Contents

1 The Lyα transmission probability distribution function from the E-XQR-30 sample **1**
1.1 Introduction . . . 2
1.2 The Data Set . . . 4
1.3 Methods . . . 8
1.3.1 Quasar continua prediction and transmission profiles . 8
1.3.2 Mock spectra generation . . . 11
1.3.3 PDF intervals determination . . . 13
1.4 Results . . . 15
1.4.1 Transmission PDF and residuals . . . 15
1.4.2 Covariance and correlation matrices . . . 18
1.4.3 Chi-square distribution . . . 20
1.4.4 Discussion on the results . . . 29
1.5 Summary . . . 30
1.6 Appendix: Remaining quasar spectra and continua . . . 31
1.7 Appendix: Correlation matrices example . . . 34
1.8 Appendix: Non-Gaussian mock PDF distribution . . . 34

1 The Lyα transmission probability distribution function from the E-XQR-30 sample

Work in preparation in **S. Onorato**, T. Kist, J. F. Hennawi, F. B. Davies, Overleaf draft, almost ready to be published on *Monthly Notices of the Royal Astronomical Society*. Reprinted here in its entirety.

Abstract

We present a statistical analysis of the Lyα transmission probability distribution function (PDF) from the spectra of 21 E-XQR-30 quasars at $5.76 < z < 6.59$ grouped in redshift bins, to study the ionization and thermal state of the intergalactic medium (IGM) during the epoch of reionization. Using a forward-modeling approach, we generate thousands of mock spectra for each quasar, incorporating uncertainties in the intrinsic continuum and instrumental noise, and compute transmission PDFs in two rest-frame wavelength intervals: the Lyα forest ($1170-1190$ Å) and the proximity zone ($1190-1218$ Å). For each redshift bin, we compare the observed and mock PDFs using χ^2 statistics and the full covariance of the mock ensembles to assess goodness of fit. We find that transmission PDFs in the proximity zones are generally well reproduced by the model, supporting consistency between data and simulations in quasar-dominated regions. In contrast, we observe significant discrepancies in the Lyα forest for several redshift bins, suggesting limitations in the assumed IGM model. These mismatches may arise from idealized assumptions such as a cold, fully reionized, and spatially homogeneous IGM, which may not hold at $z \sim 6$. Our findings highlight the Lyα transmission PDF as a powerful tool for probing the IGM and motivate future models that incorporate patchy reionization, higher temperatures, and ultraviolet background (UVB) fluctuations during this critical phase of cosmic history.

1.1 Introduction

One of the central aims of observational cosmology is to understand and characterize the Epoch of Reionization (EoR), the era during which the first luminous sources started reionizing the intergalactic medium (IGM), making it transition from a neutral to a highly ionized state. Although recent years have witnessed significant progress in tracing the timing and progression of reionization, suggesting that it probably concludes by redshift $z \sim 5.3$ (e.g., McGreer et al. 2015; Bosman et al. 2022), many questions remain regarding its topology, the physical processes driving it, and its connection to the formation of galaxies and supermassive black holes (SMBHs). Quasars at $z \gtrsim 6$ offer a unique observational window into this era, providing some of the most powerful constraints on the state of the IGM during reionization, particularly through the analysis of the resonant Lyman-α (Lyα) absorption line and the Lyα forest (Fan et al. 2006b; Becker et al. 2015).

The IGM is composed primarily of hydrogen and helium, and its evolution reflects the integrated history of structure formation, feedback processes, and cosmic ionizing radiation. At redshifts $z \lesssim 5.3$, the IGM is almost completely ionized, and absorption in the Lyα forest traces large-scale fluctuations in density and temperature under the assumption of photoionization equilibrium. At higher redshifts ($5.3 \lesssim z \lesssim 7.5$), however, the rapid increase in Lyα opacity, large sightline-to-sightline scatter, and the emergence of extended Gunn–Peterson troughs (Gunn & Peterson 1965) all suggest a patchy, inhomogeneous, and still partially neutral IGM, consistent with the final stages of reionization (e.g., Fan et al. 2006b; Becker et al. 2015; Eilers et al. 2018; Yang et al. 2020; Bosman et al. 2022).

Quasar spectra provide multiple avenues for probing the IGM, among which we highlight two key diagnostics. First, damping wings, i.e., broad, redward-extended, off-resonant Lyα absorption features, are produced by the presence of neutral hydrogen at $z \gtrsim 6.5$, and can provide direct evidence for a significantly neutral IGM ($\langle x_{\rm HI} \rangle \gtrsim 0.1$, e.g., Miralda-Escudé 1998; Davies et al. 2018a; Greig et al. 2022; Ďurovčíková et al. 2024; Kist et al. 2025b; Hennawi et al. 2025). Second, proximity zones (or near zones), appear as regions of enhanced transmission immediately blueward of the Lyα emission line and arise due to the intense radiation field of the quasar ionizing the surrounding IGM at $5.5 \lesssim z \lesssim 7.5$ (Cen & Haiman 2000; Haiman & Cen 2001; Bolton & Haehnelt 2007a,b). The size of this region is sensitive to the quasar's ionizing luminosity, lifetime $t_{\rm Q}$, and the fraction of neutral hydrogen in its environment (see e.g., Fan et al. 2006b; Eilers et al. 2017, 2020; Davies et al. 2020; Satyavolu et al. 2023; Onorato et al. 2025a).

The study of damping wings and proximity zones in high-z quasars is currently one of the most powerful tools for probing the IGM during the EoR. Yet, state-of-the-art analyses are still limited by the small number of

available high signal-to-noise ratio (SNR) spectra and by challenges in modeling both the intrinsic quasar continuum and the complex IGM absorption features (Hennawi et al. 2025). In particular, systematic uncertainties due to continuum modeling, instrumental resolution, and noise must be carefully accounted for to draw robust inferences from the observed spectra (Kist et al. 2025a,b).

In this paper, we aim to address these challenges and push beyond traditional proximity zone measurements by studying a more global observable: the Lyα transmission probability distribution function (PDF). The transmission PDF, first studied by Jenkins & Ostriker (1991), provides a statistical description of the flux values in the Lyα region, and is sensitive to both the distribution of neutral hydrogen in the IGM and the local radiation field of the quasar. Importantly, it can also capture non-trivial effects such as spatial fluctuations in the IGM temperature, or patchy reionization morphology. Among the key factors shaping the transmission PDF are the pixel-level noise, the spectral resolution, and uncertainties in the quasar continuum reconstruction. In this work, we carefully account for each of these elements, treating them with a level of precision commensurate with the quality of our dataset and the demands of the analysis.

This method has previously been applied at lower redshifts using the Sloan Digital Sky Survey (SDSS-III; Eisenstein et al. 2011) Baryon Oscillation Spectroscopic Survey (BOSS) DR9 quasars (Ahn et al. 2012) by Lee et al. (2015), who modeled the transmission PDF in the Lyα forest at $z \sim 2-3$. Their work relied on forward-modeling large samples of mock spectra to match the resolution, noise properties, and continuum uncertainties of the data, and demonstrated the ability of the PDF to constrain IGM parameters such as temperature and photoionization rate. However, BOSS spectra are limited by their moderate data quality (e.g., velocity pixel scale $dv = 69$ km s^{-1}, and median SNR $\gtrsim 6-10$ per pixel), and probe the post-reionization Universe. By contrast, our work focuses on the EoR itself, using high-quality spectra of $z \gtrsim 5.7$ quasars to measure the Lyα transmission PDF in the proximity zones, where the ionizing radiation from the quasar still dominates the ionization state of the IGM, and in the Lyα forest, targeting the ambient IGM.

In particular, we analyze a subset of 21 quasars in the redshift range $5.76 < z < 6.59$, drawn from the extended XQR-30 (E-XQR-30) sample (D'Odorico et al. 2023). These spectra are continuum-modeled with the inference framework of Hennawi et al. (2025), which infers the average fraction of neutral hydrogen $\langle x_{\rm HI} \rangle$ and quasar lifetimes $t_{\rm Q}$, and jointly reconstructs the intrinsic quasar continuum shape together with IGM absorption parameters. For each quasar, we generate thousands of mock spectra that resemble the corresponding object, incorporating both continuum uncertainty and instrumental noise. These mock spectra are constructed by combining density, temperature, and velocity skewers from the `Nyx` hydrodynamical

simulations (Almgren et al. 2013; Lukić et al. 2015) with neutral hydrogen fraction fields generated using `21cmFAST` (Mesinger et al. 2011; Davies & Furlanetto 2022), followed by one-dimensional (1D) radiative transfer along each skewer to compute Lyα transmission. We then convolve the resulting spectra with the instrumental line-spread function of the data to match their exact spectral resolution. By comparing the observed transmission PDF to the distribution of mock PDFs in the selected rest-frame wavelength intervals, we are able to statistically assess the compatibility of the data with our assumed IGM model.

This paper is organized as follows. In Section 1.2, we describe the dataset and sample selection. Section 1.3 outlines the continuum inference framework, the construction of mock spectra, and the method for computing transmission PDFs. In Section 1.4, we compare the real and mock PDFs, evaluate the goodness of fit, and discuss the implications for the ionization state of the IGM. Section 1.5 summarizes our main conclusions.

1.2 The Data Set

This study is based on a sub-sample of 21 quasars selected from the 42 objects in the E-XQR-30 dataset presented by D'Odorico et al. (2023). We refer the reader to that work for a comprehensive description of the data reduction and quasar properties. Here, we summarize the key details relevant to our analysis.

The primary motivation for using the E-XQR-30 sample is its exceptional spectral quality. All spectra are obtained with VLT/X-Shooter (Vernet et al. 2011), using both the optical (VIS) and near-infrared (NIR) arms. They are rebinned to a velocity pixel scale of $dv = 50$ km s^{-1}, and the median SNR per 10 km s^{-1} pixel is ≈ 29 (D'Odorico et al. 2023). As already mentioned in Section 1.1, high and uniform spectral resolution is crucial for our work, as it directly influences the shape of the transmission PDF. Furthermore, ensuring consistent spectral properties across the sample enables a fair comparison with theoretical models.

To maximize redshift accuracy, we include only those quasars with host galaxy redshifts determined from either the [C II] 158 μm line or the narrow [O III] λ5008.24 Å emission line, observed with JWST/NIRCam WFSS (*James Webb Space Telescope* Near-InfraRed Camera Wide Field Slitless Spectroscopy; Kashino et al. 2023; Wang et al. 2023; Yang et al. 2023). We assign uncertainties of $\Delta v = 100$ km s^{-1} to both measurements (see Onorato et al. 2025a). Quasars with redshifts based solely on the Mg II line are excluded from our analysis because of the larger associated errors ($\Delta v = 390$ km s^{-1}).

We also exclude broad absorption line (BAL) quasars, as their absorption troughs, with full width at half maximum FWHM $\gtrsim 2000$ km s^{-1},

can complicate the continuum determination. Similarly, we remove quasars affected by strong absorption from optically thick associated systems, such as proximate damped Lyα absorbers (pDLAs), which can mimic the damping wing signature of a partially neutral IGM. These classifications follow Table 1 of Greig et al. (2024), although we also perform a visual inspection of all spectra to assess ourselves the potential contamination. The cases where mini-BAL or pDLA features are present, but not significant enough to warrant exclusion, are noted in Table 1.1.

The final sample spans the redshift range $5.76 < z < 6.59$, with absolute magnitudes M_{1450} at 1450 Å between -26.40 and -29.14, converting to the ionizing photon emission rates[1] $\log_{10} Q$ between 57.02 and 58.11 (see Table 1.1 and Fig. 1.1). In particular, we compute all M_{1450} following the method described in Onorato et al. (2025b), and we compare them with those calculated from the apparent magnitude M_{1450} at 1450 Å reported in D'Odorico et al. (2023) as a sanity check, finding that they are in perfect agreement.

For the purposes of our analysis, we divide the sample into seven redshift bins (see dashed lines in Fig. 1.1), each containing at least two quasars (most bins include three objects, with one bin containing two and another containing four). The lowest-z bin is often taken as an example along this work, e.g., to visualize the quasar spectra with their continuum prediction (see Section 1.3.1) in Fig. 1.2, or the mock quasar spectra in Fig. 1.3. The spectra and continua of the sources in the remaining z bins are all displayed in Fig. 1.6. This binning strategy ensures that the stacked transmission PDF represents quasars residing at similar cosmic times and are therefore expected to probe regions of the IGM with comparable properties. Moreover, this approach allows us to retain sufficient statistical power within each bin for meaningful comparisons with theoretical models.

All relevant information on the quasars in this work are listed in Table 1.1 as follows: the name of the quasar, its redshift z, the reference for this measurement, its magnitude M_{1450}, the logarithm of the ionizing photon emission rate $\log_{10} Q$, the reference for the discovery of each quasar, some comments, and the inferred values of average fraction of neutral hydrogen $\langle x_{\rm HI} \rangle$ and logarithm of quasar lifetime $\log_{10} t_{\rm Q}/{\rm yr}$ (see Section 1.3.1). The references are listed at the bottom of the table, along with additional notes.

[1] It quantifies the number of ionizing photons that the quasar emits per second, and it is computed from M_{1450}.

Table 1.1: Information on the 21 quasars in this study, sorted by increasing redshift. The horizontal lines divide every redshift bin. The details on the columns are reported in Section 1.2.

