Title

Measuring the thermal and ionization state of the low-z IGM using likelihood free inference

Journal

Monthly Notices of the Royal Astronomical Society, 515(2)

ISSN

0035-8711

Authors

Hu, Teng

Khaire, Vikram

Hennawi, Joseph F

et al.

Publication Date

2022-07-28

DOI

10.1093/mnras/stac1865

MNRAS 000, 1–19 (2021) Preprint 18 July 2022 Compiled using MNRAS LATEX style file v3.0

Measuring the thermal and ionization state of the low-z IGM using likelihood free inference

Teng Hu,[1]★ Vikram Khaire[1,2], Joseph F. Hennawi[1,3], Michael Walther[1,4], Hector Hiss[5], Justin Alsing[6,7], Jose Oñorbe[8], Zarija Lukic[9] and Frederick Davies[5]

[1]*Physics Department, Broida Hall, University of California Santa Barbara, Santa Barbara, CA 93106-9530, USA*
[2] *Indian Institute of Space Science & Technology, Thiruvananthapuram, Kerala - 695547, INDIA*
[3]*Leiden Observatory, Leiden University, PO Box 9513, NL-2300 RA Leiden, the Netherlands*
[4]*University Observatory, Faculty of Physics, Ludwig-Maximilians-Universität München, Scheinerstr. 1, 81679 Munich, Germany*
[5]*Max-Planck-Institut für Astronomie, Königstuhl 17,69117 Heidelberg, Germany*
[6]*Oskar Klein Centre for Cosmoparticle Physics, Department of Physics, Stockholm University, Stockholm SE-106 91, Sweden*
[7]*Imperial Centre for Inference and Cosmology, Department of Physics, Imperial College London, Blackett Laboratory,Prince Consort Road, London SW7 2AZ, UK*
[8]*Facultad de Física, Universidad de Sevilla, Avda. Reina Mercedes s/n, Campus de Reina Mercedes, E-41012 Sevilla, Spain*
[9]*Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA*

Accepted XXX. Received YYY; in original form ZZZ

ABSTRACT

We present a new approach to measure the power-law temperature density relationship $T = T_0(\rho/\bar{\rho})^{\gamma-1}$ and the UV background photoionization rate $\Gamma_{\rm H\,I}$ of the intergalactic medium (IGM) based on the Voigt profile decomposition of the Lyα forest into a set of discrete absorption lines with Doppler parameter b and the neutral hydrogen column density $N_{\rm HI}$. Previous work demonstrated that the shape of the b-$N_{\rm H\,I}$ distribution is sensitive to the IGM thermal parameters T_0 and γ, whereas our new inference algorithm also takes into account the normalization of the distribution, i.e. the line-density dN/dz, and we demonstrate that precise constraints can also be obtained on $\Gamma_{\rm H\,I}$. We use density-estimation likelihood-free inference (DELFI) to emulate the dependence of the b-$N_{\rm H\,I}$ distribution on IGM parameters trained on an ensemble of 624 Nyx hydrodynamical simulations at $z = 0.1$, which we combine with a Gaussian process emulator of the normalization. To demonstrate the efficacy of this approach, we generate hundreds of realizations of realistic mock HST/COS datasets, each comprising 34 quasar sightlines, and forward model the noise and resolution to match the real data. We use this large ensemble of mocks to extensively test our inference and empirically demonstrate that our posterior distributions are robust. Our analysis shows that by applying our new approach to existing Lyα forest spectra at $z \simeq 0.1$, one can measure the thermal and ionization state of the IGM with very high precision ($\sigma_{\log T_0} \sim 0.08$ dex, $\sigma_\gamma \sim 0.06$, and $\sigma_{\log \Gamma_{\rm H\,I}} \sim 0.07$ dex).

Key words: intergalactic medium – method: statistical – quasars: absorption lines

1 INTRODUCTION

The intergalactic medium (IGM) is the largest reservoir of baryons in the Universe, which plays an essential role in its evolution and structure formation. Current theoretical models, supported by many observations, predict two major phase transition events that dominate the thermal evolution of the IGM. The first one is the reionization of hydrogen by the first galaxies at redshift $6 < z < 20$ (Madau et al. 1998; Faucher-Giguère et al. 2008; Robertson et al. 2015; McGreer et al. 2015; Fan et al. 2006). The second phase transition is the double reionization of Helium (He II→He III) driven by quasi-stellar objects (QSO)s (see e.g. Madau & Meiksin 1994; Miralda-Escudé et al. 2000; McQuinn et al. 2009; Dixon & Furlanetto 2009; Syphers & Shull 2014), which is expected to happen at $z \sim 3$, where the quasar luminosity density peaks (see e.g. Worseck et al. 2011; Khaire 2017; Worseck et al. 2018; Kulkarni et al. 2019). These two events change the ionization state of the IGM dramatically and heat it to temperatures as high as 15,000K.

After hydrogen reionization ($z < 6$), the thermal state of the IGM is determined by the balance between photoionization heating from the extragalactic UV background and various cooling processes such as cooling due to Hubble expansion, recombinations, and the excitation and inverse Compton scattering of electrons from the cosmic microwave background (CMB). As a result of these processes, the IGM is expected to follow a tight temperature-density relation:

$$T(\Delta) = T_0 \Delta^{\gamma-1}, \quad (1)$$

where $\Delta = \rho/\bar{\rho}$ is the overdensity, T_0 is the temperature at mean density, and γ is the adiabatic index (Hui & Gnedin 1997; McQuinn & Upton Sanderbeck 2016), and these two parameters characterize the thermal state of the IGM. By measuring T_0 and γ at different epochs, we are thus able to constrain the IGM thermal history (Miralda-Escudé & Rees 1994; Hui & Haiman 2003), improving our

★ E-mail: tenghu@ucsb.edu (UCSB)

knowledge of the evolution of the IGM and our understanding of the relevant heating and cooling processes responsible.

The thermal state of the IGM is encoded in the Lyα forest, a swath of Lyα absorption lines originating from a trace amount of neutral hydrogen gas in the IGM (Gunn & Peterson 1965; Lynds 1971). The Lyα forest is thus used as the premier probe of the IGM thermal history. Various statistical properties of the Lyα forest are used to measure the IGM thermal state, including the power spectrum (Theuns et al. 2000; Zaldarriaga et al. 2001; McDonald et al. 2001; Walther et al. 2017; Walther et al. 2018; Khaire et al. 2019; Gaikwad et al. 2021), the flux probability density function (PDF) (Bolton et al. 2008; Viel et al. 2009; Lee et al. 2015), the transmission curvature (Becker et al. 2011; Boera et al. 2014), the wavelet decomposition of the forest (Theuns & Zaroubi 2000; Theuns et al. 2002; Lidz et al. 2010; Garzilli et al. 2012; Wolfson et al. 2021), and the quasar pair phase angle distribution (Rorai et al. 2013, 2017). These measurements are typically performed using Lyα forest spectra from ground-based telescopes at $z > 1.6$, where the Lyα transition lies above the atmospheric cutoff ($\lambda \sim 3300$Å), explaining why there are currently very few measurements of the IGM thermal state at redshift below such limit (i.e. $z < 1.6$), which is, however, an essential epoch for galaxy formation. By far the only available direct measurements at redshift $z < 1.6$ is reported by Ricotti et al. (2000) at $z \sim 0$, which was done two decades ago using only 43 Ly-α absorption lines from HST Goddard High Resolution Spectrograph data, suggesting a need for new and precise measurements at redshift $z \sim 0$.

Long after the helium reionization ($z < 3$), the thermal state of the IGM is expected to be dominated by adiabatic cooling from Hubble expansion, where theoretical models and simulations predict such cooling leads to an IGM thermal state with $T_0 \sim 5000$K and $\gamma \sim 1.6$ at the current epoch $z = 0$ (McQuinn & Upton Sanderbeck 2016). However, to date, this predicted cooling to low temperatures has not been verified observationally. Moreover, recent studies based on the low-z Lyα forest dataset (Danforth et al. 2016) show that these lines appear broader (i.e. have larger b parameter) than numerical model predictions (Gaikwad et al. 2017; Viel et al. 2017; Nasir et al. 2017). While it has been speculated that such a discrepancy might be resolved by an additional source of turbulence (Bolton et al. 2021), an alternative explanation would be that there are additional sources of heating, and the IGM is actually hotter than expected, with T_0 conceivably approaching 10000K.

If true, such unexpected heating would change our understanding of IGM physics drastically, highlighting a severe need to investigate processes that are possibly responsible for it, such as dark matter annihilation (Araya & Padilla 2014), gamma ray sources (Puchwein et al. 2012), or feedback from galaxy formation, whose effects are not fully understood in low-z (see Springel et al. 2005; Croton et al. 2006; Sijacki et al. 2007; Hopkins et al. 2008). To this end, precise measurements of the thermal state at low-z are needed to determine whether the IGM cools down as predicted.

In this work, we follow the method for measuring the IGM thermal state based on Voigt profile decomposition of the Lyα forest (Schaye et al. 1999; Ricotti et al. 2000; McDonald et al. 2001). In this approach, a transmission spectrum is treated as a superposition of multiple discrete Voigt profiles, with each line described by three parameters: redshift $z_{\rm abs}$, Doppler broadening b, and neutral hydrogen column density $N_{\rm HI}$. By studying the statistical properties of these parameters, i.e. the b-$N_{\rm H\,I}$ distribution, one can recover the thermal information encoded in the absorption profiles. The majority of past applications of this method constrained the IGM thermal state by fitting the low-b-$N_{\rm HI}$ cutoff of the b-$N_{\rm H\,I}$ distribution (Schaye et al. 1999, 2000; Ricotti et al. 2000; McDonald et al. 2001; Rudie et al. 2012; Bolton et al. 2014; Boera et al. 2014; Garzilli et al. 2015, 2018; Rorai et al. 2018; Hiss et al. 2018). The motivation for this approach is that the Lyα lines are broadened by both thermal motion and non-thermal broadening resulting from combinations of Hubble flow, peculiar velocities and turbulence. By isolating the narrow lines in the Lyα forest that constitutes the lower-cutoff in b-$N_{\rm H\,I}$ distributions, of which the line-of-sight component of non-thermal broadening is expected to be zero, the broadening should be purely thermal, thus allowing one to constrain the IGM thermal state. However, this method has three crucial drawbacks. First, the IGM thermal state actually impacts all the lines besides just the narrowest lines. Therefore, by restricting attention to data in the distribution outskirts, this approach throws away information and reduces the sensitivity to the IGM thermal state significantly(Rorai et al. 2018; Hiss et al. 2019). Second, in practice, determining the location of the cutoff is vulnerable to systematic effects, such as contamination from the narrow metal lines (Rorai et al. 2018; Hiss et al. 2018). Lastly, the results from this approach critically depend on the choice of low-b cutoff fitting techniques, where different techniques might result in inconsistent T_0 and γ measurements (Rorai et al. 2018; Hiss et al. 2018).

To overcome these limitations, Hiss et al. (2019) developed a new approach to measure the IGM thermal state from the full b-$N_{\rm H\,I}$ distribution based on density estimation and Bayesian analysis. We further advance the b-$N_{\rm H\,I}$ distribution emulation by employing a novel density estimation technique based on machine learning, namely Density-Estimation Likelihood-Free Inference (DELFI) (see Papamakarios & Murray 2016; Alsing et al. 2018; Papamakarios et al. 2018; Lueckmann et al. 2018; Alsing et al. 2019). In addition, we augment the likelihood function to take into account the absorber number density dN/dz, making our improved method far more sensitive to the photoionization rate of hydrogen $\Gamma_{\rm H\,I}$ sourced by the UV background.

In this work, we introduce our new method, demonstrate its robustness, and perform an analysis using realistic mock datasets to illustrate the sensitivity to IGM parameters. Our inference is based on a suite of cosmological hydrodynamic simulations with different thermal parameters at redshift $z \sim 0.1$. While this method can be applied to the Lyα forest at any redshift where the opacity is low enough to make it amenable to Voigt profile decomposition (e.g. $z \lesssim 3.4$, see Hiss et al. 2018), we choose to focus on $z \sim 0.1$ because we want to quantify the sensitivity of archival *Hubble Space Telescope* spectra, so as to perform the first measurements of the IGM thermal state at $z < 1.6$ in future work. Such a measurement would directly test the prediction that the IGM cools down at low-z, which has been challenged by recent observations. To this end, we run a set of cosmological hydrodynamic simulations with different thermal parameters at redshift $z \sim 0.1$, from which we create mock datasets with the same properties as the Danforth et al. (2016) low redshift Lyα forest dataset observed with the *Cosmic Origins Spectrograph* (COS, Green et al. 2012) on the *Hubble Space Telescope* (HST). We demonstrate that our method applied to such a dataset can reliably and accurately determine the thermal state of the IGM.

This paper is structured as follows. In §2 we introduce our hydrodynamic simulations, parameter grid, and data processing procedures, which include generating Lyα forest from simulation, forward-modeling and our method to fit Voigt profiles (`VPFIT`). In §3 we present our inference algorithm, including likelihood, emulators, inference results, and a set of inference tests. Finally, we discuss these results and summarize the highlights of this study in §4. Throughout this paper, we write log in place of $\log_{10}$. Cosmology parameters used in this study ($\Omega_m = 0.319181, \Omega_b h^2 = 0.022312, h =$

$0.670386, n_s = 0.96, \sigma_8 = 0.8288$) are taken from Planck Collaboration et al. (2014) .

2 SIMULATIONS

A set of Nyx cosmological hydrodynamic simulations (see Lukić et al. 2015; Almgren et al. 2013) is used to model the low-redshift IGM. Nyx is a massively-parallel, cosmological simulation code primarily developed to simulate the IGM. [1] In Nyx simulations, the evolution of dark matter is traced by treating dark matter as self gravitating Lagrangian particles, while baryons are modeled as as ideal gas on a uniform Cartesian grid following an Eulerian approach. The Eulerian gas dynamics equations are solved following a second-order piece-wise parabolic method (PPM), which captures shock waves accurately.

Nyx includes the main physical processes relevant for modeling the Lyα forest. First of all, gas in the Nyx is assumed to have a primordial composition with a hydrogen mass fraction of 0.76, and helium mass fraction of 0.24 and zero metallicity. The recombination, collisional ionization, dielectric recombination, and cooling are implemented based on prescriptions given in Lukić et al. (2015). Nyx keeps track of the net loss of thermal energy resulting from atomic collisional processes and takes into account the inverse Compton cooling off the microwave background. Ionizing radiation is modeled by a spatially homogeneous but time-varying ultraviolet background radiation field (from Haardt & Madau 2012) that changes with redshift, while assuming all cells in the simulation are optically thin. We later make the UV background a free parameter for generating Lyα forest in post-processing (See §2.2). Since Nyx simulations are developed mainly to study the IGM, no feedback or galaxy formation processes are included, significantly reducing the computational requirement allowing us to run a large ensemble of simulations varying the thermal parameters (see 3.3).

Each Nyx simulation used in this study is initialized at $z = 159$ and evolves down to $z = 0.03$ in a $L_{\rm box} = 20$ cMpc$/h$ simulation domain, using $N_{\rm cell} = 1024^3$ Eulerian cells and 1024^3 dark matter particles. The box size is chosen as the best compromise between computational cost and the need to be converged at least to $< 10\%$ on small scales (large k). More discussion about resolutions and box sizes can be found in Lukić et al. (2015). We also performed box size convergence tests at low redshift as explained in appendix D.

2.1 Thermal parameters and simulation grid

To model the IGM with different thermal states, we use part of the publicly available Thermal History and Evolution in Reionization Models of Absorption Lines (THERMAL)[2] suite of Nyx simulations (see Hiss et al. 2018; Walther et al. 2019). We make use of in total 48 models with different thermal histories, and for each model, we generate a simulation snapshot at $z = 0.1$, from which we measure the thermal state [$\log T_0$,γ]. The thermal grid is illustrated in the left panel of Fig.1, which shows that $\log(T_0/\mathrm{K})$ spans from 3.2 to 3.95, and γ ranges from 0.86 to 2.41. Here different thermal histories are achieved by artificially changing the photoheating rates (ϵ) following the method presented in Becker et al. (2011). In this method, ϵ is treated as a function of overdensity, i.e.

$$\epsilon = \epsilon_{\rm HM12} A \Delta^B, \tag{2}$$

where $\epsilon_{\rm HM12}$ represents the photoheating rate per ion tabulated in Haardt & Madau (2012), and A and B are parameters that are varied to obtain different thermal histories. It is noteworthy that the thermal state tends to converge towards low redshifts due to the cooling and other physical processes in the evolution, and it is therefore difficult to generate models with a uniform grid of T_0 and γ (for more details, see Walther et al. 2019). Moreover, it is especially challenging to generate models with low $T_0 (< 10^{3.5}$ K) and high $\gamma (> 1.9)$ at low-z, because when one reduces the photoheating rates to obtain lower T_0, the cooling rate from Hubble expansion dominates, and γ asymptotically approaches values near 1.6 (see McQuinn & Upton Sanderbeck 2016). As a result, the T_0-γ grid has an irregular shape, and there are no models in the high γ low T_0 regions. In addition, such an irregular T_0-γ grid is also a result of the original grid of the THERMAL suite, which is driven by the high-z thermal state analysis in Walther et al. (2019) that obtains relatively high temperatures.

