



On the Physical Nature of Ly α Transmission Spikes in High-redshift Quasar Spectra

Hanjue Zhu (朱涵珏)¹ , Nickolay Y. Gnedin^{1,2,3} , and Camille Avestruz^{4,5} ¹ Department of Astronomy & Astrophysics, The University of Chicago, Chicago, IL 60637, USA; hanejuezhu@uchicago.edu² Theory Division, Fermi National Accelerator Laboratory, Batavia, IL 60510, USA³ Kavli Institute for Cosmological Physics, The University of Chicago, Chicago, IL 60637, USA⁴ Department of Physics, University of Michigan, Ann Arbor, MI 48109, USA⁵ Leinweber Center for Theoretical Physics, University of Michigan, Ann Arbor, MI 48109, USA

Received 2024 January 8; revised 2024 August 22; accepted 2024 September 8; published 2024 October 29

Abstract

We investigate Lyman-alpha (Ly α) transmission spikes at $5.2 < z < 6.8$ using synthetic quasar spectra from the “Cosmic Reionization on Computers” simulations. We focus on understanding the relationship between these spikes and the properties of the intergalactic medium (IGM). Disentangling the complex interplay between IGM physics and the influence of galaxies on the generation of these spikes presents a significant challenge. To address this, we employ Explainable Boosting machines, an interpretable machine learning algorithm, to quantify the relative impact of various IGM properties on the Ly α flux. Our findings reveal that gas density is the primary factor influencing absorption strength, followed by the intensity of background radiation and the temperature of the IGM. Ionizing radiation from local sources (i.e., galaxies) appears to have a minimal effect on Ly α flux. The simulations show that transmission spikes predominantly occur in regions of low gas density. Our results challenge recent observational studies suggesting the origin of these spikes in regions with enhanced radiation. We demonstrate that Ly α transmission spikes are largely a product of the large-scale structure, of which galaxies are biased tracers.

Unified Astronomy Thesaurus concepts: [Intergalactic medium \(813\)](#); [Reionization \(1383\)](#); [Ly \$\alpha\$ forest \(980\)](#)

1. Introduction

Observations of the Lyman-alpha (Ly α) forest in quasar spectra effectively map hydrogen gas distribution in the Universe. Simulations from the mid-1990s demonstrated that the Ly α forest at intermediate redshifts ($z \sim 2-4$) manifests the small-scale tail of the cosmic large-scale structure, challenging earlier physical models that attributed the features to astronomical objects such as pressure-confined clouds, shocks, and minihalos (R. Cen 1994; L. Hernquist et al. 1996; Y. Zhang et al. 1997). The Ly α forest has subsequently become a fundamental tool in modern astrophysics, with applications ranging from measuring cosmological parameters and matter clustering to constraining the level of turbulence in the intergalactic medium (IGM; P. Gaikwad et al. 2017; J. S. Bolton et al. 2022), as well as the thermal and ionization state of the gas (L. Hui & N. Y. Gnedin 1997; G. D. Becker et al. 2011; G. D. Becker & J. S. Bolton 2013; E. Boera et al. 2016; K. N. Telikova et al. 2019; M. Walther et al. 2019; P. Gaikwad et al. 2020, 2021).

However, the large Ly α cross section means that even a tiny hydrogen neutral fraction ($x_{\text{H I}} \sim 10^{-4}$) produces complete absorption (H. Bi & A. F. Davidsen 1997; M. Rauch 1998; D. H. Weinberg et al. 2003). As the Universe is more neutral at higher redshifts, the Ly α forest becomes denser. Eventually, somewhere around $z \sim 5$, the appearance of the quasar spectrum changes dramatically—clear absorption lines disappear, replaced by blended absorption features and sporadic “transmission spikes,” i.e., regions of incomplete absorption (see G. D. Becker et al. 2015 for a review). The quasar absorption spectrum now resembles an emission spectrum, yet the transmission spikes do not exhibit the defined shapes typical of Doppler or Voigt profiles

—in fact, the precise shapes and variations of the transmission spikes are still unknown. The origins of these spikes are also unclear: while many of the spikes likely arise from low-density regions (where there is minimum absorption) (R. A. C. Croft et al. 1998; J. Miralda-Escudé et al. 2000; M. S. Peebles et al. 2010), it is uncertain if they are exclusively associated with cosmic voids (E. Garaldi et al. 2019; P. Gaikwad et al. 2020; L. C. Keating et al. 2020; F. Nasir & A. D’Aloisio 2020). Moreover, the spatial correlation between the transmission spikes and galaxies is unclear. While galaxies enhance local ionizing radiation, leading to decreased absorption (K. L. Adelberger et al. 2003; S. Cantalupo et al. 2012), they also live in denser regions where absorption is naturally higher (H. J. Mo & S. D. M. White 1996; O. Rakic et al. 2012). Despite this, contradictory evidence exists in current observational data, which shows decreased absorption near galaxies. (e.g., K. Kakiichi et al. 2018; R. A. Meyer et al. 2019). However, it is important to note that these observations are sensitivity-limited and potentially missing low surface brightness galaxies that could significantly contribute to the ionizing photon budget.

Our understanding of the Ly α forest is rather comprehensive at $z < 5$, yet it becomes markedly limited beyond $z \sim 5$. The influx of high-quality data from JWST intensifies this discrepancy. In the near future, 30 m class telescopes will begin observing fainter quasars in fields with sufficiently deep JWST exposures, and it is anticipated that numerous synergistic observational programs will emerge (G. Becker et al. 2019; A. Cooray et al. 2019; S. Furlanetto et al. 2019; P. La Plante et al. 2019; M. Rieke et al. 2019). Such programs could provide valuable observational constraints and enhance our understanding of the connection between galaxies and the IGM. However, these detailed observational data risk being underleveraged without a proper theoretical interpretation.

Fortunately, reliable theoretical and computational tools are now available to enhance our understanding of high-redshift quasar spectra significantly. State-of-the-art cosmological

simulations of cosmic reionization produce reasonably realistic models in volumes in excess of 100 cMpc , providing the foundation for creating high-resolution synthetic quasar spectra (N. Y. Gnedin 2014; P. Ocvirk et al. 2016, 2020; R. Kannan et al. 2022; E. Garaldi et al. 2024). The ‘‘Cosmic Reionization on Computers’’ (CROC) project (N. Y. Gnedin & A. A. Kaurov 2014; N. Y. Gnedin 2014) is one of them. In this work, we rely on CROC simulations to provide a reasonably realistic model for the high-redshift IGM. Furthermore, this study incorporates the use of Explainable Boosting Machines (EBMs), an enhanced iteration of generalized additive models (GAMs) as implemented by Y. Lou et al. (2013). GAMs are built on the principle that the model’s output is the aggregate of individual contributions from each feature, offering interpretability by revealing feature significance in predicting outcomes. EBMs enhance this interpretability by integrating modern machine learning techniques to boost performance without compromising the ability to understand the models’ decisions. The synergy of advanced reionization simulations with cutting-edge machine learning tools equips us to identify the properties of the environments from which transmission spikes in $z > 5$ quasar spectra originate.

This paper is organized as follows. Section 2 describes the CROC simulations, synthetic $\text{Ly}\alpha$ spectra generation, and EBM methodology. In Section 3, we examine model performances with different sets of IGM properties as inputs, as well as the average contribution of each parameter to the target quantity ($\text{Ly}\alpha$ transmission flux). We summarize and discuss these results in Section 4.