Name	z^{a}	z_{Ref}	M_{1450}	$\log_{10} Q$	Discovery	Comments[c]	$\langle x_{\mathrm{HI}} \rangle$	$\log_{10} t_{\mathrm{Q}}/\mathrm{yr}$
SDSSJ0927+2001	5.7722	14	−26.41	57.02	13		$0.05^{+0.09}_{-0.04}$	$7.36^{+0.57}_{-1.28}$
PSOJ308−27	5.803	6	−26.73	57.15	4		$0.12^{+0.11}_{-0.08}$	$5.70^{+0.74}_{-0.88}$
PSOJ242−12	5.8468	6	−26.41	57.02	4		$0.08^{+0.11}_{-0.06}$	$7.94^{+0.04}_{-0.09}$
PSOJ108+08	5.9647	6	−27.18	57.32	4		$0.34^{+0.40}_{-0.27}$	$7.15^{+0.36}_{-1.37}$
ULASJ0148+0600	5.9896	6	−27.24	57.35	15	mini BAL	$0.19^{+0.23}_{-0.14}$	$7.49^{+0.30}_{-0.20}$
SDSSJ0818+1722	5.9991	6	−27.20	57.32	13	pDLA	$0.80^{+0.15}_{-0.57}$	$7.29^{+0.14}_{-1.64}$
VDESJ0408−5632	6.0264	6	−26.49	57.06	2	mini BAL	$0.33^{+0.18}_{-0.21}$	$4.96^{+1.59}_{-0.56}$
SDSSJ1306+0356	6.033	3	−26.77	57.16	5		$0.12^{+0.15}_{-0.09}$	$7.94^{+0.05}_{-0.08}$
PSOJ158−14	6.0687	7	−27.03	57.22	11,12		$0.89^{+0.05}_{-0.24}$	$6.03^{+0.19}_{-1.99}$
SDSSJ0842+1218	6.0754	3	−26.61	57.09	10	mini BAL	$0.83^{+0.12}_{-0.37}$	$7.67^{+0.10}_{-0.11}$
CFHQSJ1509−1749	6.1228	6	−26.60	57.10	9		$0.15^{+0.22}_{-0.12}$	$7.80^{+0.14}_{-0.20}$
ULASJ1319+0950	6.1347	3	−26.95	57.23	8	pDLA	$0.62^{+0.07}_{-0.18}$	$6.81^{+0.16}_{-0.48}$
PSOJ217−16	6.1466	1	−26.70	57.13	4		$0.13^{+0.14}_{-0.10}$	$4.18^{+0.31}_{-0.22}$
PSOJ359−06	6.1722	7	−26.60	57.09	4		$0.80^{+0.13}_{-0.06}$	$6.83^{+0.16}_{-0.19}$
PSOJ060+24	6.1793	6	−26.73	57.16	4		$0.17^{+0.18}_{-0.12}$	$7.69^{+0.22}_{-0.23}$
PSOJ065−26	6.1871	3	−27.19	57.33	4	pDLA	$0.61^{+0.07}_{-0.10}$	$5.61^{+0.72}_{-1.01}$

Table 1.1: *continued*

Name	z^{a}	$z_{\rm Ref}$	M_{1450}	$\log_{10} Q$	Discovery	Comments[c]	$\langle x_{\rm HI} \rangle$	$\log_{10} t_{\rm Q}/{\rm yr}$
SDSSJ1030+0524	6.298[b]	17	−26.54	57.10	18		$0.71^{+0.16}_{-0.47}$	$7.27^{+0.18}_{-0.44}$
SDSSJ0100+2802	6.3269	3	−29.13	58.11	16		$0.34^{+0.08}_{-0.13}$	$4.40^{+0.38}_{-0.10}$
VST-ATLASJ025−33	6.3373	3	−27.76	57.56	15		$0.92^{+0.06}_{-0.12}$	$7.61^{+0.18}_{-0.13}$
PSOJ036+03	6.5405	3	−27.19	57.35	19		$0.60^{+0.07}_{-0.17}$	$6.65^{+0.18}_{-0.19}$
PSOJ323+12	6.5872	20	−26.91	57.19	20		$0.28^{+0.40}_{-0.20}$	$7.68^{+0.21}_{-0.24}$

Notes. Ref: 1 - Decarli et al. (2018); 2 - Reed et al. (2017); 3 - Venemans et al. (2020); 4 - Bañados et al. (2016); 5 - Fan et al. (2001); 6 - Private communication with Sarah Bosman; 7 - Eilers et al. (2021); 8 - Mortlock et al. (2009); 9 - Willott et al. (2007); 10 - Jiang et al. (2015); 11 - Bañados et al. (2023); 12 - Chehade et al. (2018); 13 - Fan et al. (2006a); 14 - Carilli et al. (2007); 15 - Carnall et al. (2015); 16 - Wu et al. (2015); 17 - Kashino et al. (2023); 18 - Fan et al. (2001); 19 - Venemans et al. (2015); 20 - Mazzucchelli et al. (2017).

[a]All measurements but one come from the [C II] 158 μm line, with $\Delta v = 100$ km s^{-1}.

[b]Measurement from the narrow [O III] λ5008.24 Å line, with $\Delta v = 100$ km s^{-1}.

[c]From Greig et al. (2024).

1.3 Methods

Measuring the Lyα transmission PDFs both within and beyond the proximity zone requires an accurate reconstruction of the intrinsic, unabsorbed quasar continuum, as well as a robust method for generating realistic mock spectra that reproduce the IGM fluctuations and instrumental characteristics observed in the data. In this work, we adopt the fully Bayesian inference framework developed by Hennawi et al. (2025), and already described by Kist et al. (2025a,b), which jointly models the quasar continuum and the IGM damping wing imprint. A full description of the methodology can be found in the aforementioned works; here, we summarize the main aspects relevant to our analysis.

1.3.1 Quasar continua prediction and transmission profiles

Each quasar spectrum in our sample is analyzed with a Bayesian inference pipeline that jointly fits the intrinsic continuum and the IGM absorption. The absorption model is based on the IGM transmission skewers simulated following Davies et al. (2018a), and thus obtained by post-processing sightlines from the `Nyx` hydrodynamical simulations (Almgren et al. 2013; Lukić et al. 2015) and `21cmFAST` reionization topologies (Mesinger et al. 2011; Davies & Furlanetto 2022) with 1D ionizing radiative transfer. For each object, we generate IGM transmission profiles matched to the quasar's redshift and ionizing photon emission rate $\log_{10} Q$. These simulations also include spatial fluctuations in the ultraviolet background (UVB), unlike in previous works.

The continuum is modeled using a principal component analysis (PCA; Suzuki et al. 2005; Pâris et al. 2011; Davies et al. 2018b) basis derived from a large training set of 42 854 unabsorbed quasar spectra with $1.878 < z < 3.427$ and median SNR > 10, from the SDSS-III BOSS and SDSS-IV Extended BOSS (eBOSS). The natural logarithm of the dimensionality-reduced quasar continuum $\boldsymbol{s}_{\rm DR}(\boldsymbol{\eta})$ can be described as:

$$\log \boldsymbol{s}_{\rm DR}(\boldsymbol{\eta}) = \langle \log \boldsymbol{s} \rangle + \boldsymbol{\eta}^T \boldsymbol{A}, \tag{1.1}$$

where $\langle \log \boldsymbol{s} \rangle$ is the mean of the natural logarithm of the training continua $\boldsymbol{s}$, $\boldsymbol{A}$ is the matrix of PCA basis vectors, and $\boldsymbol{\eta}$ is the vector of PCA coefficients. The corresponding relative continuum reconstruction error is estimated from a testing set of 1733 quasar spectra and defined as:

$$\boldsymbol{\delta} = \frac{\boldsymbol{s} - \boldsymbol{s}_{\rm DR}(\boldsymbol{\eta})}{\boldsymbol{s}}, \tag{1.2}$$

with $\boldsymbol{\delta}$ following a Gaussian distribution, as shown by Hennawi et al. (2025).

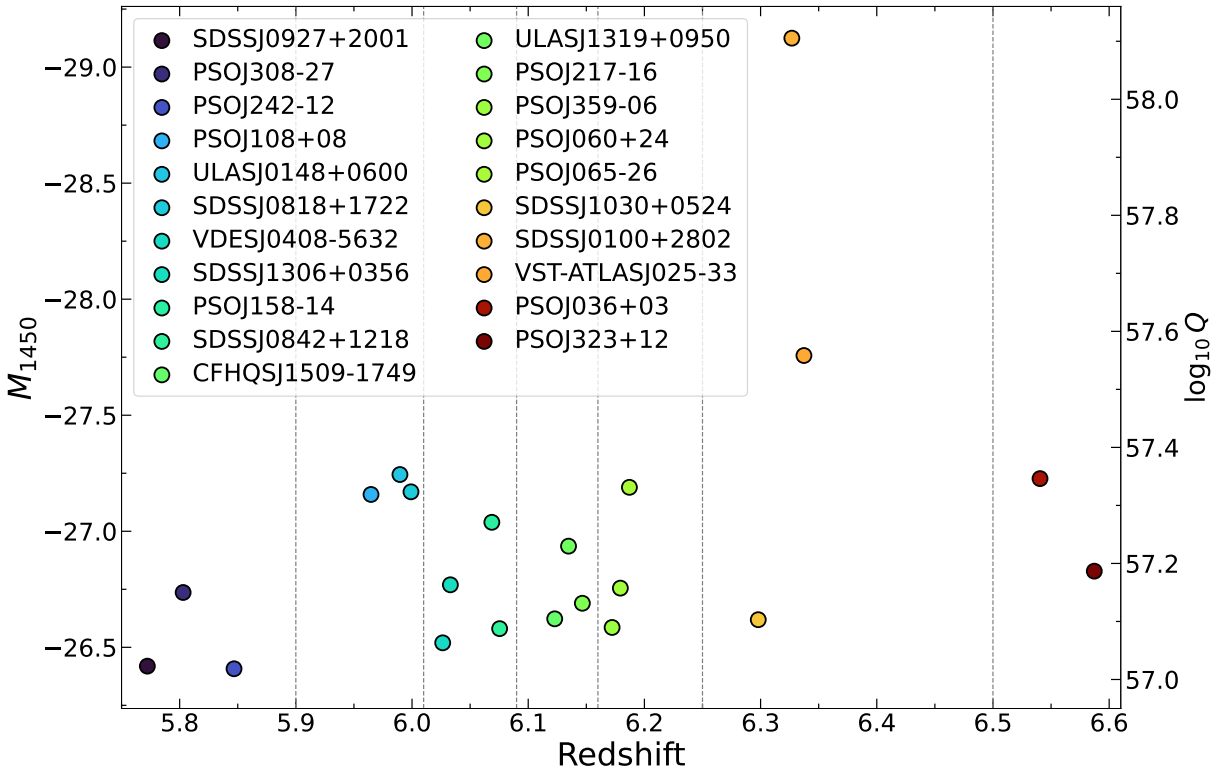


Figure 1.1: Distribution of M_{1450} (and $\log_{10} Q$, on the right-hand y-axis) as a function of z for the sub-sample of the E-XQR-30 quasars included in this work. The quasars are colour-coded with increasing z, and the grey dashed vertical lines show how they are grouped into every redshift bin.

This model expresses the rest-frame quasar spectrum as a linear combination of seven PCA coefficients $\boldsymbol{\eta}$, trained over the range $1170 < \lambda_{\rm rest}$ [Å] < 3000. They are treated as free parameters and jointly inferred with the astrophysical parameters $\boldsymbol{\theta} \equiv (\langle x_{\rm HI} \rangle, \log_{10} t_{\rm Q}/{\rm yr})$ in our Bayesian framework (see Table 1.1; they are actually inferred from the range $1190 < \lambda_{\rm rest}$ [Å] < 3000, and will be discussed in future work, i.e., Hennawi et al. in preparation). Jointly fitting the entire spectrum represents a significant improvement of the commonly adopted approach of splitting the spectral range into red and blue sides, and predicting the blue side continuum from the red one (see e.g., Davies et al. 2018b).

Posterior samples for the PCA coefficients $\boldsymbol{\eta}$ as well as the astrophysical parameters $\boldsymbol{\theta}$ are obtained by sampling the joint posterior $P(\boldsymbol{\eta}, \boldsymbol{\theta} \,|\, \boldsymbol{f}, \boldsymbol{z})$ with Hamiltonian Monte Carlo (HMC; Duane et al. 1987) using the No U-Turn Sampler (NUTS) implementation in `NumPyro`. Here, $\boldsymbol{f}$ denotes the observed quasar spectrum and $\boldsymbol{z}$ the observational noise vector. We adopt the Gaussian likelihood:

$$L(\boldsymbol{f}|\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\eta}) = \mathcal{N}\left(\boldsymbol{f}; \langle \boldsymbol{t} \rangle \circ \langle \boldsymbol{s} \rangle, \boldsymbol{Z} + \langle \boldsymbol{S} \rangle \boldsymbol{C_t} \langle \boldsymbol{S} \rangle + \langle \boldsymbol{T} \rangle \boldsymbol{C_s} \langle \boldsymbol{T} \rangle\right), \tag{1.3}$$

where $\langle \boldsymbol{t} \rangle$ and $\langle \boldsymbol{s} \rangle$ are the mean IGM transmission and continuum vectors, $\circ$ is the element-wise product, $\boldsymbol{Z} \equiv {\rm diag}(\boldsymbol{z})$, $\boldsymbol{S} \equiv {\rm diag}(\boldsymbol{s})$, $\boldsymbol{T} \equiv {\rm diag}(\boldsymbol{t})$, and $\boldsymbol{C_s}$, $\boldsymbol{C_t}$ are the corresponding covariance matrices. We impose uniform priors on the astrophysical parameters $\boldsymbol{\theta}$, i.e., $0 < \langle x_{\rm HI} \rangle < 1$, $3 < \log_{10} t_{\rm Q}/{\rm yr} < 8$, and we allow the PCA coefficients $\boldsymbol{\eta}$ to vary over the full range ${\rm Uniform}(-\infty, \infty)$.

For this inference procedure, we rebin the forward-modeled spectra to $dv = 500$ km s^{-1} over the full wavelength range because of the approximate nature of our likelihood, which does not allow us to faithfully use higher-resolution data. For our PDF analysis, instead, we adopt a dual-resolution wavelength grid to balance spectral fidelity and computational cost: a fine pixel spacing of $dv = 50$ km s^{-1} is used on the blue side ($1170 - 1218$ Å) to resolve Lyα transmission variations, while a coarser spacing of $dv = 500$ km s^{-1} is used redward of Lyα where high spectral sampling is not required.

