

Lawrence Berkeley National Laboratory

LBL Publications

Title

Fast, High-fidelity Ly α Forests with Convolutional Neural Networks

Permalink

<https://escholarship.org/uc/item/6hf02107>

Journal

The Astrophysical Journal, 929(2)

ISSN

0004-637X

Authors

Harrington, Peter
Mustafa, Mustafa
Dornfest, Max
[et al.](#)

Publication Date

2022-04-01


DOI

10.3847/1538-4357/ac5faa

Peer reviewed



Fast, High-fidelity Ly α Forests with Convolutional Neural Networks

Peter Harrington¹, Mustafa Mustafa¹, Max Dornfest¹, Benjamin Horowitz^{1,2} , and Zarija Lukic¹¹Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA 94720, USA; pharrington@lbl.gov²Department of Astronomy, Princeton University, Princeton, NJ, USA

Received 2021 July 9; revised 2022 March 3; accepted 2022 March 20; published 2022 April 26

Abstract

Full-physics cosmological simulations are powerful tools for studying the formation and evolution of structure in the universe but require extreme computational resources. Here, we train a convolutional neural network to use a cheaper N -body-only simulation to reconstruct the baryon hydrodynamic variables (density, temperature, and velocity) on scales relevant to the Ly α forest, using data from N_{yx} simulations. We show that our method enables rapid estimation of these fields at a resolution of ~ 20 kpc, and captures the statistics of the Ly α forest with much greater accuracy than existing approximations. Because our model is fully convolutional, we can train on smaller simulation boxes and deploy on much larger ones, enabling substantial computational savings. Furthermore, as our method produces an approximation for the hydrodynamic fields instead of Ly α flux directly, it is not limited to a particular choice of ionizing background or mean transmitted flux.

Unified Astronomy Thesaurus concepts: [Cosmology \(343\)](#); [Large-scale structure of the universe \(902\)](#)

1. Introduction

Understanding the distribution and evolution of the matter in the universe is at the core of modern cosmology and a main motivation for the upcoming generation of astronomical surveys. But to extract cosmological insights, observations have to be compared to theoretical predictions for different cosmological scenarios. Predicting observable quantities in general, and the Ly α forest specifically, from the underlying large-scale structure is complicated by the nonlinear mapping between baryons (responsible for most of observables) and dark matter (which dominates the gravitational dynamics). In the early days preceding “precision cosmology,” simplified models were successfully used, for example, the baryons-trace-dark-matter approximation (Petitjean et al. 1995), or modifying the matter density field (Gnedin & Hui 1996) or gravitational potential (Gnedin & Hui 1998), to mimic the effects of baryonic pressure smoothing. The gas temperature of the intergalactic medium (IGM) in those models is usually computed by imposing a polytropic temperature–density relation (Hui & Gnedin 1997).

More recently, recipes to model the Ly α forest without running full hydrodynamical simulations have been provided by Peirani et al. (2014) and Sorini et al. (2016), as well as models for specific summary statistics in Seljak (2012), but they have not really been used in practice, due to their complexity and/or applicability to only a limited range of scales. Thus, Ly α forest forward models rely on computationally expensive hydrodynamical simulations (Iršič et al. 2017; Boera et al. 2019; Walther et al. 2019; Palanque-Delabrouille et al. 2020; Rogers & Peiris 2021; Walther et al. 2021), and solving inverse problems commonly requires running dozens of simulations, at a rough cost of 2 million CPU hours per simulation (see, for example, Walther et al. 2021). Of course, if one wants to consider cosmological models beyond standard Λ CDM or to model complex astrophysical phenomena

affecting the Ly α signal like high column density absorbers (Rogers et al. 2018), UV background fluctuations (Oñorbe et al. 2019), or HeII reionization (Upton Sanderbeck & Bird 2020) the number of needed simulations, and thus the computational costs, will increase significantly. Such applications present a clear need for surrogate models that can mitigate some of the cost of running full-fledged hydrodynamic simulations for every test case.

Beyond generation of mock catalogs and hydrodynamic emulators, there is also a growing need for fast hydrodynamical mapping in the context of differentiable forward model reconstructions (Horowitz et al. 2019). There, the underlying matter distribution is reconstructed under constraints from gravitational evolution and using, for example, the three-dimensional Ly α tomographic reconstructions under the Fluctuating Gunn–Peterson Approximation (FGPA; Gunn & Peterson 1965) as done in Horowitz et al. (2019), Porqueres et al. (2019), and Horowitz et al. (2021b). Reducing the modeling cost via fast and efficient—even if approximate—methods of reproducing the Ly α signal is thus of significant interest to the community. Our paper addresses this issue using data-driven modeling trained on hydrodynamical simulations.

In recent years, deep neural networks (NN) have become a promising method to assist in solving these highly nonlinear problems, by acting as surrogate models (which are differentiable by design) for complex phenomena. In particular, Generative Adversarial Networks (Goodfellow et al. 2014; GANs) and U-Nets (Ronneberger et al. 2015) have proven useful in a variety of cosmology tasks that either directly generate synthetic hydrodynamic quantities (Zamudio-Fernandez et al. 2019), or reconstruct them from dark matter distributions. The latter approach allows one to run cheaper N -body simulations (or even surrogates for N -body evolution) to generate a dark matter distribution, then rapidly estimate the target hydrodynamic quantity corresponding to the dark matter structure. In the context of the thermal and kinematic Sunyaev–Zel’dovich effects, this approach has been shown to be useful in reconstructing electron density, pressure, and momenta in 3D (Thiele et al. 2020) as well as gas pressure in 2D (Tröster et al. 2019). These works showcase the success of NNs in



Original content from this work may be used under the terms of the [Creative Commons Attribution 4.0 licence](#). Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

producing hydrodynamic fields with superior statistical fidelity compared to competing classical or semi-analytical methods. While extremely promising, deep learning approaches are data-hungry, and require generation of training data at the desired output resolution (e.g., ~ 100 kpc in Thiele et al. 2020) to achieve maximal model fidelity. It is thus important to maintain a reasonable volume of training data when extending such models to finer-scale hydrodynamic reconstruction, and investigate the data requirements in such regimes.

In this work, we use NNs to reconstruct 3D baryon hydrodynamics from N -body dark matter simulations at ~ 20 kpc h^{-1} resolution, an order of magnitude finer than previous deep-learning-based hydrodynamic reconstructions. Given a snapshot of dark matter, our model allows rapid estimation of baryon density, temperature, and velocity fields, which can be painted in over an arbitrarily large volume at the given resolution.