To measure the thermal state for each of the 48 models, we fit temperature-density (T-Δ) relation (see Eq. 1) to the temperatures and densities in the simulation domain. While fitting the T-Δ relationship, we noticed broader distributions of the IGM temperatures in low redshift ($z < 0.5$) compared to high redshift ($z > 3$). Examples of low-z IGM temperature-density distributions are illustrated in Fig. 2, where we show 2D histograms of T-Δ of gas in each cell for two of our simulations on the THERMAL grid at $z = 0.1$. The gas cells are divided into four phases depending on the temperature and density, namely the Warm Hot Intergalactic Medium (WHIM), Diffuse Lyα, Hot Halo gas, and Condensed [3]. The density-weighted gas phase fractions are shown in the legends of the figure, where the diffuse Lyα phase representing the densities and temperatures probed by the Lyα forest occupies about 40% of the total gas mass, while this percentage can be up to about 80% or higher at high-z. Therefore at high-z most of the gas lies on or around the T-Δ power-law relation. Whereas the high-temperature low-density WHIM phase is negligible at high-z, it appears significantly at low-z, resulting in puffy-looking gas distribution around the T-Δ power-law (see Fig. 2), which makes T-Δ power-law fitting non-trivial at low-z.

We address this issue by implementing an improved fitting procedure following Villasenor et al. (2021). First, we extract the temperature T and the overdensity Δ for each cell of a simulation and then select gas with $-1.5 < \log\Delta < 0$ and $T < 10^5 K$ to avoid regions significantly deviant from the expected power-law T-Δ relationship. Afterward, we divide the selected region into 15 equal-width bins in $\log\Delta$, where the overdensity $\log\Delta_i$ for each bin i is given by the median value of overdensity in the bin. Here we define the bin temperature $\log T_i$ to be the maximum of the marginal temperature distribution $P(\log T \mid \log\Delta_i)$ and its effective 1-$\sigma_{T,i}$ interval to be 1/2 of the temperature range containing the 68% (16% ~ 84%) highest probability density. The temperature-density relationship Eq. (1) is then fitted using a least squares linear fit on these $(\log\Delta_i, \log T_{\rm i})$ pairs weighted by $1/{\sigma_{\rm T,i}}^2$. Examples of temperature-density relationships for two models in our thermal grid are illustrated in Fig.2. Power-law fits of the T-Δ relationship of our simulations are shown as white dashed lines while their values are given in the legends

[1] Nyx simulation is able to run with Adaptive Mesh Refinement (AMR). However, the AMR feature is not used in this work, since this work focus on the Lyα forest, which distribute nearly the entire simulation domain rather than isolated concentrations of matter where AMR is more effective.

[2] For details of THERMAL suite, see http://thermal.joseonorbe.com

[3] Here we follow the definition used in Davé et al. (2010), where the cutoffs are set to be $T = 10^5$K and $\Delta = 120$, more discussion about the different cutoff used in literature can be found in Gaikwad et al. (2017).

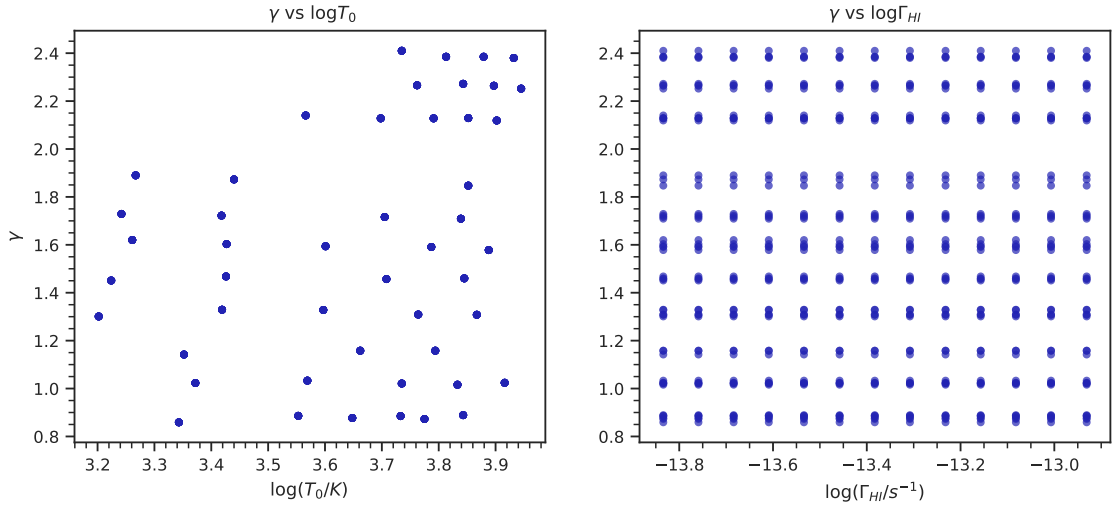

Figure 1. Thermal grid (blue circles) from snapshots of hydrodynamic simulations of the THERMAL suite at $z = 0.1$. The left-hand panel is the γ - T_0 grid, whose shape is determined by the parameters of thermal grid at and the evolution of the thermal state of the IGM. The right-hand panel is γ - $\Gamma_{\rm H\,I}$ grid, showing the thirteen $\Gamma_{\rm H\,I}$ values for each point on the 2D γ - T_0 grid.

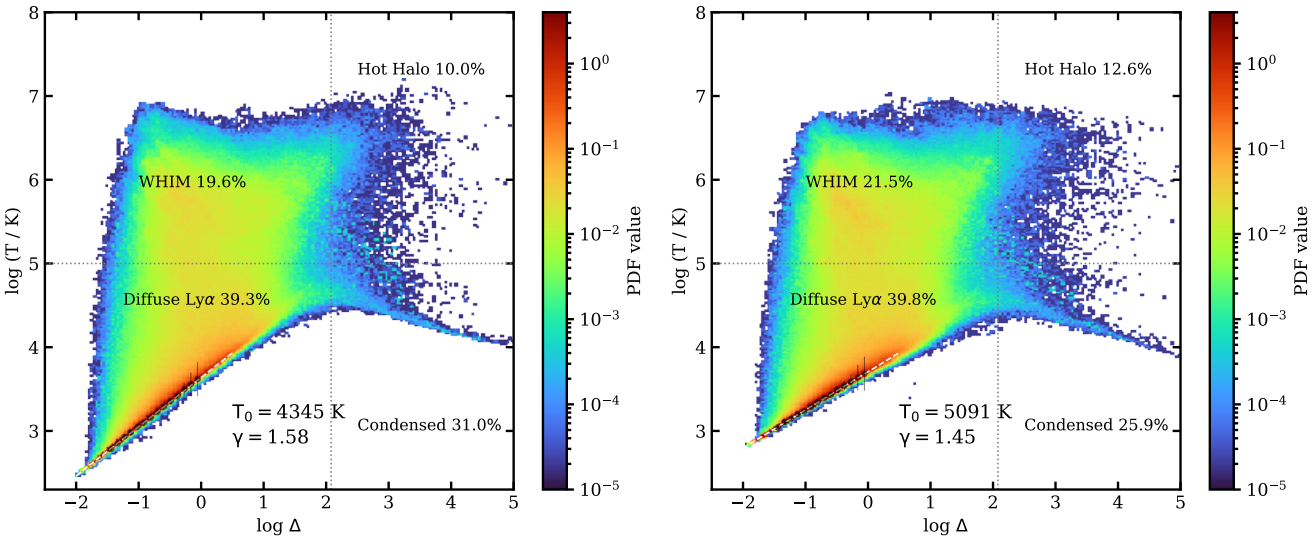

Figure 2. Temperature-density (T -Δ) distribution for the IGM gas in two different models from Nyx simulation. White dashed lines is the power-law fit to the T -Δ relation, and legends show the best fit values of T_0 and γ. Dotted T = 10^5 K lines divide the phase diagram into hot and cold region, while only cold gas is used for the fitting. Density peaks ($\log T_{\rm peak,i}$, $\log \Delta_i$) for each bin are plotted as black dot, and 1-$\sigma_{T,i}$ error bars are shown as black bars. The left-hand panel shows a model with T_0 = 4345 K and γ = 1.58, and the right-hand panel shows a model with T_0 = 5091 K and γ = 1.45. The density weighted gas phase fractions are shown in the annotation.

texts. The peak temperature in each $\log\Delta$ bin ($\log T_{\rm peak,i}$,$\log\Delta_i$) are plotted as black dots and 1-$\sigma_{T,i}$ error bars are also shown. Left panel shows a model with T_0 = 4353 K and γ = 1.58, while right panel shows another model with T_0 = 5091 K and γ = 1.45. Finally, as will be discussed later in §2.2, we let the H I photoionization rate $\Gamma_{\rm H\,I}$ be a free parameter when generating Lyα forest skewers from our simulations. As such, we add an additional parameter $\log\Gamma_{\rm H\,I}$ to our thermal grid, extending it to [$\log T_0$, γ, $\log\Gamma_{\rm H\,I}$]. The value of $\Gamma_{\rm H\,I}$ we used in this study spans from $\log(\Gamma_{\rm H\,I}/{\rm s}^{-1})$ = -13.834 to −12.932 in logarithmic steps of 0.075 dex, which gives 13 values in total (see right-hand panel of Fig.1). In total, the 3D thermal grid consists of $48 \times 13 = 624$ models.

2.2 Skewers

We generate simulated Lyα spectra by calculating the Lyman-α optical depth (τ) array along the line-of-sight, which hereafter will be referred as skewers for simplicity. For each model on the thermal grid, a set of 60,000 skewers are constructed parallel to the x, y, z axes of the simulation box (20,000 skewers in each direction). For each cell on these skewers, we extract properties needed for optical depth calculation, including temperature T, overdensity Δ, and the velocity along the line-of-sight v_z. The hydrogen neutral fraction $x_{\rm H\,I}$, which is also needed to generate the synthetic Lyα forest skewers, is calculated by assuming ionization equilibrium while considering both collisional ionization and photoionization. Here the collisional

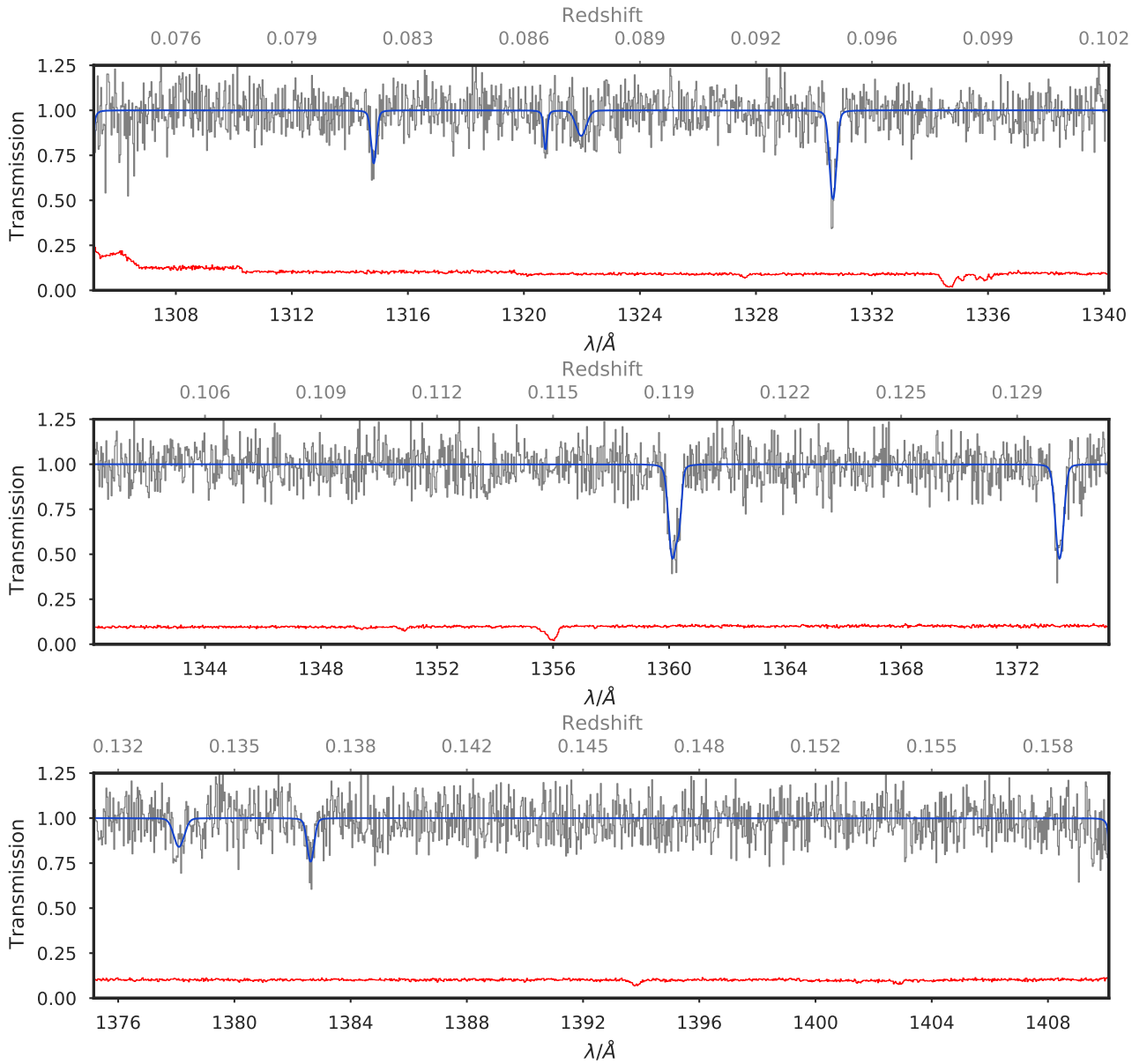

Figure 3. Illustration of one of our forward modeled spectra from Nyx simulation. The simulated raw spectrum is shown in gray, while a model spectrum based on `VPFIT` line fitting (described in § 2.4) is shown in blue and the noise vector is plotted in red. This particular spectrum is forward-modeled in order to model the instrumental effect and noise properties of one of the HST COS spectra in Danforth et al. (2016) low redshift dataset.

ionization rate is computed based on the gas temperature T. Whereas the photoionization rate $\Gamma_{\rm H\,I}$ is set to be a free parameter in the post-processing of the simulation. Since Nyx does not model radiative transfer, we approximate the self-shielding of the UV background for optically thick gas following the method given by Rahmati et al. (2013), which amounts to attenuating $\Gamma_{\rm H\,I}$ for cells containing dense gas.

Given $x_{\rm H\,I}$, T, Δ, v_z, and $\Gamma_{\rm H\,I}$, we then calculate the optical depth τ in redshift space by summing contributions from all cells in real space along the line-of-sight following the full Voigt profile approximation by Tepper-García (2006). Then $F = e^{-\tau}$ gives us continuum normalized flux of Lyα forest along with skewers. Lastly, we redo the procedure described above for each $\Gamma_{\rm H\,I}$ value to recalculate the skewers. More specifically , we do not re-scale the τ to obtain skewers for a different $\Gamma_{\rm H\,I}$, which is the standard procedure at higher redshifts. This is because, whereas the high-z IGM is predominantly photoionized, there is significantly more shock-heated WHIM gas at low-z, rendering the contribution from collisional ionization important as shown by Khaire et al. (2019) for the case of Lyα flux power spectrum. Although, it may not be essential for studying Lyα forest absorption lines, to be more precise we recalculate skewers for each value of $\Gamma_{\rm H\,I}$.

2.3 Forward Modeling of Noise and Resolution

As discussed in §1, we are interested in understanding the constraints on the IGM achievable with realistic data. To this end, we generate mock datasets with properties consistent with the Danforth et al. (2016) low redshift Lyα forest dataset, which comprises 82 unique quasar spectra with S/N $>$ 5 observed with the *Cosmic Origins*

Spectrograph (COS) on the HST. To avoid proximity regions and contamination from lower Ly-β, we use rest-frame wavelength range 1050 – 1180 Å to identify Lyα forest for each of these spectra. As a result, we select 34 of Danforth et al. (2016) quasar spectra covering the redshift range $0.06 < z < 0.16$ of our interest for the study, comprising a total redshift pathlength of $\Delta_{\rm ob} = 2.136$, which corresponds to our observational dataset for forward modelling. We choose this redsfhit bin to be the same as the redshift bin used for power spectrum calculation by Khaire et al. (2019) at $z = 0.1$ so that we can compare our future analysis with the results obtained with power spectrum measurements.

The COS has a nominal resolution $R \sim 20000$, which corresponds to roughly 15 km/s, and a non-Gaussian line spread function (LSF) exhibiting significantly broad Lorentzian wings, which could alter the shape of absorption lines on velocity scales larger than the resolution quoted above. For low-z IGM with temperatures at mean density $T_0 \sim 5000$ K, the b-values for pure thermal broadening (i.e. the narrowest lines in the Lyα forest) are about $10 \sim 20$ km/s, which means that the corresponding absorption features can not be fully resolved by COS. Thus, it is crucial to treat the instrumental effect carefully, including the peculiar shape of COS LSF. Therefore, we forward model noise and resolution to make our simulation results statistically comparable with the observation data.