Before starting the HMC sampling, we perform an initial maximum-likelihood optimization to find a best-fit model and then apply iterative sigma-clipping to mitigate the influence of outlier pixels on the red side ($\lambda_{\rm rest} > 1260$ Å). After masking, we run eight HMC chains with 1000 steps each, yielding 8000 posterior draws per quasar. For each source, we evaluate the dimensionality-reduced continuum model $\boldsymbol{s}_{\rm DR}(\boldsymbol{\eta})$ at the posterior draws of the PCA coefficients $\boldsymbol{\eta}$ to produce an ensemble of continuum realizations that reconstructs the continuum and its uncertainty across the full spectral range, including the Lyα forest where the intrinsic continuum cannot be directly observed.

Telluric absorption features and, more importantly, absorption features associated with intervening systems identified in the catalog of Davies et al.

(2023), are masked before inference, thus eliminating possible contamination from systems that are not related to the surrounding IGM of the quasar. All model transmission profiles are forward-modeled to resemble the instrumental properties of the data by convolving with the instrument line-spread function appropriate for each spectrograph. Finally, to ensure that the observed spectra are evaluated on the same grid as our simulated models, we rebin each observed spectrum onto the simulation dual-resolution wavelength grid, introducing only a negligible change in pixel spacing and not altering the effective spectral resolution in the studied interval.

From our continuum realizations, we compute the median continuum and its 16^{th} and 84^{th} percentile bounds (see Fig. 1.2 and 1.6). We then get Lyα transmission profiles by computing the ratio between the observed flux and the median continuum, taking the 16^{th} and 84^{th} percentiles as transmission uncertainty bands (see bottom panels of Fig. 1.5). To facilitate interpretation of the spectral features in physical space, we provide a secondary x-axis at the top left sub-panels of Fig. 1.2 and 1.6, showing the proper distance from the quasar in megaparsecs (pMpc). This conversion between rest-frame wavelength and physical distance is calculated using the measured systemic redshift of each quasar and the Planck Collaboration et al. (2020) cosmological parameters, allowing direct comparison between observed spectral features and their physical location relative to the ionizing source.

1.3.2 Mock spectra generation

We generate mock spectra for each quasar to statistically compare the observed Lyα transmission PDF to theoretical expectations. For each object, we use the 8000 posterior samples produced by the HMC inference described in Section 1.3.1 as input to the forward model to generate a *noiseless* mock spectrum. Thus, such a spectrum is entirely based on the sampled PCA coefficients $\boldsymbol{\eta}$ and astrophysical parameters $\boldsymbol{\theta} \equiv (\langle x_{\text{HI}} \rangle, \log_{10} t_{\text{Q}}/\text{yr})$, which define the intrinsic IGM state around every quasar. We ensure that the mock spectra have the same spectral resolution as the observed data ($dv = 50$ km s^{-1} in the wavelength range $1170-1218$ Å, and $dv = 500$ km s^{-1} elsewhere) to guarantee consistency in the analysis.

To render the mock spectra realistic, we incorporate the main observational sources of uncertainty as follows:

1. We apply a set of continuum perturbations drawn from the precomputed continuum error moments (equation 1.2; using the testing set of quasar spectra), with one perturbation $\boldsymbol{\delta}$ per posterior sample $\boldsymbol{\eta}$, thereby producing 8000 perturbed continua per quasar.

2. We then add wavelength-dependent Gaussian noise, constructed by scaling random normal deviates by the pixel-wise flux uncertainty of

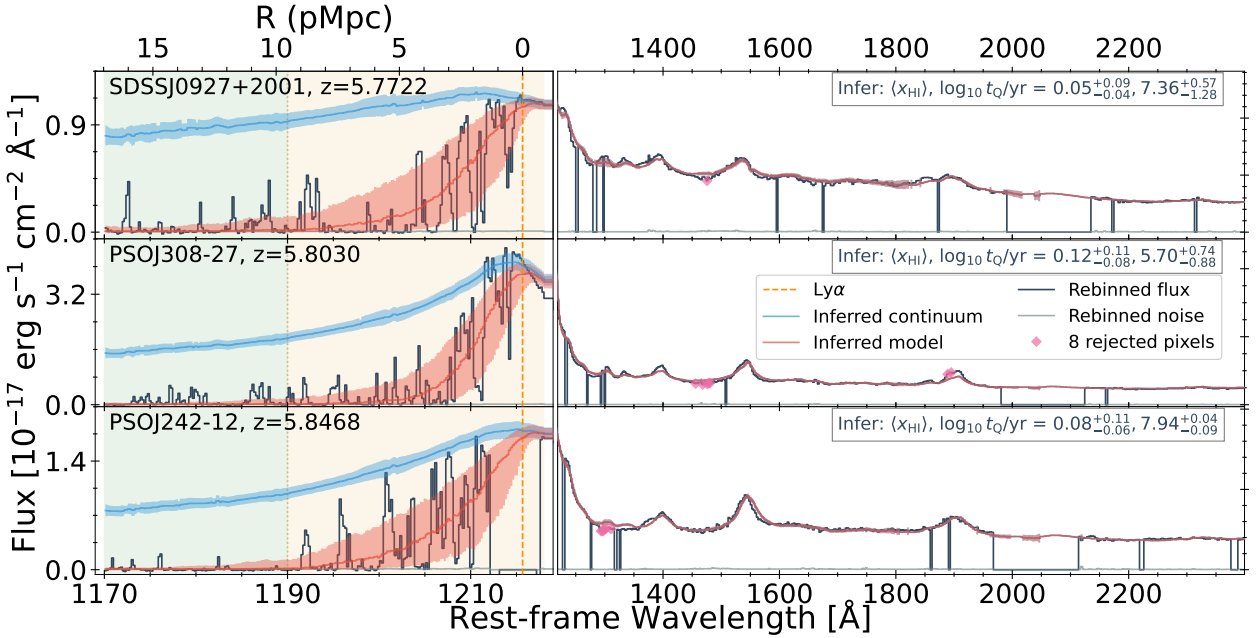


Figure 1.2: Example case for the quasars in the lowest-z bin, showing the spectra rebinned into the simulation's wavelength grid at 50 km s^{-1} for $\lambda_{\rm rest} < 1218$ Å and 500 km s^{-1} for $\lambda_{\rm rest} > 1218$ Å (black), their associated noise vector (grey), their median flux model with 16$^{\rm th}$ and 84$^{\rm th}$ percentiles (red), and their median continuum model with 1σ uncertainties computed as 16$^{\rm th}$ and 84$^{\rm th}$ percentiles of the distribution of model realizations (blue). In every sub-panel, we report the name and the redshift of the quasar on the left, and the inferred astrophysical parameters ($\langle x_{\rm HI}\rangle, \log_{10} t_{\rm Q}/{\rm yr}$) on the right. The pink diamonds are the pixels rejected from sigma-clipping. The orange dashed line marks the position of the Lyα line. The green and gold shaded regions highlight the two distinct wavelength ranges (1170 − 1190 Å and 1190 − 1218 Å, respectively) in which we study the Lyα transmission PDF (see Section 1.3.3). On the top left x-axis, we show the distance from the quasar (pMpc) converted from the rest-frame wavelengths using the systemic redshift of the quasars. The spectra and continuum predictions for the remaining sources included in this work are shown in Fig. 1.6.

the observed spectrum. This yields 8000 fully realistic mock spectra per quasar.

The final *noisy* mock spectra (see examples in Fig. 1.3) are analyzed identically to the data, allowing us to robustly estimate statistical uncertainties and perform forward-modeling comparisons.

1.3.3 PDF intervals determination

To systematically characterize the ionization structure around high-z quasars, we define two distinct rest-frame wavelength intervals for our Lyα transmission PDF analysis. These intervals are chosen to isolate the regions representing the ambient IGM from those dominated by quasar radiation. Their boundaries are established through inspection of mock transmission profiles generated under controlled conditions.

We perform this test on the first three lowest-z quasars in the sample, and hence the models adopted have the systemic redshift and ionizing photon emission rate of the corresponding sources. We first simulate mock quasar spectra using the same procedure described in Section 1.3.2, adopting as astrophysical parameters $\langle x_{\rm HI} \rangle = 0.0$ and $\log_{10} t_{\rm Q}/{\rm yr} = 6.0$ to represent a fully ionized IGM around long-lived quasars. This configuration suppresses damping wing absorption and highlights the extent of the ionized proximity zone. For each quasar in our lowest-z bin, we compute the mock Lyα transmission and average the results to construct a composite transmission profile as a function of rest-frame wavelength (see Fig. 1.4).

This averaged profile reveals a clear transition between two regimes:

1. *Lyα forest* ($1170 - 1190$ Å): a lower-transmission region located in an interval $\simeq 9 - 17$ pMpc from the quasar, where the ionizing influence of the quasar is minimal. This interval reflects the ambient IGM and serves as a baseline for comparison.

2. *Proximity zone* ($1190 - 1218$ Å): a high-transmission region located approximately in a range $0 - 9$ pMpc from the quasar. Transmission here is elevated due to the quasar's ionizing radiation. We choose the upper limit at 1218 Å to avoid contamination from the intrinsic Lyα emission line and because it is the wavelength limit of our high-resolution model, while the lower limit at 1190 Å marks the point where the transmission gradient flattens.

To validate this demarcation, we compare the mock transmission profile in the Lyα forest with the mean Lyα flux measured by Bosman et al. (2022, blue curve in Fig. 1.4), interpolated at the corresponding absorption redshifts. The agreement between our mock transmission and the empirical measurements confirms that the $1170 - 1190$ Å interval reliably traces the ambient IGM unaffected by quasar radiation.

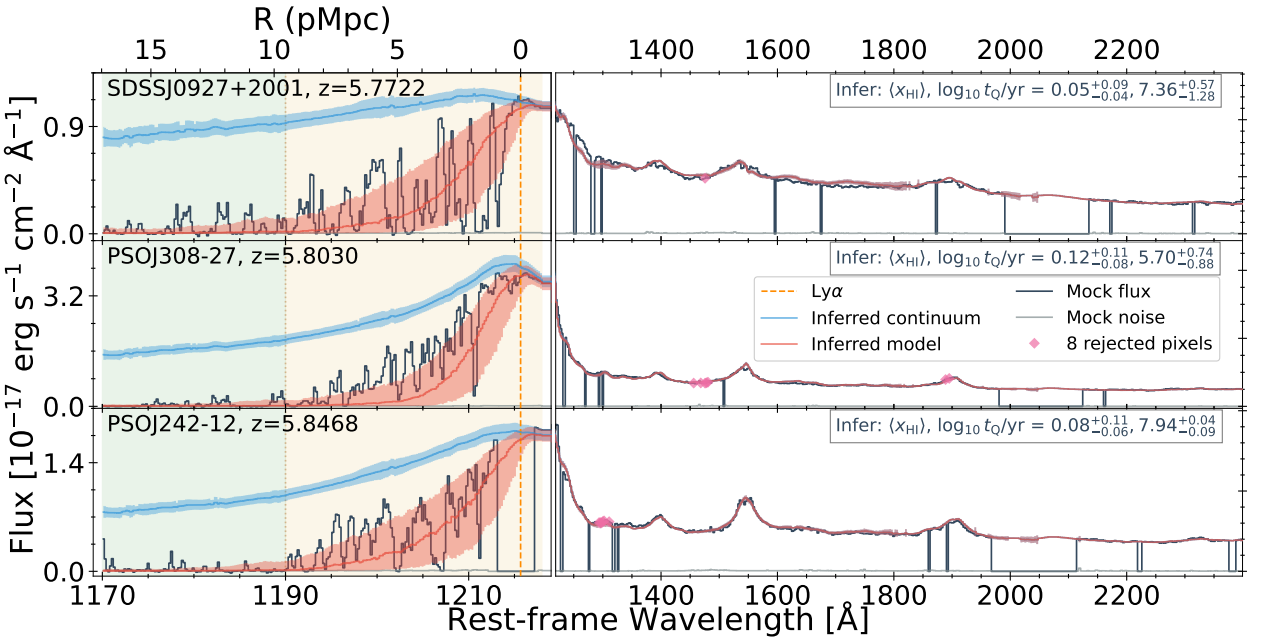


Figure 1.3: Example case for the realistic mock quasars in the lowest-z bin, showing a randomly extracted noisy mock spectrum (black), its associated noise vector (grey), its median flux model with $16^{\rm th}$ and $84^{\rm th}$ percentiles (red), and median continuum model with 1σ uncertainties computed as $16^{\rm th}$ and $84^{\rm th}$ percentiles of the distribution of model realizations (blue). These mock spectra have $dv = 50$ km s^{-1} for $\lambda_{\rm rest} < 1218$ Å and $dv = 500$ km s^{-1} for $\lambda_{\rm rest} > 1218$ Å. In every sub-panel, we report the name and the redshift of the simulated quasar on the left, and the inferred astrophysical parameters ($\langle x_{\rm HI}\rangle, \log_{10} t_{\rm Q}/{\rm yr}$) on the right. The pink diamonds are the pixels rejected from sigma-clipping. The orange dashed line marks the position of the Lyα line. The green and gold shaded regions highlight the two distinct wavelength ranges ($1170-1190$ Å and $1190-1218$ Å, respectively) in which we study the Lyα transmission PDF (see Section 1.3.3). On the top left x-axis, we show the distance from the quasar (pMpc) converted from the rest-frame wavelengths using the systemic redshift of the quasars.

These two wavelength intervals are used consistently across our analysis. By measuring the Lyα transmission PDF separately in the Lyα forest and the proximity zone, we can isolate the imprint of the quasar's ionizing field and compare it to the surrounding medium, enabling more robust constraints on our IGM models.

1.4 Results

In this section, we describe our real and mock transmission PDFs, select informative histogram bins, and assess residuals between observations and theoretical predictions. We then compute their covariance matrix, which is essential for the χ^2 analysis that we adopt to quantify the agreement between real and simulated data.

1.4.1 Transmission PDF and residuals

We compute the Lyα transmission PDF for both the real and mock data within the two rest-frame wavelength intervals defined in Section 1.3.3: the Lyα forest ($1170 - 1190$ Å) and the proximity zone ($1190 - 1218$ Å). We construct the transmission PDF by binning the transmitted flux, obtained by dividing the observed rebinned flux by the modeled continuum (see Section 1.3.1), into normalized histograms (see upper panels of Fig. 1.5). The procedure is described in detail below.