2. Methodology

2.1. CROC Simulations

In this study, we employ simulations from the CROC project, a suite of cosmological hydrodynamic simulations of cosmic reionization. For detailed information on the simulations, we direct readers to the CROC methods paper (N. Y. Gnedin 2014). We utilize three independent realizations of the $40 h^{-1} \text{ Mpc}$ comoving (cMpc) simulation boxes, each offering a spatial resolution of 100 pc in proper units. This setup enables the accurate modeling of the IGM properties and yields ample independent LOS data for our analyses. By adopting different ‘‘DC-mode’’ values (N. Y. Gnedin et al. 2011) for different independent realizations, the CROC simulations model varied reionization histories in separate simulation boxes, which allows us to investigate the impact of reionization history on our findings.

In this paper, we use three different CROC simulations that we label ‘‘early,’’ ‘‘intermediate,’’ and ‘‘late reionization.’’ These three simulations sample the full range of reionization histories from six independent random realizations of initial conditions. Hence, the ‘‘late’’ and ‘‘early’’ models roughly correspond to a $\pm 1\sigma$ spread in possible reionization histories, while the ‘‘intermediate reionization’’ history is close to the cosmic mean. In Figure 1, we show the distribution of mean opacities obtained from $50 h^{-1} \text{ cMpc}$ LOSs from all three simulations and compare them to observational data. Collectively, these simulations provide a marginal fit to the observational data—unfortunately, no other reionization simulation currently offers a better fit. The CROC simulations also match the observed distribution of long gaps in quasar spectra (N. Y. Gnedin 2022), but only under the ‘‘late reionization’’ model. In other words, while neither CROC

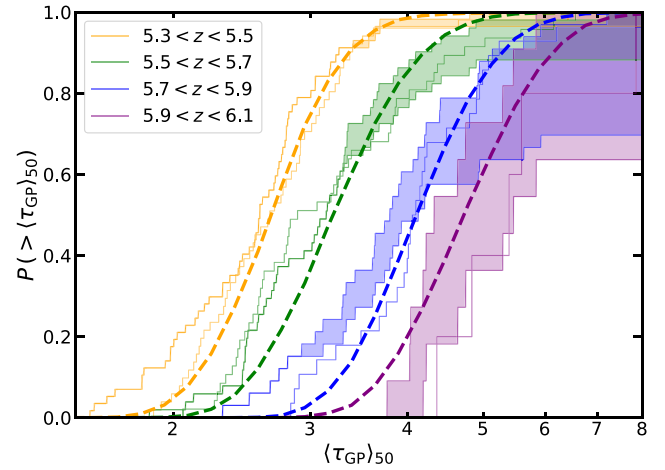


Figure 1. Distribution of mean opacities in $50 h^{-1} \text{ cMpc}$ LOSs for the combination of the three CROC simulations (thick dashed lines). Thin solid lines and bands show the observational data from G. D. Becker et al. (2015) and S. E. I. Bosman et al. (2018). The three CROC simulations together marginally match the distribution of opacities (at least at $z > 5.5$).

nor any other existing reionization simulation can match all the observational data, some of the CROC boxes do meet certain observational constraints. Hence, we have some confidence that CROC captures the key physical processes, justifying our use of these simulations despite known limitations in how precisely we capture all details of reionization.

2.2. Synthetic $\text{Ly}\alpha$ Absorption Spectra

In comoving space, $\text{Ly}\alpha$ optical depth is obtained by integrating along the LOS:

$$\tau(\lambda) = \sigma_0 \int n_{\text{HI}}(x) \frac{c}{\sqrt{\pi} b_x} e^{-\frac{(u_\lambda - u_x)^2}{b_x^2}} \frac{dx}{1 + z_x}, \quad (1)$$

where σ_0 is the cross section, n_{HI} is the H I number density, c is the speed of light, b_x is the Doppler parameter at position x along the LOS, u_λ is the velocity corresponding to the observed wavelength λ , u_x is the gas velocity at x , and z_x is the redshift at x . Then, the flux is

$$F(\lambda) = \exp(-\tau(\lambda)). \quad (2)$$

We note that u_λ and u_x include both the Hubble flow and gas peculiar velocity.

We adopt two critical simplifications in this first study. Peculiar velocities induce a surjective but noninjective mapping from comoving space to velocity space, thereby obscuring the direct association between gas properties (sampled in comoving space) and observed flux (in velocity space). To avoid this additional complication and focus on the fundamental physical connection between gas properties and transmission spikes, our first simplification is to use synthetic $\text{Ly}\alpha$ absorption spectra in ‘‘real space,’’ i.e., assuming a zero peculiar velocity ($v_{\text{pec}} = 0$). This approach is equivalent to using the ‘‘fluctuating Gunn–Peterson approximation’’ (L. Hui et al. 1997; R. A. C. Croft et al. 1998; D. H. Weinberg et al. 2003), which has been instrumental in understanding the $\text{Ly}\alpha$ forest at lower redshifts. While it is crucial to account for peculiar velocities for interpreting real observational data, our current focus is on the theoretical investigation of the physical causes behind transmission spikes. Therefore, we defer the inclusion of peculiar velocities to follow-up works.

Our second simplification is using a Doppler profile instead of the correct Voigt profile for convolving spectra. This choice effectively excludes nonlocal effects from high-column density damped Ly α systems (DLAs). The Lorentzian wings of a DLA can impact absorption far from its actual location in velocity space, thereby obscuring the relationship between the transmitted flux and the physical properties of the IGM. Similarly, applying the correct profile is essential for correctly interpreting observational data, a step we plan for future work.

As our data sample, we generate 100,000 LOSs at each redshift, randomly oriented to sample the full volume of the simulation boxes. We note that the results shown in this paper converge when using only a tenth of these data. Along these LOS, we record the Ly α flux and the physical properties of the gas, as well as the intensity of the radiation field. We provide the details in Section 3.1.

2.3. EBMs

EBMs provide a fitted relationship between the target quantity y and the features $\theta \in \mathbb{R}^n$. They are specifically designed to be interpretable, i.e., the dependence of the target quantity on the features is given in explicit functional forms. Mathematically, we have

$$y(\theta) = y_0 + \sum_{i=0}^{n_p-1} f_y^i(\theta_i) + \sum_{i=0, i \neq j}^{n_p-1} \sum_{j=0}^{n_p-1} f_y^{ij}(\theta_i, \theta_j), \quad (3)$$

where $y(\theta)$ is the predicted value of the target quantity y given n_p features θ , y_0 is the baseline (see the mean) value of the target quantity y , and f_y^i and f_y^{ij} are piecewise one- and two-dimensional functions, respectively. The magnitudes of f_y^i and f_y^{ij} indicate the relative importance of each feature in predicting the target quantity. We refer to f_y^i as feature functions and f_y^{ij} as interaction functions. An example of employing EBMs in astrophysics is presented in R. Hausen et al. (2023), where EBM models reveal the relative importance of different dark matter halo properties in setting galaxy stellar mass and star formation rate. We evaluate the EBM performance by computing the r^2 variance metric:

$$r^2 = 1 - \frac{\sum_{i=0}^{N-1} (y_i - \hat{y}_i)^2}{\sum_{i=0}^{N-1} (y_i - \bar{y})^2}, \quad (4)$$

where N is the number of objects, y_i is the true value of the target quantity for object i , and \hat{y}_i is the predicted value from the model for object i . r^2 measures the extent to which the variance in the actual outcomes is captured by the model. A higher r^2 value indicates that the model's predictions more closely match the actual data.

In terms of the training procedure, we use the InterpretML (H. Nori et al. 2019) implementation of EBMs, and adopt most of the default hyperparameter values for InterpretML version 0.4.4., except for ‘‘max bins,’’ which we set to 4096, after performing a grid search for optimal model performance using the allocated subset of data for training.

We adopt a standard 80/20 train-test split, allocating 80% of our data set for training the models and the remaining 20% for assessing their performance.

