Producing High-fidelity Flux Fields From N-body Simulations Using Physically Motivated Neural Networks

Peter Harrington
Lawrence Berkeley National Laboratory
1 Cyclotron Road, Berkeley, CA, 94720, USA
`pharrington@lbl.gov`

Mustafa Mustafa
Lawrence Berkeley National Laboratory
1 Cyclotron Road, Berkeley, CA, 94720, USA
`mmustafa@lbl.gov`

Max Dornfest
Lawrence Berkeley National Laboratory
1 Cyclotron Road, Berkeley, CA, 94720, USA
`dornfest@berkeley.edu`

Wahid Bhimji
Lawrence Berkeley National Laboratory
1 Cyclotron Road, Berkeley, CA, 94720, USA
`wbhimji@lbl.gov`

Zarija Lukić
Lawrence Berkeley National Laboratory
1 Cyclotron Road, Berkeley, CA, 94720, USA
`zarija@lbl.gov`

Abstract

Cosmological simulations are a powerful tool for predicting observable characteristics and inferring physical parameters of our universe, but that capability comes at extreme CPU-time cost. Here we present a multi-stage, physically motivated deep learning pipeline that translates 3D dark matter distributions, which can be obtained using inexpensive N-body simulations, into corresponding Lyman-α redshift-space flux fields, which commonly require running expensive multi-physics simulations. Our pipeline consists of two independent networks: the first maps a 3D dark matter distribution into a real-space Lyman-α flux field, while the second is conditioned on the gas velocity field along the z direction, and warps 2D real-space flux fields into their corresponding redshift-space fields. We demonstrate that the pipeline reproduces both the probability distribution and power spectrum of the flux field with high accuracy, improving on the relative error of the standard method by an order of magnitude.

1 Introduction

Cosmological simulations — modeling the evolution of structure in the universe — are among the most expensive simulations run at supercomputing centers, with cost going into tens of millions of CPU hours. These simulations are a necessary component for answering fundamental physics questions from sky surveys, for example: the nature of dark matter and dark energy, the mass of neutrino particles, and how inflation and reionization happened in the early universe. They can accurately predict observable quantities but only when they retain the high spatial resolution necessary to resolve density fluctuations and model complex physical phenomena. The overarching aim of the work in this paper is to build data-driven machine-learning techniques that achieve the structure, exceptional fidelity and interpretability of fully science-based models, with the compute speed of approximate methods, through exploiting a structure motivated by those science models.

Second Workshop on Machine Learning and the Physical Sciences (NeurIPS 2019), Vancouver, Canada.

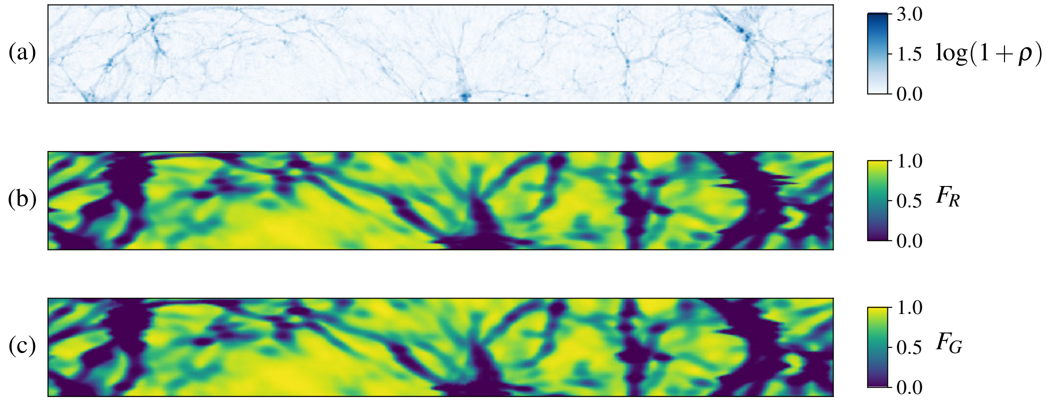


Figure 1: Visualization of our pipeline output. A 3D dark matter distribution (a 2D slice of which is shown in panel a) is the principal input to the workflow, which tries to produce the corresponding Lyα flux field F_R (b). The prediction F_G is shown in panel (c). Generally, structures at both large and small scales, as well as the distortions which warp them in redshift space, are captured well.