We analyze the transmission in the interval $[-0.2, 1.5]$, with bin width $\Delta\mathrm{B} = 0.025$ for the Lyα forest, and $\Delta\mathrm{B} = 0.05$ for the proximity zone, resulting in $B = 67$ bins in the first case and $B = 33$ in the second one. We define the minimum possible non-zero probability density value $p_{\min}$ in a normalized histogram as:

$$p_{\min} = \frac{1\,\mathrm{sample}}{\mathrm{N_{samples}} \cdot \Delta\mathrm{B}}, \tag{1.4}$$

where $\mathrm{N_{samples}}$ is the total number of data points in the real transmission sample and $\Delta\mathrm{B}$ is the width of each histogram bin. The value of $p_{\min}$ corresponds to the height that a single bin would reach in a normalized PDF if it contained exactly one data point. We use it as a reference threshold in the histogram to show the minimum meaningful signal level (dashed black line in the upper panels of Fig. 1.5), as it helps identify bins with too few samples.

To quantify the statistical uncertainties in the transmission PDF due to finite sampling, we generate an ensemble of mock PDFs (named as $\boldsymbol{p}_{\mathrm{mock},i}$, with i the i-th mock PDF realization, in Section 1.4.2) via stochastic subsampling from the total set of mock quasars. Specifically, we first compute

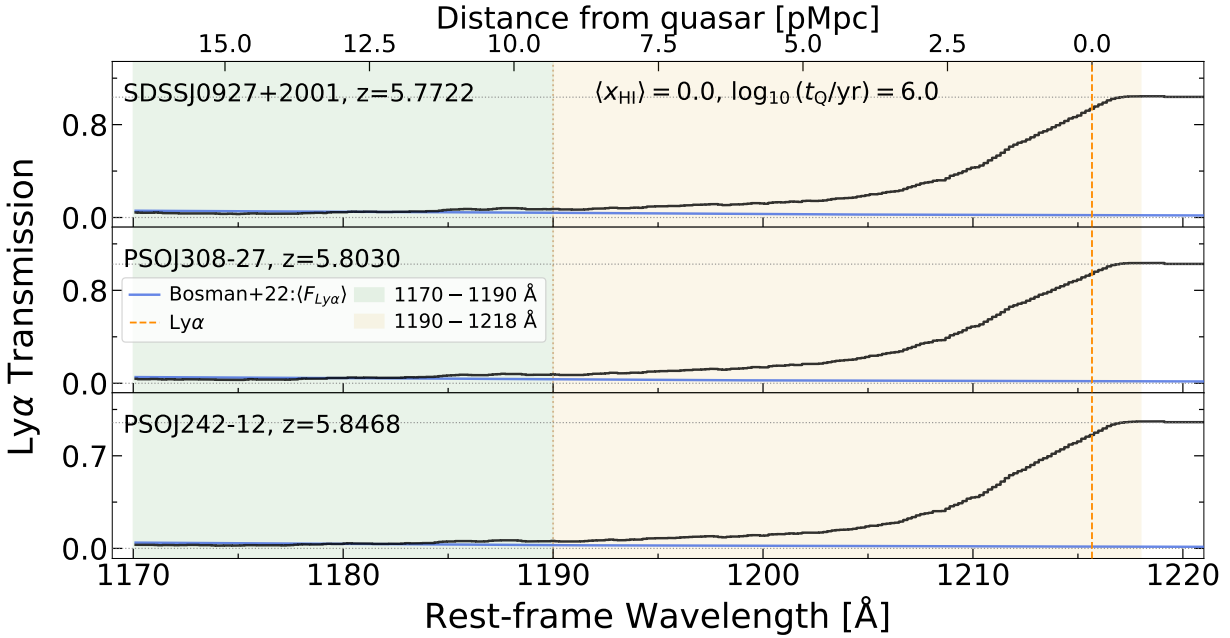


Figure 1.4: Average mock Lyα transmission profiles for the quasars in the lowest-z bin (black), generated using our forward-modeling approach with parameters inferred from HMC implementation (see Sections 1.3.1 and 1.3.2) assuming $\langle x_{\rm HI}\rangle = 0.0$ and $\log_{10} t_{\rm Q}/{\rm yr} = 6.0$. In every sub-panel, we report the name and the redshift of the quasar whose mean mock transmission is plotted. The horizontal dotted grey lines indicate where the transmissions are 0 and 1. The orange dashed line marks the position of the Lyα line. On the upper x-axis, we show the distance from the quasar [pMpc] converted from the rest-frame wavelengths using the systemic redshift of the quasars. The green and gold shaded regions highlight the two distinct wavelength ranges ($1170 - 1190$ Å and $1190 - 1218$ Å, respectively) in which we study the Lyα transmission PDF (see Section 1.3.3). The first interval is representative of the ambient IGM, while the second one is defined such that the transmission profile is mostly dominated by the quasar radiation. The blue curve displays the empirical measurements of the mean Lyα flux from Bosman et al. (2022), showing that our transmission in the Lyα forest converges to the expected mean flux at the corresponding redshifts, and further validating our interval choices.

normalized transmission histograms for each individual mock quasar realization using a fixed set of flux bins. We then perform $N = 8000$ iterations of resampling, where in each iteration we randomly select a subset of mock quasars equal in number to the real quasar sample, and compute the average transmission PDF across that subset. This yields an ensemble of 8000 mock PDFs, which collectively represent the stochastic variance expected from a sample of our size.

This ensemble is used to assess the robustness of features in the observed PDF, distinguishing statistically significant structure from noise due to limited sample size. Furthermore, the ensemble forms the basis for calculating the PDF covariance and correlation matrices (Section 1.4.2), which are critical for the subsequent χ^2 comparison between real and mock data (Section 1.4.3).

The number of histogram bins ultimately used for the statistical analysis is smaller than the total initially defined. In fact, many of them are populated with PDF values at or near $p_{\min}$ and are therefore uninformative. Consequently, we retain only bins that satisfy one or more of the following inclusion criteria:

- *Statistical significance filter*: we select bins where the average of simulated PDF exceeds $90\%\, p_{\min}$ (defined in equation 1.4). This ensures we only analyze bins with statistically significant data in the mock samples.

- *Real data filter*: we select bins that contain at least one real data point. This way, we ensure we do not exclude bins with actual observations, even if mock data is sparse there.

- *Continuous range creation*: we create a contiguous range of bins from the combined selection defined by the previous requirements. This ensures continuity in the subsequent analysis without any gaps among the bins selected with the first two criteria.

Applying these conditions yields k *non-zero* bins, which are the only ones included in the subsequent statistical analysis, and we mark them as light green bands in the upper panels of Fig. 1.5.

In the second rows of Fig. 1.5, we display how significantly the observed data PDF differs from the mock simulation PDF in each transmission bin, normalized by the uncertainty σ. We define σ as the standard deviation in every bin, computed as $\sigma = \sqrt{\mathrm{diag}(\mathbf{\Sigma})}$, where $\mathbf{\Sigma}$ is the covariance matrix calculated in Section 1.4.2 (see equation 1.6). Bars near zero indicate good agreement between data and mocks, while values beyond ± 3 would typically be considered significant deviations.

1.4.2 Covariance and correlation matrices

For each wavelength region, we quantify the covariance and correlation between transmission histogram bins by computing the covariance and correlation matrices from our ensemble of mock PDFs generated through stochastic subsampling (see Section 1.4.1).

Our set of mock PDF can be represented by a 2D array $\boldsymbol{P}_{\rm mock} \in \mathbb{R}^{N\times B}$, where N is the number of mock realizations ($N = 8000$), and B is the total number of histogram bins. Each row of $\boldsymbol{P}_{\rm mock}$ corresponds to a single mock PDF realization, and each column corresponds to a specific bin of the transmission PDF. We compute the sample mean PDF across all mock realizations as:

$$\langle \boldsymbol{p}_{\rm mock} \rangle = \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{p}_{\rm mock,i}, \tag{1.5}$$

where $\boldsymbol{p}_{\rm mock,i}$ denotes the i-th row of $\boldsymbol{P}_{\rm mock}$, i.e., the i-th mock PDF realization. The unbiased[2] sample covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{B\times B}$ is then calculated as:

$$\boldsymbol{\Sigma} = \frac{1}{N-1} \sum_{i=1}^{N} \left(\boldsymbol{p}_{\rm mock,i} - \langle \boldsymbol{p}_{\rm mock} \rangle\right) \left(\boldsymbol{p}_{\rm mock,i} - \langle \boldsymbol{p}_{\rm mock} \rangle\right)^{\top}, \tag{1.6}$$

which captures the bin-to-bin covariances of the transmission PDF due to sampling variance in the mocks. To facilitate the interpretation of bin correlations independent of amplitude, we also compute the correlation matrix $\boldsymbol{C} \in \mathbb{R}^{B\times B}$, derived from $\boldsymbol{\Sigma}$ as:

$$\boldsymbol{C}_{ij} = \frac{\boldsymbol{\Sigma}_{ij}}{\sqrt{\boldsymbol{\Sigma}_{ii}\boldsymbol{\Sigma}_{jj}}}, \qquad \forall, i, j = \{1, ..., B\}. \tag{1.7}$$

This normalized matrix has unit diagonal elements ($\boldsymbol{C}_{ii} = 1$) and values in the range $[-1, 1]$ elsewhere, quantifying the linear correlation between fluctuations in different bins. Our final $\boldsymbol{\Sigma}$ and $\boldsymbol{C}$ matrices are obtained after applying the three conditions listed in Section 1.4.1, and have shape $(k \times k)$, with k being equal to the number of *non-zero* bins. We show two examples of $\boldsymbol{C}$ in Fig. 1.7; they are used as a diagnostic tool to identify which bins in the PDF are statistically correlated, and to assess whether any features in the observed PDF might be due to correlations induced by the finite sampling.

The resulting covariance matrix $\boldsymbol{\Sigma}$ is used to compute the χ^2 statistic for the comparison between the observed and mock PDF.

[2]The sample covariance is unbiased when computed with a denominator of $N-1$ (instead of N), meaning its expectation value equals the true population covariance. This correction accounts for the additional variance introduced by estimating the mean from the same sample.

Table 1.2: Summary of the χ^2 analysis results for each redshift bin and wavelength interval. The first column indicates the z bin considered, along with the number of quasars in it ($N_{\rm QSOs}$). The second column lists the rest-frame wavelength interval considered, corresponding to the Lyα forest ($1170 - 1190$ Å) and the proximity zone ($1190 - 1218$ Å). The third column reports the number of transmission bins k that satisfy our selection criteria and are retained for statistical comparison (see Section 1.4.1). The fourth and fifth columns show the $\chi^2_{\rm real}$ statistic and the associated CDF($< \chi^2_{\rm real}$), computed via comparison with the multivariate distribution of mock PDF (see Section 1.4.3). High CDF($< \chi^2_{\rm real}$) values, e.g., ≈ 1, indicate tension between the observed transmission PDF and the predictions from our IGM model.

z bin ($N_{\rm QSOs}$)	$\lambda_{\rm rest}$ [Å] range	k	$\chi^2_{\rm real}$	CDF($< \chi^2_{\rm real}$)
$5.7722 - 5.8468$ (3)	$1170 - 1190$	23	30.87	0.79
	$1190 - 1218$	23	19.42	0.38
$5.9647 - 5.9991$ (3)	$1170 - 1190$	21	28.24	0.76
	$1190 - 1218$	23	12.23	0.06
$6.0264 - 6.0754$ (4)	$1170 - 1190$	28	130.90	0.97
	$1190 - 1218$	23	43.04	0.97
$6.1228 - 6.1466$ (3)	$1170 - 1190$	11	16.69	0.82
	$1190 - 1218$	23	17.46	0.29
$6.1722 - 6.1871$ (3)	$1170 - 1190$	17	20.84	0.71
	$1190 - 1218$	26	> 1000	1.00
$6.298 - 6.3373$ (3)	$1170 - 1190$	17	79.99	0.95
	$1190 - 1218$	22	30.76	0.87
$6.5405 - 6.5872$ (2)	$1170 - 1190$	12	> 1000	1.00
	$1190 - 1218$	26	> 1000	1.00

1.4.3 Chi-square distribution

To quantitatively compare the observed Lyα transmission PDF to theoretical expectations, we compute χ^2 statistic that measures the likelihood of the observed PDF under the multivariate distribution defined by the mocks. Specifically, we model the mock transmission PDF as samples drawn from a multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ (computed in equation 1.7). To do so, we take into account the following quantities:

- $\boldsymbol{p}_{\text{real}}$: the observed transmission PDF (with size k *non-zero* bins) from our quasar spectra.
- $\boldsymbol{p}_{\text{mock,i}}$: the ensemble of mock PDF (N samples, each with k *non-zero* bins) generated from our forward-modeling pipeline, representing the expected statistical variation.
- $\boldsymbol{\mu}$: the mean mock PDF (with size k *non-zero* bins), calculated by creating a histogram of all transmission values aggregated from our mock spectra, and serving as our theoretical model prediction.
- $\boldsymbol{\Sigma}$: the final covariance matrix of the mock PDF with shape $(k \times k)$, capturing correlations between transmission values.

We have that the PDF for the multivariate normal model is given by:

$$\boldsymbol{P}(\boldsymbol{p}) = \frac{1}{\sqrt{(2\pi)^k \det \boldsymbol{\Sigma}}} e^{-\frac{1}{2}(\boldsymbol{p}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{p}-\boldsymbol{\mu})}, \tag{1.8}$$

where $-\frac{1}{2}(\boldsymbol{p}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{p}-\boldsymbol{\mu}) = \chi^2(\boldsymbol{p})$, and $\boldsymbol{p}$ represents either $\boldsymbol{p}_{\text{real}}$ or $\boldsymbol{p}_{\text{mock,i}}$. From this, the χ^2 statistic is derived as:

$$\chi^2(\boldsymbol{p}) = -2\left[\log \boldsymbol{P}(\boldsymbol{p}) + \frac{k}{2}\log(2\pi) + \frac{1}{2}\log\det\boldsymbol{\Sigma}\right]. \tag{1.9}$$

This approach accounts for the full covariance of the mock distribution, giving a rigorous metric for evaluating the consistency of the observed PDF with model expectations, provided that our assumption of Gaussianity is correct (see Appendix 1.8).