3. Results

3.1. Predictive Power of Models

Our first objective is to train the EBM models to accurately predict the Ly α flux. Given that EBMs reveal the relative contribution of each input on the prediction outcome, we can quantify the importance of individual physical properties in generating the spectra. We first investigate a range of physical properties as inputs for training our EBM models based on the physical relationships between the inputs. In Section 3.4, we use EBM to explicitly quantify the relative importance of each feature.

At first glance, Equations (1) and (2) suggest a direct dependence of flux on HI number density (n_{HI}) and temperature (T), as the Doppler parameter is a function of T . Consequently, we anticipate that using n_{HI} and T as inputs for the EBMs will yield a highly accurate prediction of the flux. We convolve n_{HI} along the LOS using the Gaussian profile e^{-x^2/b_x^2} , where $b_x(T)$ is the temperature-dependent Doppler width. This convolution accounts for local temperature variations, employing a Gaussian profile specific to the temperature at each point along the LOS. In Figure 2, we show the $1 - r^2$ values on a logarithmic scale as a function of redshift. Uncertainties associated with these values, obtained after training with three different sets of LOSs, are comparable to the width of the line on the plot for $z \leq 6.4$, and slightly larger than the width of the plotted line beyond this redshift. The three panels represent different boxes with different reionization histories. We first only focus on red lines, which show the model predictions with $\{n_{\text{HI}}, T\}$ as inputs. Consistent with our expectations, in all three boxes, the model predictions with $\{n_{\text{HI}}, T\}$ inputs are the best performing. A $1 - r^2$ value of 10^{-3} means that 0.1% of the total variance in the target quantity is not captured by the model. We find larger discrepancies between our predictions and the test data primarily at the peak regions of transmission spikes. These discrepancies stem from the limitations of the EBM algorithm, which models the target quantity using piecewise constant functions and thus loses accuracy in regions with sparse data points.

3.2. Role of Ionizing Radiation

The HI number density is affected by the baryon number density (n_b) and the prevailing ionizing radiation in the Universe. Similar to what we did to n_{HI} , we also convolve n_b along the LOS, applying a temperature-dependent Gaussian profile that varies based on the temperature at each point. As a baseline, we first use the baryon number density (n_b) and temperature (T) as inputs. The light blue lines in Figure 1 represent models trained using Gaussian convolved $\{n_b, T\}$ inputs. The quality of the predictions is much lower compared to the $\{n_{\text{HI}}, T\}$ input case. This difference in r^2 values illustrates the important role of ionizing radiation in producing the Ly α flux. This leads to the question: How can we quantify ionizing radiation at each point along the LOS? In the CROC simulations, the radiative transfer is implemented using the OTVET method (N. Y. Gnedin 2014). To summarize, in the simulations, the radiation energy density E_ν is computed as

$$E_\nu = \bar{E}_\nu \mathcal{F}_\nu + (\mathcal{G}_\nu - \bar{\mathcal{G}}_\nu \mathcal{F}_\nu), \quad (5)$$

where \bar{E}_ν , a constant, is the spatial average (i.e., the cosmic background). The term \mathcal{G}_ν accounts for the local ionizing sources inside the simulation box, and can be written as the

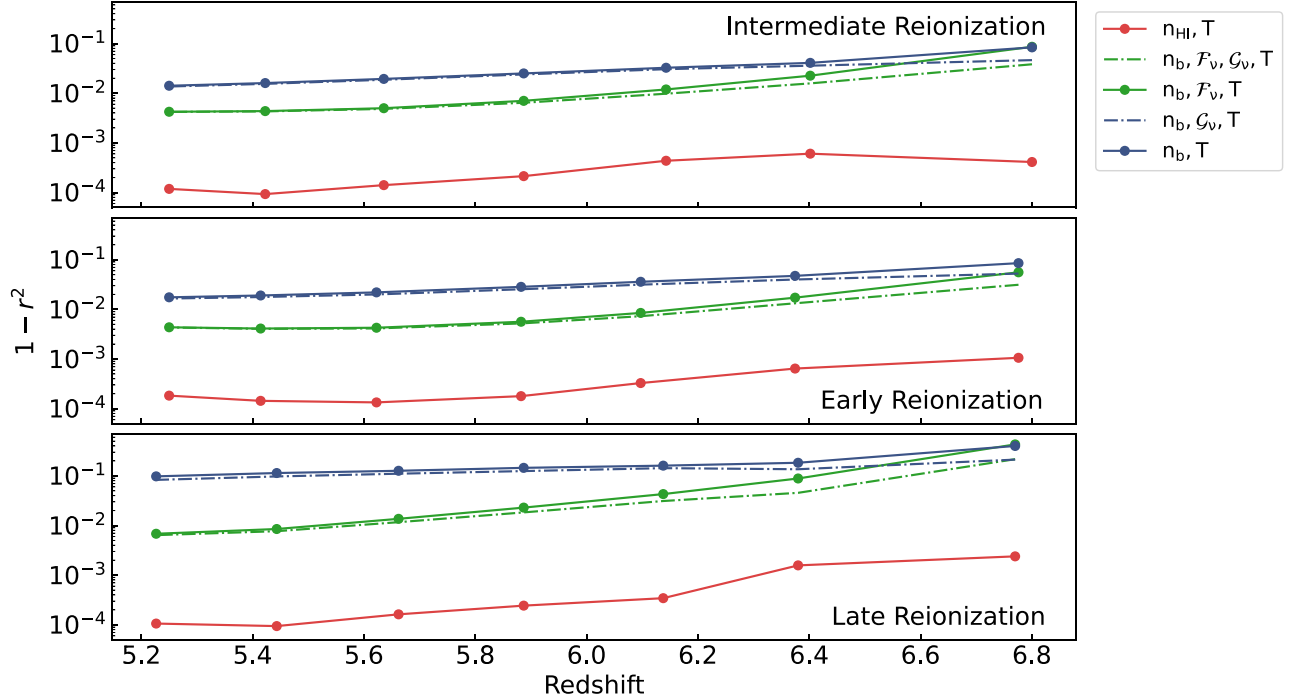


Figure 2. EBM model performance, measured using the $1 - r^2$ “missing-variance” metric, as a function of redshift for models trained with different input combinations. Each panel corresponds to a different reionization history. As a baseline, with H I number density (n_{HI}) and temperature (T) as inputs, the model predictions are highly accurate. Then n_{HI} is replaced by combinations of baryon number density (n_b) with ionization parameters, \mathcal{F}_v (representing the cosmic background) and \mathcal{G}_v (representing the local ionizing sources). The results indicate that \mathcal{F}_v is the predominant component of ionizing radiation.

sum of fluxes from all sources inside the box,

$$\mathcal{G}_v(\mathbf{x}) = \frac{1}{4\pi c a^2} \sum_i \frac{L_{i,v}}{|\mathbf{x} - \mathbf{x}_i|^2} e^{-\tau_v(\mathbf{x}, \mathbf{x}_i)},$$

where \mathbf{x}_i and $L_{i,v}$ are the comoving location and the luminosity of source i inside the simulation box, and $\tau_v(\mathbf{x}, \mathbf{x}_i)$ is the optical depth between the spatial locations \mathbf{x} and \mathbf{x}_i .

$\mathcal{F}_v(\mathbf{x})$ is the angle average of $f_v(\mathbf{x}, \mathbf{n})$ that satisfies the following equation:

$$\frac{a}{c} \frac{\partial f_v}{\partial t} + \vec{n} \frac{\partial f_v}{\partial \vec{x}} = -k_v f_v + f_v \langle k_v f_v \rangle, \quad (6)$$

where $k_v(\mathbf{x})$ is the absorption coefficient and

$$\mathcal{F}_v(\mathbf{x}) \equiv \frac{1}{4\pi} \int d\Omega f_v(\mathbf{x}, \mathbf{n}). \quad (7)$$

Physically, $\mathcal{F}_v - 1$ can be interpreted as the fluctuation in the “far radiation” (i.e., radiation outside the simulation volume) and $\langle \mathcal{F}_v \rangle = 1$.