The paper is organized as follows. We first describe the simulation data used to train our model and detail how we compute FGPA estimates and Ly α forest quantities in Section 2. Then, we describe the model design and training process in Section 3. Our results are presented in Section 4, where we show sample output from our model and make statistical comparisons against the true hydrodynamic fields. Finally, we present conclusions and discussion in Section 5.

2. Simulations

We construct our training and testing data sets from pairs of cosmological simulations run with the `Nyx` code (Almgren et al. 2013; Sexton et al. 2021). Each pair of simulations share identical initial conditions and cosmology, with one simulation modeling the full-physics problem (dark matter particle N -body dynamics as well as baryon hydrodynamics), while the other just models the evolution of dark matter particles (N -body-only). Initial conditions were produced by our in-house code `CosmicIC`, using a single transfer function for the total matter content, i.e., mass-weighted average of transfer functions for dark matter and baryons. The N -body and hydrodynamical simulations use the same initial conditions, thus the distribution of baryons follows the distribution of dark matter in the initial conditions. In `Nyx`, the dark matter particles are evolved with a particle-mesh scheme, while the additional baryon hydrodynamics are modeled as an inviscid ideal fluid on a set of Eulerian grids. We neglect physics related to galaxy formation, but we do include the main physical processes relevant for the Ly α forest. We consider the chemistry of the gas as having a primordial composition of hydrogen and helium, and we include inverse Compton cooling off the microwave background and keep track of the net loss of thermal energy resulting from atomic collisional processes (Lukić et al. 2015). All cells are assumed to be optically thin to ionizing radiation, and radiative feedback is accounted for via a spatially uniform, time-varying UV background radiation given to the code as a list of photoionization and photoheating rates (Haardt & Madau 2012). We note that this is an extremely common approach in Ly α cosmological simulations (Iršič et al. 2017; Boera et al. 2019; Walther et al. 2019; Pedersen et al. 2021; Rogers & Peiris 2021; Walther et al. 2021). The main reason for this is that regions transparent to Ly α photons at redshifts $z \gtrsim 2$ are in low-density regions (Lukić et al. 2015; McQuinn 2016), and poorly understood physical processes

related to galaxy formation only play a minor role in those regions (Desjacques et al. 2006; Kollmeier et al. 2006).³

With this paired data set, we can train our model to learn a mapping from the cheaper N -body-only simulation into the hydrodynamic fields of the full-physics simulation in a supervised fashion. We use one pair of simulations for training, and an independent pair, with identical cosmology but different initial conditions, for testing. By restricting our training set to just a single pair of simulations, we are investigating a more realistic use-case scenario for our neural-network-based approach. Training on one small cosmological volume and then applying the trained model to large cosmological volumes would allow one to more easily test different scenarios for gas physics, cosmological parameters, or other astrophysical effects. While our results would likely improve if we used a larger training set, doing so would yield diminishing returns, as the computational cost of generating data and training on it would grow.

For both the training and testing pairs of simulations, the cosmological parameters are $\Omega_b = 0.05$, $\Omega_M = 0.31$, $\Omega_L = 0.69$, and $h = 0.675$. The physical fields of interest are defined on a 3D uniform 1024^3 grid, spanning a cube of $L = 20$ Mpc h^{-1} per side, with periodic boundary conditions. For the hydrogen and helium mass abundances, we adopted values consistent with the CMB observations and Big Bang nucleosynthesis (Coc et al. 2013): $X_p = 0.76$ and $Y_p = 0.24$. Dark matter N -body dynamics are evolved with 1024^3 particles, then density and velocities are deposited on the grid using Cloud-In-Cell (CIC) interpolation. We select the $z = 3$ snapshot in each simulation, as it lies in the center of the range of redshifts most relevant for Ly α analysis.

Data scales in `Nyx` files can span many orders of magnitude, so we normalize the data to order unity for training stability. For all velocity fields, we scale linearly to $\hat{\mathbf{v}} = \mathbf{v}/(9 \times 10^7 \text{ cm} \cdot \text{s}^{-1})$. The baryon density and temperature fields have a large dynamic range, so we use

$$\begin{aligned}\hat{\rho} &= \log(\rho/14), \\ \hat{T} &= \log(T)/8 - 1.5, \\ \hat{\rho}_{\text{DM}} &= \log(1 + \rho_{\text{DM}})/12.\end{aligned}\tag{1}$$

where ρ and T are the baryon overdensity and temperature (in Kelvin), respectively, while ρ_{DM} is the overdensity of dark matter. These choices ensure all target fields lie in the range $(-1, 1)$, as is typically chosen for generative models. Without these normalizations in place, attempting to train on the native data scales (or a linear transformation of them) resulted in unstable training and poor performance. All network outputs are transformed back into their original units before performing analysis on statistical quantities.

2.1. FGPA Estimates

The long-standing method to reconstruct hydrodynamic quantities for purposes of Ly α modeling is the FGPA (Gunn & Peterson 1965; Sorini et al. 2016). This procedure takes the dark matter overdensity ρ^{DM} from the N -body simulation, with or without artificial smoothing, to produce estimates for the baryon density ρ and temperature T , given by a power-law

³ For a recent work done at higher precision and challenging this view, see Chabanier et al. (2020).

relation (Hui & Gnedin 1997)

$$T = T_0(\rho/\langle\rho\rangle)^{\gamma-1}, \quad (2)$$

where $\langle\rangle$ denotes averaging over volume. The estimate for ρ is usually attained by applying Gaussian smoothing (with smoothing length λ_G roughly given by the Jeans filtering scale) to ρ^{DM} to simulate baryonic pressure smoothing. In this work, we empirically determine the best-fit T_0 , γ , and λ_G for our simulation. We best-fit the Gaussian smoothing scale using least-square differences between the smoothed dark matter density from the N -body simulation and the actual baryon density field from the hydrodynamical simulation. We then fit density and temperature in the hydrodynamical simulations to the power-law relation in Equation (2) as described in Lukić et al. (2015), arriving for our particular simulation and redshift $z=3$ at values:

$$\begin{aligned} T_0 &= 12, 300 \text{ K} \\ \gamma &= 1.49 \\ \lambda_G &= 43.5 \text{ kpc } h^{-1}, \end{aligned} \quad (3)$$

which are in good agreement with observational constraints (Walther et al. 2019). We use these values to compute FGPA estimates for the hydrodynamic fields in the test simulation and use the resulting Ly α flux as a baseline to compare against the accuracy of our neural network approach. Additionally, FGPA is a useful starting point for the neural network training process, providing an initial guess upon which the network can improve.