Before computing χ^2, we perform an eigenvalue decomposition of the covariance matrix $\boldsymbol{\Sigma}$ to ensure numerical stability. If any eigenvalues fall below a small threshold (set to 10^{-8} times the largest eigenvalue), we regularize them by setting a minimum floor. This step ensures the invertibility of the covariance matrix, preventing numerical instabilities in the evaluation of $\chi^2(\boldsymbol{p})$, and can be written as:

$$\boldsymbol{\Sigma}_{\text{reg}} = \boldsymbol{V} \cdot \text{diag}(\max(\lambda_i, \varepsilon)) \cdot \boldsymbol{V}^\top, \tag{1.10}$$

where $\boldsymbol{V}$ are the eigenvectors, λ_i are the original eigenvalues, and ε is the threshold value.

Finally, we compute the χ^2 statistic of the real PDF relative to the mock distribution ($\chi^2_{\rm real}$, see Table 1.2) and the χ^2 statistic of each mock PDF with respect to the same distribution ($\chi^2_{\rm mock,i}$). The distribution of $\chi^2_{\rm mock}$ values allows us to put the observed $\chi^2_{\rm real}$ into context, effectively yielding a significance level for the consistency between the observed transmission PDF and the model predictions. To quantify this, we compute the empirical:

$$p_{\rm value} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\left(\chi^2_{{\rm mock},i} \geq \chi^2_{\rm real}\right), \tag{1.11}$$

where $\mathbb{I}$ is the indicator function, equal to 1 if the condition is true and 0 otherwise, and N is the number of mock realizations. The $p_{\rm value}$ estimates the probability of obtaining a χ^2 value as extreme as, or more extreme than, the real one under the null hypothesis that the model describes the data well. For further clarity, we also define the cumulative distribution function (CDF) of the $\chi^2_{\rm mock}$ distribution, which quantifies the percentile rank of $\chi^2_{\rm real}$. Given these quantities, we use ${\rm CDF}(< \chi^2_{\rm real}) = 1 - p_{\rm value}$ to highlight the deviation from the model prediction: a value of ${\rm CDF}(< \chi^2_{\rm real}) \approx 1$ would indicate a strong tension between the data and the model, while ${\rm CDF}(< \chi^2_{\rm real}) \lesssim 0.5$ would suggest good agreement.

To visualize this comparison, we construct a histogram-based PDF of the χ^2 values from the mocks, and compare it to the theoretical χ^2 distribution with $\nu = k - 1$ degrees of freedom (dof), where k is the number of *non-zero* bins, and we subtract one due to the normalization constraint on the PDF:

$${\rm PDF}_{\chi^2}(\chi^2;\nu) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)}(\chi^2)^{\nu/2-1}e^{-\chi^2/2}, \tag{1.12}$$

here, Γ is the gamma function. The position of the observed $\chi^2_{\rm real}$ statistic relative to both the distribution of $\chi^2_{\rm mock}$ values and the theoretical χ^2 PDF serves as a diagnostic tool for assessing the goodness of fit between the data and the IGM model. If the mock transmission PDF are consistent with being drawn from a multivariate Gaussian distribution, as assumed in our likelihood formulation, then the histogram of $\chi^2_{\rm mock}$ values should closely follow the theoretical χ^2 PDF.

We illustrate this comparison in the third rows of Fig. 1.5. The orange bars show the empirical distribution of $\chi^2_{\rm mock}$ computed from the mock ensemble, the blue curve represents the theoretical χ^2 PDF for the corresponding number of dof, and the red dashed line indicates the value of $\chi^2_{\rm real}$ measured from the data. For additional insight, we include a secondary y-axis showing the CDF of the $\chi^2_{\rm mock}$ distribution (solid green curve). The horizontal green dotted line corresponds to ${\rm CDF}(< \chi^2_{\rm real})$, intersecting the CDF at the location of $\chi^2_{\rm real}$.

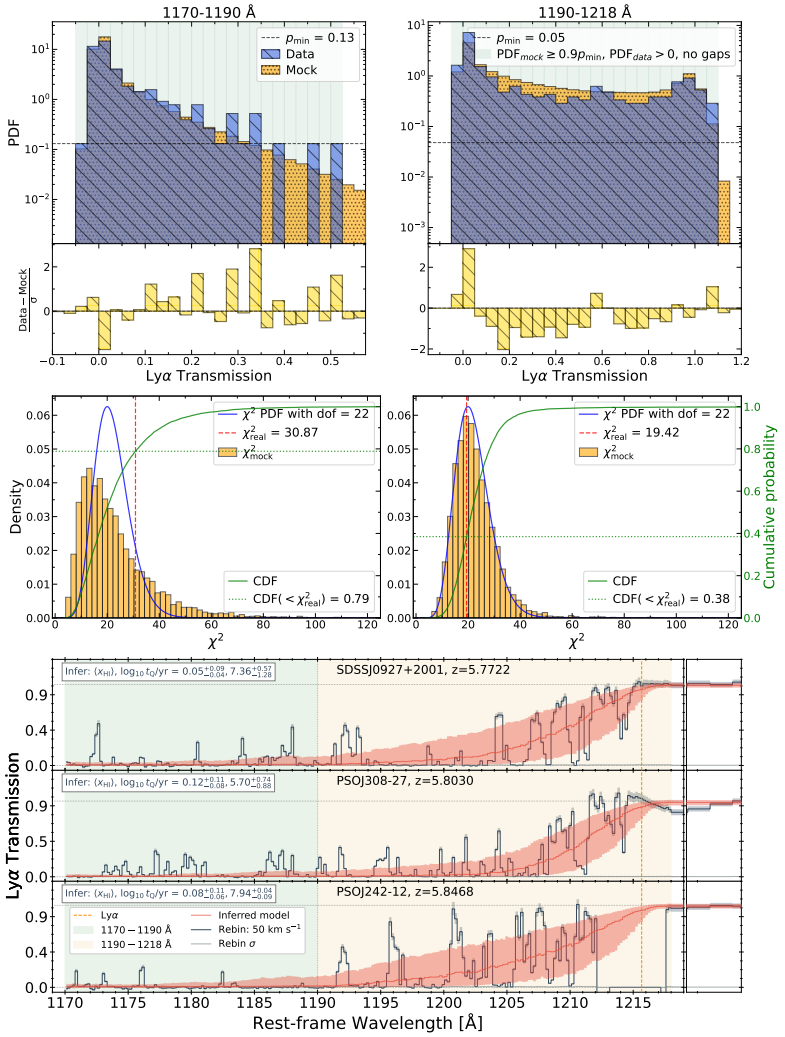


Figure 1.5: *Upper panels*: Lyα transmission PDF in the Lyα forest (left-hand side, $1170-1190$ Å) and in the proximity zone (right-hand side, $1190-1218$ Å). The blue histograms show PDF from the real data, while the orange histograms the mean PDF from the mock data. The dashed black line indicates the minimum non-zero probability $p_{\rm min}$, and the shaded green regions highlight the k *non-zero* bins used for statistical analysis. *Second rows*: Residuals between data and mean mock PDF, normalized by the standard deviation in each bin. *Third rows*: Distribution of χ^2 values from the mock ensemble (orange histogram) compared with the theoretical χ^2 PDF with $\nu = k - 1$ dof (blue curve). The red dashed line marks $\chi^2_{\rm real}$, and the secondary y-axis shows the cumulative distribution function (CDF, in green), with the green dotted line showing where CDF($< \chi^2_{\rm real}$). *Bottom panels*: transmission profiles rebinned into the simulation's wavelength grid at 50 km s^{-1} for $\lambda_{\rm rest} < 1218$ Å and 500 km s^{-1} for $\lambda_{\rm rest} > 1218$ Å with shaded bands indicating the 1σ uncertainty range due to noise and continuum realizations (black), their associated noise vector (grey), and their median transmission model with 16$^{\rm th}$ and 84$^{\rm th}$ percentiles (red). In every sub-panel, we report the name and the redshift of the quasar at the center, and the inferred astrophysical parameters ($\langle x_{\rm HI}\rangle, \log_{10} t_{\rm Q}/{\rm yr}$) on the left. The horizontal dotted grey lines indicate where the transmissions are 0 and 1. The orange dashed line marks the position of the Lyα line. The green and gold shaded regions highlight the two distinct wavelength ranges ($1170-1190$ Å and $1190-1218$ Å, respectively) in which we study the Lyα transmission PDF (see Section 1.3.3).

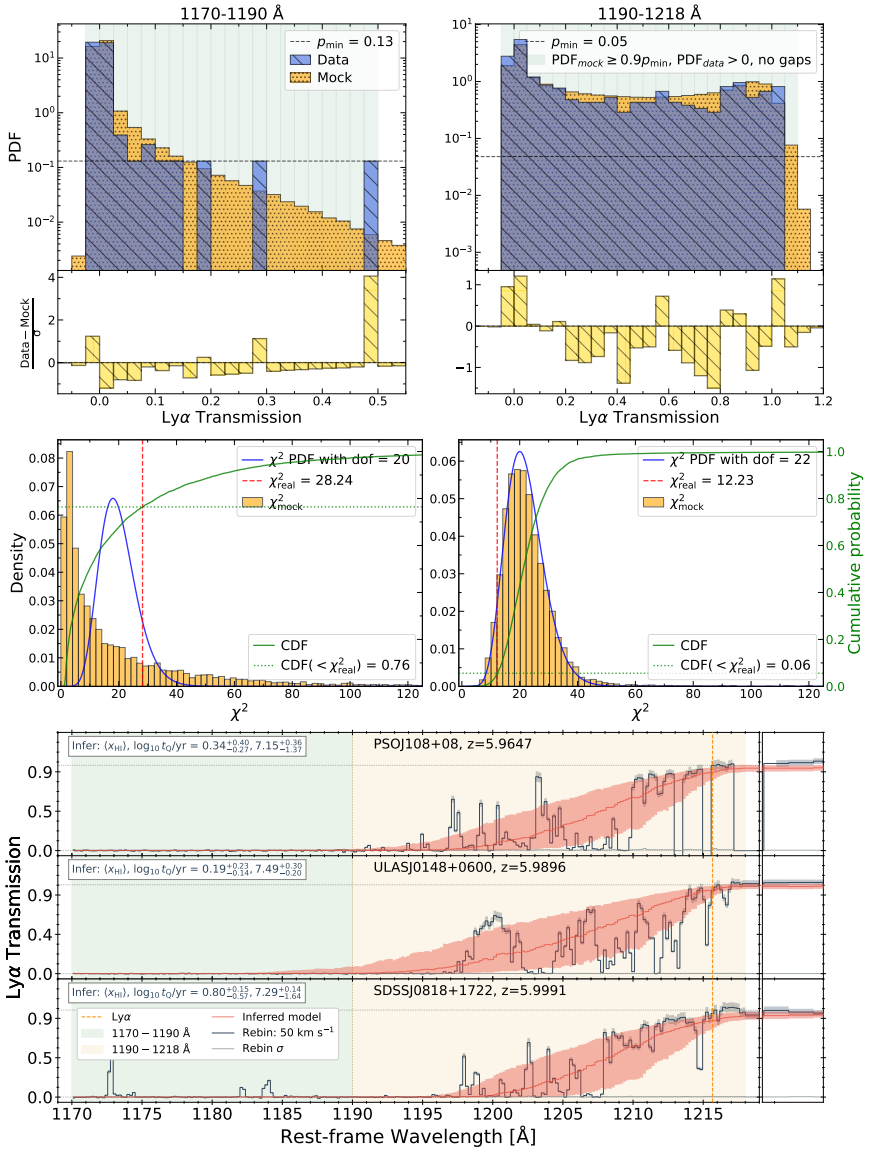


Figure 1.5: (Continued)

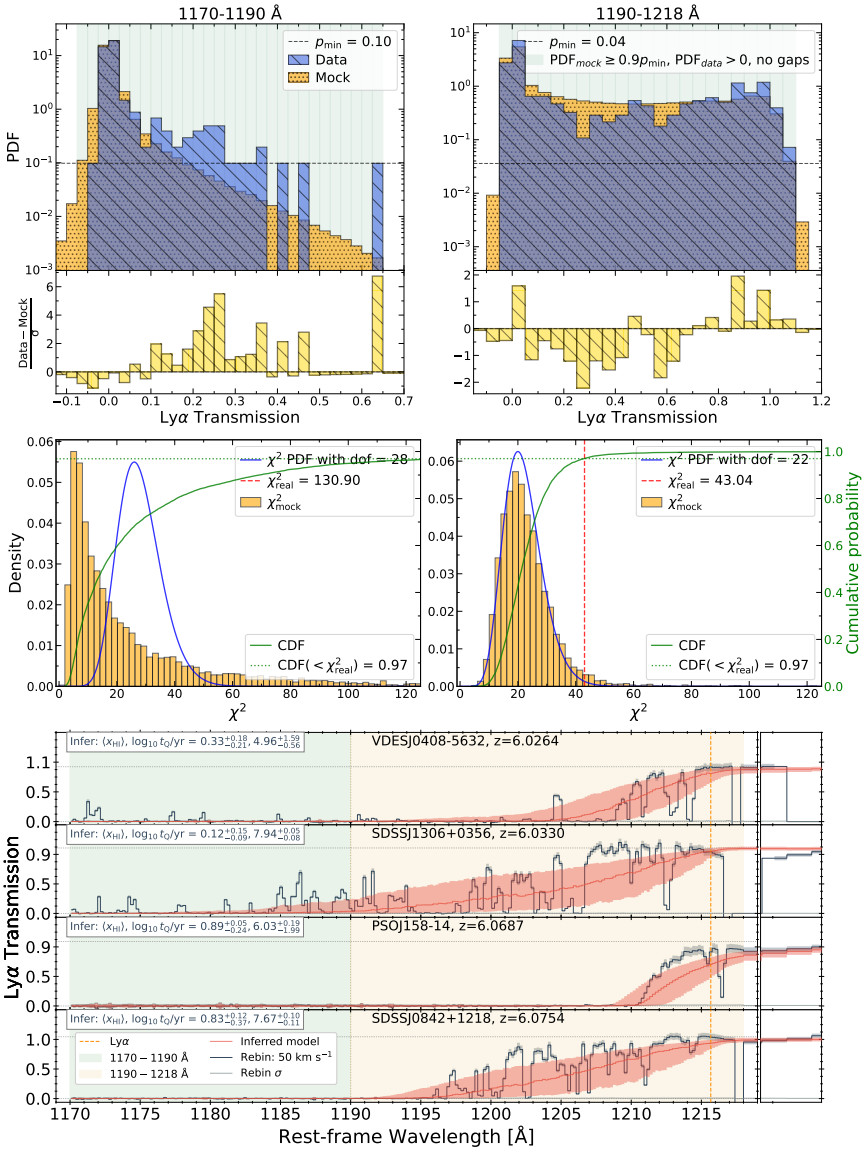


Figure 1.5: (Continued)