In practice, we make use of tabulated COS LSF and noise vectors from Danforth et al. (2016) data. For any individual quasar spectrum from the observation dataset, we first stitch randomly selected simulated skewers without repetition to cover the same wavelength (in the rest frame 1050 – 1180 Å) of that quasar and then rebin the skewers onto the pixels of the observed spectra. Then we convolve the simulated spectra with the HST COS line spread function (LSF) while taking into account the grating and life-time positions used for that specific data spectrum. Here the COS LSF is obtained from `linetools`[4] and is tabulated for up to 160 pixels in each direction. We interpolate the LSF onto the wavelengths of the mock spectrum (segment) to obtain a wavelength dependent LSF. Each output pixel is then modeled as a convolution between the input stitched skewers and the interpolated LSF for the corresponding wavelength. Afterward, the newly generated spectrum is interpolated to the wavelength of the selected COS spectra. The noise vector of the quasar spectrum is propagated to our simulated spectrum pixel-by-pixel by sampling from a Gaussian with $\sigma = \psi_i$, with ψ_i being the data noise vector value at the i$^{\rm th}$ pixel. In the end, a fixed floor is added to the error vector for all simulated spectra to avoid an artificial effect in post-processing, which will be discussed later in §2.4.

For each model, we generated 2000 forward-modeled spectra, corresponding to a total pathlength $\Delta z_{\rm tot} \sim 125$, from the 60,000 raw skewers[5], and fit voigt profiles to each line in the spectra to obtain the $\{b, N_{\rm H\,I}\}$ pairs for our dataset (as described in section § 2.4). For the purpose of illustration, an example of a forward-modeled spectrum is shown in Fig.3 where the simulated spectrum is shown in gray, the model spectrum based on `VPFIT` line fitting (see § 2.4) is in blue, and the noise vector in red.

2.4 VPFIT

To perform the analysis based on the b-$N_{\rm H\,I}$ distributions, we have to fit the Lyα lines in our simulated spectra to obtain a set of $\{b, N_{\rm H\,I}\}$ pairs for each model. To this end, we run a line-fitting program on our forward-modeled mock spectra to obtain a set of b-$N_{\rm HI}$ pairs for each simulation model in our thermal grid. In this work, we use the line-fitting program `VPFIT`, which fits a collection of Voigt profiles convolved with the instrument LSF to spectroscopic data (Carswell & Webb 2014)[6]. Here, we employ a fully automated `VPFIT` wrapper adapted from Hiss et al. (2018), which is built on the `VPFIT` version 10.2. The wrapper routine controls `VPFIT` with the help of the `VPFIT` front-end/back-end programs `RDGEN` and `AUTOVPIN` and fit our simulated spectra automatically.

`VPFIT` identifies lines automatically and fits each line with three parameters: the absorption redshift $z_{\rm abs}$ of the line, its Doppler parameter b, and column density $N_{\rm H\,I}$. `VPFIT` obtains these parameters for a collection of lines by minimizing the χ^2 between the data and the model spectrum generated from all the fitted lines. While fitting, `VPFIT` restrict b and $N_{\rm HI}$ to $1 \leq b({\rm km/s}) \leq 300$ and $11.5 \leq \log(N_{\rm H\,I}/{\rm cm}^{-2}) \leq 18$, respectively. Our `VPFIT` wrapper allows us to fit spectra with a custom LSF[7]. Since we are working at $0.06 \leq z \leq 0.16$, the Ly α forest lies completely in the wavelength range covered exclusively by the COS G130M grating having a central wavelength 1300 Å. We fit our forward-modeled spectra with the same G130M LSF. Furthermore, the effective resolution of the grating also depends on the COS lifetime position during the observations, and they are also taken into account while running `VPFIT` as well as in forward modelling. An example of model spectrum generated by combining lines fitted using `VPFIT` is shown in Fig.3 as blue lines.

Moreover, we notice the presence of a significant number of absorption lines with very low Doppler-b parameters and low column densities $N_{\rm H\,I}$ after fitting mock as well as real data with high signal-to-noise ratios (SNR). These weak narrow absorption lines, however, are not seen in our simulated and forward-modeled spectra. Visual inspection of these lines indicates that they are spurious and introduced by `VPFIT` while attempting to fit artifacts due to flat-fielding, continuum placement, or errors in the data reduction. These lines are only introduced in spectra of the highest quality, where the extremely high signal-to-noise ratio (SNR) leads to over-fitting by `VPFIT`. To avoid this problem, a fixed floor of value 0.02 is added in quadrature to the error vector of the continuum normalized flux for all simulated spectra without adding additional noise to the normalized flux. With such a noise 'floor', these weak features are essentially removed from the `VPFIT` output. We find this floor value 0.02 via trial and error. In practice, this additional noise floor mainly removes lines with $\log(N_{\rm H\,I}/{\rm cm}^{-2}) < 12.5$ from our dataset, which is outside our limits used in likelihood calculations (which will be discussed in §3.2) and therefore not used in this study.

Furthermore, we follow the convention and apply another filter for both b and $N_{\rm H\,I}$ in this study, using only b-$N_{\rm H\,I}$ pairs in region $12.5 \leq \log(N_{\rm H\,I}/{\rm cm}^{-2}) \leq 14.5$ and $0.5 \leq \log(b/{\rm km\ s}^{-1}) \leq 2.5$ in our analysis (Schaye et al. 2000; Rudie et al. 2012; Hiss et al. 2018). Such an limitation is chosen to include the b-$N_{\rm H\,I}$ distributions for all of our Nyx models while guaranteeing that the absorbers are not strongly saturated, which maximizes the sensitivity to IGM thermal

[4] For more information, visit https://linetools.readthedocs.io

[5] For each Nyx model, 2000 spectra needs about 20,000 raw skewers, i.e, we randomly pick 20,000 skewers from 60,000.

[6] VPFIT: `http://www.ast.cam.ac.uk/~rfc/vpfit.html`

[7] Although our `VPFIT` wrapper allows us to implement an LSF in `VPFIT`, only a single LSF can be used at once, i.e. the wavelength dependence can not be taken into accout. As such, for the input into `VPFIT` we use the LSF at the lifetime of the data and evaluated it at the central wavelength of the spectrum that we are trying to fit. Such treatment is applied to both observed (mock) spectra and stimulated spectra so as to make sure our statistics are not biased.

state and minimizes the impact of poorly understood strong absorbers arising from the circumgalactic medium of galaxies.

3 INFERENCE ALGORITHM

Hiss et al. (2019) introduced a Bayesian method to estimate the IGM thermal parameters from the joint b-$N_{\rm H\,I}$ distribution. In this paper, we adopt a similar approach while employing a new method for b-$N_{\rm H\,I}$ distribution emulation, namely Density-Estimation Likelihood-Free Inference (DELFI). In addition, we also include the absorber number density along the line-of-sight $\mathrm{d}N/\mathrm{d}z$ in our analysis, i.e. the number of absorption lines (in some range of b and $N_{\rm HI}$) per unit path-length along the line-of-sight, which helps us to better constrain the UV background photoionization rate $\Gamma_{\rm H\,I}$. The reason behind this is that the b-$N_{\rm H\,I}$ distribution is less sensitive to $\Gamma_{\rm H\,I}$ compared with thermal parameters T_0 and γ (see Fig.5 and §3.3), whereas the number density of absorbers (see Fig.4) depends strongly on $\Gamma_{\rm H\,I}$. It is analogous to the fact that the mean flux of the Lyα forest is sensitive to $\Gamma_{\rm H\,I}$. In this work, we emulate the $\mathrm{d}N/\mathrm{d}z$ using a Gaussian process emulator based on our simulations and employ it as a normalization factor in our likelihood function. More discussion about this modification is presented in §3.2 and Appendix A.

This section is organized as follows, we first introduce our new b-$N_{\rm H\,I}$ distribution emulator and then discuss the modifications to the likelihood function. Afterward, we investigate the relationship between thermal parameters and b-$N_{\rm H\,I}$ distribution in §3.3. Finally, we present our inference results in §3.4 and apply a series of inference tests to evaluate the statistical validity of our method in §3.5.

3.1 Emulating the b-$N_{\rm H\,I}$ distribution with DELFI

In this work, we build our b-$N_{\rm H\,I}$ distribution emulator following the density-estimation likelihood-free inference (DELFI) method (Papamakarios & Murray 2016; Alsing et al. 2018; Papamakarios et al. 2018; Lueckmann et al. 2018; Alsing et al. 2019), which turns inference into a density estimation task by learning the sampling distribution of the data as a function of the parameters. Compared with the previously used Kernel Density Estimation (KDE) method in Hiss et al. (2019), this method provides a flexible framework for conditional density estimation and does not implicitly apply a smoothing kernel to the training data. It hence is able to deliver higher-fidelity conditional density estimators given the same training data.

We make use of `pydelfi`[8] – the publicly available `python` implementation of DELFI based on neural density estimators (NDE)s and active learning (Alsing et al. 2019). `pydelfi` makes use of NDEs to learn the sampling conditional probability distribution $P(\mathbf{d}\,|\,\boldsymbol{\theta})$ of the data summaries $\mathbf{d}$, as a function of parameters $\boldsymbol{\theta}$, from a training set of simulated data summary-parameter pairs $\{\mathbf{d}, \boldsymbol{\theta}\}$. In this work, the parameters $\boldsymbol{\theta}$ are $\log T_0$, γ and $\log\Gamma_{\rm H\,I}$, and the data summaries $\mathbf{d}$ are $\log N_{\rm H\,I}$ and $\log b$[9], and the b-$N_{\rm H\,I}$ distribution is considered as a conditional probability distribution $P(b\,, N_{\rm H\,I}\mid T_0, \gamma, \Gamma_{\rm H\,I})$ learned from our simulations. More specifically, the b-$N_{\rm H\,I}$ distribution is modeled as a Masked Autoregressive Flow (MAF; Papamakarios et al. 2017) neural density estimator, which is constructed as a stack of five Masked Autoencoders for Density Estimation, (MADE; Germain et al. 2015), each with two hidden layers with 50 units each and tanh activation functions. The NDEs are trained by stochastic gradient descent. For more technical details about MAF and MADE neural network architectures see Germain et al. (2015), Papamakarios et al. (2017) and Alsing et al. (2019). To prevent over-fitting, the NDEs are weighted by their relative cross-validation losses and are trained with early-stopping (see Alsing et al. 2019 for details). For convenience, in this paper we will refer to the b-$N_{\rm H\,I}$ distribution emulator discussed above as the DELFI emulator.

As mentioned above, the DELFI emulator is trained on the data summary-parameter pairs $\{[\log T_0, \gamma, \log\Gamma_{\rm H\,I}], [b, \log N_{\rm H\,I}]\}$. For each model, we fit (VPFIT) 2000 simulated spectra, corresponding to a total pathlength $\Delta z_{\rm tot} \sim 123$, to get $\{b, N_{\rm H\,I}\}$ pairs for the model, and label these $\{b, N_{\rm H\,I}\}$ pairs with their simulation parameters $[\log T_0, \gamma, \log\Gamma_{\rm H\,I}]$. Our training set therefore consists of all these labeled $\{b, N_{\rm H\,I}\}$ pairs for all models on the thermal grid. Here we quantify the size of data by its total pathlength rather than number of lines[10], because the latter depends on the $\mathrm{d}N/\mathrm{d}z$ that varies among different models.

3.2 Likelihood function

Hiss et al. (2019) used only the shape of b-$N_{\rm H\,I}$ distribution to constrain IGM thermal parameters, but ignored the normalization, which can be thought of as the total number of absorption lines in the dataset or equivalently as the line density dN/dz. Here we generalize the likelihood formalism introduced in Hiss et al. (2019) to include the information contained in the absorber density $\mathrm{d}N/\mathrm{d}z$ (see also Hiss 2019). Our goal is to find the likelihood of observing a set of absorption lines $\{b_i, N_{{\rm H\,I},i}\}$ given a model with a set of thermal parameters $[\log T_0{}', \gamma', \log\Gamma_{\rm H\,I}{}']$. We first assume that the probability density function (PDF)s are normalized such that

$$\iint P(b, N_{\rm H\,I})\,\mathrm{d}N_{\rm H\,I}\,\mathrm{d}b = 1, \tag{3}$$

where $P(b, N_{\rm H\,I})$ is the conditional probability distribution function $P(b\,, N_{\rm H\,I}\mid T_0{}', \gamma', \Gamma_{\rm H\,I}{}')$, for simplicity we write it as $P(b, N_{\rm H\,I})$ in the rest of this subsection. We imagine dividing the b-$N_{\rm HI}$ into a set of infinitesimally fine grid cells, such that the occupation number of each grid cell is either one or zero. Knowing that our set of observational/mock dataset $\{b_i, N_{{\rm H\,I},i}\}$ is comprised of n lines, and assuming that there are $N_{\rm g}$ grid cells in total, the likelihood for a model with thermal parameters $[\log T_0{}', \gamma', \log\Gamma_{\rm H\,I}{}']$ can thus be written as the following product of Poisson probabilities[11]

$$\begin{aligned}\mathcal{L} =& P(\mathrm{data}|\mathrm{model}) \\ =& \left(\prod_{i=1}^{n}\mu_i\, e^{-\mu_i}\right)\left(\prod_{j\neq i}^{N_{\rm g}} e^{-\mu_j}\right),\end{aligned} \tag{4}$$

where the first product is over the occupied cells, and the second

[8] See https://github.com/justinalsing/pydelfi

[9] `pydelfi` also has the option to apply different data compression methods (e.g., Alsing & Wandelt 2018) and active learning methods to optimize the data and parameter space sampling. Here we do not exploit these features since we have pre-chosen our summary statistics and simulation grid (the b-$N_{\rm H\,I}$ distribution) at a fixed grid of thermal parameters.

[10] It means that the learned b-$N_{\rm H\,I}$ distribution has a resolution that depends on the $\mathrm{d}N/\mathrm{d}z$ of the model. We could instead set the number of $\{b, N_{\rm H\,I}\}$ pairs to be fixed while using different total panthlength for each model. However such a change does not affect the results of our inference method.

[11] In assuming the probability distribution for each grid cell is Poisson, we are implicitly assuming each b-$N_{\rm HI}$ pair is an uncorrelated draw from the b-$N_{\rm H\,I}$ distribution. This assumption, also made by Hiss et al. (2019), amounts to ignoring the spatial correlations between absorption lines. Hiss et al. (2019) showed that this is a very good approximation and yields unbiased inference as we will also demonstrate in § 3.5)

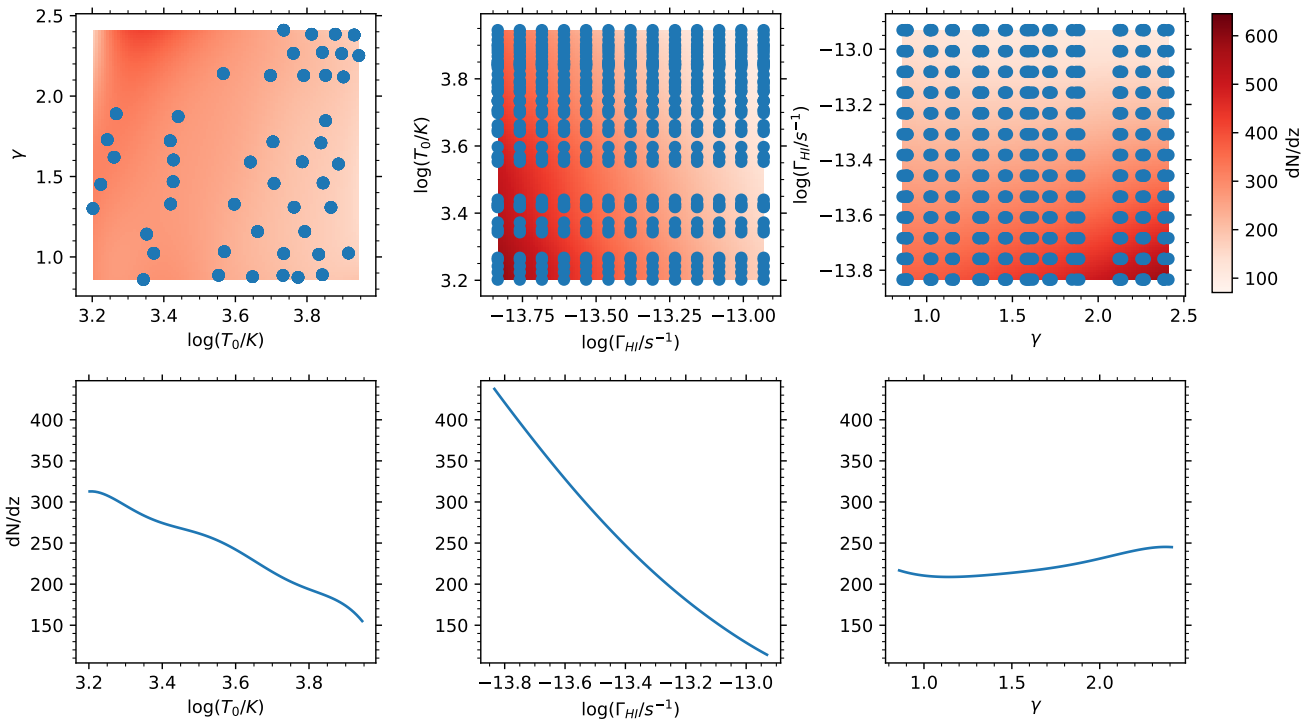

Figure 4. An example of emulation of the absorber density $\mathrm{d}N/\mathrm{d}z$ generated by the Gaussian emulator sliced at the median value of the posterior from the MCMC process, where $\log(T_0/\mathrm{K}) = 3.69$, $\gamma = 1.55$, $\log(\Gamma_{\mathrm{H\,I}}/\mathrm{s}^{-1})$ = -13.30. Top panels are the 2D $\mathrm{d}N/\mathrm{d}z$ distributions where Nyx models are shown in blue circle. The top left panel is the $\mathrm{d}N/\mathrm{d}z$ on γ-$\log T_0$ plane at $\log(\Gamma_{\mathrm{H\,I}}/\mathrm{s}^{-1})$= -13.30. The top middle is $\log T_0$-$\log\Gamma_{\mathrm{H\,I}}$ plane at $\gamma = 1.55$. The top right is $\log\Gamma_{\mathrm{H\,I}}$-$\gamma$ plane at $\log(T_0/\mathrm{K}) = 3.69$. Bottom panels are marginalized 1D $\mathrm{d}N/\mathrm{d}z$ distributions at the thermal parameters mentioned above. From left to right: $\mathrm{d}N/\mathrm{d}z$ vs $\log T_0$, $\mathrm{d}N/\mathrm{d}z$ vs $\log\Gamma_{\mathrm{H\,I}}$, and $\mathrm{d}N/\mathrm{d}z$ vs γ.

product is over the empty cells. Here the μ_i is the Poisson rate of occupying a cell in the b-N_{HI} plane with area $\Delta\mathrm{N}_{\mathrm{H\,I}i} \times \Delta b_i$, i.e.

$$\mu_i = \left(\frac{\mathrm{d}N}{\mathrm{d}z}\right)_{\mathrm{model}} P(b_i, N_{\mathrm{H\,I},i})\, \Delta N_{\mathrm{H\,I}}\, \Delta b\, \Delta z_{\mathrm{data}}, \tag{5}$$

where $P(b_i, N_{\mathrm{H\,I}i})$ is the probability distribution function evaluated at the point $(b_i, N_{\mathrm{H\,I},i})$ using the DELFI b-$N_{\mathrm{H\,I}}$ distribution emulator described in § 3.1, and Δz_{data} is the total redshift path covered by the data spectra from which we obtain our data set $\{b, N_{\mathrm{H\,I}}\}$, whereas $(\mathrm{d}N/\mathrm{d}z)_{\mathrm{model}}$ is the absorber density of the model which will be further discussed later in this subsection.