In this work we focus on an important cosmological observable, the Lyman-α (Lyα) flux: the characteristic pattern in the absorption spectra of quasars imprinted by the neutral hydrogen in the intergalactic medium (for a recent review, see [7]). It represents an extraordinary cosmological probe, capable of tracing density fluctuations as far away as redshift of $z \sim 6$ when the universe was less than a billion years old. The large number of quasars discovered to date enables statistical analyses of the absorption spectra by considering the transmitted Lyα flux along many different lines of sight, called "skewers". The measured statistical properties of the flux, like the probability density function (PDF) or the power spectrum ($P(k)$), can be compared to theoretical models of structure formation, constraining cosmological parameters as well as the thermal history and reionization process of the universe. Unfortunately, there is no viable analytical solution of these theoretical models, so to generate the required high-fidelity simulations, the problem currently has to be treated numerically (see for example [6] and references therein). As a result the full simulations, such as those used for training and validation in this paper, require upwards of 10^5 CPU hours to produce. Finding an inexpensive way to obtain an approximate (yet accurate enough) Lyman-α flux field has been a research topic for decades, starting with the pioneering work of Gunn and Peterson [3], and is critical to exploiting the full potential of future sky surveys to uncover new fundamental physics.

We propose to tackle this critical challenge through the use of a novel multi-stage deep learning pipeline that directly reconstructs a Lyα flux field from a given dark matter distribution, which is considerably cheaper to produce (sample visualizations of these fields can be see in Figure 1). This pipeline, if trained to a sufficient level of accuracy, can be used in conjunction with much faster dark-matter-only N-body simulations in order to bypass the need for the resource-intensive full-physics methods. In the following sections, we describe our pipeline and demonstrate its ability to capture features across a range of scales with a much higher degree of accuracy than current approximate approaches.

2 Dataset and Models

2.1 Dataset

We construct our training and validation datasets from a cosmological simulation run with the `Nyx` code [1, 6]. The cosmological parameters in the simulation are $\Omega_b = 0.05$, $\Omega_M = 0.32$, $\Omega_L = 0.68$, $h = 0.67$, $\sigma_8 = 0.83$, and $n_s = 0.97$, but this choice of parameters does not affect any of the conclusions presented here. The physical fields of interest are defined on a 3D uniform 1024^3 mesh, spanning a cube of $L = 20$ Mpc/h on a side, with periodic boundary conditions. Due to the motion of the gas, the observable Lyα flux is distorted along the line of sight (an effect called "redshift-space distortion"), and we choose this line-of-sight to be the z-axis. We set aside one eighth of the full 1024^3 domain as our validation dataset, and use the rest for training the network.

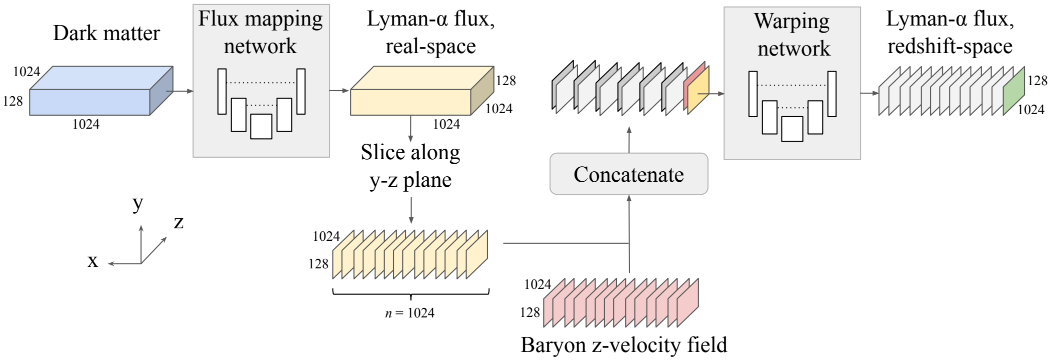


Figure 2: Schematic diagram of our full inference pipeline. The flux mapping network M takes the dark matter density field as input and produces a real-space Lyα flux field, which is then concatenated with the baryon z-velocity field and fed into the warping network W to produce the redshift-space Lyα flux.