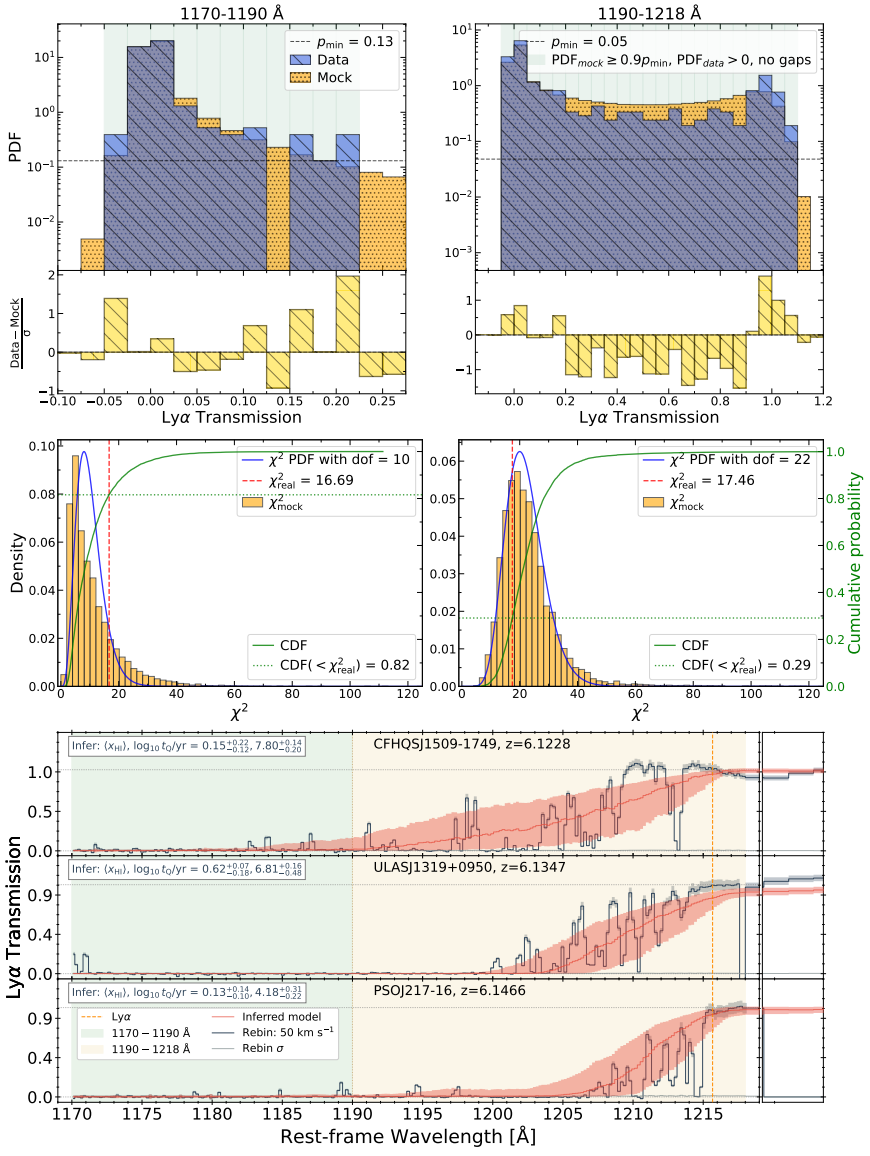


Figure 1.5: (Continued)

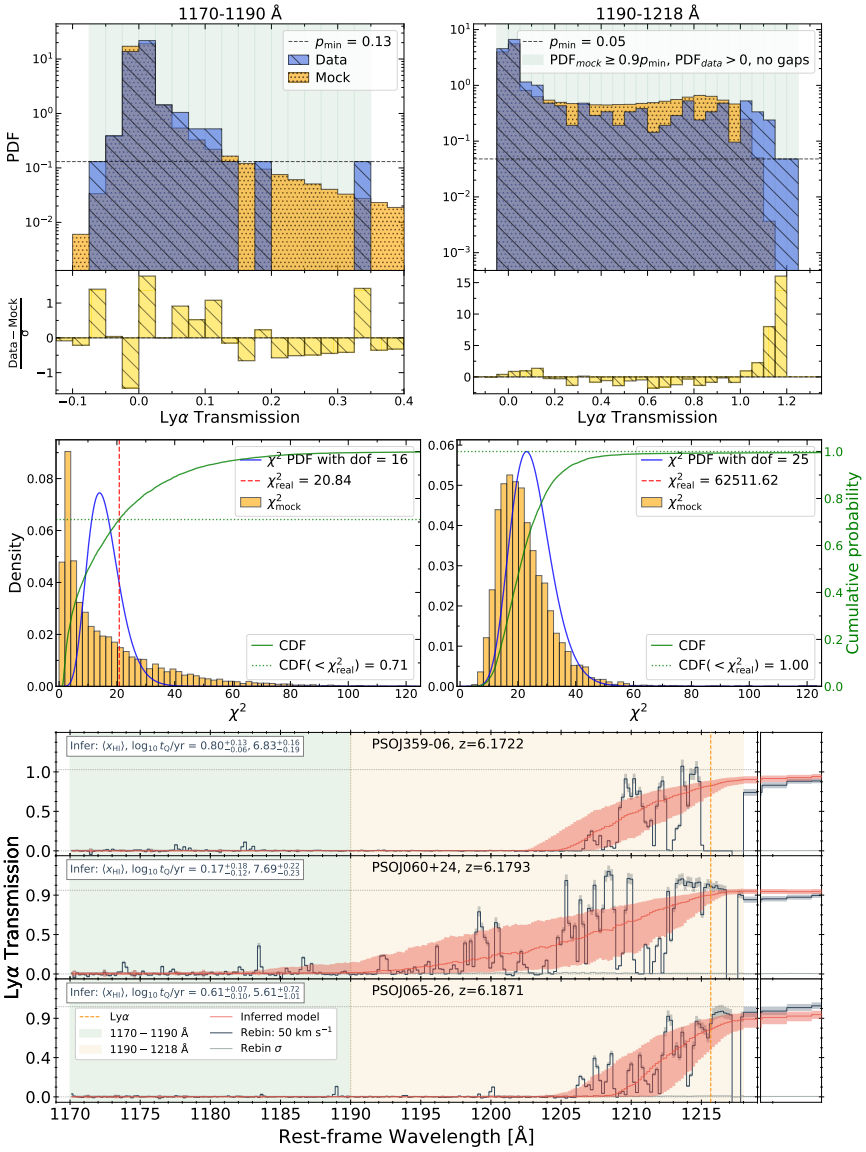


Figure 1.5: (Continued)

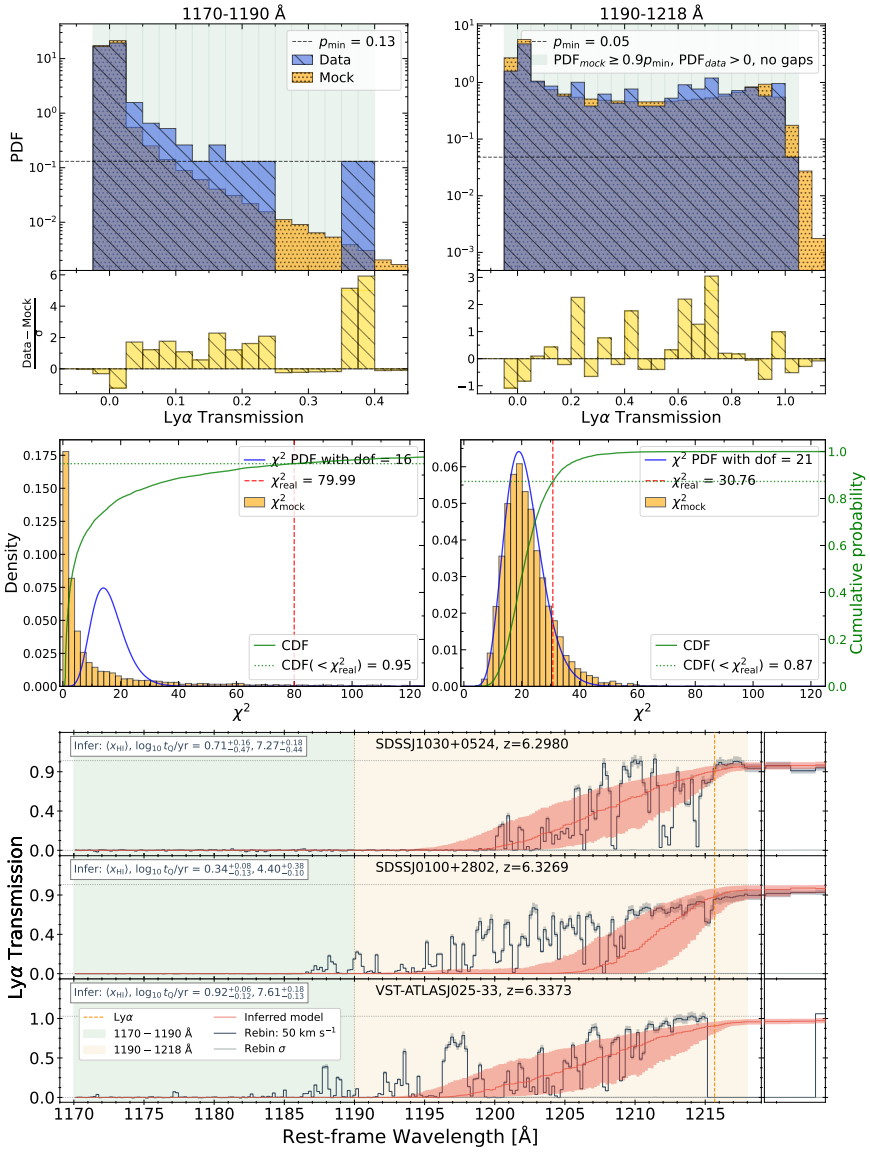


Figure 1.5: (Continued)

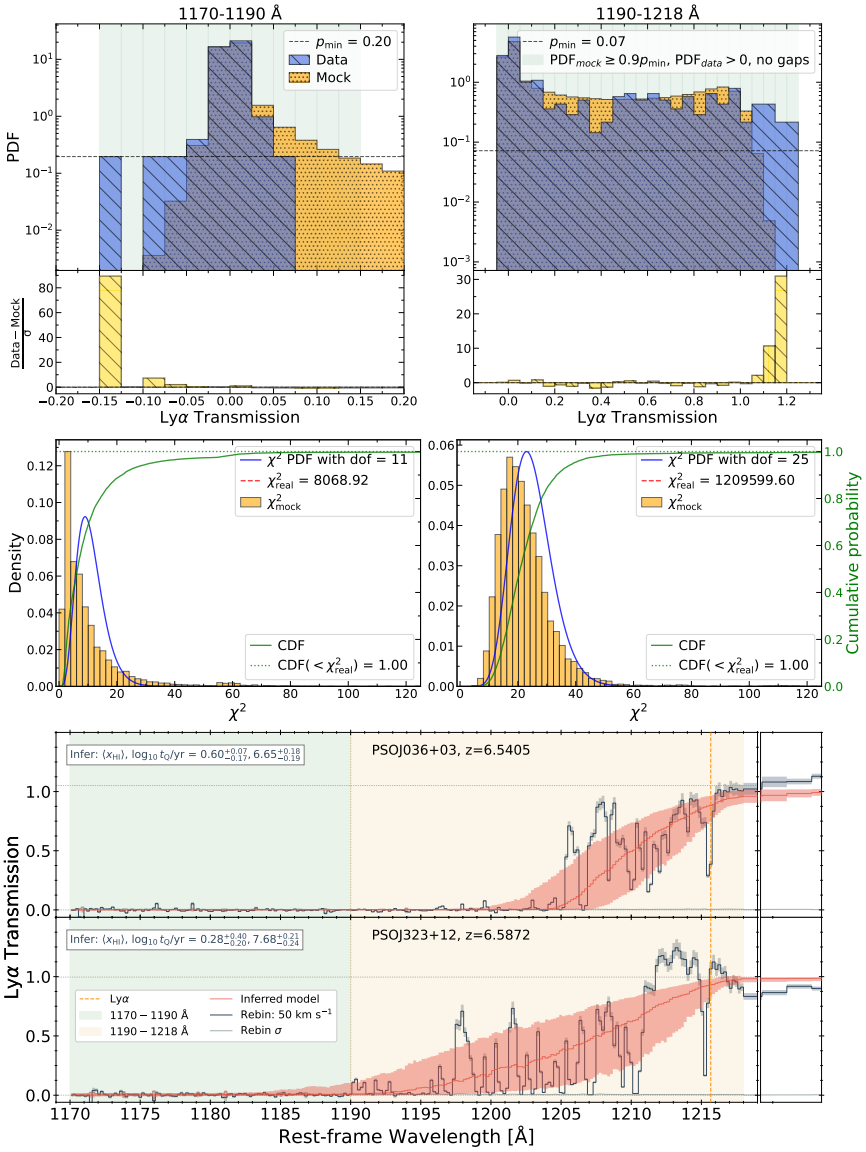


Figure 1.5: (Continued)

1.4.4 Discussion on the results

Looking at the third rows of Fig. 1.5, we can see that in the proximity zone ($1190 - 1218$ Å) the histogram of $\chi^2_{\rm mock}$ is generally well described by the theoretical χ^2 distribution in nearly all z bins. Exceptions, such as the intervals $z = 6.1722 - 6.1871$ and $z = 6.5405 - 6.5872$, exhibit mild shifts due to the relatively large number of dof, resulting from a higher number of retained transmission bins that are poorly populated (see Section 1.4.1). In contrast, the behavior is markedly different in the Lyα forest ($1170 - 1190$ Å), where the empirical $\chi^2_{\rm mock}$ distributions often diverge from the theoretical prediction, particularly in $z = 6.0264 - 6.0754$, $z = 6.2980 - 6.3373$, and $z = 6.5405 - 6.5872$. We investigate the origins of this non-Gaussian behavior in Appendix 1.8.