To account for ionizing radiation in our model, we have included \mathcal{F}_v and \mathcal{G}_v in the input parameters. This is depicted in Figure 2, where the $\{\mathcal{F}_v, \mathcal{G}_v, n_b, T\}$ input case is represented by dark green lines. To separate the contribution from \mathcal{F}_v and \mathcal{G}_v , we also use $\{\mathcal{F}_v, n_b, T\}$ and $\{\mathcal{G}_v, n_b, T\}$ as inputs to train the models, illustrated in light green and dark blue lines, respectively. Our analysis indicates that \mathcal{F}_v is the predominant component of ionizing radiation; \mathcal{G}_v moderately improves model performance at early times. This finding is significant, as previous studies primarily linked transmission spikes to local ionizing sources such as galaxies. In contrast, our study, which integrates numerical simulations with an interpretable machine learning algorithm, suggests that density is a key factor in generating these spikes, and that background radiation plays a

more crucial role than local ionizing sources. We also note that as redshift increases, the performance of the EBM models decreases. This trend is linked to the increasing IGM opacity at higher redshifts, which makes transmission spikes less frequent and requires a larger number of sightlines for effective model training. However, we are already using more sightlines than could be observed even with future 30 m class telescopes. We therefore limit this analysis to our existing number of sightlines to focus on physical interpretation instead of generating more for the sake of improved model performance.

Additionally, we note variations in the performance of the $\{n_b, \dots\}$ models trained and evaluated on data from boxes with different reionization histories. The models based on the intermediate and early reionization boxes yield comparable results across the entire redshift range considered. However, the late reionization box shows relatively worse model performance, particularly at higher redshifts. This trend is consistent with our expectations: in the late reionization scenario, particularly at $z > 6$, the Universe is still going through rapid reionization, presenting a more complex environment for EBMs to model accurately. Regardless of the reionization scenario, the effects of the cosmic background radiation on Ly α transmission peaks are dominant to those from local ionizing sources.

3.3. Nonequilibrium Effects

In theory, the three physical quantities \mathcal{F}_v , \mathcal{G}_v , and n_b should collectively determine n_{HI} . Therefore, we would expect a model trained with $\{\mathcal{F}_v, \mathcal{G}_v, n_b, T\}$ to have the same level of performance as one trained with $\{n_{\text{HI}}, T\}$ inputs. However, as shown in Figure 2, the $\{\mathcal{F}_v, \mathcal{G}_v, n_b, T\}$ model consistently exhibits lower r^2 values across all redshifts. Moreover, the ratio of r^2 values between the $\{\mathcal{F}_v, \mathcal{G}_v, n_b, T\}$ model and $\{n_{\text{HI}}, T\}$

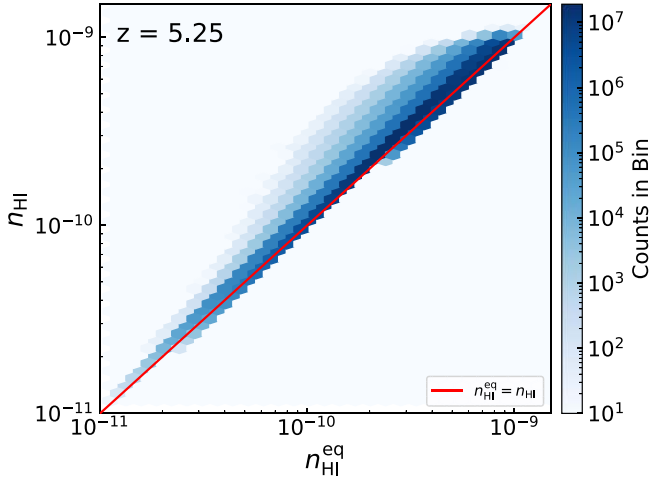


Figure 3. n_{HI} , sampled along LOSs vs. $n_{\text{HI}}^{\text{eq}}$, calculated with Equation (8) assuming ionization equilibrium at $z = 5.25$. Colors show the number counts in each hexbin. We observe significant deviations of many data points from the equilibrium line.

model decreases with increasing redshift, indicating a decline in the $\{\mathcal{F}_\nu, \mathcal{G}_\nu, n_b, T\}$ model performance at higher redshifts.

We posit that the decrease in model prediction power is due to nonequilibrium effects that the EBM model cannot capture. Under the assumption of ionization equilibrium, we can relate n_{HI} to n_b using the following equation:

$$n_{\text{HI}}^{\text{eq}} = \frac{R(T)n_e n_{\text{HII}}}{\Gamma}, \quad (8)$$

where $R(T)$ is the recombination coefficient, n_e ($\propto n_b$) is the electron number density, n_{HII} is the H II number density, and Γ is the photoionization rate. However, the straightforward mathematical relation does not hold when residual nonequilibrium effects from cosmic reionization influence the neutral hydrogen number density. When the relationship between n_{HI} and n_e does not simply follow Equation (8), the accuracy of the EBM in mapping n_b , \mathcal{F}_ν , and \mathcal{G}_ν to n_{HI} decreases. This argument is consistent with our finding that $\{\mathcal{F}_\nu, \mathcal{G}_\nu, n_b, T\}$ model performance decreases when the correlation coefficient between n_{HI} and $n_{\text{HI}}^{\text{eq}}$ is smaller.

We demonstrate the nonequilibrium effects in Figure 3, which shows the simulated n_{HI} versus $n_{\text{HI}}^{\text{eq}}$ computed using Equation (8). There is a notable deviation in many data points from the equilibrium line. Figure 4 presents an example of nonequilibrium effects along a random LOS. We observe a distinct deviation from equilibrium, highlighted by the ratio of n_{HI} and $n_{\text{HI}}^{\text{eq}}$ displayed in the bottom panel of Figure 4, which coincides with a drop in n_b . This pattern suggests the possible presence of a shock at this location. We expect nonequilibrium effects to be important in shock regions; as the temperature jumps across the shock, it takes time for the neutral fraction to decrease to a new equilibrium value, corresponding to the lower recombination rate at a higher temperature.

A similar comparison at other redshifts does not show any clear redshift trend. At higher redshifts, the reionization is incomplete, so one might expect nonequilibrium effects to become relatively more important with increasing redshift. On the other hand, shocks are stronger at lower redshifts as larger wavelengths become nonlinear, so there is also an argument for the importance of the nonequilibrium effects to increase with

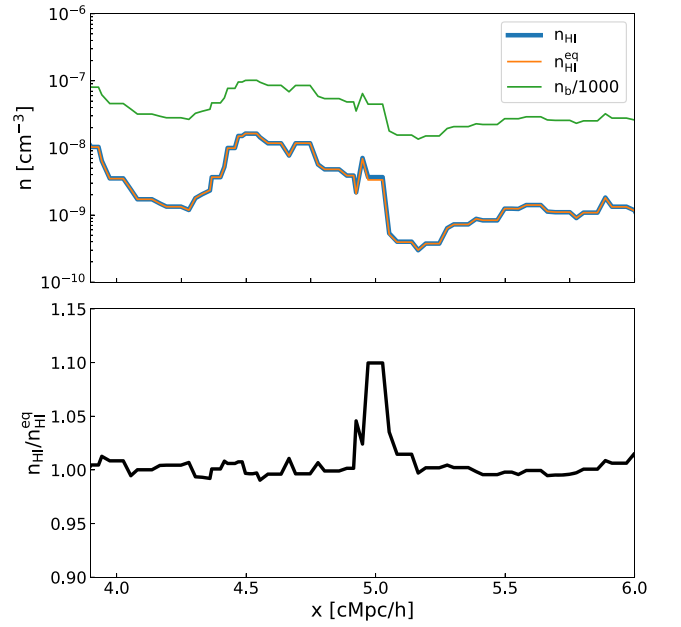


Figure 4. An example of nonequilibrium effects along a random LOS. Top: n_{HI} , $n_{\text{HI}}^{\text{eq}}$ calculated assuming ionization equilibrium and rescaled n_b as a function of LOS position. Bottom: ratio of n_{HI} and $n_{\text{HI}}^{\text{eq}}$ as a function of LOS position. A noticeable deviation from equilibrium coincides with a drop in n_b , suggesting the presence of a shock at this location.

decreasing redshift. In fact, the behavior of the red curve in Figure 2 appears to show that both trends take place.