Afterwards, it is easy to show that Eq. (4) implies

$$\ln\mathcal{L} = \sum_{i=1}^{n} \ln(\mu_i) - \sum_{k=1}^{N_g} \mu_k. \tag{6}$$

Above, the second sum over k is simply an integral of Eq. (5) over the b-$N_{\mathrm{H\,I}}$ plane, while the integral of $P(b, N_{\mathrm{H\,I}})\, \mathrm{d}N_{\mathrm{H\,I}}\, \mathrm{d}b$ over the plane is unity according to Eq. (3). As a result, we can write our likelihood function as

$$\ln\mathcal{L} = \sum_{i=1}^{n} \ln(\mu_i) - \left(\frac{\mathrm{d}N}{\mathrm{d}z}\right)_{\mathrm{model}} \Delta z_{\mathrm{data}}. \tag{7}$$

Since Hiss et al. (2019) did not consider the absorber density, the likelihood in their analysis is simply given by $\ln\mathcal{L}_{\mathrm{Hiss}} = \sum_{i=1}^{n} \ln P(b_i, N_{\mathrm{H\,I},i})$. In comparison, our likelihood function Eq. (7) can be written as

$$\ln\mathcal{L} = \sum_{i=1}^{n} \ln P(b_i, N_{\mathrm{H\,I},i}) + n\ln\xi - \xi, \tag{8}$$

where $\xi = (\mathrm{d}N/\mathrm{d}z)_{\mathrm{model}}\Delta z_{\mathrm{data}}$. We can see that the first term remains the same, and our modification (the implementation of absorber density $\mathrm{d}N/\mathrm{d}z$) can be considered as a correction term based on the absorber density of the model, the number of lines observed, and the pathlength of the data set Δz_{data}.

As a result of our modification, the likelihood of observing a line with certain line parameter $(b, N_{\mathrm{H\,I}})$ now depends not only on the b-$N_{\mathrm{H\,I}}$ distributions of models but also on absorber densities of the models. Consequently, to evaluate the likelihood $\mathcal{L}$ on the parameter space, we need the ability to evaluate $(\mathrm{d}N/\mathrm{d}z)_{\mathrm{model}}$ at an arbitrary location on the parameter space. To this end, a Gaussian process emulator (based on `George`, see Ambikasaran et al. 2016) is employed to emulate $(\mathrm{d}N/\mathrm{d}z)_{\mathrm{model}}$ by interpolating the $\mathrm{d}N/\mathrm{d}z$ of models from Nyx simulations based on their $N_{\mathrm{model}}/\Delta z_{\mathrm{model}}$, where $\Delta z_{\mathrm{model}}$ is the total pathlength of simulated spectra that are fed into `VPFIT`, and N_{model} is the total number of lines identified by `VPFIT` from these spectra. The Gaussian process emulator is constructed with smoothing lengths of 40% of our thermal grid length[12] in $\log T_0$ and $\log\Gamma_{\mathrm{H\,I}}$ and a smoothing length of 80% of thermal grid length in γ. The longer smoothing length in γ is set to prevent the emulator from over-fitting the noise, considering that γ has less effect on the absorber density $\mathrm{d}N/\mathrm{d}z$ compared with T_0 and $\Gamma_{\mathrm{H\,I}}$ (see Fig.4), which makes small fluctuations induced by noise more significant.

The results of our $\mathrm{d}N/\mathrm{d}z$ emulation are shown in Fig.4, where

[12] The smoothing length is input as initial guess, which is then refined later in the routine. In addition, all dimensions in the thermal grid are rescaled to unity in the Gaussian process emulator.

both $\log T_0$ and $\log \Gamma_{\rm H\,I}$ (left and middle column) have negative correlations with absorber density dN/dz. This dependence can be explained qualitatively by the fluctuating Gunn-Peterson approximation (FGPA, see Weinberg et al. 1997)

$$\tau_{{\rm Ly}\alpha} \propto n_{\rm HI} \propto x_{\rm HI} n_{\rm H} \propto \frac{n_{\rm H}^2 T^{-0.7}}{\Gamma_{\rm HI}}, \tag{9}$$

where the $\tau_{{\rm Ly}\alpha}$ denotes the Lyα optical depth and the $n_{\rm H}$ is the hydrogen number density. This equation implies that both higher temperatures and higher photoionization rates reduce the Lyα optical depth of gas absorbers in the IGM, leading to lower absorber density. The wiggles shown in dN/dz vs T_0 plot (bottom left panel of Fig.4) are effects of poor interpolation due to lack of models at $\gamma \sim 1.5$ (see top left panel). Moreover, we notice a weak correlation between γ and dN/dz (see the bottom right panel of Fig.4). However, such γ dependence is relatively weak compared with T_0 and $\Gamma_{\rm HI}$ dependencies, and is likely caused by artifacts due to the emulation. As shown in the top left-hand panel, we do not have models in low T_0 high γ region, the absorber density dN/dz could thus be over-extrapolated in these regions, further biasing the γ dependence on the whole parameter space. We performed some tests and found that the weak correlations in γ - dN/dz vanishes if we do not include the high γ simulations. Therefore, in conclusion, the marginalized γ - dN/dz correlation shown in Fig. 4 is an artifact introduced by our Gaussian emulator, however, it is too weak to affect our inference results.

3.3 Parameter study

A new feature of the DELFI b-$N_{\rm H\,I}$ distribution emulator is its ability to emulate b-$N_{\rm H\,I}$ distributions continuously on the parameter space. With such a feature, we are now able to illustrate the parameter dependence of the b-$N_{\rm H\,I}$ distribution and investigate the physics behind these dependence. Fig.5 shows emulated b-$N_{\rm H\,I}$ distributions with different values of thermal parameters [$\log T_0$, γ , $\log \Gamma_{\rm H\,I}$]. The top panel shows b-$N_{\rm H\,I}$ distributions with increasing T_0, where $\log(T_0/{\rm K})$ = 3.25 (left), 3.60 (middle) and 3.95 (right) respectively, while γ=1.55 and $\log(\Gamma_{\rm H\,I}/{\rm s}^{-1})$=-13.36 for all three plots. Increasing T_0 results in the upward shifting of the b-$N_{\rm H\,I}$ distributions, which can be explained by the thermal component of the b parameter and the T-Δ relationship Eq. (1), i.e.

$$b_T \propto (2kT/m)^{1/2} \propto (T_0 \Delta^{\gamma-1})^{1/2}, \tag{10}$$

where higher T_0 results in higher IGM temperature, leading to larger b parameters. In addition, we notice that as the T_0 goes up, the b-$N_{\rm H\,I}$ distribution becomes more concentrated, i.e. the distribution becomes tighter, and the pdf values increases. Such behavior might be explained as follows. There are two components contributing to b parameter, namely thermal motion and non-thermal broadening. The thermal component is associated with the IGM temperature and thus follows a distribution determined by T_0. On the other hand, as a result of the small-scale motion of the gas, the non-thermal component is independent of the temperature and has a large dispersion, leading to broader distribution. At low temperatures, where the thermal contribution is weak, the b parameter is dominated by the non-thermal component, resulting in broad distribution. As the temperature goes up, the thermal component dominates over non-thermal broadening, and the b parameter thus concentrates on a central value of b determined by the IGM temperature.

The middle panel of Fig.5 shows the b-$N_{\rm H\,I}$ distribution with increasing γ, where γ =1.15 (left), 1.55 (middle), and 1.95 (right), respectively, while $\log(T_0/{\rm K})$ =3.65 and $\log(\Gamma_{\rm H\,I}/{\rm s}^{-1})$=-13.36 are fixed. These plots indicate that there are degeneracies between γ and T_0, where an increasing γ also shifts b-$N_{\rm H\,I}$ distributions upwards, which can be understood from Eq. (10) and the fact that at low-z, the Lyα lines originate predominantly from gas with $\Delta_{\rm abs} > 1$ ($\Delta_{\rm abs} \sim 10$, see Gaikwad et al. 2017), which results in higher temperatures at densities of absorbers for models with larger γ. The concentration effect is also seen in the middle panel, which can be explained in the same way as the upward shifting of the b-$N_{\rm H\,I}$ distribution due to increasing γ. It can also be seen from the middle panel that the γ is correlated with the slope of the low-b cutoff of the b-$N_{\rm H\,I}$ distribution, which is consistent with the analytical fit of the low-b cutoff, where the slope can be approximated by $\Delta \log b / \Delta \log N = (\gamma - 1)/3$ (see Rudie et al. 2012).

The bottom panel of Fig.5 shows b-$N_{\rm H\,I}$ distributions with increasing photoionization rate $\Gamma_{\rm H\,I}$, where $\log(\Gamma_{\rm H\,I}/{\rm s}^{-1})$= -13.66 (left), -13.36 (middle) and -13.06 (right), while $\log T_0$ and γ remain unchanged. We observe that increasing $\Gamma_{\rm H\,I}$ results in a similar but much weaker effect compared with increasing T_0, i.e. the b-$N_{\rm H\,I}$ distribution slightly shifts upward and becomes more concentrated with increasing $\Gamma_{\rm H\,I}$. Such effects are because the photoionization rate $\Gamma_{\rm H\,I}$ alters the Lyα optical depth of the IGM. Since the Lyα forest typically probes regions with optical depth $\tau_{{\rm Ly}\alpha} \sim 1$, given higher $\Gamma_{\rm H\,I}$, it probes regions with higher temperatures and densities, which can be derived from Eq.(9), causing effects similar to increasing T_0. However, such effects are relatively weak, making the b-$N_{\rm H\,I}$ distribution less sensitive to the photoionization rate $\Gamma_{\rm H\,I}$.

All these aforementioned parameter dependences (except $\Gamma_{\rm H\,I}$, which is not considered in previous works) of the b-$N_{\rm H\,I}$ distribution are consistent with previous works that measure the IGM thermal state based on the full b-$N_{\rm H\,I}$ distribution[13] (Hiss et al. 2019) and low-b cutoff (Schaye et al. 1999; Rudie et al. 2012; Bolton et al. 2014; Rorai et al. 2018; Hiss et al. 2018), indicating that our DELFI emulator successfully reproduce the parameter dependences of the b-$N_{\rm H\,I}$ distribution. Furthermore, it also implies that our understanding of the b-$N_{\rm H\,I}$ distribution agrees with the physics prediction.

3.4 Inference results

Sets of mock spectra are created from our Nyx simulations to test the performance of our inference algorithm under realistic conditions. These mock spectra set are generated from a set of simulated spectra following a forward-modeling approach designed to match the pathlength, resolution, and noise properties of the Danforth et al. (2016) low-redshift quasar spectra in one-to-one correspondence as described in §2.3. Consequently, each mock spectra set consists of 34 forward-modeled spectra, which has exactly the same noise vectors, instrumental effects, and total pathlength ($\Delta z_{\rm data}$=2.136) as the real observed dataset, which ensures that the accuracy of our analysis is realistic and achievable when the method is applied to real data. A set of $\{b, N_{\rm HI}\}$ pairs, obtained by fitting these spectra using `VPFIT` (see §2.4), is then used as the 'data' in the likelihood function (see Eq.4) to infer the posterior distribution for IGM thermal parameters for this mock dataset.

In this work, we perform inference via Markov chain Monte Carlo (MCMC) sampling using the `python` package `emcee` (Foreman-Mackey et al. 2013), which implements the affine-invariant sampling

[13] In Hiss et al. (2019), at $z \sim 2$, the Lyα lines originate predominantly from gas with $\Delta < 0$, causing different effects when changing γ. However, the physics explanations behind the effect are coherent.

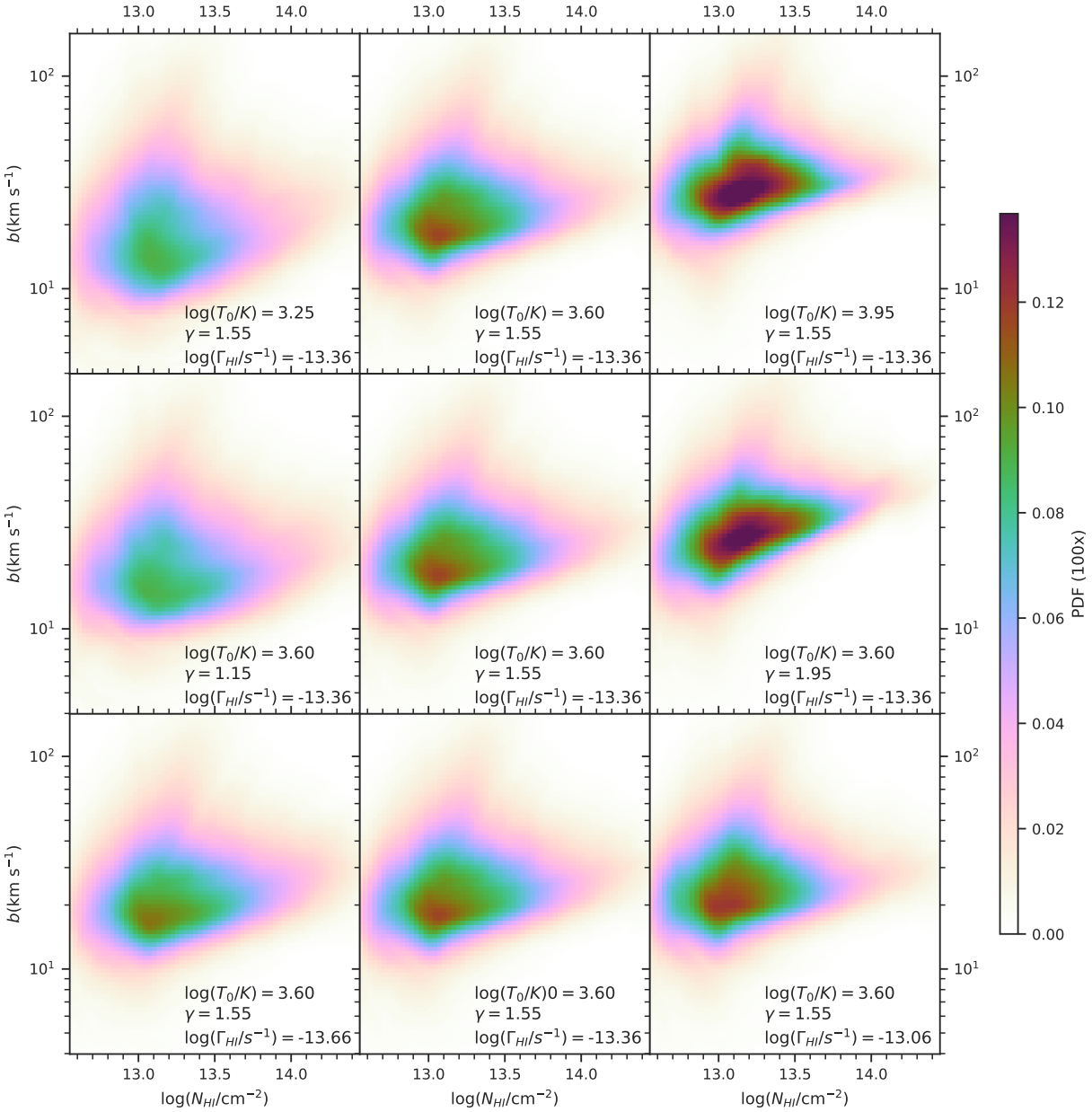

Figure 5. Comparisons of b-$N_{\rm H\,I}$ distributions modeled by DELFI emulator with different thermal parameters. Top panel shows changes in the b-$N_{\rm H\,I}$ distribution with increasing $\log T_0$, where $\log(T_0/{\rm K})$ = 3.25 (left), 3.60 (middle) and 3.95 (right) respectively, while γ=1.55 and $\log(\Gamma_{\rm H\,I}/{\rm s}^{-1})$=-13.36 for all three plots. The middle panel shows changes of the b-$N_{\rm H\,I}$ distribution where γ =1.15 (left), 1.55 (middle) and 1.95 (right) respectively, while $\log(T_0/{\rm K})$=3.60 and $\log(\Gamma_{\rm H\,I}/{\rm s}^{-1})$=-13.36 are fixed. The bottom panel shows b-$N_{\rm H\,I}$ distributions with decreasing UV background. $\log(\Gamma_{\rm H\,I}/{\rm s}^{-1})$= -13.66 (left), -13.36 (middle) and -13.06 (right), while $\log T_0$ and γ remain unchanged. All pdfs here are normalized to unity. For illustration purposes, values of pdf are multiplied by 100 in the color bar.