Large values of $\chi^2_{\rm real}$ in some bins indicate significant discrepancies between the observed transmission PDF and those predicted by our IGM model. These deviations are typically driven by individual transmission bins that the simulations fail to reproduce accurately, suggesting tension between the data and the underlying physical assumptions. The values of $\chi^2_{\rm real}$ and the associated CDF($< \chi^2_{\rm real}$) offer a quantitative diagnostic of where and how the current approach falls short, providing valuable guidance for refining the IGM parameters or improving the theoretical prescriptions used to describe reionization-era environments.

These observed discrepancies may arise from limitations in our current modeling framework. For instance, the UVB fluctuations incorporated in our simulations were calibrated using the large-scale Lyα forest flux distribution under the simplifying assumptions of a spatially uniform temperature field and a fully reionized IGM. However, these assumptions may not hold at $z \sim 6$, where the Universe is still undergoing the final stages of reionization. In particular, the hydrodynamical simulations used in our radiative transfer modeling (e.g., `Nyx`) assume an IGM that is colder than what may be physically realistic at these redshifts. A hotter IGM would modify the thermal broadening and small-scale structure of absorption features, altering the shape of the transmission PDF. Furthermore, recent reionization activity could introduce significant spatial temperature fluctuations, leading to additional variance in the observed transmission not captured by our current models. These caveats highlight that deviations between the real and mock PDF are not merely statistical fluctuations, but may reflect physically meaningful shortcomings in our assumptions about the thermal and ionization state of the IGM during the EoR. Consequently, they provide a valuable opportunity to refine our understanding of early structure formation, quasar radiative feedback, and the timing and morphology of cosmic reionization.

1.5 Summary

In this work, we present a detailed analysis of the Lyα transmission PDF of 21 high-z quasars ($5.76 < z < 6.59$), drawn from the E-XQR-30 dataset. By employing a forward-modeling framework that simultaneously infers the intrinsic quasar continuum and the surrounding IGM absorption, we construct high-fidelity measurements of the Lyα transmission profiles for each quasar, and compare them against realistic mock spectra generated under a physically motivated IGM model.

We define two rest-frame wavelength intervals, $1170 - 1190$ Å (the Lyα forest) and $1190 - 1218$ Å (the proximity zone), through mock spectra simulations designed to isolate regions representing the ambient IGM from those dominated by quasar radiation. The transmission PDFs are then constructed in each interval for a set of z bins, and their agreement with theoretical expectations is assessed through a rigorous comparison using χ^2 statistics derived from the full covariance structure of the mocks.

Our main findings are as follows:

- The observed transmission PDF in the proximity zone is generally well described by the IGM model across most redshift bins. This is supported by $\mathrm{CDF}(< \chi^2_{\mathrm{real}}) \lesssim 0.5$ from the χ^2 analysis, indicating statistical consistency between data and model predictions. The only exceptions occur for $z = 6.1722 - 6.1871$ and $z = 6.5405 - 6.5872$, where we also take into account a few transmission bins poorly populated by mock PDFs.

- In contrast, the transmission PDF in the Lyα forest often shows mild to significant deviations from the model. Some redshift bins (i.e., $z = 6.0264 - 6.0754$, $z = 6.2980 - 6.3373$, and $z = 6.5405 - 6.5872$) exhibit $\mathrm{CDF}(< \chi^2_{\mathrm{real}}) \approx 1.0$, suggesting tension between the data and the assumed IGM properties in these regions.

- The χ^2 distributions derived from the mocks are well approximated by the theoretical χ^2 PDF distribution in the proximity zone but show notable deviations in the Lyα forest. This likely reflects non-Gaussian features in the transmission PDF induced by low transmission values and small sample sizes, a point explored further in Appendix 1.8.

- Our methodology robustly accounts for observational uncertainties, continuum modeling errors, and intrinsic variance in the IGM, providing a principled framework for evaluating the statistical significance of deviations from theoretical expectations.

These results highlight the utility of the transmission PDF as a sensitive probe of the ionization state and thermal structure of the IGM during the

EoR. While current models perform well in proximity-dominated regions, the discrepancies observed in the Lyα forest point to possible refinements in our understanding of the IGM. In particular, the assumptions of a cold, fully reionized IGM and a spatially homogeneous UVB may not be valid at $z \sim 6$. Realistically, a significant fraction of the IGM may still be neutral, with elevated temperatures and large-scale temperature fluctuations induced by recent, patchy reionization events. These factors can strongly affect the Lyα transmission profile and may drive the mismatches we observe.

Building on this foundation, future efforts can explore a broader range of reionization models, including alternative thermal histories, inhomogeneous UVB fluctuations, and improved prescriptions for the timing and patchiness of reionization. By doing so, we can aim to better capture the complex astrophysics shaping the IGM in the final stages of the EoR.

Acknowledgements

We acknowledge D'Odorico et al. (2023) for their public data release, which made this work possible. We acknowledge helpful conversations with the ENIGMA group at UC Santa Barbara and Leiden University. SO is grateful to Elia Pizzati for comments on an early version of the manuscript. JFH acknowledges support from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No 885301), from the National Science Foundation (NSF) under Grant No. 2307180, and from NASA under the Astrophysics Data Analysis Program (ADAP, Grant No. 80NSSC21K1568).

Data Availability

The data and code underlying this paper will be shared on reasonable request to the corresponding author.

1.6 Appendix: Remaining quasar spectra and continua

Due to the large sample size, for visual choices, we report here the remaining quasar spectra, along with their median continuum model prediction (see Fig. 1.6). The description of how we obtain them is provided in Section 1.3.1.

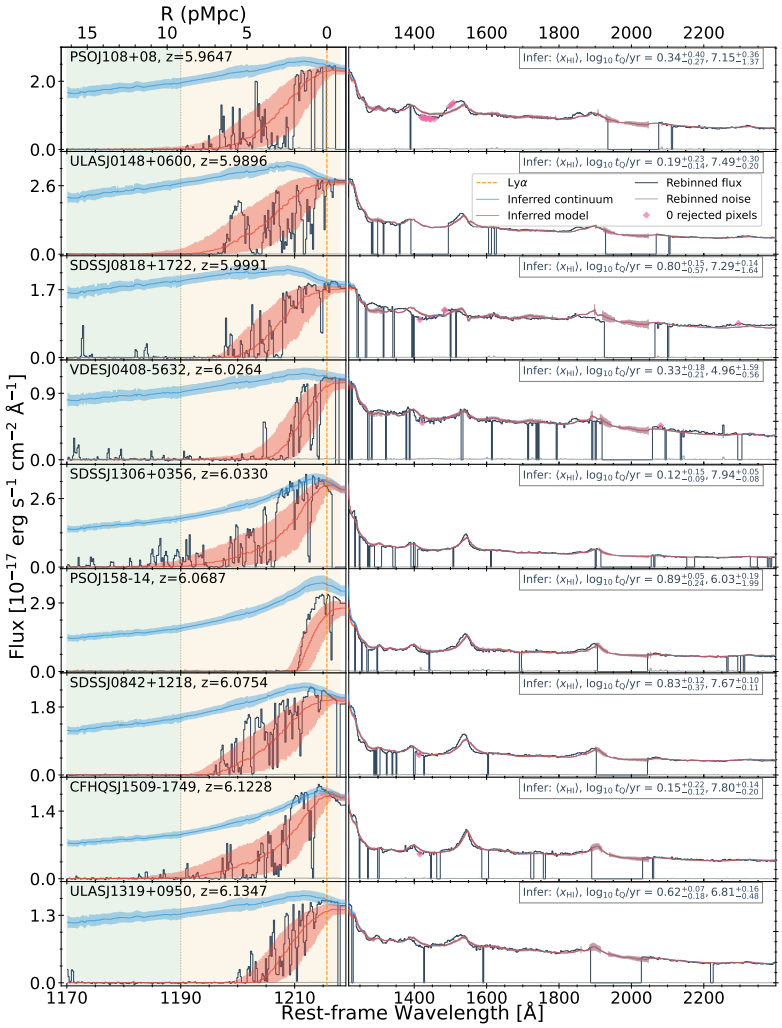


Figure 1.6: Quasars in the remaining z bins, showing the the spectra rebinned into the simulation's wavelength grid at 50 km s^{-1} for $\lambda_{\rm rest} < 1218$ Å and 500 km s^{-1} for $\lambda_{\rm rest} > 1218$ Å (black), their associated noise vector (grey), their median flux model with 16$^{\rm th}$ and 84$^{\rm th}$ percentiles (red), and their median continuum model with 1σ uncertainties computed as 16$^{\rm th}$ and 84$^{\rm th}$ percentiles of the distribution of model realizations (blue). In every sub-panel, we report the name and the redshift of the quasar on the left, and the inferred astrophysical parameters $(\langle x_{\rm HI}\rangle, \log_{10} t_{\rm Q}/{\rm yr})$ on the right. The pink diamonds are the pixels rejected from sigma-clipping. The orange dashed line marks the position of the Lyα line. The green and gold shaded regions highlight the two distinct wavelength ranges ($1170-1190$ Å and $1190-1218$ Å, respectively) in which we study the Lyα transmission PDF (see Section 1.3.3). On the top left x-axis, we show the distance from the quasar (pMpc) converted from the rest-frame wavelengths using the systemic redshift of the quasars. The spectra and continuum predictions for the sources in the lowest-z bin are shown in Fig. 1.2.

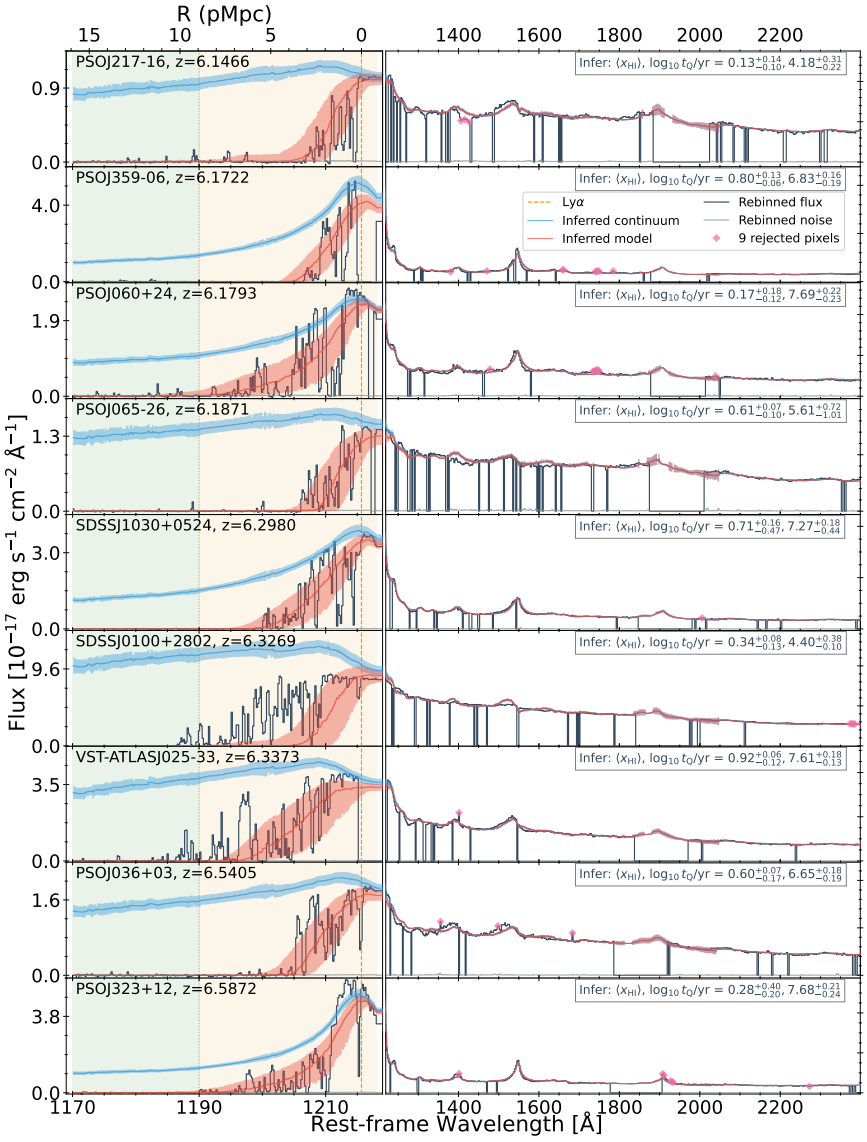


Figure 1.6: (Continued)

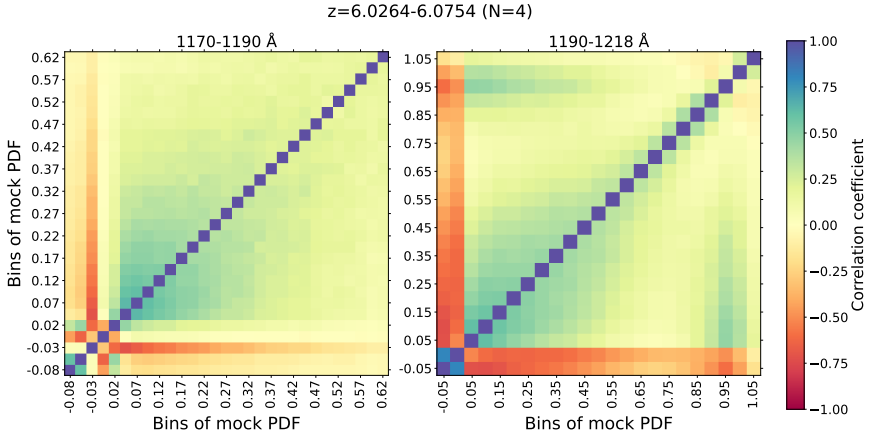


Figure 1.7: Correlation matrices $\boldsymbol{C}$ of the mock transmission PDFs in the redshift bin $z = 6.0264 - 6.0754$ (see Section 1.4.2), highlighting correlations between the k *non-zero* bins for the Lyα forest ($1170 - 1190$ Å, left-hand panel), and the proximity zone ($1190 - 1218$ Å, right-hand panel).