2.2. Ly α Computation

Given the true or estimated baryon hydrodynamics fields, we can then compute Ly α flux analytically. We use the `Gimlet` code (Friesen et al. 2016) to do this step. We compute the optical depth, τ , of neutral hydrogen at a fixed redshift, which is related to the transmitted flux fraction with $F = \exp(-\tau)$. We do not account for the finite speed of light when we cast rays in the simulation, but we use the thermodynamical state of baryons at a fixed cosmic time. The simulated spectra are thus not meant to fully reproduce observed Ly α forest spectra, but to recover the flux statistics in a given redshift window. Our calculation of the spectra accounts for Doppler shifts due to bulk flows of the gas and for the thermal broadening of the Ly α line. We have neglected noise and metal contamination, but this is common practice in the field and they are not relevant for this paper. We refer to Lukić et al. (2015) for specific details of these calculations. Finally, we rescale the UV background intensity so that the mean flux of all spectra from the network model predictions and hydrodynamical simulation match. Note that, with our NNs, we predict the hydrodynamic fields (ρ , T , v) directly, and thus we use identical pipelines and assumptions when calculating Ly α forest quantities from `Nyx` simulations and NN models. This makes our approach orthogonal to any particular choices made in modeling Ly α forest from basic thermodynamical quantities.

3. Model Design

3.1. Evaluation Metrics

We evaluate the accuracy of our model primarily by comparing the PDF (Rauch et al. 1997) and 1D power spectra (Croft et al. 1999) of our model’s Ly α fields against those of

the test simulation. The Ly α flux variation along some line-of-sight axis is defined with respect to fluctuations about the global mean, \bar{F} , as $\delta_F = (F - \bar{F})/\bar{F}$. From this, the 1D power spectrum $P(k)$ is given by

$$P(k) = \langle \tilde{\delta}_F(k) \tilde{\delta}_F^*(k) \rangle, \quad (4)$$

where $\tilde{\delta}_F(k)$ is the Fourier transform of δ_F and the mean $\langle\rangle$ is taken over all modes with magnitude k along the given line-of-sight axis (one flux skewer per spatial location i, j). From this, we can compute the transfer function $T(k)$, defined by

$$T(k) = \sqrt{\frac{P^{\text{U-Net}}(k)}{P^{\text{True}}(k)}}, \quad (5)$$

to assess how closely our predicted Ly α fields match the test simulation in power spectra.

During training, we use early stopping to select the best model according to how closely the PDFs of the hydrodynamic reconstructions for $\hat{\rho}$, \hat{T} , and \hat{v} match those of the ground truth fields. This is sufficient for evaluating most of our hyperparameter settings. To select our final model from the best configurations, the hydrodynamic reconstruction is computed for the entire test simulation at once by applying our network to the full 1024^3 volume.⁴ We choose our final model based on the accuracy of the PDF and $P(k)$ for the Ly α fields computed from the hydrodynamic reconstruction.

3.2. Network Architecture

As the target baryon hydrodynamical fields each have their own unique features, we partition the problem and train one generator network each for $\hat{\rho}$, \hat{T} , and \hat{v} . For all generator networks, we use the U-Net architecture (Ronneberger et al. 2015). U-Nets are convolutional neural networks, consisting of several downsampling blocks that reduce spatial dimensions as they extract features, then several upsampling blocks that build back up to the output dimensionality. At each spatial feature scale in the upsampling path, skip connections concatenate the corresponding features from the downsampling path along the channel dimension. This design is efficient in extracting high-level, long-range features through the sequential downsampling, while maintaining the ability to resolve fine-grained details via skip connections. The fully convolutional nature of the U-Net allows us to train on smaller subvolumes and then predict on larger volumes without issue.

We base our U-Net architecture loosely on the `pix2pix` model (Isola et al. 2016), and on the architectures used in our initial investigation (Harrington 2019). We did not do an exhaustive hyperparameter sweep, due to the computational costs of training and evaluating our networks, but we tested numerous settings for major design parameters like the number of down- and upsampling blocks, learning rate, loss function weights, and input crop size.

Each downsampling block in our U-Nets consists of a 3D convolution with stride 2, followed by a leaky ReLU activation (Maas et al. 2013). There are a total of six downsampling blocks, and a corresponding six upsampling blocks. Each upsampling block consists of a 3D transposed convolution with stride 1/2,

⁴ This step is done on CPU rather than GPU due to the memory requirement, but only takes several minutes to complete. Alternatively, one can reduce memory requirements by iteratively predicting on smaller subvolumes and stitching them together post hoc by averaging overlapping edges. We have confirmed this approach does not noticeably degrade the quality of results.

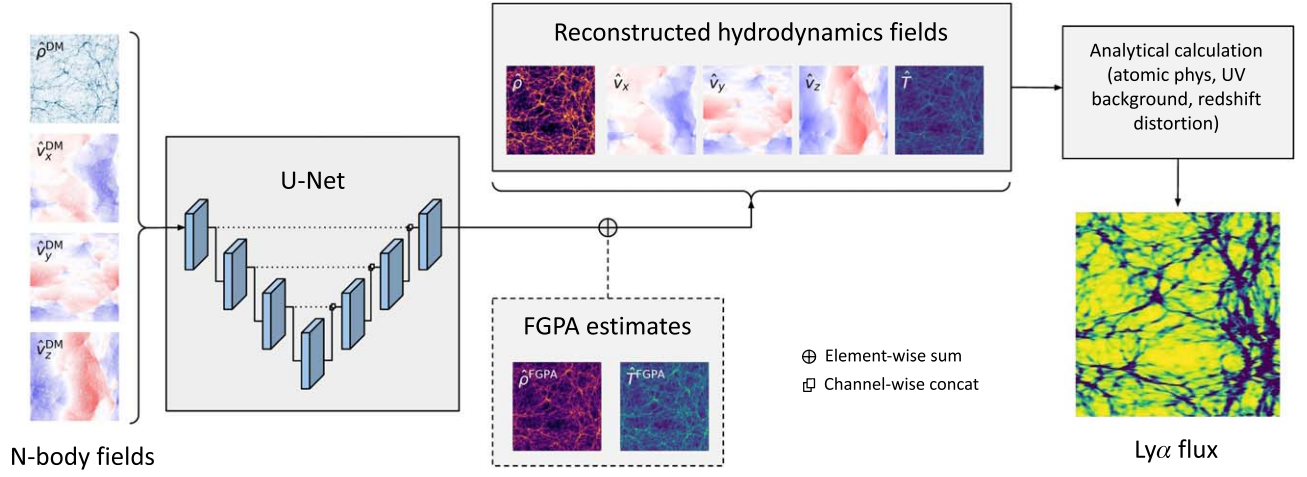


Figure 1. Schematic depicting our model, which predicts the hydrodynamic fields from an input n -body simulation. We train separate networks for $\hat{\rho}$, \hat{T} , and $\hat{\mathbf{v}} = (\hat{v}_x, \hat{v}_y, \hat{v}_z)$, and for $\hat{\rho}$ and \hat{T} we predict the correction to the FGPA approximations $\hat{\rho}^{\text{FGPA}}$, \hat{T}^{FGPA} rather than the full fields. As the model reconstructs the raw hydrodynamic fields, it is independent of Ly α -specific physical details.