3.4. Contribution of Each Physical Quantity

As discussed in Section 2.3, the univariate and bivariate functions f_y^i and f_y^{ij} from Equation (3) quantify the contribution of each feature. We can then define a summary statistic, the average contribution of each feature (denoted as \bar{f}_y^i) such that

$$\bar{f}_y^i = \frac{\sum_{j=0}^{N_{\text{bin}}-1} f(\theta_{i,j}) |N_j|}{\sum_{j=0}^{N_{\text{bin}}-1} N_j}, \quad (9)$$

where f represents the feature or interaction function (f_y^i or f_y^{ij}), $\theta_{i,j}$ the value of feature θ_i in the j th bin, N_j the number of samples in bin j , and N_{bin} the total number of bins.

In Figure 5, we revisit the $\{\mathcal{F}_\nu, \mathcal{G}_\nu, n_b, T\}$ input case and show the normalized average contributions (scaled to sum to 1) across three simulation boxes with different reionization histories. Consistently across all redshifts, n_b emerges as a dominant factor in predicting the Ly α flux. In boxes of intermediate and early reionization, the contribution of n_b decreases with increasing redshift, while the influence of \mathcal{G}_ν (the local sources) increases. This pattern is consistent with the expectation that ionizing radiation from local sources has a greater impact at higher redshifts when the Universe is still undergoing substantial reionization. In the late reionization scenario, a similar but nonmonotonic trend is evident, reflecting phenomena (e.g., overlapping ionizing bubbles) characteristic of a universe still undergoing rapid reionization.

The specific forms of functions f_y^i and f_y^{ij} are shown in the Appendix.

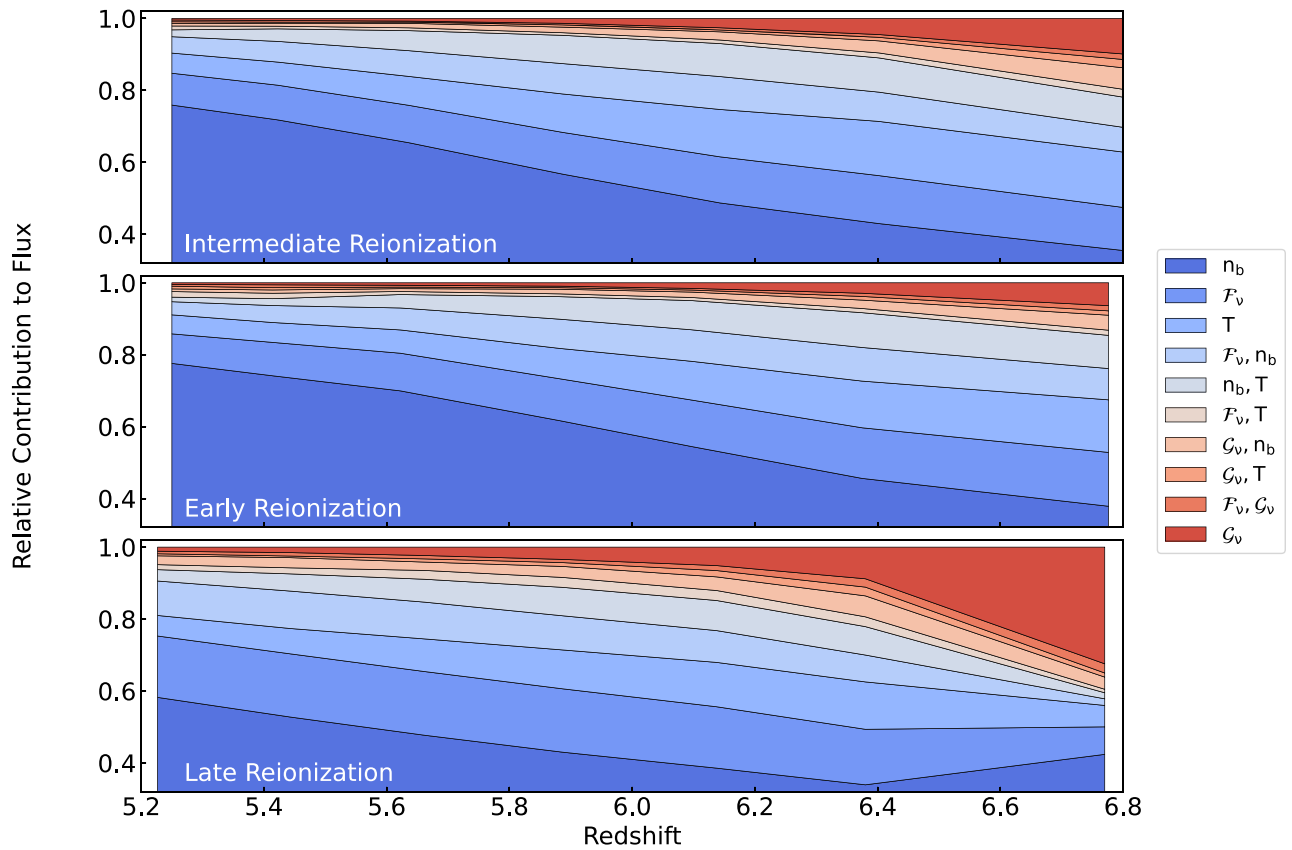


Figure 5. Normalized average contributions (scaled to sum to one) of IGM properties to $\text{Ly}\alpha$ flux prediction. Each panel corresponds to a different reionization history. In intermediate and early reionization scenarios, the contribution of n_b decreases with increasing redshift, while the impact of G_v grows, aligning with the expected greater influence of local ionization at higher redshifts. A similar, yet nonmonotonic trend is observed in the late reionization scenario, indicative of a universe still undergoing rapid reionization.

3.5. Interpreting Observations

Several observational studies have measured the average $\text{Ly}\alpha$ transmission flux as a function of the distance from nearby galaxies (e.g., K. Kakiichi et al. 2018; R. A. Meyer et al. 2019, 2020; H. M. Christenson et al. 2023; D. Kashino et al. 2023). Overall, observations paint a complex and sometimes contradictory picture. The most clearly established trend across all redshifts $z > 5$ is the decrease in the transmitted flux within about 1 pMpc from galaxy locations. The same trend is found in CROC (E. Garaldi et al. 2019) and another similar large-scale simulation project, “THESAN” (E. Garaldi et al. 2022).

Taken at face value, the observed decrease in the transmitted flux is inconsistent with our finding that the radiation from nearby galaxies has a negligible impact on the transmitted flux at the late stages of cosmic reionization. In order to understand this apparent discrepancy, we also generate $\text{Ly}\alpha$ transmission spectra assuming a spatially uniform photoionization rate (Γ), using the global mean of each box at every redshift. These spectra allow us to control for local ionizing sources. Since galaxies influence IGM properties through photoionization, the constant- Γ spectra effectively remove the impact of galaxies on the IGM. If flux suppression still occurs at small distances in our control sample, it would indicate that galaxies are not the primary cause.