1.7 Appendix: Correlation matrices example

In Fig. 1.7, we show an example case for the correlation matrices in the redshift bin $z = 6.0264 - 6.0754$, computed as described in Section 1.4.2.

1.8 Appendix: Non-Gaussian mock PDF distribution

In Section 1.4.3, we observe that the $\chi^2_{\rm mock}$ distribution computed in the proximity zone ($1190 - 1218$ Å) closely follows the expected theoretical χ^2 PDF with $\nu = k - 1$ dof. However, this behavior is not replicated in the Lyα forest ($1170 - 1190$ Å), where significant deviations from Gaussianity emerge. In this section, we investigate the origin of this discrepancy and examine whether it is reasonable to assume that the $\chi^2_{\rm mock}$ statistic should follow a Gaussian distribution.

Following the approach of the right-hand panel of Fig. 8 in Hennawi et al. (2025), we test the assumption that the transmission PDF measurements, modeled by $\boldsymbol{p}_{{\rm mock},i}$, follow a multivariate Gaussian distribution. To do this, we define the residual vector:

$$\boldsymbol{\chi} = \mathbf{B}_{\rm reg}^{-1} \left(\boldsymbol{p}_{\rm mock,i} - \langle \boldsymbol{p}_{\rm mock} \rangle \right) , \tag{1.13}$$

where $\langle \boldsymbol{p}_{\rm mock} \rangle$ is the mean mock transmission PDF across all samples (see

Section 1.4.2), and $\boldsymbol{B}_{\rm reg}$ is the regularized square root of the covariance matrix $\boldsymbol{\Sigma}$. In the same way as detailed in Section 1.4.3, we regularize the eigenvalue decomposition of $\boldsymbol{\Sigma}$ to ensure numerical stability, defining $\boldsymbol{B}_{\rm reg} = \boldsymbol{V} \cdot \mathrm{diag}(\sqrt{\max(\lambda_i, \varepsilon)})$, where $\boldsymbol{V}$ and λ_i are the eigenvectors and eigenvalues of $\boldsymbol{\Sigma}$, and ε is the regularization threshold.

By construction, the transformed residuals $\boldsymbol{\chi}$ are uncorrelated and have unit variance. If the distribution of $\boldsymbol{p}_{{\rm mock},i}$ is truly multivariate Gaussian, then the elements of $\boldsymbol{\chi}$ should follow a standard normal distribution, i.e., $\mathcal{N}(0, 1)$. In Fig. 1.8, we compare the histogram of $\boldsymbol{\chi}$ to a standard Gaussian for the example case of the redshift bin $z = 6.0264 - 6.0754$, in which we do not observe a particularly good agreement between the PDFs in any of the wavelength intervals. We only report this case, but we see a similar behavior in all redshift bins. In the proximity zone, the match is usually excellent, while in the Lyα forest, the distribution deviates significantly from Gaussianity. They can both show an excess near zero, but it is always more pronounced in the Lyα forest.

This deviation originates from the sparsity of data in the selected transmission bins. With limited sampling, the statistical fluctuations in individual bins are no longer well-approximated by the central limit theorem, and the resulting PDF histogram becomes skewed or peaked. Despite this, the χ^2 statistic remains a useful approximate summary metric. However, the theoretical χ^2 PDF with ν dof becomes unreliable as a reference distribution, particularly because insufficient averaging causes the PDF distribution to retain Poisson-like features.

To investigate this further, we also plot the distribution of deviations from the mean, i.e., the set of differences $\boldsymbol{p}_{{\rm mock},i} - \langle \boldsymbol{p}_{\rm mock} \rangle$, for each transmission bin, in the same redshift bin $z = 6.0264 - 6.0754$. The stacked histograms shown in Fig. 1.9 illustrate that a small number of bins dominate the contribution to the total residual, and that many bins are sparsely populated, leading to the observed spike near zero.

This behavior is strongly dependent on the number and quality of the selected transmission bins identified in Section 1.4.1. By applying more stringent selection criteria, e.g., requiring higher mock sample significance, we can reduce the number of bins and achieve a more Gaussian-like residual distribution. However, this introduces a tradeoff: fewer bins result in more averaging (and thus a better-behaved Gaussian), but at the cost of losing information about the detailed shape of the transmission PDF. At the same time, overly coarse binning may obscure key features of the IGM transmission.

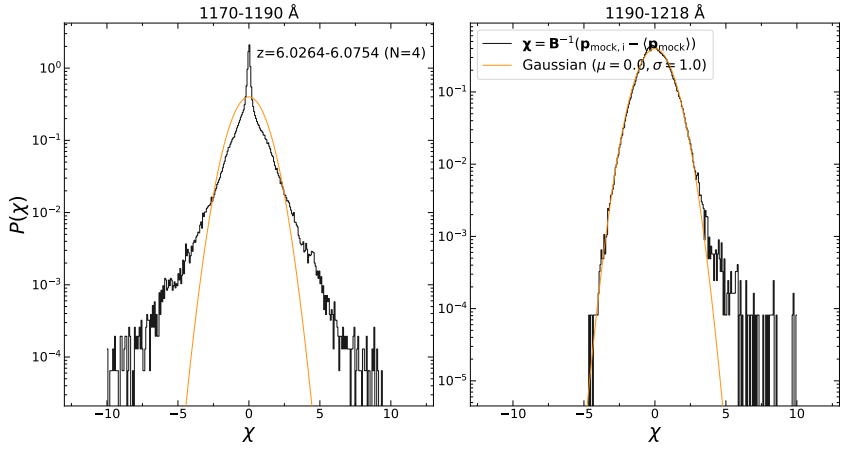


Figure 1.8: Histogram of the transformed residuals $\boldsymbol{\chi}$ for mock transmission PDF (see Appendix 1.8), computed using the regularized covariance matrix $\boldsymbol{\Sigma}$, in the redshift bin $z = 6.0264 - 6.0754$. The black curve shows the distribution of $\boldsymbol{\chi}$ for the Lyα forest ($1170 - 1190$ Å, left-hand panel), and the proximity zone ($1190 - 1218$ Å, right-hand panel), compared to a standard normal distribution (orange curve). While the proximity zone residuals are well described by a Gaussian, the Lyα forest distribution deviates significantly, indicating non-Gaussian behavior in the mock ensemble. In the Lyα forest, there is a noticeable excess around zero.

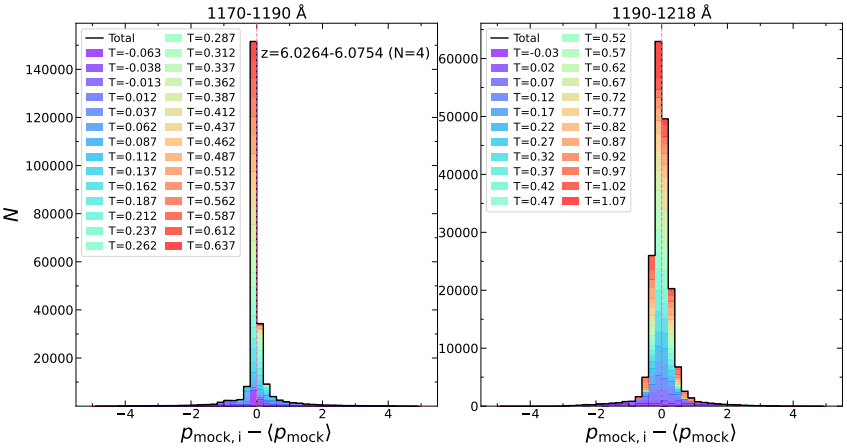


Figure 1.9: Stacked histograms of the deviations between each mock transmission PDF and the mean PDF ($\boldsymbol{p}_{\mathrm{mock},i} - \langle \boldsymbol{p}_{\mathrm{mock}} \rangle$) for all retained transmission bins (colored segments), in the redshift bin $z = 6.0264 - 6.0754$. The left-hand panel shows the Lyα forest ($1170 - 1190$ Å), while the right-hand panel shows the proximity zone ($1190 - 1218$ Å). The total histogram is shown as a black stepped line. A strong spike near zero emerges due to the small number of mocks contributing significantly in each bin. This highlights how insufficient bin averaging leads to deviations from Gaussianity and can result in a breakdown of the theoretical χ^2 PDF assumptions.

Bibliography

Ahn C. P., et al., 2012, ApJS, 203, 21

Almgren A. S., Bell J. B., Lijewski M. J., Lukić Z., Van Andel E., 2013, ApJ, 765, 39

Bañados E., et al., 2016, ApJS, 227, 11

Bañados E., et al., 2023, ApJS, 265, 29

Becker G. D., Bolton J. S., Madau P., Pettini M., Ryan-Weber E. V., Venemans B. P., 2015, MNRAS, 447, 3402

Bolton J. S., Haehnelt M. G., 2007a, MNRAS, 374, 493

Bolton J. S., Haehnelt M. G., 2007b, MNRAS, 381, L35

Bosman S. E. I., et al., 2022, MNRAS, 514, 55

Carilli C. L., et al., 2007, ApJ, 666, L9

Carnall A. C., et al., 2015, MNRAS, 451, L16

Cen R., Haiman Z., 2000, ApJ, 542, L75

Chehade B., et al., 2018, MNRAS, 478, 1649

D'Odorico V., et al., 2023, MNRAS, 523, 1399

Davies F. B., Furlanetto S. R., 2022, MNRAS, 514, 1302

Davies F. B., et al., 2018a, ApJ, 864, 142

Davies F. B., et al., 2018b, ApJ, 864, 143

Davies F. B., Hennawi J. F., Eilers A.-C., 2020, MNRAS, 493, 1330

Davies R. L., et al., 2023, MNRAS, 521, 289

Decarli R., et al., 2018, ApJ, 854, 97

Duane S., Kennedy A. D., Pendleton B. J., Roweth D., 1987, Physics Letters B, 195, 216

Eilers A.-C., Davies F. B., Hennawi J. F., Prochaska J. X., Lukić Z., Mazzucchelli C., 2017, ApJ, 840, 24

Eilers A.-C., Davies F. B., Hennawi J. F., 2018, ApJ, 864, 53

Eilers A.-C., et al., 2020, ApJ, 900, 37

Eilers A.-C., et al., 2021, ApJ, 914, 74

Eisenstein D. J., et al., 2011, AJ, 142, 72

Fan X., et al., 2001, AJ, 122, 2833

Fan X., et al., 2006a, AJ, 131, 1203

Fan X., et al., 2006b, AJ, 132, 117

Greig B., Mesinger A., Davies F. B., Wang F., Yang J., Hennawi J. F., 2022, MNRAS, 512, 5390

Greig B., et al., 2024, MNRAS, 530, 3208

Gunn J. E., Peterson B. A., 1965, ApJ, 142, 1633

Haiman Z., Cen R., 2001, in Umemura M., Susa H., eds, Astronomical Society of the Pacific Conference Series Vol. 222, The Physics of Galaxy Formation. p. 101

Hennawi J. F., Kist T., Davies F. B., Tamanas J., 2025, MNRAS, 539, 2621

Jenkins E. B., Ostriker J. P., 1991, ApJ, 376, 33

Jiang L., McGreer I. D., Fan X., Bian F., Cai Z., Clément B., Wang R., Fan Z., 2015, AJ, 149, 188
Kashino D., Lilly S. J., Matthee J., Eilers A.-C., Mackenzie R., Bordoloi R., Simcoe R. A., 2023, ApJ, 950, 66
Kist T., Hennawi J. F., Davies F. B., 2025a, arXiv e-prints, p. arXiv:2508.21812
Kist T., Hennawi J. F., Davies F. B., 2025b, MNRAS, 538, 2704
Lee K.-G., et al., 2015, ApJ, 799, 196
Lukić Z., Stark C. W., Nugent P., White M., Meiksin A. A., Almgren A., 2015, MNRAS, 446, 3697
Mazzucchelli C., et al., 2017, ApJ, 849, 91
McGreer I. D., Mesinger A., D'Odorico V., 2015, MNRAS, 447, 499
Mesinger A., Furlanetto S., Cen R., 2011, MNRAS, 411, 955
Miralda-Escudé J., 1998, ApJ, 501, 15
Mortlock D. J., et al., 2009, A&A, 505, 97
Onorato S., Hennawi J. F., Pizzati E., Venemans B. P., Eilers A.-C., 2025a, arXiv e-prints, p. arXiv:2505.09676
Onorato S., et al., 2025b, MNRAS, 540, 1308
Pâris I., et al., 2011, A&A, 530, A50
Planck Collaboration et al., 2020, A&A, 641, A6
Reed S. L., et al., 2017, MNRAS, 468, 4702
Satyavolu S., et al., 2023, MNRAS, 522, 4918
Suzuki N., Tytler D., Kirkman D., O'Meara J. M., Lubin D., 2005, The Astrophysical Journal, 618, 592
Venemans B. P., et al., 2015, ApJ, 801, L11
Venemans B. P., et al., 2020, ApJ, 904, 130
Vernet J., et al., 2011, A&A, 536, A105
Wang F., et al., 2023, ApJ, 951, L4
Willott C. J., et al., 2007, AJ, 134, 2435
Wu X.-B., et al., 2015, Nature, 518, 512
Yang J., et al., 2020, ApJ, 904, 26
Yang J., et al., 2023, ApJ, 951, L5
Ďurovčíková D., et al., 2024, arXiv e-prints, p. arXiv:2401.10328