technique (Goodman & Weare 2010) to sample the posterior probability distribution. Here the posterior is calculated based on the likelihood in Eq. (6), which takes into account the absorber density dN/dz as described in §3.2, while assuming uniform (flat) priors for $\log T_0$, γ and $\log\Gamma_{\rm H\,I}$, where the boundaries are chosen to be the range of each respective parameter in 1D. MCMC posteriors obtained from the aforementioned mock datasets ($\{b, N_{\rm HI}\}$ pairs) are shown in Fig.6. We obtain $\log(T_0/{\rm K}) = 3.668^{+0.075}_{-0.080}$, $\gamma = 1.611^{+0.060}_{-0.055}$ and $\log(\Gamma_{\rm H\,I}/{\rm s}^{-1}) = -13.497^{+0.065}_{-0.066}$ from the marginalized distributions, whereas the true parameters are: $\log(T_0/{\rm K}) = 3.643$, $\gamma = 1.591$ and $\log(\Gamma_{\rm H\,I}/{\rm s}^{-1}) = -13.458$ (red dot and red vertical lines). We recover the input parameters in very high precision with errors $\Delta\log(T_0/{\rm K}) = +0.025$dex, $\Delta\gamma = +1.3\%$, and $\Delta\log(\Gamma_{\rm H\,I}/{\rm s}^{-1}) = -0.039$dex, while true parameters (red dot/solid lines) are all in the 1-σ interval (inner black contours/ black dashed lines) of the posterior. Here the degeneracy between T_0 and γ can be quantitatively understood by the T-Δ relationship Eq. (1) and the typical overdensity of absorbers $\Delta_{\rm abs} \sim 10$. More specifically, both higher T_0 and γ result in higher temperature of the absorbers, shifting the b-$N_{\rm H\,I}$ distribution upward (see Fig.5 and relevant discussion

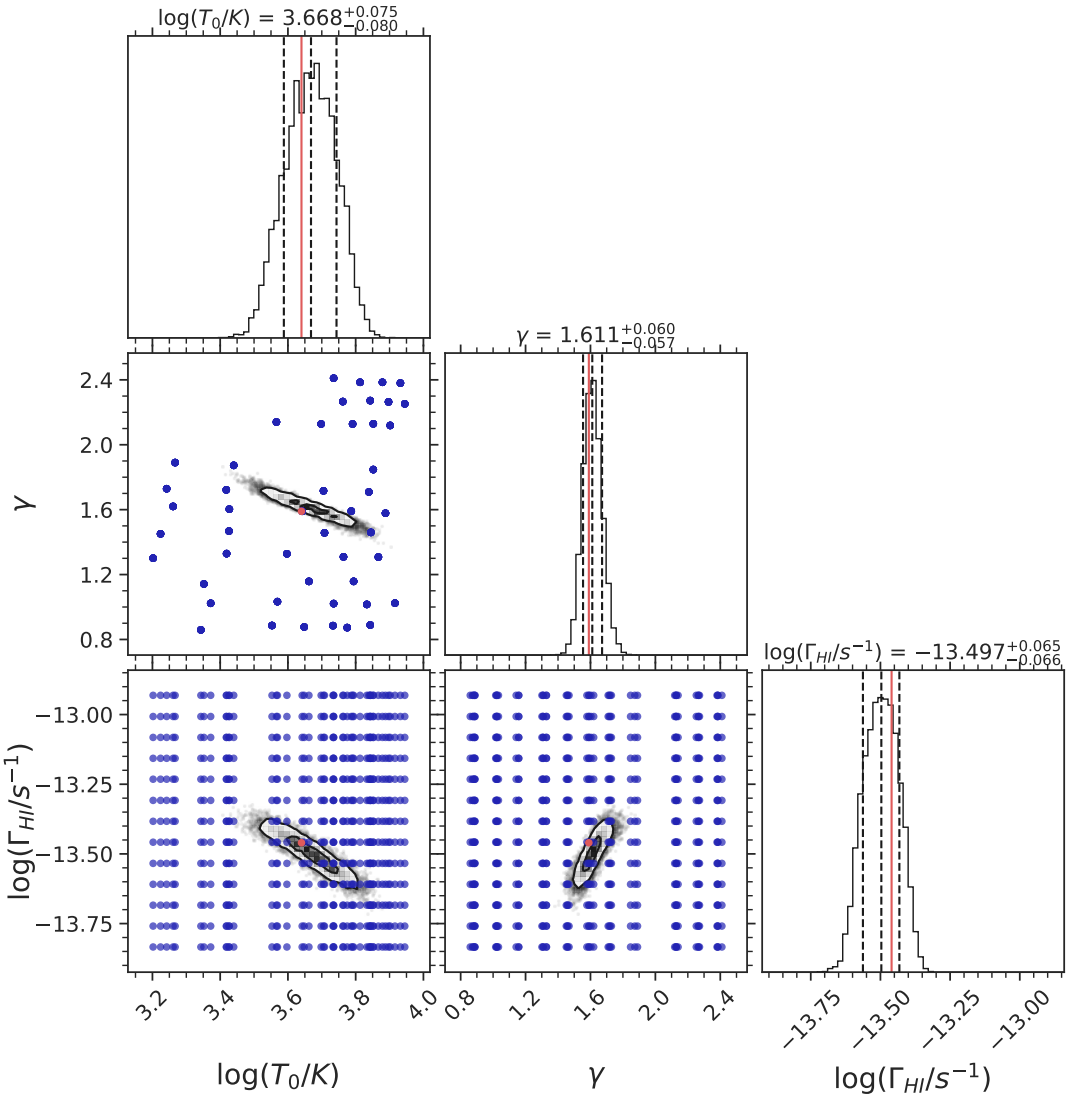

Figure 6. MCMC posterior for one of the models from Nyx simulation (absorbers shown in Fig.7) using the likelihood function Eq. 7. Projections of the thermal grid used for generating models are shown as blue dots, while the true model is shown as red dot. Inner (outer) black contour represents the projected 2D 1(2)-sigma interval. The parameters of true model are indicated by red lines in the marginal distributions, while the dashed black lines indicates the 16, 50, and 84 percentile values of the posterior. The true parameters are: $\log(T_0/\mathrm{K}) = 3.643$, $\gamma = 1.591$ and $\log(\Gamma_{\mathrm{H\,I}}/\mathrm{s}^{-1}) = -13.458$.

in §3.3). The degeneracy between T_0 and $\Gamma_{\mathrm{H\,I}}$ is mainly a result of the degeneracy in the absorber density $\mathrm{d}N/\mathrm{d}z$ with respect to the two parameters (see Fig.4 and Fig.A1 as comparison), which is explained in §3.2. It is noteworthy that our inference algorithm provides pre-eminent accuracy for all three parameters even under a very realistic condition, where the resolution of spectra is rather low (with lines not fully solved), and the number of data is limited (with a total pathlength $\Delta_z = 2.136$). Such a high sensitivity and precision makes our inference method a powerful tool in the study of the low-z IGM and Lyα forest.

Fig.7 shows the full b-$N_{\mathrm{H\,I}}$ distribution recovered from the mock dataset, which is emulated by our DELFI emulator based on the best-fit parameters (median values of the marginalized MCMC posterior). It appears that the PDF (color map) successfully represents the density distribution of the data points. Furthermore, marginalized 1D distributions of b and $N_{\mathrm{H\,I}}$ are given in Fig.8 for both the mock dataset (black dots) and random samples from the emulated b-$N_{\mathrm{H\,I}}$ distribution (blue bars). It can be seen that our emulator successfully reproduces the 1D marginalized b and $N_{\mathrm{H\,I}}$ distribution, though there is a fluctuations in $N_{\mathrm{H\,I}}$ for the mock dataset at around $\log(N_{\mathrm{H\,I}}/\mathrm{cm}^{-2}) \sim 13.5$. We figured out that such fluctuation is caused by the random error during the generation of the mock dataset, which can be reduced by increasing the size of the mock datasets. However, to test the performance of our inference method under realistic conditions, we fix the size of the mock datasets and bear with such fluctuation in this work.

3.5 Inference test

As discussed above, the likelihood function used in our inference algorithm involves several approximations and emulation/interpolation procedures. Most importantly, our inference ignores correlations be-

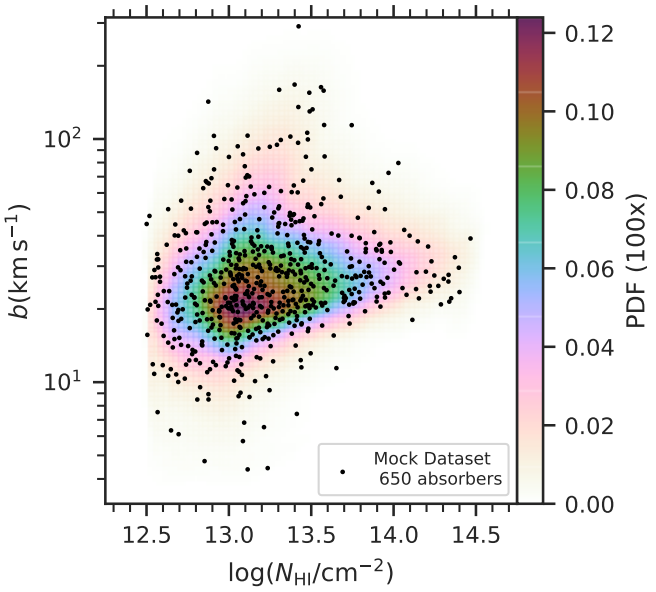

Figure 7. The color map is the full b-$N_{\mathrm{H\,I}}$ distribution recovered from the Nyx mock dataset, which is emulated by our DELFI emulator based on the best-fit parameters (median values of the marginalized MCMC posterior), where $\log(T_0/\mathrm{K}) = 3.668$, $\gamma = 1.611$ and $\log(\Gamma_{\mathrm{H\,I}}/\mathrm{s}^{-1})$ =-13.498. Black dots are the mock datasets we used in the inference. For illustration purposes, values of pdf are multiplied by 100 in the color bar.

tween the lines (see the discussion in Hiss et al. 2019), and we emulate the b-$N_{\mathrm{H\,I}}$ distribution and the dN/dz with our DELFI and Gaussian emulators respectively, while both emulations involve interpolations. These procedures might induce additional uncertainties that are counted in our error budget[14], we hence want to make sure our inference results are valid under these assumptions, and our interpolation procedures work correctly. Therefore, we perform a series of inference tests to evaluate the robustness of the entire inference method. An inference test is to carry out a set of realizations of the inference algorithm based on the mock dataset and inspect the results to reveal if the inference method returns valid posterior probability distributions, i.e. whether the 'true model' is included in a set of probability contours following the ratio indicated by the posterior.

The inference test is done as follows. First of all, we adopt the same prior as described in 3.4, and construct a regular uniform grid in the parameter space spanning the range set by our prior. For each realization, we pick a model (set of parameters) on the above grid, which we refer to as the 'true model'. and we refer its thermal parameters as 'true parameters' $\boldsymbol{\theta}_{\mathrm{true}}$, We then create a corresponding mock dataset following the prescription described in §3.4. Given the mock dataset, since our priors are flat, we can determine the corresponding posterior probability distribution by evaluating the likelihood function $\mathcal{L} = P(\mathrm{data}|\mathrm{model})$ on the whole parameter space. We then normalize the posterior function to unity and determine 3D posterior

[14] The uncertainty of the b-$N_{\mathrm{H\,I}}$ distributions emulated by DELFI is also ignored in our analysis. Such uncertainty is caused by the randomness in the training process, and has not been included in the results. But since our inference method (and the toy model) does well in the inference test, such randomness should be smaller than stochastic error shown in our analysis, and should not dominate our error budget.

Table 1. Table of results of the inference test

models	Total	68(%)	95 (%)
random models	480	290 (60.42 ± 2.29%)	439 (94.67 ± 1.25%)
single model	200	134 (67.00 ± 3.50%)	190 (95.00 ± 1.50%)

probability contours based on the posterior (likelihood) distribution. Knowing that the likelihood function is continuous on the whole domain, the 3D volume integral can hence be substituted by a 1D integral over the sorted likelihood function. Here we define the probability contours C_P and the likelihood thresholds $\mathcal{L}_P$ in the following way,

$$\iiint_{C_P} \mathcal{L}\mathrm{d}V = \int_{\mathcal{L}_P}^{\infty} \mathcal{L}\mathrm{d}\mathcal{L} = P, \tag{11}$$

such that a probability contour C_P is simply where $\mathcal{L} = \mathcal{L}_{\mathrm{P}}$, and any 'model' with parameter $\boldsymbol{\theta}$ being inside a contour C_P thus becomes equivalent to $\mathcal{L}(\boldsymbol{\theta}) > \mathcal{L}_{\mathrm{P}}$. We further define the effective 1σ (68%) and 2σ (95%) intervals as the volume between contour pairs $(C_{0.16}, C_{0.84})$ and $(C_{0.025}, C_{0.975})$ respectively. Finally, we judge the performance of our inference method based on how often the parameters of the 'true model' $\boldsymbol{\theta}_{\mathrm{true}}$ falls in these 1(2)-σ interval contour pairs compared to the expectation based on the corresponding probabilities, i.e. if our posterior distribution is perfect, the true model should land within the 1σ (2σ) contours 68% (95%) of the time. An example of the distribution of the likelihood function is shown in Fig.B1, and more details about the calculation of the likelihood distribution is presented in Appendix B.

In practice, we perform an inference test on a set of random models on the thermal grid to test the overall performance of our inference algorithm. We pick 12 models and execute 40 realizations per model. The result shows that the true values are within the 1-σ (68%) interval for 60.42 ± 2.29 % (290/480) of the time, and in the 2σ(95%) interval for 94.67 ± 1.25 % (439/480) of the time, while the upper and lower limits are given by the $\pm 1\sigma_{\mathrm{bi}}$ error for corresponding binomial distributions. In addition, we carry out a cross-validation test to ensure our emulators are not affected by over-fitting problem. Here we select a single model near the center of the parameter space ($\log(T_0/\mathrm{K}) = 3.643$, $\gamma = 1.591$, and $\log(\Gamma_{\mathrm{H\,I}}/\mathrm{s}^{-1}) = -13.458$.), and exclude the model[15] from the training dataset. We train our emulators (both b-$N_{\mathrm{H\,I}}$ distribution and dN/dz) based on the new dataset, and run 200 realizations of our inference method. We observe that the true values are inside the 1σ (68%) interval for 67.00 ± 3.50% (134/200) of the time, and inside the 2σ (95%) interval for 95.0 ± 1.50% (190/200) of the time. Results are presented in Table 1. The overall performance indicates that our algorithm passes the inference[16].

In the end, to further demonstrate and elaborate on the effectiveness of our inference algorithm, we created a toy model, which

[15] In practice we exclude all models with the same T_0 and γ ($\log(T_0/\mathrm{K}) = 3.643$ and $\gamma = 1.591$), since we mostly want to test the performance of the b-$N_{\mathrm{H\,I}}$ distribution emulator on the T_0-γ plane.

[16] Our inference method performs better when the model is close to the center of the grid. This might be because our emulators, both DELFI and Gaussian process emulator, perform better at the center of the grid where the interpolation is more accurate. Besides, our thermal grid has an irregular shape on the T_0-γ plane, and might thus make the interpolation even harder or distorted when there are no or only a few models around. Such a problem might be addressed by adding more simulation models, extending the thermal grid to make sure the region we are interested in always lies at the center of the grid.