2.2 Pipeline description & network architectures

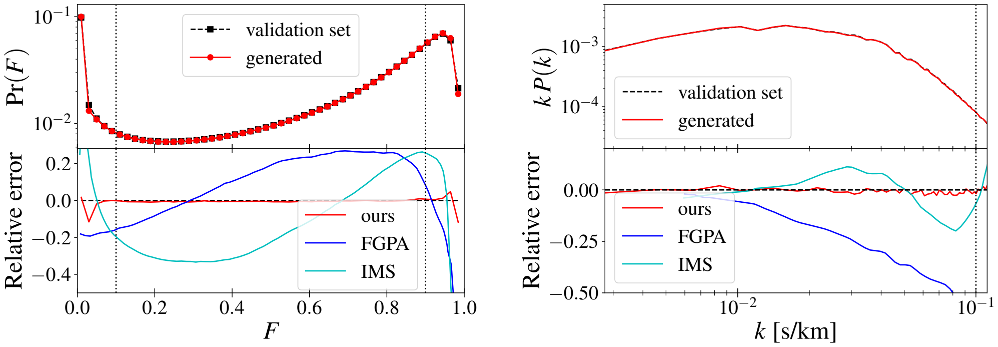
Motivated by the physical relationship between dark matter distributions and Lyα redshift-space flux fields, we split the problem into two sequential steps. Since the gas responsible for Lyα flux has a distribution largely dependent on the 3D structures encoded in the dark matter density field ρ, we first train a deep network M to translate ρ into corresponding real-space Lyα flux fields F^*. Then, noting that the physical transformation from real-space flux to redshift-space flux is dependent only on the velocity of the gas along the z-direction (the line-of-sight in our training data), we train a second network W to translate 2D slices of the real-space flux field F^* into corresponding slices of the redshift-space flux field F, by conditioning W with the z-component of the gas velocity. We choose to train W using 2D slices because the physical transformation from real-space flux to redshift-space flux is inherently one-dimensional, and we have already tasked M with generating the important 3D structures, so training W with 3D fields does not supply additional useful information. The networks M and W are trained independently, then during inference they are chained together to produce the generated redshift-space flux field $F_G = W \circ M(a(\rho))$, where a is a normalization function. A diagram of this inference pipeline is given in Figure 2.

Both M and W are fully convolutional, and very similar in design. M is a V-Net [8], and uses 3D convolutions, while W is a U-Net [12], using 2D convolutions[1]. We investigated the results of training M and W with a combination of $\mathcal{L}1$ loss and an adversarial loss (given by unique discriminator networks for M and W), and found that the bulk of the useful gradient signal comes from $\mathcal{L}1$ loss alone. We apply the adversarial loss as a supplementary loss term to train M, which provides slight refinements to the summary statistics, but not to W, where no improvement from adding this loss was observed. During the training of M, sample sub-cubes of size 128^3 are randomly cropped from the training set region of the full simulation. Similarly, during the training of W, sample slices of size 128x1024 are randomly cropped from the training set region. In the 2D slices, the longer dimension (of length 1024) is always the z-axis of the original simulation.

3 Results

After training both M and W, we test our pipeline by generating predictions for the redshift-space Lyα flux F_G over the entire validation set region of the simulation. Since M is fully convolutional, it generalizes well to larger input sizes, and we are able to generate the real-space Lyα flux F_G^* for the entire validation set region (of size 1024x128x1024) at once. Once we have F_G^*, we pass slices of it to W, along with the z-component velocity field, to generate predictions for the redshift-space flux field F_G.

[1] The code and full architecture details are available at `https://github.com/pzharrington/Lya_demo`



(a)

(b)

Figure 3: Statistical comparisons between the validation set and the pipeline output. In (a), we show the flux PDF $\Pr(F)$, for the generated flux fields F_G and the ground truth from the validation set F_R, as well as the relative error $\Pr(F_G)/\Pr(F_R) - 1$. In (b), we show the 1D power spectrum $P(k)$ of the two fields F_G and F_R, along with the relative error $P_G(k)/P_R(k) - 1$. To demonstrate the improvement that our new method brings, we also show in ratio panels the most standard physics-based method (FGPA) as well as the latest and much more complex model (IMS), as published in Sorini et al. [14] (their figure 7).

A sample visualization of our results is given in Figure 1. Qualitatively, it is clear that the pipeline does well in capturing large-scale structures and the distortions which warp them in redshift-space, as well as some of the smaller-scale, more fine-grained variations within the filaments and voids. The details which are harder for the networks to capture almost always involve sharp variations in the flux field – these can be the sharp transitions between filaments and voids, as well as the more extreme redshift-space distortions caused by abnormally dense superclusters.