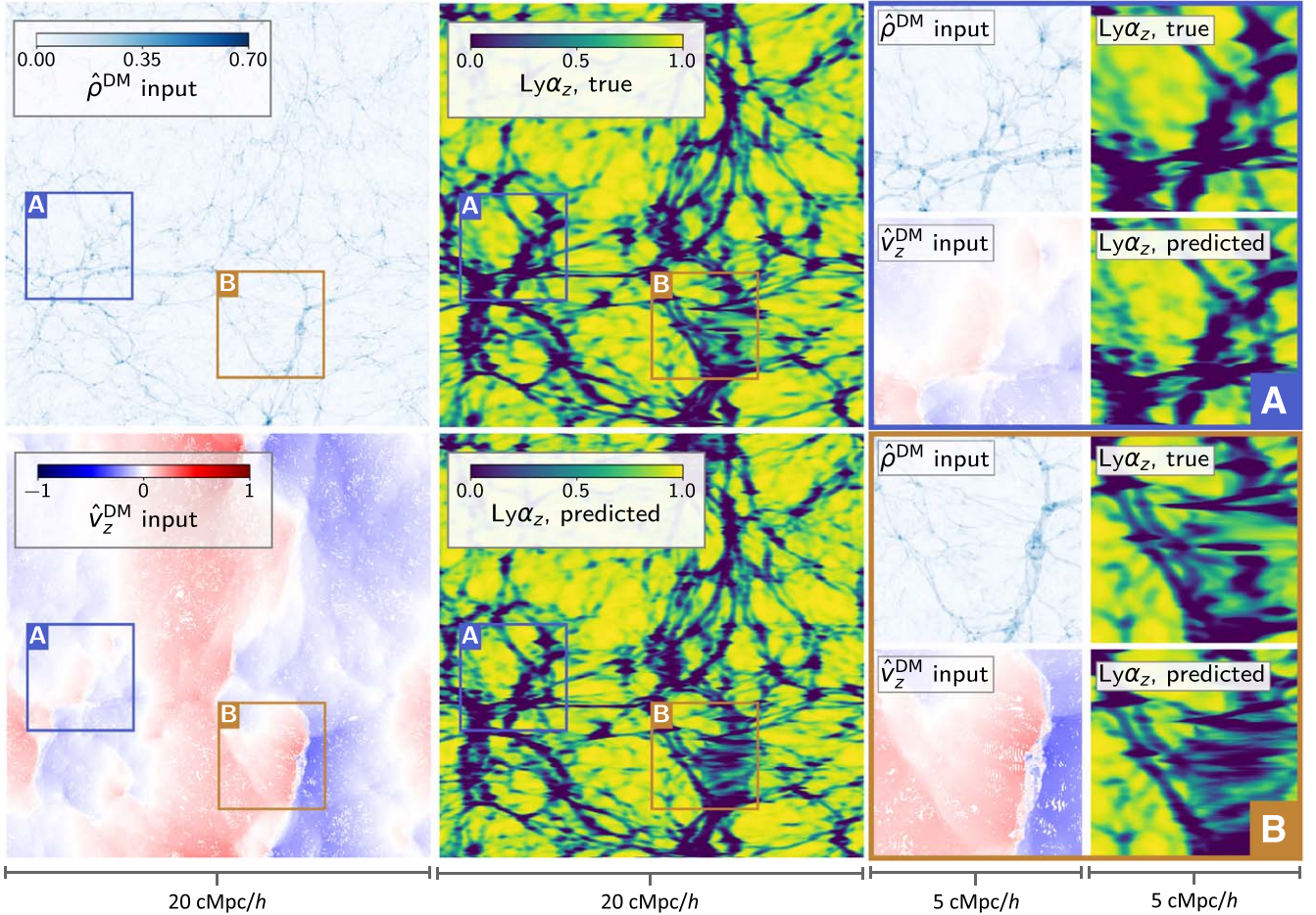


Figure 2. Slices of the test simulation, with the true and predicted Ly α fields along the z -axis line-of-sight. At left, we show a slice through the full volume of the input dark matter density and the line-of-sight dark matter velocity v_z for reference. The corresponding true and predicted Ly α_z fields for this slice are shown in the center panels, and at left we show zoom-ins to regions A and B to better contrast our model against the ground truth.

followed by a ReLU activation (Nair & Hinton 2010), except the final upsampling convolution, which has a tanh nonlinearity.⁵

⁵ We follow the pix2pix model in using Leaky ReLU activations in the downsampling path and ReLU activations in the upsampling path.

After each upsampling block, the incoming skip connections from the corresponding downsampling block are concatenated to the features along the channel axis.

The input to the network consists of the normalized dark matter density and velocity fields ($\hat{\rho}^{\text{DM}}$, \hat{v}_x^{DM} , \hat{v}_y^{DM} , \hat{v}_z^{DM}),

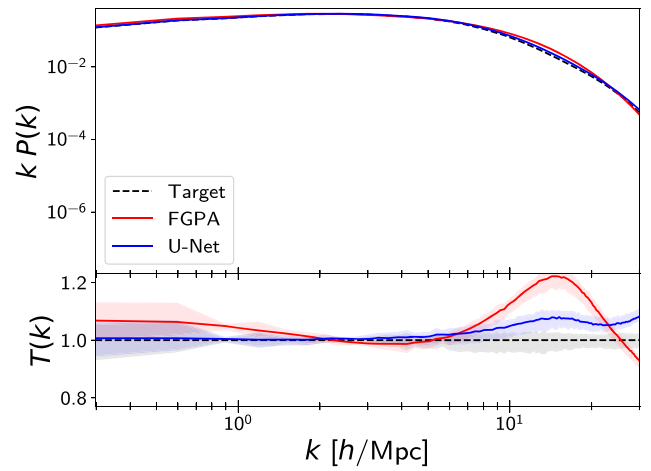
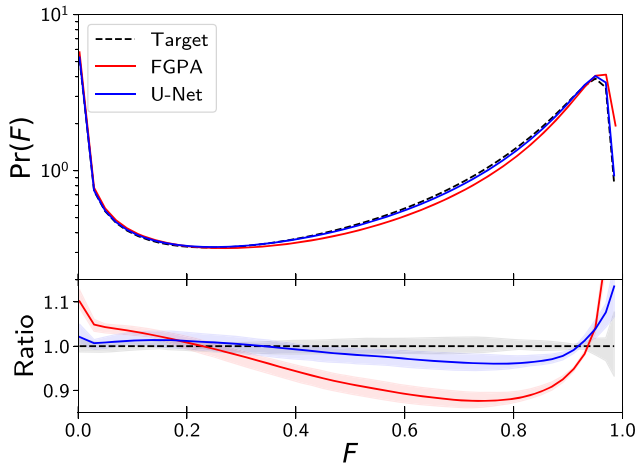