Figure 6 shows our measurement of the average $\text{Ly}\alpha$ transmission flux as a function of distance to galaxies, comparing simulated $\text{Ly}\alpha$ transmission spectra (in black) with spectra generated assuming a constant Γ (in red). The left panel

considers galaxies only if their virial radii (R_{vir}) exceed the radial distance from the dark matter halo center to the LOS, whereas in the right panel, we relax this criterion to $2 \times R_{\text{vir}}$. The black and red lines in Figure 6 show the same level of anticorrelation at distances less than 2 pMpc. Hence, in the simulations, the observed flux suppression is entirely due to the large-scale correlation between halos and density, and galaxies simply serve as biased tracers of the large-scale structure.

For illustration, we also show the observed points from R. A. Meyer et al. (2019). The observations appear to be offset from zero for all distances above 2 pMpc, which may indicate a potential bias in the observationally determined value for \bar{f} . Such a bias in observations is possible since R. A. Meyer et al. (2019) compute \bar{f} as the mean flux within 7.5 pMpc around each galaxy, and at such small distances, the halo-density correlation remains nonnegligible. In fact, the average halo-mass bias over the distance of 7.5 pMpc at $z = 5.4$ is 0.14, 0.074, and 0.045 for halos of mass 10^{12} , 10^{11} , and $10^{10} M_{\odot}$, respectively. To account for the potential bias, we also rescale \bar{f} by 13% to show how the observational data would appear if the mean flux in the observations were computed differently. The correspondence between rescaled observational data and the CROC data lends support to our proposition that the observed flux suppression is entirely due to the halo-density correlation.

4. Summary and Discussion

There are two main conclusions from this work. First, somewhat unexpectedly, we find that radiation from nearby

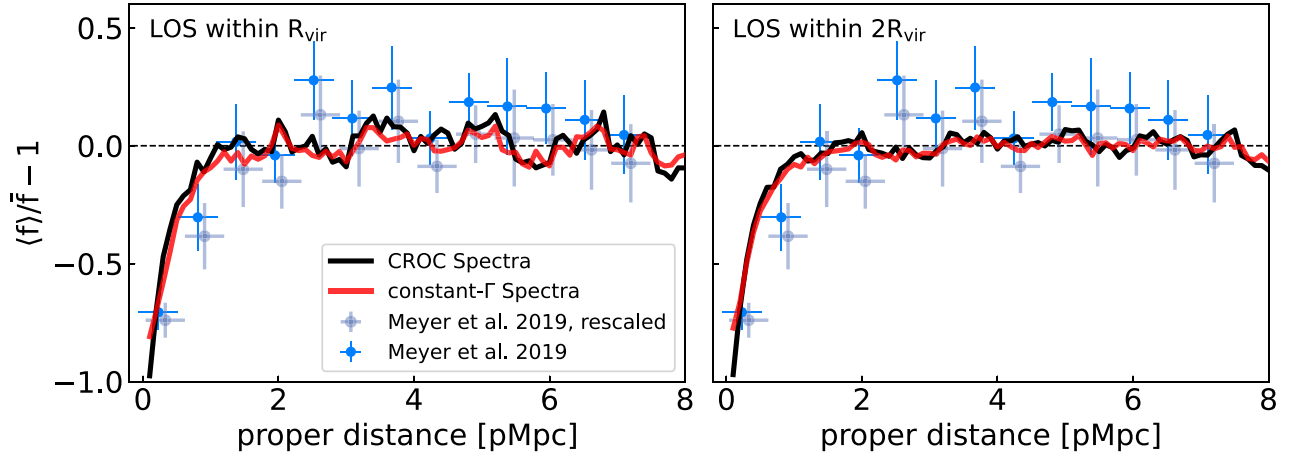


Figure 6. Normalized average transmitted flux as a function of distance from galaxies. We show measurements using both the simulated spectra and spectra generated, assuming a spatially uniform photoionization rate (Γ). Points with error bars reference measurements from R. A. Meyer et al. (2019); brighter blue points show the original measurement and lighter blue points show the measurement with the rescaled mean flux (see the text for details). The observed flux suppression at small distances in both data sets suggests that galaxies are not the primary cause since galaxies influence IGM properties through Γ .

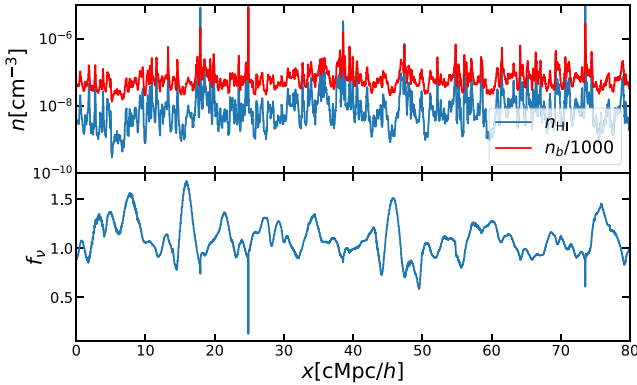


Figure 7. Baryon and H I number densities (top) and \mathcal{F}_ν (bottom) along a random LOS. \mathcal{F}_ν shows an anticorrelation with density, albeit not a strong one.

galaxies plays a negligible role in controlling the transmitted flux in quasar absorption spectra. The observed strong correlation between the transmitted flux and galaxy locations is explained entirely by the correlation of the transmitted flux with cosmic large-scale density distribution, of which galaxies are just a biased tracer.

Second, we find that in addition to the obvious factor determining the transmitted flux—the local gas density, the flux is also affected by a component of the radiation field that describes “cosmic background,” \mathcal{F}_ν from Equations (6) and (7). The complete physical interpretation of that result is elusive and will require substantial additional effort, and we delegate it to future work. What can be said immediately is that \mathcal{F}_ν does not depend on the local distribution of sources (Equation (6)) but does depend on the large-scale cosmic density (via the absorption coefficient k_ν), and hence encodes information about local variations in the photon free path (of which the commonly known photon mean free path is the mean). Figure 7 shows the baryon density, the neutral hydrogen density, and \mathcal{F}_ν along a random LOS. While some anticorrelation of \mathcal{F}_ν with density smoothed on ~ 1 Mpc scale is apparent from the figure, the two are not equivalent.

In fact, we find no smoothing scale R_S such that the baryon density smoothed on scale R_S along the LOS (i.e., in 1D) approximates \mathcal{F}_ν . It might be possible to find a smoothing scale

for the 3D baryon density that offers a better match to \mathcal{F}_ν . We leave such an investigation to future work since a more complex and laborious exploration may be required to come up with a better physical interpretation of \mathcal{F}_ν . At this point, it is sufficient for our purpose that we can present an equation for \mathcal{F}_ν and a plausible physical interpretation as a variation in the photon free path.

Acknowledgments

This work was supported in part by the NASA Theoretical and Computational Astrophysics Network (TCAN) grant 80NSSC21K0271. This manuscript has also been coauthored by Fermi Research Alliance, LLC under contract No. DE-AC02-07CH11359 with the United States Department of Energy. This work used resources of the Argonne Leadership Computing Facility, which is a DOE Office of Science User Facility supported under Contract DE-AC02-06CH11357. An award for computer time was provided by the Innovative and Novel Computational Impact on Theory and Experiment (INCITE) program. This research is also part of the Blue Waters sustained-petascale computing project, which is supported by the National Science Foundation (awards OCI-0725070 and ACI-1238993) and the State of Illinois. Blue Waters is a joint effort of the University of Illinois at Urbana-Champaign and its National Center for Supercomputing Applications. C.A. acknowledges support from DOE grant DE-SC009193 and the Leinweber Foundation at the University of Michigan.