We quantitatively evaluate the performance of our pipeline with the two most standard statistics used in the analysis of Lyα flux, which are shown in Figure 3. First, we inspect the flux PDF, which in the range $F \in [0.1 - 0.9]$ (vertical dashed lines in Fig. 3) can be used to infer cosmological and thermodynamical properties of the universe [5, 15]. We find excellent agreement between the flux PDF of our generated field and that of the ground truth, with a mean absolute relative error of less than 1%. As a more detailed probe, we compute 1D power spectra $P(k)$ of 50000 skewers aligned with the z-axis, randomly sampled from the generated flux field F_G, and compare to the spectra of the ground truth skewers. This power spectrum is a summary statistic of the Lyα flux field which measures the Fourier-space analogue of 2-point correlations, and can be used to measure "standard" cosmological parameters [13, 9], constrain neutrino properties [9, 18], probe dark matter models [16, 4, 2], or measure thermal properties of the intergalactic medium [17]. We find a tight agreement across all relevant length scales ($k \leq 0.1$ s/km), with a mean absolute relative error of 1.1%.

The accuracy of our pipeline in capturing these two crucial statistics of the Lyα field is a remarkable improvement over the $\sim 20 - 30\%$ relative error incurred by using the Fluctuating Gunn-Peterson Approximation (FGPA [3]), the de-facto standard physics-based method. We report significantly improved results even compared to quite modern and complex techniques like Iteratively Matched Statistics (IMS) which — like our neural network — relies on the existence of hydrodynamical simulations onto which the statistical matching is performed [14].

4 Conclusions

The results presented here represent significant improvement over the current state-of-the-art [11, 14] in reconstructing the Lyα flux on inexpensive N-body simulations. Moreover, the few percent accuracy we are achieving in the power spectrum approaches the accuracy of current multi-physics simulations, which are themselves converged to about 1% [6] precision. This result opens the door to using multi-fidelity methods [10] when inferring cosmological parameters, where parameters' posterior probability is first reconstructed via an approximate method and then confirmed or corrected using only a few expensive, high-fidelity simulations. In this context, a potential extension of our work would be to experiment with conditional training of our networks using simulations at different parameter points, which would enable sensible interpolation of Lyα flux at locations in parameter space where a full simulation does not exist. Finally, we want to emphasise that the power of the presented method also lies in the fact that training was done on smaller simulated volumes, while inference was done on a much larger volume (the full validation set) at once. Thus, our method allows for calibration on small-volume hydrodynamical simulations, then application on large-volume N-body simulations, just like previous physics-driven methods [11, 14]. Our pipeline is thus able to achieve a much more precise reconstruction of the Lyα flux field without the loss of any capability of the existing physics-based methods.

Acknowledgments

This research used resources of the National Energy Research Scientific Computing Center (NERSC), a U.S. Department of Energy Office of Science User Facility operated under Contract No. DE-AC02-05CH11231. PH is supported by the Exascale Computing Project (17-SC-20-SC), a joint project of the U.S. Department of Energy's Office of Science and National Nuclear Security Administration, responsible for delivering a capable exascale ecosystem, including software, applications, and hardware technology, to support the nation's exascale computing imperative. MD and ZL were supported by the Scientific Discovery through Advanced Computing (SciDAC) program funded by U.S. Department of Energy Office of Advanced Scientific Computing Research and the Office of High Energy Physics.

References

[1] A. S. Almgren, J. B. Bell, M. J. Lijewski, Z. Lukić, and E. Van Andel. Nyx: A Massively Parallel AMR Code for Computational Cosmology. *Astroph. J.* , 765(1):39, Mar 2013. doi: 10.1088/0004-637X/765/1/39.

[2] E. Armengaud, N. Palanque-Delabrouille, C. Yèche, D. J. E. Marsh, and J. Baur. Constraining the mass of light bosonic dark matter using SDSS Lyman-α forest. *Mon. Not. Royal Astro. Soc.* , 471(4):4606–4614, Nov 2017. doi: 10.1093/mnras/stx1870.

[3] J. E. Gunn and B. A. Peterson. On the Density of Neutral Hydrogen in Intergalactic Space. *Astroph. J.* , 142:1633–1636, Nov 1965. doi: 10.1086/148444.

[4] V. Iršič, M. Viel, M. G. Haehnelt, J. S. Bolton, S. Cristiani, G. D. Becker, V. D'Odorico, G. Cupani, T.-S. Kim, T. A. M. Berg, S. López, S. Ellison, L. Christensen, K. D. Denney, and G. Worseck. New constraints on the free-streaming of warm dark matter from intermediate and small scale Lyman-α forest data. *Phys. Rev. D.* , 96(2):023522, Jul 2017. doi: 10.1103/PhysRevD.96.023522.