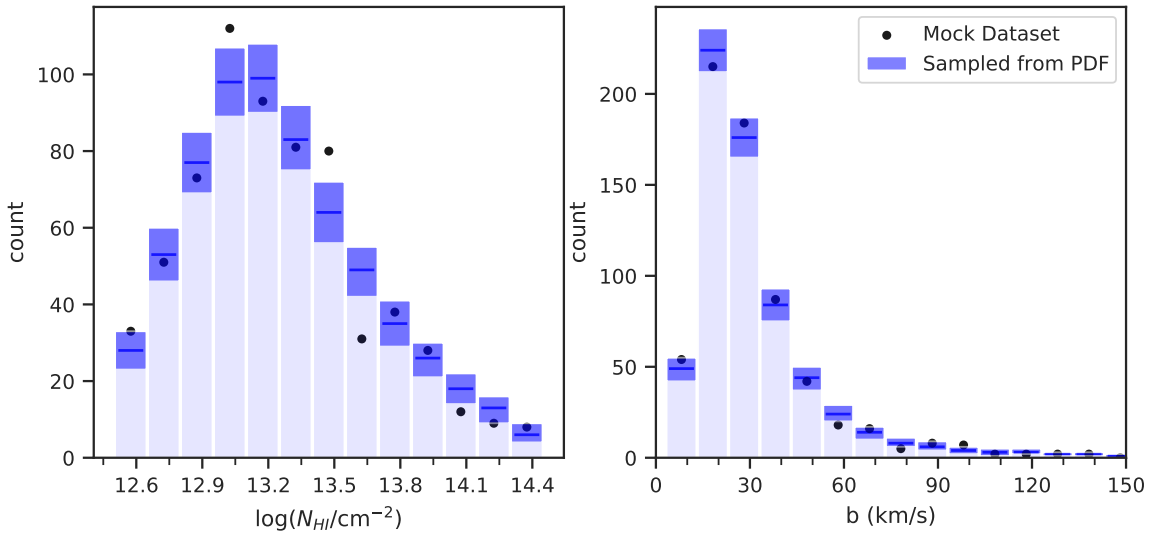

Figure 8. Marginalized 1D distributions of $N_{\rm H\,I}$ (*left-hand panel*) and b (*right-hand panel*) for the mock dataset (black dots) and the sampling from emulated b-$N_{\rm H\,I}$ distribution (blue bars) at $\log(T_0/{\rm K}) = 3.699$, $\gamma = 1.549$ and $\log(\Gamma_{\rm H\,I}/{\rm s}^{-1})$ =-13.506. Blue bars show the average of 5000 sampling from the emulated b-$N_{\rm H\,I}$ distribution using MCMC, while the (dark) blue shaded regions represent the 1-σ fluctuation (16%-84% percentile among 5000 samples).

involves entire inference pipeline, (in Appendix C) to test the whole inference algorithm under more controlled conditions, where the toy b-$N_{\rm H\,I}$ distribution is analytical, and the parameter dependence is known. Here the toy b-$N_{\rm H\,I}$ distribution consists of a multivariate Gaussian distribution parameterized by three mock parameters following the parameter dependence discussed in § 3.3. Moreover, these mock parameters also control the line density dN/dz of the model based on the dN/dz map generated by the Gaussian emulator from our Nyx simulation models (see Appendix C for more details). As a result of this toy model and also the inference test, we conclude that our inference algorithm is sound.

4 SUMMARY AND CONCLUSIONS

In this study, we have presented and evaluated our new method of measuring the thermal state $[T_0, \gamma]$ and the photoionization rate $\Gamma_{\rm H\,I}$ of the low redshift IGM using its b-$N_{\rm H\,I}$ distribution and absorber density dN/dz. We made use of a novel machine learning technique DELFI to build a b-$N_{\rm H\,I}$ distribution emulator and used a Gaussian process emulator to simulate the absorber density dN/dz. We trained both emulators on a dataset generated from a set of Nyx simulations on a large parameter grid. To test the performance of our inference algorithm under realistic conditions, we applied forward modeling techniques to model the noise and instrumental effects based on the HST COS quasar spectra from Danforth et al. (2016). We showed using extensive tests that our inference method is proficient and reliable. Here we conclude by discussing the performance and summarizing the essential elements of our new algorithm.

- We used mock datasets to simulate the measurement of the thermal state $[T_0, \gamma]$ of the low redshift IGM from the full joint b-$N_{\rm H\,I}$ distribution, for the first time taking the absorber density dN/dz into account. The latter enables us to constrain the photoionization rate $\Gamma_{\rm HI}$, since only the shape of the b-$N_{\rm H\,I}$ distribution is insensitive to this parameter (see Fig.5). We also confirm that the dN/dz term we introduced is consistent with our inference based on the b-$N_{\rm H\,I}$ distribution alone, and improves the performance of our inference method (see Appendix A).
- Our new inference method successfully recovers thermal parameters of models from the Nyx simulation with small uncertainties (in our example, $\sigma_{\log T_0} \sim 0.08$ dex, $\sigma_\gamma \sim 0.06$, and $\sigma_{\log \Gamma_{\rm H\,I}} \sim 0.07$ dex), using a relatively small dataset with $\Delta z = 2.316$. Furthermore, these results are obtained under realistic conditions as we forward-model the observational effects and noise from the Danforth et al. (2016) low-z COS quasar spectra while setting the size of our mock datasets to be the same as the observational dataset (i.e. having the same total pathlength $\Delta z_{\rm ob}$). Considering all these factors, the accuracy and sensitivity we attained in this study should be achievable when our inference method is applied to real observational data, making it a powerful tool for studying the Lyα forest.
- Our algorithm passes the inference test (see §3.5), indicating that our approximation and emulation/interpolation are reliable. We also demonstrate the robustness of our inference method by testing the entire inference pipeline, including emulation and interpolation procedures on a toy model under better-controlled conditions(see Appendix C).
- The b-$N_{\rm H\,I}$ distribution (DELFI) emulator successfully emulates both the 2D b-$N_{\rm H\,I}$ distributions and 1D marginalized distributions of b and $N_{\rm H\,I}$. We find that the 2D b-$N_{\rm H\,I}$ distribution shifts upward (towards higher b values) with increasing T_0 and γ, while larger γ also tilts up the low-b cut off. We explain these effects qualitatively in section § 3.3 and show that they are consistent with previous work.

Moreover, previous work (Viel et al. 2017; Gaikwad et al. 2017; Nasir et al. 2017) reported a discrepancy in the 1D marginalized b distribution for low redshift IGM between the observation and current simulations, implying the existence of additional heating or turbulence that is stronger than expected (Bolton et al. 2021). While these works mainly focus on 1D marginalized distributions of b and column density distribution function (CDDF), our new inference algorithm, which successfully emulate both 1D marginalized and 2D joint b-$N_{\rm H\,I}$ distribution, would allow us to investigate such problem using the joint distribution together with dN/dz statistics. We aim to investigate this problem by applying our inference method to observational data in future works, which we expect would provide an accurate measurement of the thermal state of the low-z IGM and

possibly solve this discrepancy. In addition, we also look forward to applying our method to other recent cosmological galaxy formation simulations like Illustris (TNG) (Genel et al. 2014; Weinberger et al. 2017), to study the effect of feedback on the Lyα forest which is not yet completely understood (see for e.g, Gurvich et al. 2017; Christiansen et al. 2020; Burkhart et al. 2022).

ACKNOWLEDGEMENTS

We thank the members of the ENIGMA[17], Siang Peng Oh, Timothy Brandt, and K.G. Lee for helpful discussions and suggestions. Thanks also to Ilya Khrykin for useful feedback as well as contributions to the inference code.

Calculations presented in this paper used the hydra and draco clusters of the Max Planck Computing and Data Facility (MPCDF, formerly known as RZG). MPCDF is a competence center of the Max Planck Society located in Garching (Germany). This research also used resources of the National Energy Research Scientific Computing Center (NERSC), a U.S. Department of Energy Office of Science User Facility located at Lawrence Berkeley National Laboratory, operated under Contract No. DE-AC02-05CH11231. In addition, we acknowledge Partnership for Advanced Computing in Europe (PRACE) for awarding us access to JUWELS hosted by JSC, Germany.

Justin Alsing was supported by research project grant Fundamental Physics from Cosmological Surveys funded by the Swedish Research Council (VR) under Dnr 2017-04212.

DATA AVAILABILITY

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

REFERENCES

Almgren A. S., Bell J. B., Lijewski M. J., Lukić Z., Van Andel E., 2013, ApJ, 765, 39
Alsing J., Wandelt B., 2018, Monthly Notices of the Royal Astronomical Society: Letters, 476, L60
Alsing J., Wandelt B., Feeney S., 2018, MNRAS, 477, 2874
Alsing J., Charnock T., Feeney S., Wandelt B., 2019, MNRAS, 488, 4440
Ambikasaran S., Foreman-Mackey D., Greengard L., Hogg D. W., O'Neil M., 2016, IEEE Transactions on Pattern Analysis and Machine Intelligence, 38, 252
Araya I. J., Padilla N. D., 2014, MNRAS, 445, 850
Becker G. D., Bolton J. S., Haehnelt M. G., Sargent W. L. W., 2011, MNRAS, 410, 1096
Boera E., Murphy M. T., Becker G. D., Bolton J. S., 2014, MNRAS, 441, 1916
Bolton J. S., Viel M., Kim T. S., Haehnelt M. G., Carswell R. F., 2008, MNRAS, 386, 1131
Bolton J. S., Becker G. D., Haehnelt M. G., Viel M., 2014, MNRAS, 438, 2499
Bolton J. S., Gaikwad P., Haehnelt M. G., Kim T.-S., Nasir F., Puchwein E., Viel M., Wakker B. P., 2021, arXiv e-prints, p. arXiv:2111.09600
Burkhart B., Tillman M., Gurvich A. B., Bird S., Tonnesen S., Bryan G. L., Hernquist L. E., Somerville R. S., 2022, arXiv e-prints, p. arXiv:2204.09712
Carswell R. F., Webb J. K., 2014, VPFIT: Voigt profile fitting program (ascl:1408.015)

[17] http://enigma.physics.ucsb.edu/

Christiansen J. F., Davé R., Sorini D., Anglés-Alcázar D., 2020, MNRAS, 499, 2617
Croton D. J., et al., 2006, MNRAS, 365, 11
Danforth C. W., et al., 2016, VizieR Online Data Catalog, p. J/ApJ/817/111
Davé R., Oppenheimer B. D., Katz N., Kollmeier J. A., Weinberg D. H., 2010, MNRAS, 408, 2051
Dixon K. L., Furlanetto S. R., 2009, ApJ, 706, 970
Fan X., et al., 2006, AJ, 132, 117
Faucher-Giguère C.-A., Lidz A., Hernquist L., Zaldarriaga M., 2008, ApJ, 688, 85
Foreman-Mackey D., Hogg D. W., Lang D., Goodman J., 2013, PASP, 125, 306
Gaikwad P., Srianand R., Choudhury T. R., Khaire V., 2017, MNRAS, 467, 3172
Gaikwad P., Srianand R., Haehnelt M. G., Choudhury T. R., 2021, MNRAS, 506, 4389
Garzilli A., Bolton J. S., Kim T. S., Leach S., Viel M., 2012, MNRAS, 424, 1723
Garzilli A., Theuns T., Schaye J., 2015, MNRAS, 450, 1465
Garzilli A., Theuns T., Schaye J., 2018, preprint, (arXiv:1808.06646)
Genel S., et al., 2014, MNRAS, 445, 175
Germain M., Gregor K., Murray I., Larochelle H., 2015, in International Conference on Machine Learning. pp 881–889
Goodman J., Weare J., 2010, CAMCoS, 5, 65
Green J. C., et al., 2012, ApJ, 744, 60
Gunn J. E., Peterson B. A., 1965, ApJ, 142, 1633
Gurvich A., Burkhart B., Bird S., 2017, ApJ, 835, 175
Haardt F., Madau P., 2012, ApJ, 746, 125
Hiss H., 2019, PhD thesis, Dekanat der Fakultät für Physik und Astronomie, https://doi.org/10.11588/heidok.00027299
Hiss H., Walther M., Hennawi J. F., Oñ orbe J., O'Meara J. M., Rorai A., Lukić Z., 2018, ApJ, 865, 42
Hiss H., Walther M., Oñorbe J., Hennawi J. F., 2019, ApJ, 876, 71
Hopkins P. F., Hernquist L., Cox T. J., Kereš D., 2008, ApJS, 175, 356
Hui L., Gnedin N. Y., 1997, MNRAS, 292, 27
Hui L., Haiman Z., 2003, ApJ, 596, 9
Khaire V., 2017, MNRAS, 471, 255
Khaire V., et al., 2019, MNRAS, 486, 769
Kulkarni G., Worseck G., Hennawi J. F., 2019, MNRAS, 488, 1035
Lee K.-G., et al., 2015, The Astrophysical Journal, 799, 196
Lidz A., Faucher-Giguère C.-A., Dall'Aglio A., McQuinn M., Fechner C., Zaldarriaga M., Hernquist L., Dutta S., 2010, ApJ, 718, 199
Lueckmann J.-M., Bassetto G., Karaletsos T., Macke J. H., 2018, arXiv preprint arXiv:1805.09294
Lukić Z., Stark C. W., Nugent P., White M., Meiksin A. A., Almgren A., 2015, MNRAS, 446, 3697
Lynds R., 1971, Apj, 164, L73
Madau P., Meiksin A., 1994, ApJ, 433, L53
Madau P., Pozzetti L., Dickinson M., 1998, ApJ, 498, 106
McDonald P., Miralda-Escudé J., Rauch M., Sargent W. L. W., Barlow T. A., Cen R., 2001, ApJ, 562, 52
McGreer I. D., Mesinger A., D'Odorico V., 2015, MNRAS, 447, 499
McQuinn M., Upton Sanderbeck P. R., 2016, MNRAS, 456, 47
McQuinn M., Lidz A., Zaldarriaga M., Hernquist L., Hopkins P. F., Dutta S., Faucher-Giguère C.-A., 2009, Apj, 694, 842
Miralda-Escudé J., Rees M. J., 1994, MNRAS, 266, 343
Miralda-Escudé J., Haehnelt M., Rees M. J., 2000, ApJ, 530, 1
Nasir F., Bolton J. S., Viel M., Kim T.-S., Haehnelt M. G., Puchwein E., Sijacki D., 2017, MNRAS, 471, 1056
Papamakarios G., Murray I., 2016, in Advances in Neural Information Processing Systems. pp 1028–1036
Papamakarios G., Pavlakou T., Murray I., 2017, arXiv e-prints, p. arXiv:1705.07057
Papamakarios G., Sterratt D. C., Murray I., 2018, arXiv preprint arXiv:1805.07226
Planck Collaboration et al., 2014, A&A, 571, A16
Puchwein E., Pfrommer C., Springel V., Broderick A. E., Chang P., 2012, MNRAS, 423, 149

Rahmati A., Pawlik A. H., Raičević M., Schaye J., 2013, MNRAS, 430, 2427
Ricotti M., Gnedin N. Y., Shull J. M., 2000, ApJ, 534, 41
Robertson B. E., Ellis R. S., Furlanetto S. R., Dunlop J. S., 2015, ApJL, 802, L19
Rorai A., Hennawi J. F., White M., 2013, ApJ, 775, 81
Rorai A., et al., 2017, Science, 356, 418
Rorai A., Carswell R. F., Haehnelt M. G., Becker G. D., Bolton J. S., Murphy M. T., 2018, MNRAS, 474, 2871
Rudie G. C., Steidel C. C., Pettini M., 2012, ApJ, 757, L30
Schaye J., Theuns T., Leonard A., Efstathiou G., 1999, MNRAS, 310, 57
Schaye J., Theuns T., Rauch M., Efstathiou G., Sargent W. L. W., 2000, MNRAS, 318, 817
Sijacki D., Springel V., Di Matteo T., Hernquist L., 2007, MNRAS, 380, 877
Springel V., et al., 2005, Nature, 435, 629
Syphers D., Shull J. M., 2014, ApJ, 784, 42
Tepper-García T., 2006, MNRAS, 369, 2025
Theuns T., Zaroubi S., 2000, MNRAS, 317, 989
Theuns T., Schaye J., Haehnelt M. G., 2000, MNRAS, 315, 600
Theuns T., Schaye J., Zaroubi S., Kim T.-S., Tzanavaris P., Carswell B., 2002, The Astrophysical Journal, 567, L103
Viel M., Bolton J. S., Haehnelt M. G., 2009, Monthly Notices of the Royal Astronomical Society: Letters, 399, L39
Viel M., Haehnelt M. G., Bolton J. S., Kim T.-S., Puchwein E., Nasir F., Wakker B. P., 2017, MNRAS, 467, L86
Villasenor B., Robertson B., Madau P., Schneider E., 2021, ApJ, 912, 138
Walther M., Hennawi J. F., Hiss H., Oñorbe J., Lee K.-G., Rorai A., O'Meara J., 2017, The Astrophysical Journal, 852, 22
Walther M., Hennawi J. F., Hiss H., Oñorbe J., Lee K.-G., Rorai A., O'Meara J., 2018, ApJ, 852, 22
Walther M., Oñorbe J., Hennawi J. F., Lukić Z., 2019, ApJ, 872, 13
Weinberg D. H., Hernsquit L., Katz N., Croft R., Miralda-Escudé J., 1997, in Petitjean P., Charlot S., eds, Structure and Evolution of the Intergalactic Medium from QSO Absorption Line System. p. 133 (arXiv:astro-ph/9709303)
Weinberger R., et al., 2017, MNRAS, 465, 3291
Wolfson M., Hennawi J. F., Davies F. B., Oñorbe J., Hiss H., Lukić Z., 2021, MNRAS, 508, 5493
Worseck G., et al., 2011, ApJl, 733, L24
Worseck G., Davies F. B., Hennawi J. F., Prochaska J. X., 2018, preprint, (arXiv:1808.05247)
Zaldarriaga M., Hui L., Tegmark M., 2001, The Astrophysical Journal, 557, 519

APPENDIX A: INFERENCE WITHOUT ABSORBER DENSITY

In this section we provide more details about the inference without using the absorber density. In such a case, the likelihood function would simply be the first term of Eq.(8), i.e.

$$\ln \mathcal{L} = \sum_{i=1}^{n} \ln P(b_i, N_{\mathrm{H\,I},i}). \tag{A1}$$

Such likelihood function is evaluated based on our b-$N_{\mathrm{H\,I}}$ distribution emulator solely. To make better comparison, we use the same mock dataset and training dataset as used in §3.4. The MCMC posterior is given in Fig.A1, where we obtain $\log(T_0\mathrm{K}) = 3.709^{+0.058}_{-0.073}$, $\gamma = 1.550^{+0.066}_{-0.068}$ and $\log(\Gamma_{\mathrm{H\,I}}/\mathrm{s}^{-1}) = 13.401^{+0.097}_{-0.090}$ from the marginalized distributions, whereas the true parameters are: $\log(T_0\mathrm{K}) = 3.643$, $\gamma = 1.591$ and $\log(\Gamma_{\mathrm{H\,I}}/\mathrm{s}^{-1}) = 13.458$. In comparison, the posterior obtained using Eq.(6), which takes into account the $\mathrm{d}N/\mathrm{d}z$, is shown in blue in Fig.A1. As we show here, the two inference results are coherent, but our modified inference algorithm (green posteriors) perform better. By implementing the $\mathrm{d}N/\mathrm{d}z$ feature, our modified inference algorithm provides more accurate results in both T_0 and $\Gamma_{\mathrm{H\,I}}$, and reduce the uncertain in $\Gamma_{\mathrm{H\,I}}$ significantly. Furthermore, the inference without absorber density dose not pass the inference where the true model falls in the 1-σ (68%) interval for about 50% of the time.