Figure 3. The PDF (left) and power spectrum (right) of full-volume Ly α flux fields from our model and FGPA, compared against the test simulation. Lines and shaded regions represent the mean and min/max per bin across Ly α fields with lines of sight along the x , y , and z axes.

which are supplied as different feature channels over a 3D volume. During training, we generate samples by randomly cropping subvolumes from the training simulations. We find best performance with crops of size 128^3 for the density and temperature fields, and 256^3 for the velocity fields. As additional augmentations, we randomly apply rotations and transposes from the octahedral group to these training samples. Together, the random crops and rotations implicitly enforce the large-scale homogeneity and isotropy of the universe, and help extend the utility of our single pair of training simulations.

For the baryon density and temperature targets, we observe a performance benefit when incorporating the FGPA estimates into the training workflow, as depicted in Figure 1. In this setup, the network predictions are added to the FGPA estimate to produce the final field, so the model is trained to predict a correction term to the FGPA estimate rather than the full hydrodynamic field at once. This accelerates convergence and improves the statistical fidelity of our final fields.

3.3. Loss Functions

The main component of our loss function is the \mathcal{L}_1 distance between the generated and target fields. The velocity U-Net is trained entirely with this loss function, but for the baryon density and temperature fields, we observe an inability to adequately capture high-density/temperature features training only with \mathcal{L}_1 loss. This is consistent with the well-known tendency of \mathcal{L}_1 to seek a conservative (maximum-likelihood) estimate and average over fine-scale features, causing noticeable blurring in the network output (Isola et al. 2016).

Thus, for the density and temperature fields, we also employ an adversarial loss, given by a discriminator D . Contrary to a standard adversarial setup, our discriminator operates in Fourier space rather than the original data space. This is motivated by the observation that taking a sample (cube with side length N) and removing the $N/2$ smallest-scale Fourier modes along each axis retains much of the important detail, while reducing dimensionality by a roughly an order of magnitude (a factor of 2 along each dimension). With our large 3D training samples, this reduction of dimensionality helps compensate for the added overhead of training an additional discriminator network.

We thus use the following procedure when training our spectral discriminator D . The discriminator’s input fields x are crops (with

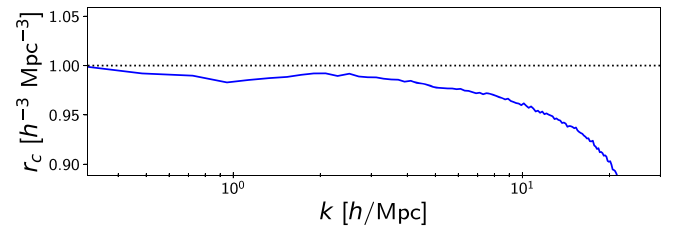


Figure 4. The cross-correlation coefficient r_c of our predicted Ly α field over the relevant length scales. We used the z -axis line of sight for this plot.

side length $N = 128$) of the baryon density or temperature field, coming from either the training simulation or our generator network. We compute the discrete Fourier transform of these crops (with wavenumbers $\{-N/2, \dots, 0, \dots, N/2-1\}$) along each dimension, then drop modes with $|k| \geq N/4$ to get the truncated Fourier transform \tilde{x}_r . Then, we feed the truncated Fourier coefficients $f = \tilde{x}_r \tilde{x}_t^*$ to the spectral discriminator, which tries to classify the sample as real or fake. Our discriminator has four downsampling convolutional layers, followed by three fully connected layers that end in a single output classifying the input as real or fake. To give the spectral discriminator additional context, we concatenate the dark matter density input as an additional feature channel before applying the Fourier truncation.

Following the standard non-saturating GAN formulation (Goodfellow et al. 2014), the adversarial loss for the discriminator is

$$\mathcal{L}_D = -\log D(f_{\text{real}}) - \log(1 - D(f_{\text{fake}})), \quad (6)$$

while the total loss for the $\hat{\rho}$ and \hat{T} U-Nets is

$$\mathcal{L}_{\text{U-Net}} = \lambda \mathcal{L}_1 + \log D(f_{\text{fake}}), \quad (7)$$

where $\lambda = 500$ is a hyperparameter to up-weight the importance of the \mathcal{L}_1 loss.

4. Results

In the following sections, we will visualize our Ly α and corresponding hydrodynamic fields, and detail how closely they match the test simulation in the summary statistics. We compare against the FGPA Ly α estimates as a baseline. Then, we will analyze the hydrodynamic reconstructions and examine some failure modes.