[5] K.-G. Lee, J. F. Hennawi, D. N. Spergel, D. H. Weinberg, D. W. Hogg, M. Viel, J. S. Bolton, S. Bailey, M. M. Pieri, W. Carithers, D. J. Schlegel, B. Lundgren, N. Palanque-Delabrouille, N. Suzuki, D. P. Schneider, and C. Yèche. IGM Constraints from the SDSS-III/BOSS DR9 Lyα Forest Transmission Probability Distribution Function. *Astroph. J.* , 799(2):196, Feb 2015. doi: 10.1088/0004-637X/799/2/196.

[6] Z. Lukić, C. W. Stark, P. Nugent, M. White, A. A. Meiksin, and A. Almgren. The Lyman α forest in optically thin hydrodynamical simulations. *Mon. Not. Royal Astro. Soc.* , 446:3697–3724, Feb. 2015. doi: 10.1093/mnras/stu2377.

[7] M. McQuinn. The Evolution of the Intergalactic Medium. *Annual Review of Astronomy and Astrophysics*, 54:313–362, Sept. 2016. doi: 10.1146/annurev-astro-082214-122355.

[8] F. Milletari, N. Navab, and S.-A. Ahmadi. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. *arXiv e-prints*, art. arXiv:1606.04797, Jun 2016.

[9] N. Palanque-Delabrouille, C. Yèche, J. Baur, C. Magneville, G. Rossi, J. Lesgourgues, A. Borde, E. Burtin, J.-M. LeGoff, J. Rich, M. Viel, and D. Weinberg. Neutrino masses and cosmology with Lyman-alpha forest power spectrum. *J. Cosmology Astropart. Phys.* , 2015(11):011, Nov 2015. doi: 10.1088/1475-7516/2015/11/011.

[10] B. Peherstorfer, K. Willcox, and M. Gunzburger. Survey of multifidelity methods in uncertainty propagation, inference, and optimization. *SIAM Review*, 60(3):550–591, 2018. doi: 10.1137/16M1082469. URL `https://doi.org/10.1137/16M1082469`.

[11] S. Peirani, D. H. Weinberg, S. Colombi, J. Blaizot, Y. Dubois, and C. Pichon. LyMAS: Predicting Large-scale Lyα Forest Statistics from the Dark Matter Density Field. *Astroph. J.* , 784(1):11, Mar 2014. doi: 10.1088/0004-637X/784/1/11.

[12] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[13] U. Seljak, A. Slosar, and P. McDonald. Cosmological parameters from combining the Lyman-α forest with CMB, galaxy clustering and SN constraints. *Journal of Cosmology and Astro-Particle Physics*, 2006: 014, Oct 2006. doi: 10.1088/1475-7516/2006/10/014.

[14] D. Sorini, J. Oñorbe, Z. Lukić, and J. F. Hennawi. Modeling the Lyα Forest in Collisionless Simulations. *Astroph. J.* , 827(2):97, Aug 2016. doi: 10.3847/0004-637X/827/2/97.

[15] M. Viel, J. S. Bolton, and M. G. Haehnelt. Cosmological and astrophysical constraints from the Lyman α forest flux probability distribution function. *Mon. Not. Royal Astro. Soc.* , 399(1):L39–L43, Oct 2009. doi: 10.1111/j.1745-3933.2009.00720.x.

[16] M. Viel, G. D. Becker, J. S. Bolton, and M. G. Haehnelt. Warm dark matter as a solution to the small scale crisis: New constraints from high redshift Lyman-α forest data. *Phys. Rev. D.* , 88(4):043502, Aug 2013. doi: 10.1103/PhysRevD.88.043502.

[17] M. Walther, J. Oñorbe, J. F. Hennawi, and Z. Lukić. New Constraints on IGM Thermal Evolution from the Lyα Forest Power Spectrum. *Astroph. J.* , 872(1):13, Feb 2019. doi: 10.3847/1538-4357/aafad1.

[18] C. Yèche, N. Palanque-Delabrouille, J. Baur, and H. du Mas des Bourboux. Constraints on neutrino masses from Lyman-alpha forest power spectrum with BOSS and XQ-100. *J. Cosmology Astropart. Phys.* , 2017 (6):047, Jun 2017. doi: 10.1088/1475-7516/2017/06/047.