Appendix Univariate and Bivariate Functions

For illustrative purposes, we use the functions at $z=5.25$ from the intermediate reionization case. The forms of the feature and interaction functions naturally vary when trained on different data sets. In Figure 8, we observe an apparent anticorrelation between flux and n_b , and positive correlations between flux and \mathcal{F}_ν , T , and \mathcal{G}_ν , respectively. These correlations paint a clear physical picture: Ly α transmission spikes are predominately found in regions of lower density, higher ionization, and higher temperatures. The magnitude of the flux

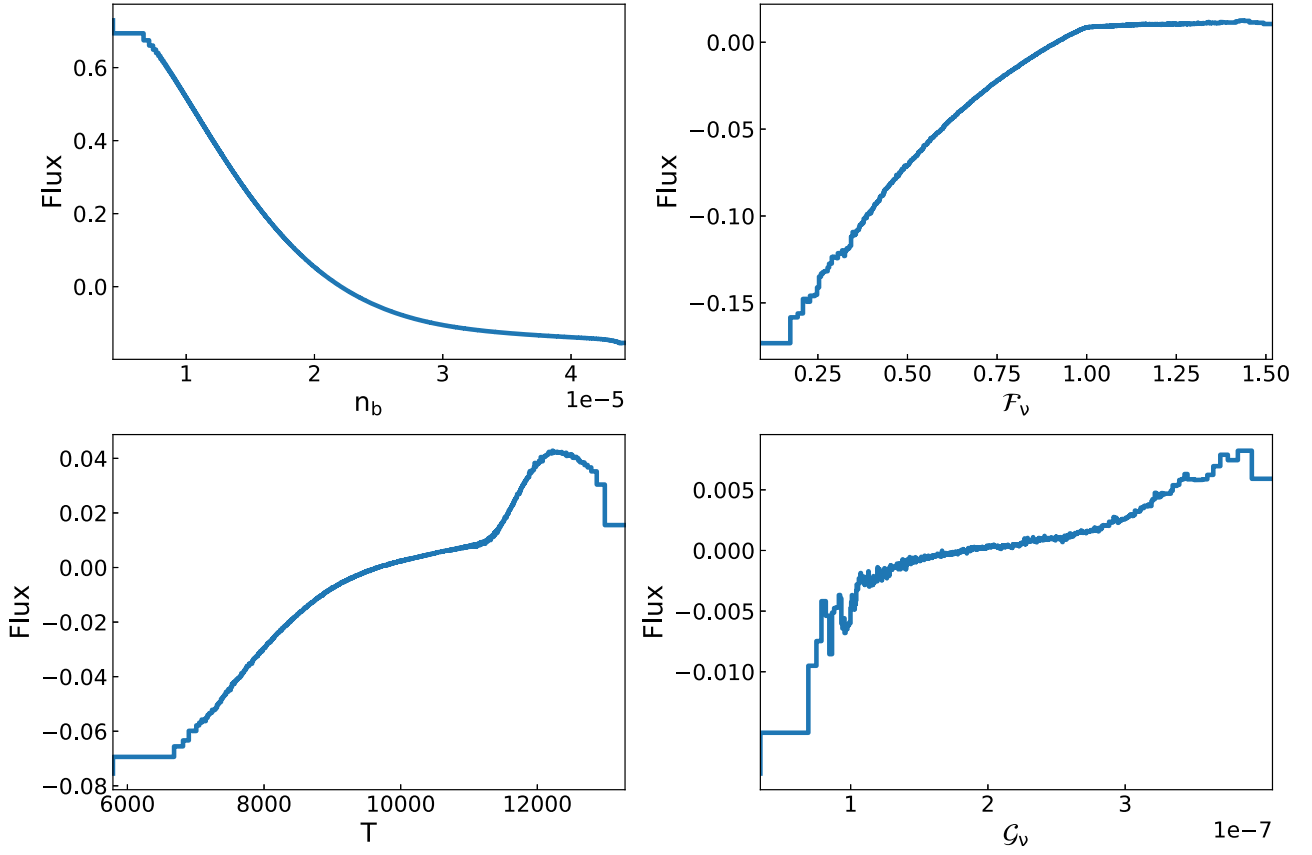


Figure 8. Univariate functions at $z = 5.25$ from the intermediate reionization case. There is an apparent anticorrelation between flux and n_b , and positive correlations between flux and \mathcal{F}_v , T , and \mathcal{G}_v , respectively. These trends suggest that Ly α transmission spikes are predominately found in regions of lower density, higher ionization, and higher temperatures.

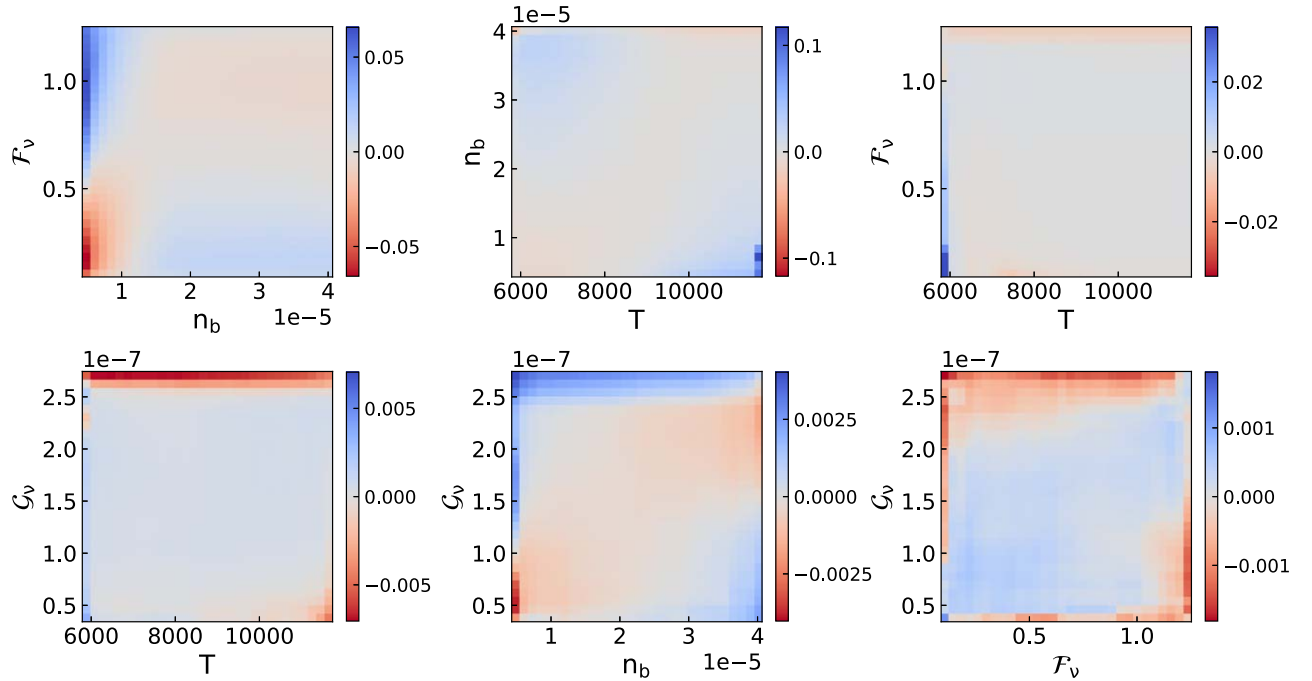


Figure 9. Bivariate functions at $z = 5.25$ from the intermediate reionization case.

in each panel indicates the relative importance of gas density, ionization, and temperature, in line with our previously presented results in Figure 5.