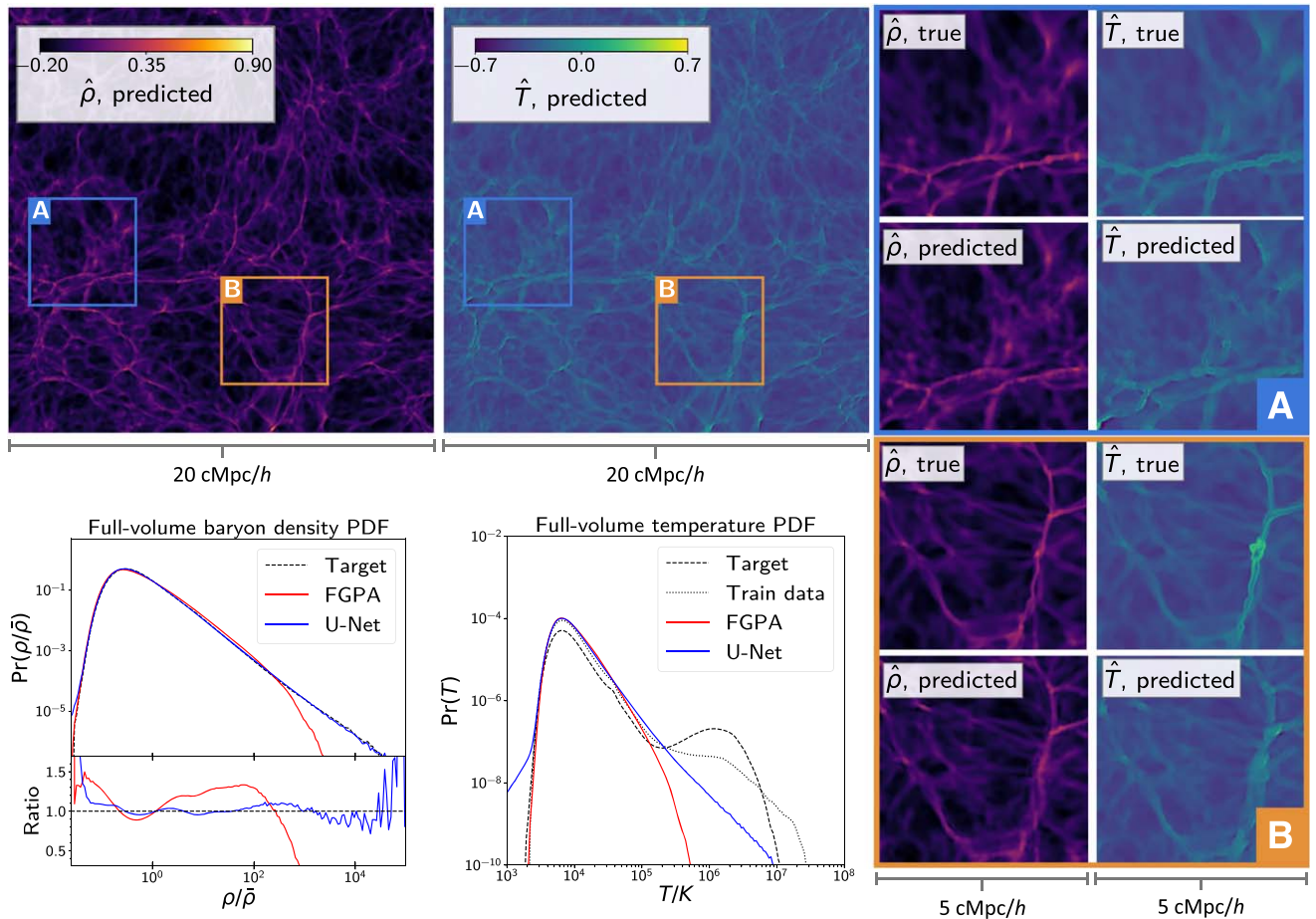


Figure 5. Predicted baryon density and temperature fields, sliced at the same location as plotted in Figure 2. In the right-side panels, we show zoom-ins to regions A and B as before. In the lower left panels, we plot the full-volume PDFs for these fields (after transforming back to original simulation units) compared to the ground truth and FGPA. The reconstructed baryon density field has captured the full range of scales to within $\sim 10\%$, while the temperature field has overfit to the distribution of the train simulation, which differs significantly from the test simulation.

4.1. Ly α Fields

A visualization of our results is shown in Figure 2, where we render some of the input fields and corresponding Ly α fields at a randomly chosen slice in the x - z plane (z was chosen as the line of sight here). The slices run the entire length of the simulation box, but several zoomed-in regions are also shown for increased detail on the right-hand side.

The predicted Ly α field has a high degree of visual fidelity, and closely matches the ground truth in diffuse IGM regions where densities are lower or where there is not much gas motion (see zoom-in panel A). By contrast, regions where gas is collapsing onto a filament structure show more of a discrepancy (zoom-in panel B), as the redshift distortion from the velocity field effectively stretches the size of the filament in the redshifted Ly α field and amplifies any discrepancies between the true and reconstructed hydrodynamic fields along the filament. In general, dense filaments and clusters make up an exceedingly small fraction of the total volume, so the model has more difficulty capturing the fine-scale features in these regions, due to limited training examples.

Because at $z = 3$ the Ly α transmission drops to zero at an overdensity of a few (see Figure 7 in Lukić et al. 2015), we can still obtain statistically accurate Ly α fields from our hydrodynamic reconstructions without necessarily capturing the full

range of scales in the gas physics. This is demonstrated in Figure 3, where we show the PDF and 1D power spectrum of our Ly α estimates and compare against the ground truth. We compute the metrics for Ly α fields with lines of sight along all three axes of our box, and plot the mean, minimum, and maximum per F and k bin across the three fields.

In both statistical measures, the Ly α estimates from our model are within a few percent of the ground truth across the relevant F and k bins, significantly improving on the FGPA Ly α estimates. Our model begins to accumulate an excess of power in $P(k)$ near $k \sim 10 h \text{ Mpc}^{-1}$, as does FGPA. Beyond $10 h \text{ Mpc}^{-1}$, both our model and FGPA begin to diverge from the truth, with our model having an excess of power at high wavenumbers while FGPA diverges in the other direction. However, $k \sim 10 h \text{ Mpc}^{-1}$ is typically the highest mode considered in Ly α spectroscopic studies, limited by contamination from metal lines (see Sorini et al. 2016 and references therein), so we truncate the plot near that wavenumber and do not consider the discrepancies beyond it to be detrimental to our approach.

To further evaluate our model, we compute the cross-correlation coefficient r_c , defined as