In short, by employing the absorber density we not only evidently reduce the uncertainty in $\Gamma_{\mathrm{H\,I}}$ but also increase the accuracy in other parameters since the modification adds more information to the Bayesian analysis by matching the absorber density.

APPENDIX B: INFERENCE TEST LIKELIHOOD CALCULATION

To calculate contours of cumulative probability distribution with high dimensionality is challenging in computation power. In our case, the parameter grid size is 100^3 and we have to compute the probability density function $P(b, N_{\mathrm{H\,I}} \mid T_0, \gamma, \Gamma_{\mathrm{H\,I}})$ many hundreds times (i.e. the number of lines in the data set) to evaluate the likelihood function on a single point on the parameter grid (see Eq. 6). However, due to the structure of the b-N_{HI} PDF calculated by our DELFI emulator, we are able to save time by computing the likelihood function on the whole grid simultaneously, with help of vector operations implemented in `python`, though such treatment requires reconstruction of the likelihood function and needs extra amounts of memory. In comparison, our code is much faster than the MCMC prescription which would require a very long chain to interpolate the likelihood function on the whole grid to achieve the same precision. An example of the distribution of the likelihood function is shown in Fig.B1.

APPENDIX C: TOY MODEL

To verify the performance of our emulators in a clean environment, we build a toy model with a mock data set which roughly simulates the behavior of our real model. Here the toy b-$N_{\mathrm{H\,I}}$ distributions consist of 2D Gaussian distributions parameterized by T_{mock} and γ_{mock}, Γ_{mock} in analogy with thermal parameters T_0, γ and $\Gamma_{\mathrm{H\,I}}$. Here we follow the parameter dependence discussed in §3.3, i.e. both T_{mock} and γ_{mock} sets the y-axis location of the center of the Gaussian, while γ_{mock} also sets the tilted angle of the Gaussian, and the Γ_{mock} controls the density of data points for each model, in analogy with the $\Gamma_{\mathrm{H\,I}}$ which determines the absorber density $\mathrm{d}N/\mathrm{d}z$. For convenience, we set these mock parameters to be dimensionless. We tune these parameters in a way that the 'b-$N_{\mathrm{H\,I}}$ distribution' of our toy model falls roughly in the same range as the Nyx simulation, and we adopt absorber density emulated by our $\mathrm{d}N/\mathrm{d}z$ emulator based on our Nyx simulations, so that the mock $\mathrm{d}N/\mathrm{d}z$ follows the relationship between thermal parameters and absorber density in our Nyx simulation. We in total generate 7x7x7 = 343 (see Fig.C2) models spanning the thermal grid. An example of the b-$N_{\mathrm{H\,I}}$ distribution of a toy model is shown in Fig.C1, which is generated based on the Kernel Density Estimation (KDE) of the mock dataset using a smoothing bandwidth $(\sigma_{\log N_{\mathrm{mock}}}, \sigma_{\log b_{\mathrm{mock}}}) = (0.08, 0.32)$. Such choice of bandwidth is taken from Hiss et al. (2019).

For each toy model with different mock thermal parameters, we first generate a set of 2000 'imaginary' pathlength Δz_i, each of which equals to a randomly chosen observation spectra in Danforth et al. (2016) low-z Lyα dataset (i.e. for each model we generate a set of 2000 Δz_i but without actual spectra). For each 'imaginary' pathlength Δz_i we generate a set of mock 'b-N' pairs (lines), sampling from the b-$N_{\mathrm{H\,I}}$ distribution, while the number of lines N_i follows a Poisson distribution $\mathrm{Pois}(\lambda_i)$ with Poisson rate $\lambda_i = \Delta z_i \times (\mathrm{d}N/\mathrm{d}z)_{\mathrm{model}}$,

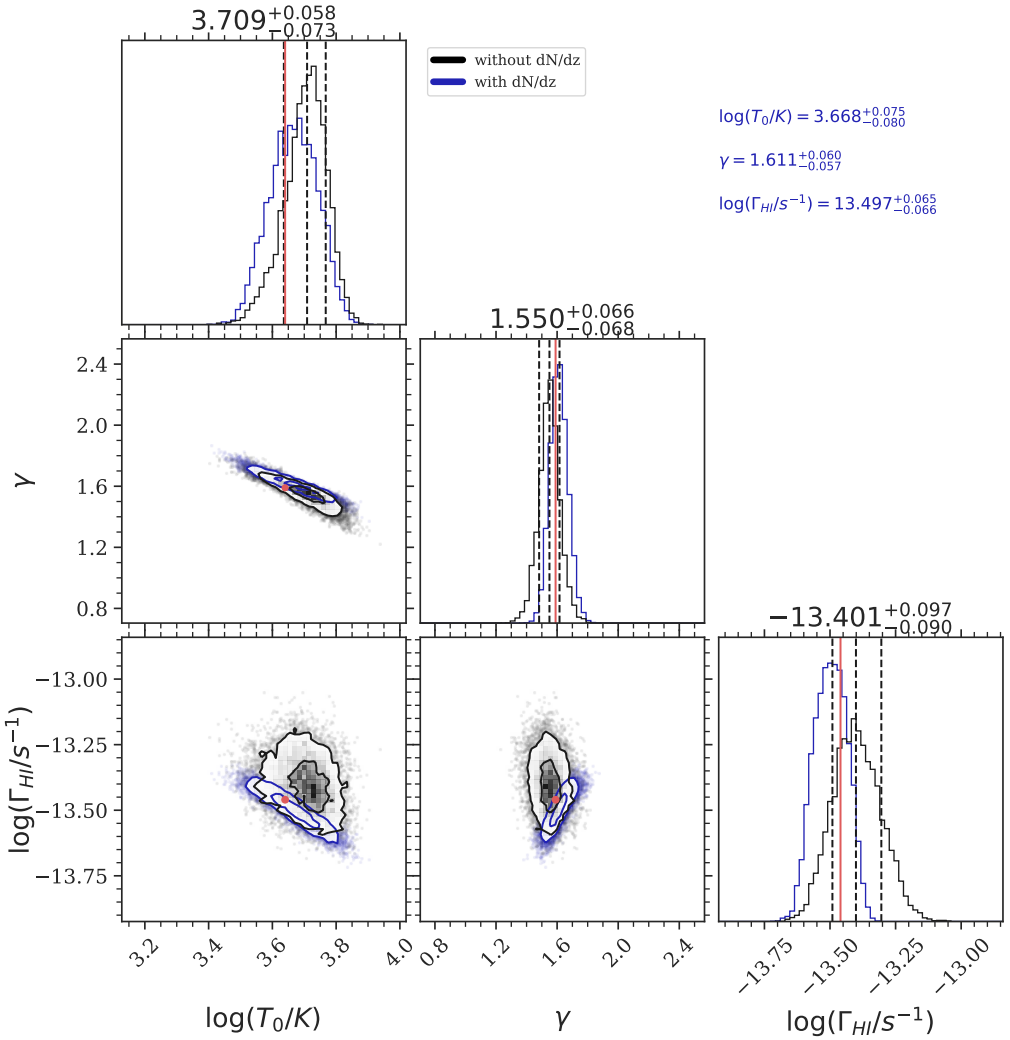

Figure A1. MCMC posterior (black) for the Nyx model discussed in §3.4 based on the likelihood function without the absorber density Eq.(A1). Projections of the true model is shown as red dot. Inner(outer) contours represents the projected 2D 1(2)-sigma interval. The parameters of true model are indicated by red lines in the marginal distributions, while the dashed black lines indicates the 16, 50, and 84 percentile values of the posterior. The true parameters are: $\log(T_0/\mathrm{K})$ = 3.643, γ = 1.591 and $\log(\Gamma_{\mathrm{H\,I}}/\mathrm{s}^{-1})$ = -13.458. In comparison, the posterior obtained using Eq.(6), which takes into account the $\mathrm{d}N/\mathrm{d}z$, is shown in blue, while the medians of the posterior are shown in blue on the top right.

where the $(\mathrm{d}N/\mathrm{d}z)_{\mathrm{model}}$ is the absorber density of that model. The total number of lines for the model is thus $N_{\mathrm{tot}} = \sum_i^{2000} N_i$. At this point we obtain a training dataset with the same structure as the one described in §3.1, which consists of 'b-N' pairs labeled by thermal parameters. We then train the DELFI (b-$N_{\mathrm{H\,I}}$ distribution) and Gaussian ($\mathrm{d}N/\mathrm{d}z$) emulators based on the above dataset, and test our whole inference algorithm on the toy model following the prescription given in §3.4. An example of the inference result is shown below, including the MCMC posteriors (Fig.C3) and the 'best fit' b-$N_{\mathrm{H\,I}}$ distribution recovered from mock dataset (Fig.C4). As a comparison, the KDE based PDF of the b-$N_{\mathrm{H\,I}}$ distribution of the model is shown in Fig.C1.

In the end, we perform inference test on our toy model for both 3D and 2D (without $\Gamma_{\mathrm{H\,I}}$) models to test the robustness of our whole inference pipeline following the method discussed in §3.5, and the results are given in table C1, showing that our inference algorithm passes the inference test perfectly for an idealized model. Moreover, the inference on toy model of b-$N_{\mathrm{H\,I}}$ distribution performs slightly better than on Nyx simulation (see Appendix C). The reason could be that the toy model b-$N_{\mathrm{H\,I}}$ distributions are 2D Gaussian distributions that solely depends on the thermal parameters T_{mock} and γ_{mock}, which is equivalent to say that the b-$N_{\mathrm{H\,I}}$ distribution fully preserved the thermal information of the IGM, however, in the Nyx simulation the b-$N_{\mathrm{H\,I}}$ distributions are affected by the complex astrophysical processes in the diffuse IGM, resulting in the loss of the thermal information.

Combining all results shown above, we conclude that our inference algorithm is able to recover the mock parameters with extraordinary

Table C1. Table of results of the inference test for the toy model

models	Total	68(%)	95 (%)
3D toy model	240	165 (68.75 ± 2.92%)	225 (93.75 ± 1.67%)
2D toy model	300	199 (66.33 ± 2.67%)	284 (94.67 ± 1.33%)

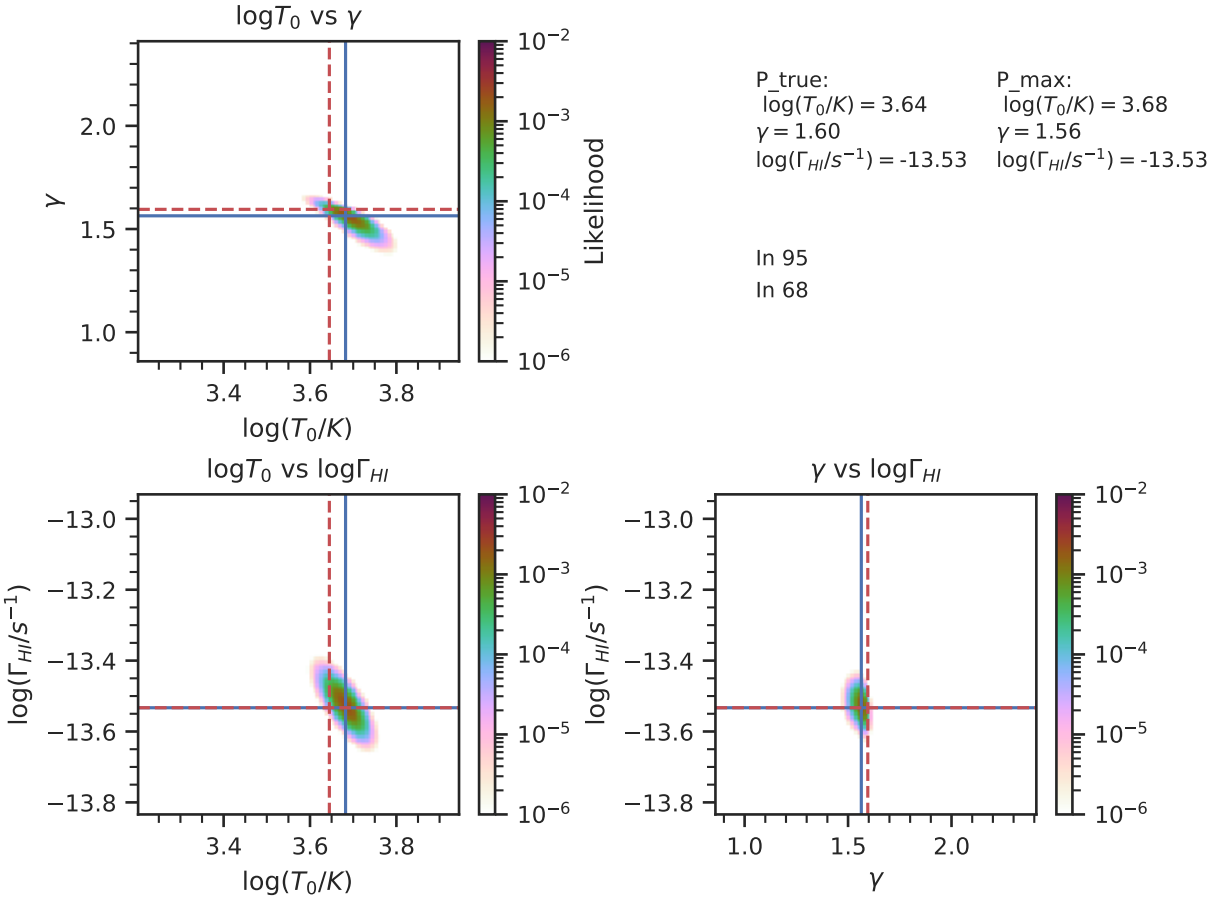

Figure B1. Example of the distribution of the likelihood function sliced at the location of the true parameters ($P_{\rm true}$, indicated by red dashed lines). The parameters corresponding to the maximum likelihood model $P_{\rm max}$ are indicated by blue solid lines. Values of both $P_{\rm true}$ and $P_{\rm max}$ are given in the up right. Calculation implies that the true parameters are in both the effective 1σ 68%) and 2σ (95%) intervals.

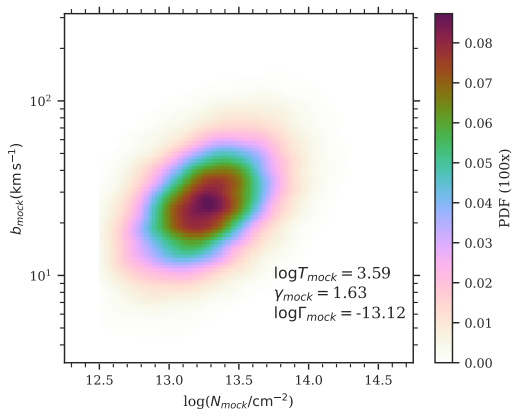

Figure C1. The KDE based PDF of b-$N_{\rm H\,I}$ distribution of one of the toy models which is a 2D Gaussian distribution parameterized by $T_{\rm mock}$, $\gamma_{\rm mock}$ and $\Gamma_{\rm mock}$ in analogy with thermal parameters T_0, γ , $\Gamma_{\rm H\,I}$. The parameters of the toy model is shown in the right bottom corner of the plot. For illustration purposes, values of pdf are multiplied by 100 in the color bar.

accuracy under idealized condition, and our entire pipeline including b-$N_{\rm H\,I}$ distribution emulation, dN/dz emulation, likelihood function and inference pipeline is robust.