Additionally, we show the two-dimensional interaction functions in Figure 9. We note that as shown in Figure 5, the interaction terms are not the dominant features. Physically, in

our case, the interaction terms can be thought of as second-order expansions of the first-order terms. Taking the \mathcal{F}_ν versus n_b interaction in Figure 9 as an example, \mathcal{F}_ν (which essentially represents Γ , given the minimal contribution from \mathcal{G}_ν) and n_b are expected to dictate n_{HI} . Referring back to Equation (8), we derive

$$\begin{aligned} n_{\text{HI}} &\sim \frac{n^2}{\Gamma} = \frac{(\langle n \rangle + \delta n)^2}{(\langle \Gamma \rangle + \delta \Gamma)} \\ &\approx \frac{\langle n \rangle^2}{\langle \Gamma \rangle} \left(1 + 2 \frac{\delta n}{\langle n \rangle} - \frac{\delta \Gamma}{\langle \Gamma \rangle} - 2 \frac{\delta n \delta \Gamma}{\langle n \rangle \langle \Gamma \rangle} \right) \\ &\quad + \text{higher-order terms.} \end{aligned}$$

What EBM picks out as the interaction term should be $-\frac{\langle n \rangle^2}{\langle \Gamma \rangle} \left(2 \frac{\delta n \delta \Gamma}{\langle n \rangle \langle \Gamma \rangle} \right)$. Indeed, if we divide the \mathcal{F}_ν versus n_b plot into four quadrants (with division lines at $\langle \mathcal{F}_\nu \rangle$ and $\langle n_b \rangle$), quadrants I and III have negative signs and quadrants II and IV have positive signs, which is consistent with the expression above.

ORCID iDs

Hanjue Zhu (朱涵珏)  <https://orcid.org/0000-0003-0861-0922>
 Nickolay Y. Gnedin  <https://orcid.org/0000-0001-5925-4580>
 Camille Avestruz  <https://orcid.org/0000-0001-8868-0810>

References

- Adelberger, K. L., Steidel, C. C., Shapley, A. E., & Pettini, M. 2003, *ApJ*, **584**, 45
- Becker, G., D’Aloisio, A., Davies, F. B., Hennawi, J. F., & Simcoe, R. A. 2019, *BAAS*, **51**, 440
- Becker, G. D., & Bolton, J. S. 2013, *MNRAS*, **436**, 1023
- Becker, G. D., Bolton, J. S., Haehnelt, M. G., & Sargent, W. L. W. 2011, *MNRAS*, **410**, 1096
- Becker, G. D., Bolton, J. S., & Lidz, A. 2015, *PASA*, **32**, e045
- Bi, H., & Davidsen, A. F. 1997, *ApJ*, **479**, 523
- Boera, E., Murphy, M. T., Becker, G. D., & Bolton, J. S. 2016, *MNRAS*, **456**, L79
- Bolton, J. S., Gaikwad, P., Haehnelt, M. G., et al. 2022, *MNRAS*, **513**, 864
- Bosman, S. E. I., Fan, X., Jiang, L., et al. 2018, *MNRAS*, **479**, 1055
- Cantalupo, S., Lilly, S. J., & Haehnelt, M. G. 2012, *MNRAS*, **425**, 1992
- Cen, R., Miralda-Escudé, J., Ostriker, J. P., & Rauch, M. 1994, *ApJL*, **437**, L9
- Christenson, H. M., Becker, G. D., D’Aloisio, A., et al. 2023, *ApJ*, **955**, 138
- Cooray, A., Aguirre, J., Ali-Haimoud, Y., et al. 2019, *BAAS*, **51**, 48
- Croft, R. A. C., Weinberg, D. H., Katz, N., & Hernquist, L. 1998, *ApJ*, **495**, 44
- Furlanetto, S., Beardsley, A., Carilli, C. L., et al. 2019, *BAAS*, **51**, 142
- Gaikwad, P., Khaire, V., Choudhury, T. R., & Srianand, R. 2017, *MNRAS*, **466**, 838
- Gaikwad, P., Rauch, M., Haehnelt, M. G., et al. 2020, *MNRAS*, **494**, 5091
- Gaikwad, P., Srianand, R., Haehnelt, M. G., & Choudhury, T. R. 2021, *MNRAS*, **506**, 4389
- Garaldi, E., Gnedin, N. Y., Madau, P., et al. 2019, *ApJ*, **876**, 31
- Garaldi, E., Kannan, R., Smith, A., et al. 2022, *MNRAS*, **512**, 4909
- Garaldi, E., Kannan, R., Smith, A., et al. 2024, *MNRAS*, **530**, 3765
- Gnedin, N. Y. 2014, *ApJ*, **793**, 29
- Gnedin, N. Y. 2022, *ApJ*, **937**, 17
- Gnedin, N. Y., & Kaurov, A. A. 2014, *ApJ*, **793**, 30
- Gnedin, N. Y., Kravtsov, A. V., & Rudd, D. H. 2011, *ApJS*, **194**, 46
- Hausen, R., Robertson, B. E., Zhu, H., et al. 2023, *ApJ*, **945**, 122
- Hernquist, L., Katz, N., Weinberg, D. H., & Miralda-Escudé, J. 1996, *ApJL*, **457**, L51
- Hui, L., & Gnedin, N. Y. 1997, *MNRAS*, **292**, 27
- Hui, L., Gnedin, N. Y., & Zhang, Y. 1997, *ApJ*, **486**, 599
- Kakiichi, K., Ellis, R. S., Laporte, N., et al. 2018, *MNRAS*, **479**, 43
- Kannan, R., Garaldi, E., Smith, A., et al. 2022, *MNRAS*, **511**, 4005
- Kashino, D., Lilly, S. J., Matthee, J., et al. 2023, *ApJ*, **950**, 66
- Keating, L. C., Weinberger, L. H., Kulkarni, G., et al. 2020, *MNRAS*, **491**, 1736
- La Plante, P., Alvarez, M., Fialkov, A., et al. 2019, *BAAS*, **51**, 394
- Lou, Y., Caruana, R., Gehrke, J., & Hooker, G. 2013, Proc. 19th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (New York: ACM), 623
- Meyer, R. A., Bosman, S. E. I., Kakiichi, K., & Ellis, R. S. 2019, *MNRAS*, **483**, 19
- Meyer, R. A., Kakiichi, K., Bosman, S. E. I., et al. 2020, *MNRAS*, **494**, 1560
- Miralda-Escudé, J., Haehnelt, M., & Rees, M. J. 2000, *ApJ*, **530**, 1
- Mo, H. J., & White, S. D. M. 1996, *MNRAS*, **282**, 347
- Nasir, F., & D’Aloisio, A. 2020, *MNRAS*, **494**, 3080
- Nori, H., Jenkins, S., Koch, P., & Caruana, R. 2019, arXiv:1909.09223
- Ocvirk, P., Aubert, D., Sorce, J. G., et al. 2020, *MNRAS*, **496**, 4087
- Ocvirk, P., Gillet, N., Shapiro, P. R., et al. 2016, *MNRAS*, **463**, 1462
- Peeples, M. S., Weinberg, D. H., Davé, R., Fardal, M. A., & Katz, N. 2010, *MNRAS*, **404**, 1281
- Rakic, O., Schaye, J., Steidel, C. C., & Rudie, G. C. 2012, *ApJ*, **751**, 94
- Rauch, M. 1998, *ARA&A*, **36**, 267
- Rieke, M., Arribas, S., Bunker, A., et al. 2019, *BAAS*, **51**, 45
- Telikova, K. N., Shternin, P. S., & Balashev, S. A. 2019, *ApJ*, **887**, 205
- Walther, M., Oñorbe, J., Hennawi, J. F., & Lukić, Z. 2019, *ApJ*, **872**, 13
- Weinberg, D. H., Davé, R., Katz, N., & Kollmeier, J. A. 2003, in AIP Conf. Ser. 666, The Emergence of Cosmic Structure, ed. S. H. Holt & C. S. Reynolds (Melville, NY: AIP), 157
- Zhang, Y., Anninos, P., Norman, M. L., & Meiksin, A. 1997, *ApJ*, **485**, 496