$$r_c = \frac{P_{rt}}{\sqrt{P_{rr}P_{tt}}}, \quad (8)$$

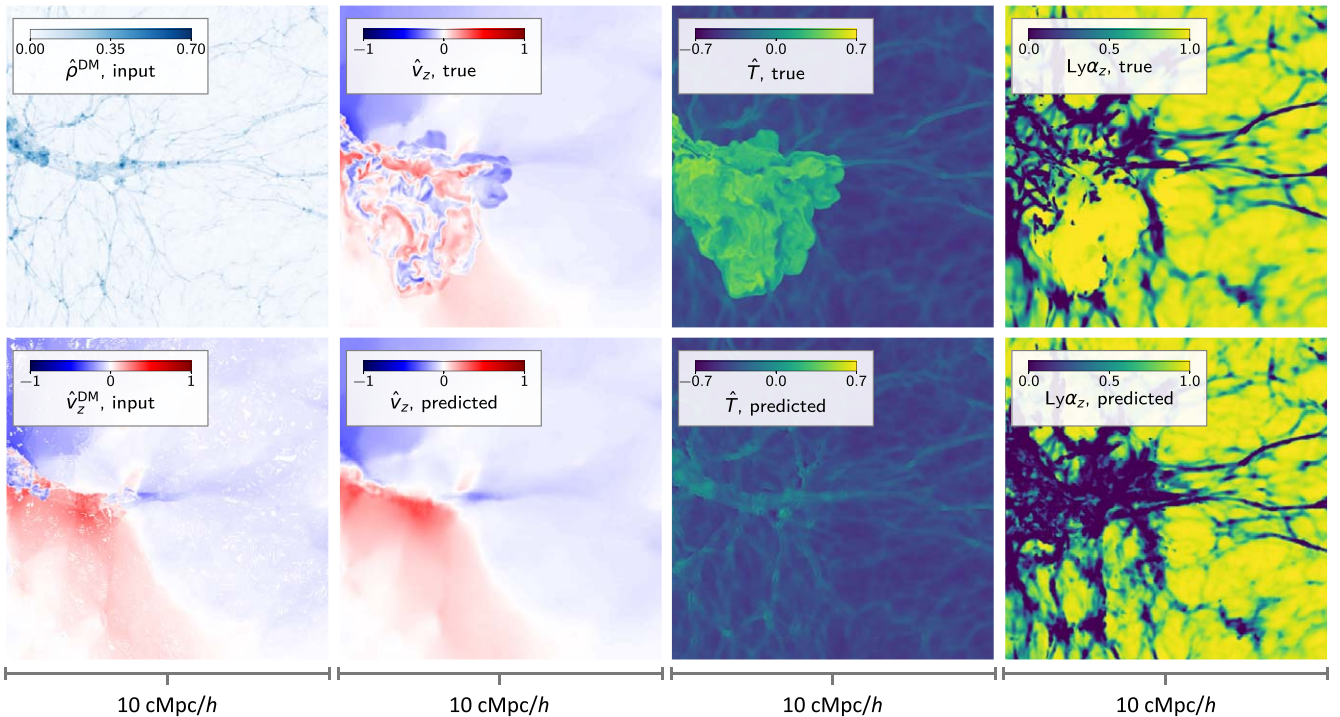


Figure 6. The widespread (~ 5 cMpc) shocked region in the test simulation our model fails to capture. The left column shows some of the input dark matter fields, which indicate a dense cluster but do not otherwise contain information about the exact location and shape of a shock front. The middle panels show the baryon velocity and temperature fields at the same location, and reveal that the model has failed to capture the shock features. On the far right, we show the Ly α slices corresponding to this shocked region.

where P_{rr} is the cross-power and P_{rr} are P_{rr} are the auto-powers of the reconstructed and true Ly α fields, respectively. The cross-correlation for the z -axis line of sight is shown in Figure 4. We see good agreement, with $r_c > 0.95$ across the relevant k range.

4.2. Hydrodynamic Fields

Inspecting the direct output of our model, the hydrodynamic fields, gives more insight into the difficulty of our task. In Figure 5, we visualize slices of the density and temperature fields at the same location as in Figure 2, comparing our reconstructions against the ground truth fields. While there is good agreement by eye, the zoomed-in panels show discrepancies between the true and generated fields, particularly in how sharply resolved the edges of filament structures are. There are also some small-scale artifacts, especially in the temperature field, which possibly contribute to the excess of power seen at small scales in the Ly α $P(k)$.

In the lower left panels, we also plot the PDFs of the two fields, computed over the entire 1024^3 volume of the test simulation. The full-volume PDF for the density field actually shows remarkable agreement between our model’s reconstruction and the true field, with the full range of scales captured to within $\sim 10\%$. This is in contrast to FGPA, which completely fails in estimating densities beyond $(\rho/\bar{\rho}) \gtrsim 10^3$.

In the temperature field, the reconstructions from FGPA and our model both differ significantly from the truth. For reference, we have also plotted the full-volume PDF of the temperature field from the training simulation, and we see that FGPA and our model both closely track the train distribution for $T/K < 10^5$, indicating overfitting. However, there is a

clear distributional difference between the training and testing simulations, with the latter having almost an order-of-magnitude higher contribution from temperatures $T/K > \sim 10^5$ and a subsequent deficit in the PDF at lower temperatures. This suggests we are in the low-data regime by training on a single $20 \text{ Mpc } h^{-1}$ at $\sim 20 \text{ kpc}$ resolution, as the training and testing data are not identically distributed despite having identical cosmologies and physical models. Compounding this discrepancy between training and testing, the clear lack of high temperatures above $T/K = 10^5$ confirms that our model is failing to correctly resolve the sharp boundaries of high-temperature shocks forming at the edges of filament structures.

Inspecting the test simulation more closely, we find a rare widespread shocked region of size ~ 5 cMpc, a significant fraction of the length of the entire box. In Figure 6, we show slices of this region for the true fields and our model’s reconstructions. It is clear that the model does not reproduce the shock structure, instead painting in a fairly smooth rendering of the filament structures in the area. Given that the model is trained on randomly cropped subvolumes of size 2.5 cMpc, it is unreasonable to expect coherent shock structures as large as this to be well-represented. Furthermore, as such structures form $O(1)$ times per $20 \text{ Mpc } h^{-1}$ box, they are exceedingly rare in the training data. In fact, there is no shocked region as large as this in the training simulation, hence the clear difference between the temperature PDFs in the training and testing simulations.

5. Conclusion

In this work, we have presented a convolutional neural network approach to map from collisionless n -body simulations to fully hydrodynamical simulations, reconstructing the baryon

density, temperature, and velocity fields. We find this mapping can match the statistical properties of the true hydrodynamical simulations across a wide range of scales. When computing an observable quantity, Ly α flux, from our model output, we can outperform existing semi-analytical methods across all scales and match the true simulated power spectra within $\sim 5\%$ up to $k \sim 10 h \text{ Mpc}^{-1}$.

The range of scales accurately mapped with this method is well beyond the resolution of next-generation observations (Walther et al. 2021), as well as that of alternate machine-learning-based approaches for modeling Ly α forest (Sinigaglia et al. 2021). Our method provides an alternative to running costly hydrodynamical simulations for mock catalog generation and the construction of Ly α power-spectra emulators. The network can conveniently be trained on small simulated boxes at a given point in cosmological and astrophysical parameter space, then applied to a large dark-matter-only box at inference time. Thus, as the size of the inference box grows, this approach enables increasingly impressive reductions in computational cost.

Beyond Ly α forest statistics, there are some notable limitations to this method. Processes not well-described by the dark matter distribution are difficult to predict, particularly the high-temperature gas shocks seen in Figure 6. While these regions have limited effect on summary statistics used for Ly α cosmological analysis, they are of significant astrophysical interest, as they are associated with the physics of the Warm Hot Intergalactic Medium (Cen & Ostriker 1999; Pfrommer et al. 2006; Vazza et al. 2019). Considering our NN design, the predictions in Figure 6 are not entirely unreasonable, given the input dark matter fields from the n -body simulation (visualized in the far left panels). These inputs contain a dense cluster toward the edge, but do not have obvious indicators of the scale or detailed shape of the shock front. Thus, passing these inputs to a deterministic model such as ours yields a more conservative prediction, as the model does not have the expressive capacity to represent the many possible shapes of shocked gas. We thus propose that a variational model producing multimodal outputs would best resolve shocks like the one shown in Figure 6. Such a model would synthesize multiple realizations of the hydrodynamic variables via some latent variable, conditioned on the input n -body fields. We explore this challenge in Horowitz et al. (2021a), a companion work to this paper.