APPENDIX D: CONVERGENCE TEST

Lukić et al. (2015) demonstrated that the b parameter of Lyα forest is sensitive to the simulation resolution, and its distribution converges for simulation finer than *L10N512* simulation (i.e., box size $L = 10h^{-1}$ Mpc and $N = 512^3$ dark matter particles and baryon grids which gives the resolution of 20 h^{-1} kpc) while the box size itself does not affect line parameters of the Lyα forest. Whereas above mentioned tests are done at redshift $\sim$ 3, it is worthy to further investigate impact of the boxsize and resolution of the simulation on the Lyα forest at lower redshifts, since the nonlinear evolution at low redshift can affect the Lyα forest.

Here, we perform a convergence test at redshift $z = 0.5$ to to check if our results are independent of the simulation box-size at low redshift. To test the convergence we use two Nyx boxes; *L20N1024* (box-size = 20 h^{-1} Mpc, $N = 1024^3$ dark matter particles and baryon grids i.e resolution of 20 h^{-1} kpc), and *L100N4096* (box-size 100 h^{1-} Mpc and $N = 4096^3$ dark matter particles and baryon grids, resolution of 24 h^{-1} kpc). These two simulation boxes are ran following the same procedures given in section §2. In Fig. D1, we plot the temperature T, overdensity Δ, and velocity along line-of-sight v_{los} of these two simulations. We can see the distributions of T and Δ are alike for both while the small box *L20N1024* simulation has much smaller line-of-sight velocity. This is expected since line-of-sight velocities are dominated by the large scale modes that exist only in the large box simulations. However, these large velocities are

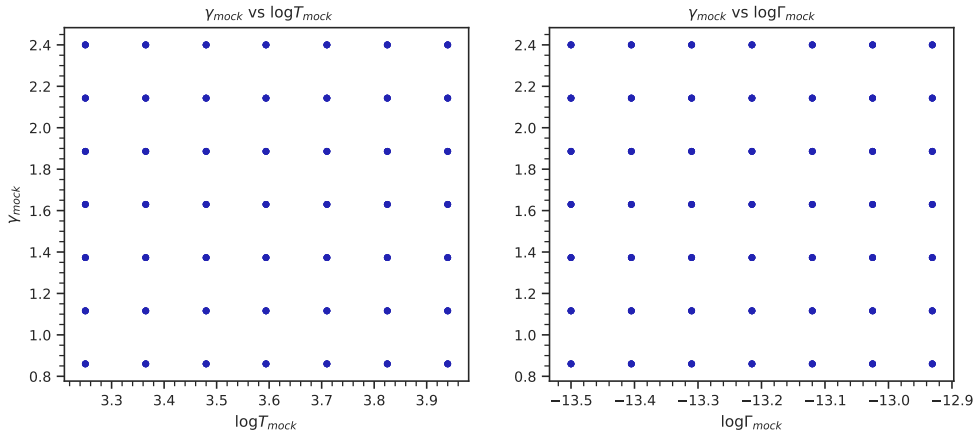

Figure C2. The thermal grid used in our toy model. The left-hand panel is the γ - T_0 grid and the right-hand panel is γ - $\Gamma_{\rm H\,I}$ slice showing the 7 $\Gamma_{\rm H\,I}$ values we have for each point on the 2D γ - T_0 grid.

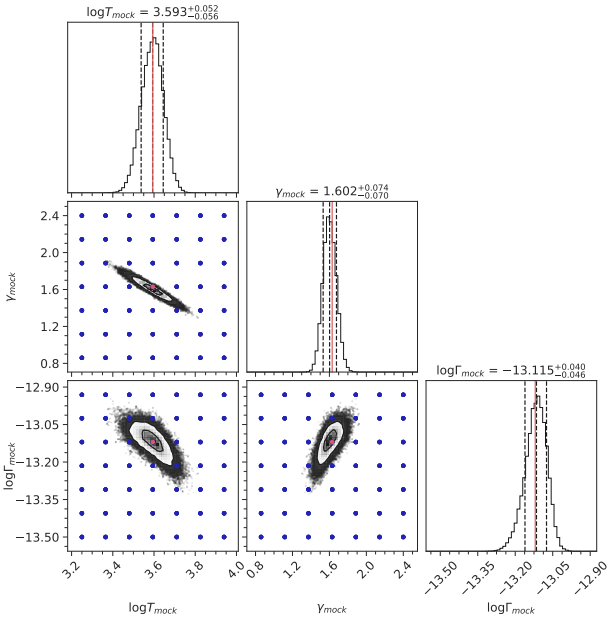

Figure C3. MCMC posterior for the fit of the b-$N_{\rm H\,I}$ distribution from one of the toy models (absorbers shown as black points in Fig.C4) using the likelihood function (Eq. 7) from DELFI and our Gaussian emulator (see § 3.2). Projections of the thermal grid used for generating models are shown as blue circles. Inner(outer) black contour represents the projected 2D 1(2)-sigma interval. The parameters of true model are indicated by red lines in the corner plot, while the dashed black lines indicates the 16, 50, and 84 percentile value of the posterior. The true parameters are: $\log T_{\rm mock}$ = 3.59, $\gamma_{\rm mock}$ = 1.63 and $\log \Gamma_{\rm mock}$ = −13.12.

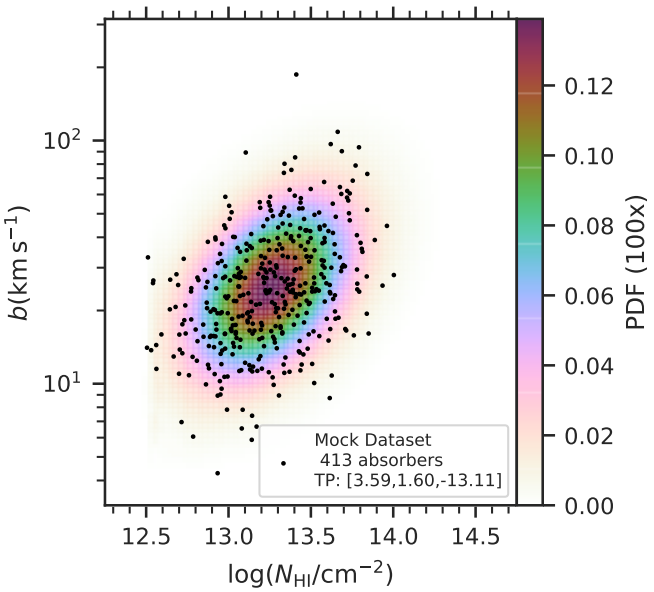

Figure C4. The 'best fit' model b-$N_{\rm H\,I}$ distribution for the Gaussian toy model emulated by DELFI. It is emulated based on the best-fit parameters (median values of the marginalized MCMC posterior), which is shown in the right bottom corner of the plot. The true parameters are: log($T_{\rm mock}$ = 3.59, $\gamma_{\rm mock}$ = 1.63 and log($\Gamma_{\rm mock}$ = -13.12. For illustration purposes, values of pdf are multiplied by 100 in the color bar.

because of bulk motion and therefore do not affect the parameters of the Lyα forest lines.

For both simulations, we follow the forward modeling and line fitting procedures discussed in Section §2, except that here we use a Gaussian LSF with fixed resolution R=3.5 km/s and assume a SNR=100. Such choices of resolution and SNR assure that the Lyα forest are fully resolved and the box-size effect are independent of resolution and instrument. For both simulations, we use the photoionization rate $\log \Gamma_{\rm H\,I}(\rm s^{-1}) = -13.308$. The 1D marginalized distributions of Doppler parameter b and column density $N_{\rm H\,I}$ of both simulations are presented in Fig. D2. The $N_{\rm H\,I}$ distribution of the two simulations are in excellent agreement with each other, with the relative difference $\Delta_{P(N)} < 10\%$. The b parameter have very similar distributions for both simulations, where the two distributions agree with each other near the peak, with relative difference $\Delta_{P(b)} < 25\%$, and the difference increases as $\log b$ becomes smaller than 1.0 or larger than 2.0, which however have very small contribution in the total cumulative distribution. The peak values of the b parameter for both simulations are given in Fig. D2, where the b distributions

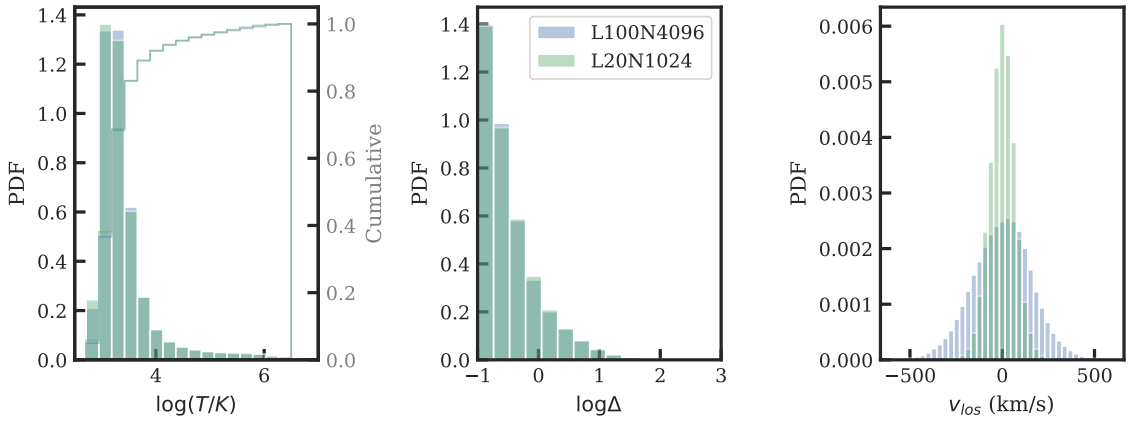

Figure D1. From left to right, the 1D marginalized distribution of the temperature T, overdensity Δ, and velocity along line-of-sight $v_{\rm los}$. The unfilled histogram in the left most panel shows the CDF of the temperature distribution. The *L100N4096* box are shown in blue, while the *L20N1024* box are shown in green.

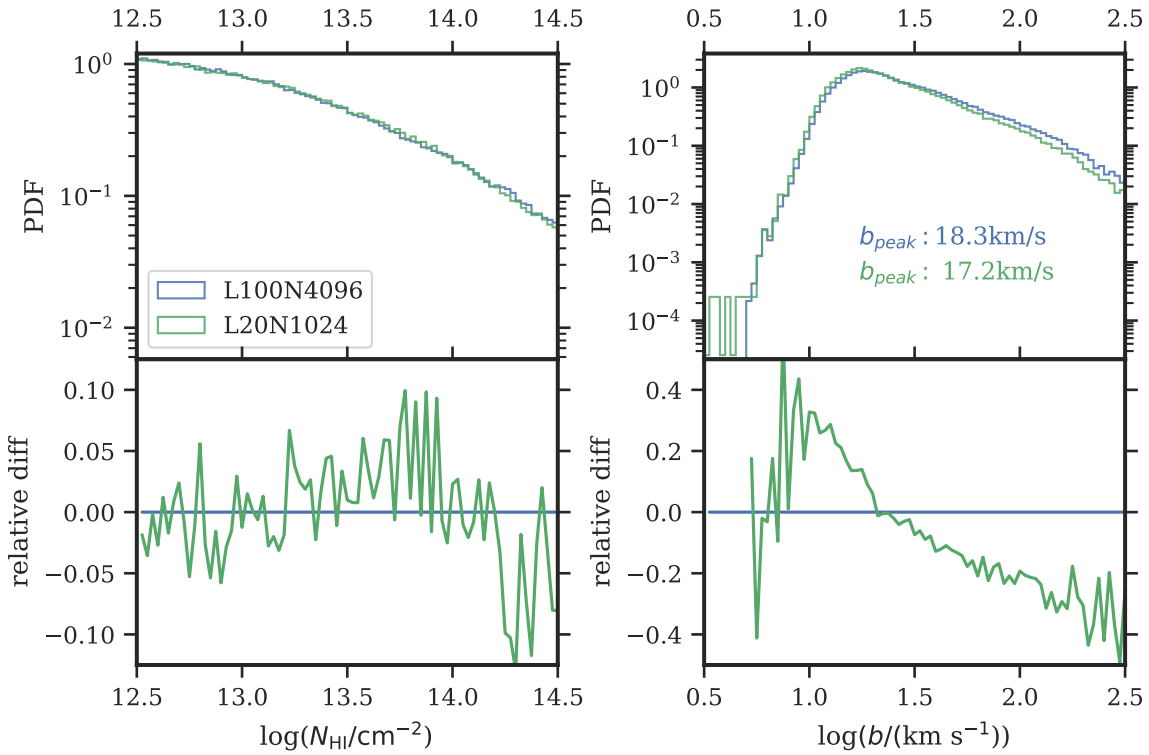

Figure D2. The 1D marginalized $N_{\rm HI}$(left) and b(right) parameters of the two simulations. The relative differences are shown in the bottom panels. The *L100N4096* box is shown in blue, while the *L20N1024* box is shown in green. The peak of b parameters are given in the text.

give $b_{\rm peak}$ = 18.3 km/s and 17.2km/s for *L100N4096* and *L20N1024* simulations respectively. We count the dN/dz for both simulations, *L20N1024* gives dN/dz= 750, and *L100N4096* gives dN/dz= 700. The difference in dN/dz is about 7%. Furthermore, we plot the 2D b-$N_{\rm H_I}$ distribution in Fig. D3 for both simulations. These are 2D KDE maps each generated by 20000 data points collected from the $\{b, N_{\rm H_I}\}$ dataset following the procedures described in Section §2. In the right most panel of Fig. D3, we plot the relative difference of the KDE map, given by $\Delta_P = (P_{L100}/P_{L20} - 1)$, where the P_{L100} and P_{L20} stand for the KDE for *L100N4096* and *L20N1024* simulations respectively. To avoid division by zero, we apply a small threshold and only include regions with $P_{L20} > P_{\rm TH}$, where $\int_{P_{\rm TH}}^{\infty} P{\rm d}P = 75\%$. We quantify the overall relative difference by calculating the root mean square and standard deviation of the Δ_P. As shown in Fig. D3, the relative differences in the 2D b-$N_{\rm H_I}$ distribution are small and only about 5%. Therefore we conclude, even at $z \sim 0.5$ box-sizes do not affect the parameters of Lyα forest significantly.

This paper has been typeset from a TEX/LATEX file prepared by the author.

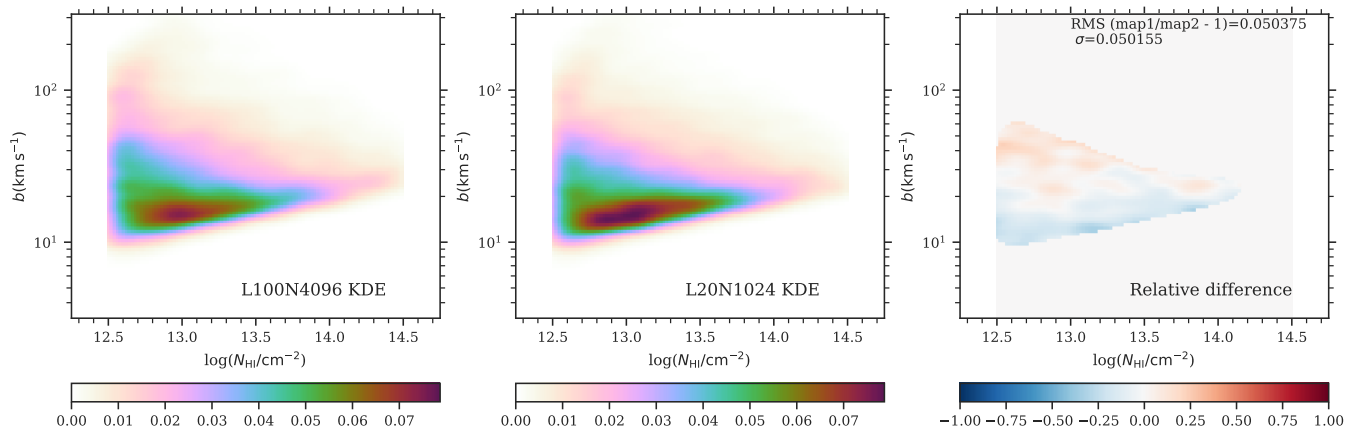

Figure D3. The 2D KDE maps of b-$N_{\rm H\,I}$ distributions for simulations *L100N4096* (left-hand panel) and *L20N1024* (middle panel). The right-hand panel shows the relative difference $\Delta_P = (P_{L100}/P_{L20} - 1)$. To avoid division by zero, we apply a threshold and only include regions integrating up to 75% for P_{L20}. The KDE maps are made from 20000 data points for each simulation, and the RMS and standard deviation of of the relative difference map are given in the right-hand panel. Details of the calculations are given in the text.