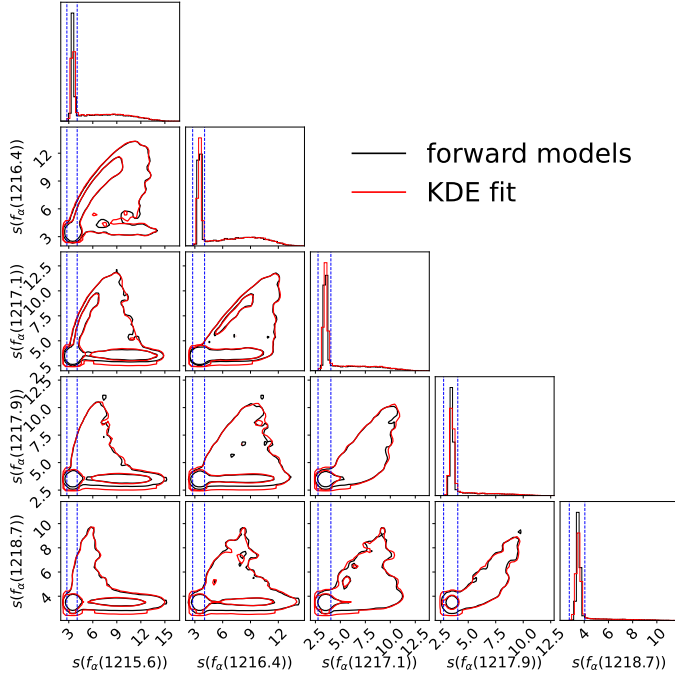


Fig. A.1. Corner plot of scaled flux spectra and the KDE fit to the forward models. Black contours represent 68% and 95% C.L. of the distribution of forward models for a galaxy located in the center of a $R_b = 10\text{Mpc}$ ionized bubble. Red contours represent same distributions for the KDE fit to the forward models. Blue dashed lines mark regions with $\pm 2\sigma$ of the noise of the spectrum. Contours outside of these dashed lines represent Lyman- α emission, with non-negligible correlation between bins.

Appendix A: Mapping the joint distribution over all flux bins with kernel density estimation

In Section 3 we showed how we generate forward models of JWST observations. Here we detail how we compute flux PDFs, i.e. likelihoods, in the space of local HII region parameters. In Fig. A.1 in black we show a corner plot of forward models of a galaxy located in the center of its local HII region with size $R_b = 10\text{Mpc}$. Black contours display 68% and 95% C.L. of the distribution respectively. We only show bins in our fiducial resolution ($R = 1000$) where the Lyman- α emission is located. In order to fit a density estimator to obtain a smooth PDF, we scale fluxes by a common function:

$$s(f_\alpha(\lambda)) = 93.5 + 5 \log(10^{-18} + 2f_\alpha(\lambda)). \quad (\text{A.1})$$

Each column and row represents one wavelength bin of the flux scaled by the function in Eq. A.1. We fit a Kernel Density Estimator to the forward models displayed in the corner plot with an exponential kernel of bandwidth $h = 0.12$. The scaling function and bandwidth were selected by optimizing the results of Section 4 (i.e., kernel density bandwidth and normalizing function that allows the inference of bubble properties for the least number of galaxies). We show contours of the fitted KDE in red in Fig. A.1.

Several features can be noted in Fig. A.1. Firstly, there is a strong peak for lower values of $s(f_\alpha(\lambda))$. This peak corresponds to the noise which is Gaussian by construction. This is further demonstrated with the blue dashed lines that show 5σ noise levels from Section. 2.4. On the other hand, there are additional features for larger values of $s(f_\alpha(\lambda))$. These features represent the Lyman- α signal that is higher than the noise. It is important

to note that there is non-negligible correlation between different bins and the distribution is highly non-trivial. This shows that a simple likelihood that uses independent Gaussians for each bin would potentially bias the inference and an approach that fits the whole distribution is necessary.

KDE fits the data for all of the bins, in the noise and signal regimes. There is a small overestimation of the width of the distribution for bins that are noise-dominated. Since these bins do not give a lot of information for Lyman- α inference, our KDE is not biasing results.