B.H. is supported by the AI Accelerator program of the Schmidt Futures Foundation. This work was partially supported by the DOE's Office of Advanced Scientific Computing Research and Office of High Energy Physics through the Scientific Discovery through Advanced Computing (SciDAC) program. This research used resources of the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility supported by the Office of Science of the U.S. Department of Energy under Contract No. DEC02-05CH11231.

ORCID iDs

Benjamin Horowitz  <https://orcid.org/0000-0001-7832-5372>

References

- Almgren, A. S., Bell, J. B., Lijewski, M. J., Lukić, Z., & Van Andel, E. 2013, *ApJ*, **765**, 39
- Boera, E., Becker, G. D., Bolton, J. S., & Nasir, F. 2019, *ApJ*, **872**, 101
- Cen, R., & Ostriker, J. P. 1999, *ApJ*, **514**, 1
- Chabanier, S., Bournaud, F., Dubois, Y., et al. 2020, *MNRAS*, **495**, 1825
- Coc, A., Uzan, J.-P., & Vangioni, E. 2013, arXiv:1307.6955
- Croft, R. A. C., Weinberg, D. H., Pettini, M., Hernquist, L., & Katz, N. 1999, *ApJ*, **520**, 1
- Desjacques, V., Haehnelt, M. G., & Nusser, A. 2006, *MNRAS*, **367**, L74
- Friesen, B., Almgren, A., Lukić, Z., et al. 2016, *ComAC*, **3**, 4
- Gnedin, N. Y., & Hui, L. 1996, *ApJL*, **472**, L73
- Gnedin, N. Y., & Hui, L. 1998, *MNRAS*, **296**, 44
- Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., et al. 2014, arXiv:1406.2661
- Gunn, J. E., & Peterson, B. A. 1965, *ApJ*, **142**, 1633
- Haardt, F., & Madau, P. 2012, *ApJ*, **746**, 125
- Harrington, P. 2019, in *NeurIPS Machine Learning & the Physical Sciences Workshop, 2019* (Red Hook, NY: Curran Associates)
- Horowitz, B., Dornfest, M., Lukić, Z., & Harrington, P. 2021a, arXiv:2106.12675
- Horowitz, B., Lee, K.-G., White, M., Krolewski, A., & Ata, M. 2019, *ApJ*, **887**, 61
- Horowitz, B., Zhang, B., Lee, K.-G., & Kooistra, R. 2021b, *ApJ*, **906**, 110
- Hui, L., & Gnedin, N. Y. 1997, *MNRAS*, **292**, 27
- Iršič, V., Viel, M., Haehnelt, M. G., et al. 2017, *PhRvD*, **96**, 023522
- Isola, P., Zhu, J.-Y., Zhou, T., & Efros, A. A. 2016, arXiv:1611.07004
- Kollmeier, J. A., Miralda-Escudé, J., Cen, R., & Ostriker, J. P. 2006, *ApJ*, **638**, 52
- Lukić, Z., Stark, C. W., Nugent, P., et al. 2015, *MNRAS*, **446**, 3697
- Maas, A. L., Hannun, A. Y., & Ng, A. Y. 2013, in *Proc. of the 30th Int. Conf. on Machine Learning* (Madison, WI: Omnipress), 3
- McQuinn, M. 2016, *ARA&A*, **54**, 313
- Nair, V., & Hinton, G. E. 2010, in *Proc. of the 27th Int. Conf. on Machine Learning* (ICML-10) (Madison, WI: Omnipress), 807
- Oñorbe, J., Davies, F. B., Lukić, Z., et al. 2019, *MNRAS*, **486**, 4075
- Palanque-Debraouille, N., Yèche, C., Schöneberg, N., et al. 2020, *JCAP*, **2020**, 038
- Pedersen, C., Font-Ribera, A., Rogers, K. K., et al. 2021, *JCAP*, **2021**, 033
- Peirani, S., Weinberg, D. H., Colombi, S., et al. 2014, *ApJ*, **784**, 11
- Petitjean, P., Mucket, J. P., & Kates, R. E. 1995, *A&A*, **295**, L9
- Pfrommer, C., Springel, V., Enßlin, T. A., & Jubelgas, M. 2006, *MNRAS*, **367**, 113
- Porqueres, N., Jasche, J., Lavaux, G., & Enßlin, T. 2019, *A&A*, **630**, A151
- Rauch, M., Miralda-Escudé, J., Sargent, W. L. W., et al. 1997, *ApJ*, **489**, 7
- Rogers, K. K., Bird, S., Peiris, H. V., et al. 2018, *MNRAS*, **474**, 3032
- Rogers, K. K., & Peiris, H. V. 2021, *PhRvL*, **126**, 071302
- Ronneberger, O., Fischer, P., & Brox, T. 2015, in *Int. Conf. on Medical Image Computing and Computer-Assisted Intervention 2015* (Berlin: Springer), 234
- Seljak, U. 2012, *JCAP*, **2012**, 004
- Sexton, J., Lukić, Z., Almgren, A., et al. 2021, *JOSS*, **6**, 3068
- Sinigaglia, F., Kitaura, F.-S., Balaguera-Antolínez, A., et al. 2021, arXiv:2107.07917
- Sorini, D., Oñorbe, J., Lukić, Z., & Hennawi, J. F. 2016, *ApJ*, **827**, 97
- Thiele, L., Villaescusa-Navarro, F., Spergel, D. N., Nelson, D., & Pillepich, A. 2020, *ApJ*, **902**, 129
- Tröster, T., Ferguson, C., Harnois-Déraps, J., & McCarthy, I. G. 2019, *MNRAS*, **487**, L24
- Upton Sanderbeck, P., & Bird, S. 2020, *MNRAS*, **496**, 4372
- Vazza, F., Ettori, S., Roncarelli, M., et al. 2019, *A&A*, **627**, A5
- Walther, M., Armengaud, E., Ravoux, C., et al. 2021, *JCAP*, **2021**, 059
- Walther, M., Oñorbe, J., Hennawi, J. F., & Lukić, Z. 2019, *ApJ*, **872**, 13
- Zamudio-Fernandez, J., Okan, A., Villaescusa-Navarro, F., et al. 2019, arXiv:1904.12846